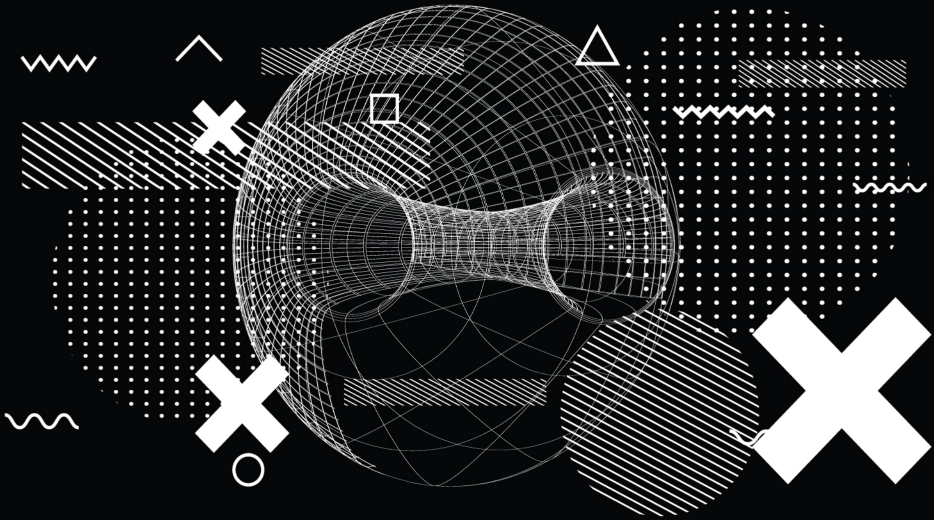
The Use of Mathematical Structures: Modelling Real Phenomena

Edited by: Olga Moreira



AP | ARCLER PRESS

The Use of Mathematical Structures: Modelling Real Phenomena

THE USE OF MATHEMATICAL STRUCTURES: MODELLING REAL PHENOMENA

Edited by:

Olga Moreira

ARCLER
PRESS

www.arclerpress.com

The Use of Mathematical Structures: Modelling Real Phenomena

Olga Moreira

Arcler Press
224 Shoreacres Road
Burlington, ON L7L 2H2
Canada
www.arclerpress.com
Email: orders@arclereducation.com

e-book Edition 2023

ISBN: 978-1-77469-697-2 (e-book)

This book contains information obtained from highly regarded resources. Reprinted material sources are indicated. Copyright for individual articles remains with the authors as indicated and published under Creative Commons License. A Wide variety of references are listed. Reasonable efforts have been made to publish reliable data and views articulated in the chapters are those of the individual contributors, and not necessarily those of the editors or publishers. Editors or publishers are not responsible for the accuracy of the information in the published chapters or consequences of their use. The publisher assumes no responsibility for any damage or grievance to the persons or property arising out of the use of any materials, instructions, methods or thoughts in the book. The editors and the publisher have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission has not been obtained. If any copyright holder has not been acknowledged, please write to us so we may rectify.

Notice: Registered trademark of products or corporate names are used only for explanation and identification without intent of infringement.

ISBN: 978-1-77469-499-2 (Hardcover)

DECLARATION

Some content or chapters in this book are open access copyright free published research work, which is published under Creative Commons License and are indicated with the citation. We are thankful to the publishers and authors of the content and chapters as without them this book wouldn't have been possible.

ABOUT THE EDITOR

Olga Moreira is a Ph.D. and M.Sc. in Astrophysics and B.Sc. in Physics/Applied Mathematics (Astronomy). She is an experienced technical writer and data analyst. As a graduate student, she held two research grants to carry out her work in Astrophysics at two of the most renowned European institutions in the fields of Astrophysics and Space Science (the European Space Agency, and the European Southern Observatory). She is currently an independent scientist, peer-reviewer and editor. Her research interest is solar physics, machine learning and artificial neural networks.

TABLE OF CONTENTS

LIST OF CONTRIBUTORS

Mary Leng
Department of Philosophy, University of York, York, UK

Arkady Plotnitsky
Theory and Cultural Studies Program, Purdue University, West Lafayette, IN, United States

Robert Pascal
Institut des Biomolécules Max Mousseron, UMR5247 CNRS–Universités Montpellier 1 & Montpellier 2, CC17006, Place E. BataillonCC17006, Place E. Bataillon, Montpellier F-34095, France.

Addy Pross
Department of Chemistry, Ben-Gurion University of the Negev, Be'er Sheva 84105, Israel.

Xiao-Feng Yang
Department of Applied Mathematics, Northwestern Polytechnical University, Xi'an, 710072, P.R. China.

Zi-Chen Deng
School of Mechanics, Civil Engineering and Architecture, Northwestern Polytechnical University, Xi'an, 710072, P.R. China.
State Key Laboratory of Structural Analysis of Industrial Equipment, Dalian University of Technology, Dalian, 116023, P.R. China.

Yi Wei
Department of Applied Mathematics, Northwestern Polytechnical University, Xi'an, 710072, P.R. China.

Sumaiya B. Islam
Department of Applied Mathematics, University of Dhaka, Dhaka 1000, Bangladesh

Suraiya A. Shefa
Department of Applied Mathematics, University of Dhaka, Dhaka 1000, Bangladesh

Tania S. Khaleque
Department of Applied Mathematics, University of Dhaka, Dhaka 1000, Bangladesh

Yonghui Huang
Helmholtz Centre for Environmental Research - UFZ, Permoserstr. 15, 04318 Leipzig, Germany.
Technical University of Dresden, Helmholtz-Strane 10, 01062 Dresden, Germany.

Olaf Kolditz
Helmholtz Centre for Environmental Research - UFZ, Permoserstr. 15, 04318 Leipzig, Germany.
Technical University of Dresden, Helmholtz-Strane 10, 01062 Dresden, Germany.

Haibing Shao
Helmholtz Centre for Environmental Research - UFZ, Permoserstr. 15, 04318 Leipzig, Germany.
Freiberg University of Mining and Technology, Gustav-Zeuner-Strasse 1, 09596 Freiberg, Germany.

Pei Wang
School of Mechanical Engineering, Yanshan University, Qinhuangdao 066004, China.

Jing Yao
School of Mechanical Engineering, Yanshan University, Qinhuangdao 066004, China.
Advanced Manufacturing Forming Technology and Equipment, Qinhuangdao 066004, China.
Hebei Provincial Key Laboratory of Heavy Fluid Power Transmission and Control, Yanshan University, Qinhuangdao 066004, China.

Baidong Feng
School of Mechanical Engineering, Yanshan University, Qinhuangdao 066004, China.

Mandi Li
School of Mechanical Engineering, Yanshan University, Qinhuangdao 066004, China.

Dingyu Wang
School of Mechanical Engineering, Yanshan University, Qinhuangdao 066004, China.

Peng Zhang
Department of Mechanical Engineering, Anhui University of Technology, Maanshan 243000, China

Bingbing Bao
Department of Mechanical Engineering, Anhui University of Technology, Maanshan 243000, China

Meng Wang
Department of Mechanical Engineering, Anhui University of Technology, Maanshan 243000, China

Sabatino Cuomo
Geotechnical Engineering Group (GEG), University of Salerno, Via Giovanni Paolo II, 132 84084 Fisciano, Italy

Mo Faheem
Department of Mathematics, Jamia Millia Islamia, New Delhi, 110025, India.

Arshad Khan
Department of Mathematics, Jamia Millia Islamia, New Delhi, 110025, India.

E.R. El-Zahar
Department of Mathematics, College of Sciences and Humanities in Al-Kharj, Prince Sattam bin Abdulaziz University, Alkharj 11942, Saudi Arabia.
Department of Basic Engineering Science, Faculty of Engineering, Menoufia University, Shebin El-Kom 32511, Egypt.

Assane Savadogo
Department of Mathematics and Informatics, UFR/ST, UNB, 01 BP 1091 Bobo Dsso 01, Bobo Dioulasso, Burkina Faso.

Boureima Sangaré
Department of Mathematics and Informatics, UFR/ST, UNB, 01 BP 1091 Bobo Dsso 01, Bobo Dioulasso, Burkina Faso.

Hamidou Ouedraogo
Department of Mathematics and Informatics, UNB, Bobo Dioulasso, Burkina Faso.

James A. R. Marshall
Department of Computer Science, University of Sheffield, Sheffield, United Kingdom

Andreagiovanni Reina
Department of Computer Science, University of Sheffield, Sheffield, United Kingdom

Thomas Bose
Department of Computer Science, University of Sheffield, Sheffield, United Kingdom

Pushpendra Kumar
Department of Mathematics, National Institute of Technology Puducherry, Karaikal, 609609, India.

V. Govindaraj
Department of Mathematics, National Institute of Technology Puducherry, Karaikal, 609609, India.

Vedat Suat Erturk
Department of Mathematics, Faculty of Arts and Sciences, Ondokuz Mayis University, Atakum, 55200, Samsun, Turkey.

Mohamed S. Mohamed
Department of Mathematics and Statistics, College of Science, Taif University, Taif, 21944, Saudi Arabia.

LIST OF ABBREVIATIONS

ADM	Adomian Decomposition Methods
BPs	Branch Points
CFC	Chlorofluorocarbon
DA	Debris Avalanche
DF	Debris Flows
DKS	Dynamic Kinetic Stability
EHD	Electrohydrodynamics
EOS	Equation of State
LPs	Limit Points
MHD	Magnetohydrodynamics
MuMoT	Multiscale Modelling Tool
NAPL	Non-Aqueous Phase Liquid
NLPDEs	Nonlinear Partial Differential Equations
ODEs	Ordinary Differential Equations
Q-numbers	Quantum Numbers
SP	Singularized Probabilistic
TDSP	Time-Dependent Singularized Probabilistic

PREFACE

The process of describing real-world problems as mathematical structures and abstract objects is often referred to as mathematical modelling. A mathematical model of a real-world problem consists of an approximate description in the form of a differential equations' system. Modern scientists and engineers strive to solve these equations to help them understand the origin and discover new features concerning a real-world problem. Then, once the physical interpretation of the mathematical solution is clearly captured, they often attempt to improve or extend these mathematical modelling approximations to more general situations by increasing the complexity of mathematical models.

This book includes several articles devoted to the mathematical modelling of real-world problems in physics, mechanical engineering, biology, and biochemistry. It is divided into four thematic sections. Each section covers a different topic in mathematical modelling for describing and understanding physical phenomena.

The first part of this book (chapters 1 to 3) reflects on mathematical modelling from a more philosophical and generic point of view. Chapter 1 inquiries about the explanatory role of mathematics in empirical science. The author attempts to answer the question: "Are there genuine mathematical explanations of physical phenomena, and if so, how can mathematical theories, which are typically thought to concern abstract mathematical objects, explain contingent empirical matters?". Chapter 2 reflects on the application of quantum mathematical modelling in the fields of psychology, economics, and decision science. The author discusses whether quantum mathematical models are necessary for dealing with specific phenomena in the aforementioned fields, or whether the classical (probabilistic or statistical) mathematical models will suffice. Chapter 3 is focused on the application of stability theory (a fundamental part of mathematical modelling of natural phenomena) in the fields of chemistry and biology. The authors propose that both fields can be conceptually connected through the concept of stability.

The second part of this book (chapters 4 to 9) is devoted to the application of mathematical modelling to specific real-world problems in fluid dynamics and mechanical engineering. It is focused on solving ordinary differential equations (ODEs) used for describing a wide variety of phenomena in physics, biology, chemistry, and several other fields. Chapter 4 describes the application of the Riccati-Bernoulli sub-ODE method for finding exact travelling wave solutions, solitary wave solutions and peaked wave solutions of nonlinear partial differential equations, which play an important role in the study of nonlinear physical phenomena. Chapter 5 describes the mathematical modelling of mantle convection at a high Rayleigh number with variable viscosity and viscous dissipation. Mantle convection is responsible for numerous physical and chemical phenomena occurring on the surface and in the interior of the Earth. The study is important for shading light on the mechanism behind this type of convection, which remains an unsolved problem since the rheology of mantle rocks. Chapter 6 describes an application of a multi-component multiphase reactive transport model for geothermal reservoir simulation. Chapter 7 describes the modelling and dynamic characteristics of a non-metal pressurized reservoir with variable volume. A closed reservoir may provide an advantage of having a smaller volume when compared to open reservoirs which are large, heavy, polluted, and threaten the operation of hydraulic systems. Chapter 8 describes the modelling and natural characteristic analysis of cycloid ball transmission using lumped stiffness method. The study is motivated by the possibility of improving the dynamic precision of robot systems. Chapter 9 describes the modelling of flowslides and debris avalanches in natural and engineered slopes. The study aims to provide a better understanding of how slope instability is affected by the rainfall from the ground surface and water springs from a bedrock.

The third part of this book (chapters 10 to 13) is devoted to the application of mathematical modelling to the fields of biology and biochemistry. Chapter 10 is focused on the Lane-Emden boundary value problem, arising in numerous real-life chemical and biochemical phenomena. Chapter 11 aims to provide a mathematical analysis and modelling of a prey-predator system to describe the effect of predation between prey and predator with a nonlinear functional response. Chapter 12 describes the mathematical modelling of collective behaviour appearing at several levels of biological complexity, from single cells to super-organisms. Chapter 13 describes a fractional mathematical model for studying the effects of greenhouse gases and hypoxia on the population of aquatic species.

Chapter

1

MODELS, STRUCTURES, AND THE EXPLANATORY ROLE OF MATHEMATICS IN EMPIRICAL SCIENCE

Mary Leng

Department of Philosophy, University of York, York, UK

ABSTRACT

Are there genuine mathematical explanations of physical phenomena, and if so, how can mathematical theories, which are typically thought to concern abstract mathematical objects, explain contingent empirical matters? The answer, I argue, is in seeing an important range of mathematical explanations as *structural explanations*, where structural explanations explain a phenomenon by showing it to have been an inevitable consequence of the structural features instantiated in the physical system under consideration. Such explanations are best cast as deductive arguments which, by virtue of their form, establish that, *given* the mathematical structure instantiated in the physical system under consideration, the explanandum *had* to occur. Against the claims of platonists such as Alan Baker and Mark Colyvan, I argue that

Citation: Citation (APA): Leng, M. (2021). Models, structures, and the explanatory role of mathematics in empirical science. Synthese, 199(3), 10415-10440. (26 pages).

formulating mathematical explanations as structural explanations in this way shows that we can accept that mathematics can play an indispensable explanatory role in empirical science without committing to the existence of any abstract mathematical objects.

Keywords: Mathematics, Models, Explanation, Structure, Indispensability

INTRODUCTION

Are there genuine mathematical explanations of physical phenomena, and if so, how can mathematical theories, which are typically thought to concern abstract mathematical objects, explain contingent empirical matters? Lange (2016), for example, argues that mathematical explanations of physical phenomena are a species of non-causal explanations that he calls *explanations by constraint*. But how can facts about spatiotemporally isolated mathematical objects can act as *constraints* on the physical world? The answer, I will argue, is in seeing an important range of mathematical explanations as *structural explanations*, where structural explanations explain a phenomenon by showing it to have been an inevitable consequence of the structural features instantiated in the physical system under consideration. Such explanations are best cast as deductive arguments which, by virtue of their form, establish that, *given* the mathematical structure instantiated in the physical system under consideration, the explanandum *had* to occur. The constraints placed on the world by the mathematical premises in these explanations are thus logical constraints: such explanations show that, given structural features of the physical system, their explananda were inevitable as a matter of logic.

Several questions arise out of this picture. First, does couching so-called mathematical explanations of physical phenomena as structural explanations establish that these are genuine *explanations*? A full answer to this question would require a full account of what it is to explain, and this is not something that I will pursue here (though I endorse much of what Lange (2016) has to say in defence of taking so-called ‘explanations by constraint’ as genuinely explanatory). My own view is there are features of these so-called ‘explanations’ that suggest that there is at least a case for including them as examples of genuine explanations. In particular, they supply important *modal* information about their explananda: they tell us why they had to occur given the structural features of the physical situation. They also offer opportunities for understanding provided by *unification*, through showing how apparently disparate phenomena are instances of a common structure.

Regardless, though, of whether what I call 'structural explanations' are genuine *explanations* or merely explanation-like (e.g. in providing some form of illumination/understanding of their target phenomena), what I am most keen to explore in this paper is a different question, that of whether the (explanatory- or explanation-like) theoretical role played by such structural 'explanations' offers support for mathematical platonism. Perhaps we might be moved to accept an account of explanation according to which all *genuine* explanations are causal. Nevertheless, as I will argue in Sect. 1, many so-called mathematical explanations of physical phenomena afford us at the very least important forms of *understanding* that are not available if focus on nominalistically-stated alternatives. So even if *supplying modal information* about an observed phenomenon, and *unifying disparate phenomena* turn out to be not enough to count as providing an explanation in a strict sense, these still remain important theoretical roles played by mathematics in science that go beyond what would be available if we confined ourselves to purely nominalistically-stated alternatives. And this raises the question of whether, if what I am calling structural 'explanations' succeed where purely non-mathematical descriptions fail in enhancing our understanding of the physical world in these kinds of ways, this amounts to an indispensable theoretical role that supports platonism. Attending to the nature of structural explanations shows that any attempt to argue from the indispensable theoretical role of structural explanations to mathematical platonism must fail, for structural explanations of physical phenomena do not require that our structure-characterising mathematical axioms are true of any *mathematical objects*, but only that they are true—or approximately true—when their non-logical terminology is interpreted to apply to systems of either actual, or idealized, *physical objects*. So admitting an indispensable theoretical role for mathematical-structural explanations does not support an inference to the existence of abstract mathematical objects.

The picture of mathematical explanations as structural explanations that I present here is sketched in Leng (2012) and Leng (2021), but it has not been developed in full detail in previously published work. This paper fills in the details of this sketch. In Sect. 1 I look at some examples of the alleged 'explanatory' role of mathematics in physical science, and agree with platonists such as Baker and Colyvan that there is important theoretical work done by mathematics in the examples they present that is not available if we focus solely on non-mathematical alternatives. I side with Baker and Colyvan there in saying that the theoretical role played by mathematics in these examples should be thought of as an 'explanatory' role, but even for those not convinced that this is genuine *explanation*, I argue that Baker and Colyvan have at the very least indicated an important theoretical role played by mathematics in physical science, and this raises the question of *how* mathematics is able to play this role, and in particular of whether the ability of mathematical theories to play this kind of role requires the existence of mathematical objects. The remainder of the paper considers the question of whether the existence of these kinds of mathematical explanations of physical phenomena supports the existence of mathematical objects. Section 2 characterises a class of mathematical explanations as structural explanations, arguing that they can be presented as deductively valid arguments whose premises include a mathematical theorem expressed modal structurally, together with empirical claims establishing that the conditions for the mathematical theorem are instantiated in the physical system under consideration. I suggest that these arguments should be thought of as genuinely explanatory by virtue of providing important modal information: they show that the phenomenon to be explained *had* to occur, given the structural features that are physically instantiated. Additionally, by identifying mathematical-structural features that necessitate the occurrence of the phenomenon to be explained, they offer opportunities for explanatory unification, showing apparently disparate phenomena to be consequences of the very same mathematical-structural features. I also show that these explanations, which can be understood in modal structural terms, involve no commitment to mathematical objects platonistically construed. In Sect. 3 I consider the application of this account to real cases where mathematical structure is instantiated not directly in physical systems, but only in an idealised model of a physical description (in what, following Bokulich, 2008 I will call 'structural model explanations'). I argue that the *explanatory* use of mathematics in these idealized model cases offers no further argument for realism than is already offered by the use of idealized models to represent physical phenomena. I also point to a helpful feature of the structural account as compared to mapping account of applications of mathematics: while it is true that in many cases the relation of mathematics to reality is of a map to a terrain, if the structural account is correct, mathematics does not explain simply by providing such a map, but by showing how mathematical-structural dependencies in mathematical models reflect mathematical-structural dependences in the physical world. I conclude, then, that viewing mathematical explanations of structural explanations provides an understanding of how mathematics can play a significant theoretical role in our understanding of physical phenomena that does not require us to adopt a platonist account of mathematical objects.

WHY THINK THAT MATHEMATICS DOES GENUINE EXPLANATORY WORK?

Since Alan Baker's (2005) paper introducing the philosophy of mathematics world to the curious case of the periodical magicicada cicadas, much has been written on the alleged existence of mathematical explanations of physical phenomena. Typically, discussion has been divided along platonist/anti-platonist lines, with most platonists agreeing that there are such explanations, and most anti-platonists disagreeing (notable exceptions are Brown (2012) on the platonist side, and Leng (2012) on the anti-platonist side). For those who reject the claim that mathematics does genuine explanatory work in our scientific theories, a standard strategy has been to point to the *nominalistic content* of putative mathematical explanations of physical phenomena, holding that while these explanations may be characterised mathematically, all the genuine explanatory work in these explanations is carried by their nominalistic content, with mathematics being used as a convenient—and perhaps indispensable—way of indexing the explanatorily relevant physical facts. (Examples of strategies along these lines include Brown, 2012; Daly & Langford, 2009; Melia, 2000; Saatsi, 2011) In Leng (2012) I side with platonists including Baker and Colyvan (2011) in suggesting that if we focus on the nominalistic content of mathematical explanations of physical phenomena, we lose explanatory power.

Take for example Brown's account of the cicada case. Brown (2012, p. 10) uses the notions of *cycle factorizability* and *non-factorizability* to pick out nominalistically characterizable features of cicada life-cycles that he takes are ultimately responsible for the prime-length period phenomenon. Although there is a clear link between these notions and the mathematical notions of 'composite' and 'prime' as applied to numbers, Brown notes that nonetheless they are intelligible in non-mathematical terms (a cicada cycle is cycle factorizable if and only if it can be broken into repeated shorter cycles of equal duration without leaving any years out). A cycle is non-factorizable if and only if its associated number (of years) is prime, hence the relevance of talk of prime numbers in indexing the standard evolutionary explanation of cicada period length. But the real explanatory work, Brown contends, is done by the nominalistically kosher feature of cycle lengths that is indexed by prime numbers (non-factorizability).

It certainly seems right that it is cycle-non-factorizability (along with the relevant evolutionary facts that the explanation presupposes about periodic predators) that is responsible for the cicada's behaviour. In that sense, the non-factorizability of the 13 and 17 year cycles does explain why those cycles were chosen. But even though an adequate explanation can be afforded in terms of the nominalistically acceptable notion of cycle-factorizability, there is at least a sense in which, by refusing to appeal to the more general notion of prime number as it relates to non-factorizable cycles, this explanation remains impoverished. By framing the explanation in terms of prime numbers (with the non-factorizable cycles being those that are indexed by prime numbers) we can make use of our knowledge of prime and composite numbers in order to understand more about the possibilities for similar periodic behaviour. For example, the fundamental theorem of algebra, which tells us that composite numbers have a unique prime decomposition, can tell us that, of composite cycles, cycle lengths with fewer distinct prime factors would be preferable. So, for example, a 4-cycle would be preferable to a 6-cycle since it has only one prime factor (2) rather than two (2, 3), so while a 4 cycle would meet 2-cycle predators every time it occurred, it would only meet 3-cycle predators every fourth cycle (once every 12 years). On the other hand, a 6-cycle creature would meet 2-cycle and 3-cycle predators every time it occurred, making that a worse choice of cycle length in conditions where 2-cycle and 3-cycle predators occur. Such extrapolations concerning potential periodic behaviour come naturally when the explanation is framed in term of prime numbers, given our familiarity with their patterns, but are lost if we drop that framing and instead focus directly on the indexed property of cycle-factorizability. The mathematical framing thus offers easy access to a range of *modal information* concerning what would have happened had different cycle lengths been chosen, that is not present if we focus solely on cycle-factorizability. Along a similar vein, the well known 'Bridges of Königsberg' explanation using Euler's theorem not only shows why a certain kind of walk is impossible, but also provides information about what kinds of bridge/landmass configurations would be required to make possible a Eulerian walk.

Focussing on cycle-factorizability also prevents us from seeing connections with other phenomena that are naturally indexed with prime numbers, but which have nothing to do with cycle lengths. A teacher may come to realise that classes of 30 students are easier to work with than classes of 25, since in splitting into groups the latter can only be split evenly into 5 groups of 5 pupils, while the former has the option of 15 pairs, 6 groups of 5, 5 groups of 6, 3 groups of 10, or 2 of 15. Better choices of cycle lengths (for the purpose of avoiding predators) turn out to be worse choices of class sizes (for the purpose of allowing maximal opportunities for group work). Of course we *could* introduce a separate notion of collection-factorizability to apply to collections of discrete individuals, where a collection is factorizable if it can be broken up into a number of smaller collections of equal size without remainder. But there is obviously a common pattern here, and we are surely best placed to appreciate and understand that common pattern once we see the natural associations between collections of individuals, repeating cycles, and the prime and composite numbers that are used to index both. Along similar lines, Baker (2017) points to another example of a use of prime vs composite cycles as part of an explanation of a physical phenomena: an explanation of why fixed gear bikes where the numbers of cogs on front and back wheel are coprime see less wear from braking than bikes where the pairs are not coprime. I agree with Baker that couching all of these explanations in mathematical terms provides them with a *topic-generality* that adds a level of explanatoriness that goes beyond what is available if we focus on the nominalistic content of each explanation. While nominalistic versions of each explanation are available that succeed in showing that the nominalistically characterizable features of the particular systems in question sufficed to guarantee that the observed phenomenon would occur, the mathematical explanations serve to add another explanatory dimension, the *ability to unify* a range of what at first glance may seem like different phenomena. This additional dimension, I would like to suggest, is a structural one: the mathematical explanations show in each case that the explanandum occurred *as a consequence of structural features of the physical system* that can be characterised mathematically. As the same theorem involving the same mathematical structure is involved in each case, the topic generality of mathematical explanations allows us to see each of these disparate phenomena as a consequence of one and the same structural feature[Footnote1].

The work done by the mathematical framing in the typical examples of candidate mathematical explanations of physical phenomena, both in *providing modal information* about the explanandum and offering possibilities of *unification* of the phenomenon to be explained with apparently disparate phenomena supports our understanding of those phenomena in such a way that suggests to me at least that it is worthy of being called *explanatory*.

In what follows, I will accept that examples such as the number theoretic explanation of cicada behaviour and the graph theoretic explanation of the impossibility of completing a Eulerian walk through Königsberg are genuine mathematical explanations of physical phenomena. I will offer an account of how these explanations work and argue that, if they do work in this way, our use of these explanations in empirical science does not commit us to mathematical platonism. Some readers may remain unconvinced, however, that the virtues I have pointed to of these so-called mathematical 'explanations' (i.e., providing modal information and unifying apparently disparate phenomena) suffice to show that these uses of mathematics are genuinely *explanatory*. But even if we agreed not to use the 'e' world to describe them, to the extent that we think that these theoretical virtues are virtues worth having, there remains a question of how mathematics can serve these functions (of providing modal information and theoretical unification), and whether using mathematics for these purposes presupposes platonism. For the reader who is not convinced my use of the 'e' word to talk about these examples, I hope the structural account I offer of how mathematics works to provide modal information and possibilities of unification will still be of interest in showing that if we wish to use mathematical theories for these purposes, doing so will not commit us to the existence of mathematical objects.

MATHEMATICAL EXPLANATIONS AS STRUCTURAL EXPLANATIONS

I propose that the mathematically couched explanations of cicada behaviour and of the impossibility of Eulerian walks through Königsberg, along with other examples that have been offered in the literature on mathematical explanation are examples of what I, following Bolulich (2008) (who herself follows Peter Railton (1980) & Hughes (1989)) will call *structural explanations*. Structural explanations explain by showing an empirical phenomenon to be a consequence of the mathematical structure of the empirical situation. According to Bokulich (2008, p. 149),

- a structural explanation is one in which the explanandum is explained by showing how the (typically mathematical) structure of the theory itself limits what sorts of objects, properties, states, or behaviors are admissible within the framework of that theory, and then showing that the explanandum is in fact a consequence of that structure.

But how does one show that an explanandum is a consequence of the mathematical structure of a theory? In answering this question, as my own interest is specifically in the status of mathematical hypotheses in structural explanations, rather than following Bokulich's discussion of this matter directly, I will focus more closely on how mathematical theories characterize structures that can be used in structural explanations, rather than discussing mathematically-structured empirical theories, which is where Bokulich's own attention lies.

In order to see what could be meant by empirical phenomena being consequences of the *mathematical structure* of an empirical set up, it will be helpful to consider an understanding of mathematical theories that is common to most forms of structuralism in the philosophy of mathematics. Consider a pure mathematical theory, presented axiomatically. These axioms will typically include logical terminology and some primitive terms. For example, in the (2nd order) Peano axioms for number theory, we have primitive terms 'zero (0)', 'number (N)' and 'successor (s)', where '0' is a singular term, 'Nx' a unary predicate, and 's(x)' a unary function. The axioms can be expressed as follows:

- $N(0)$ ('zero is a number').
- $\forall x(Nx \supset Ns(x))$ ('The successor of every number is a number').
- $(\forall x)(Nx \supset s(x) \neq 0)$ ('Zero is not the successor of any number').
- $(\forall x)(\forall y)((Nx \ \& \ Ny) \supset (x \neq y \supset s(x) \neq s(y)))$. ('distinct numbers have distinct successors').
- $(\forall F)(\forall x)((F0 \ \& \ (\forall x)(Nx \supset (Fx \supset Fs(x)))) \supset (\forall x)(Nx \supset Fx))$ ('If any property is such that it applies to 0 and, if it applies to a number it also applies to that number's successor, than that property applies to all numbers.').

We can abbreviate the conjunction of these axioms as PA $\langle 0, N, s \rangle$ (indicating the primitive terms).

A question arises of what we should make of the primitive terminology in such axiomatizations. There are two basic approaches on the table. An 'assertory' understanding of an axiom system sees its primitive terms as independently meaningful, and aiming to pick out some specific objects, predicates, and functions. The axioms are then attempts to assert basic truths about these independently meaningful primitives. On the other hand, an 'algebraic' understanding sees the primitive terminology as not having a meaning independently of the axiom system in which they occur, and (much like the ‘unknowns’ in a system of equations with various unknowns) as being given their meaning contextually by the axioms themselves[Footnote2]. Clearly an algebraic understanding is appropriate for *some* systems of axioms: the axioms for group theory, for example, can be thought of as defining what would have to be true of any collection, G, of objects with binary operator + and distinguished element 0 in order to count as a group. There is no specific intended interpretation of ‘G’, ‘ + ’ and ‘0’ about which the axioms aim to assert truths. What the leading versions of mathematical structuralism (such as Stewart Shapiro’s ante rem structuralism (Shapiro, 1997) and Geoffrey Hellman’s modal structuralism (Hellman, 1989)) have in common is that they assume an algebraic understanding of all axiomatic theories[Footnote3].

The correctness of mathematical structuralism as a picture of pure mathematics is not what is at issue here; my interest is only in the ‘algebraic’ approach to axiom systems assumed by structuralists. What is important about the algebraic understanding of mathematical theories in this context is the sense it allows us to make of the notion of mathematical structure, and in particular of the notions of a system of objects instantiating a mathematical structure, and of truths that are ‘true in virtue of’ that structure. For a particular system to instantiate an axiomatically characterized mathematical structure is simply for the axioms characterizing the structure to be true when their primitive constants, predicates, and function symbols are given an appropriate interpretation in the terms of that system. We can, for example, find particular mathematical systems instantiating axiomatically characterized mathematical structures: the natural number structure has an instantiation in the sets if we interpret 0 as ∅, s(x) as the function that takes a set A to its singleton {A}, and the predicate Nx as being true of a set A if and only if A is in the intersection of all sets containing ∅ and closed under the operation of taking successors (i.e., if and only if A ∈ {∅, {∅}, {{∅}}, …}. But we can also find ‘concrete’ systems instantiating some mathematical structures: the group axioms, for example, can be interpreted as truths about the simple system consisting of symmetric rotations of a square. Here, G is the collection of possible rotations: (id = keep as is; r_1 = rotate 90° clockwise, r_2 = rotate 180°, r_3 = rotate 270° clockwise). Of these, 0 is interpreted as the ‘id’ rotation, and the binary + operation is the result of performing two operations consecutively (so that, e.g., $r_1 + r_2 = r_3$). And, to take us back to Baker’s cicada example, if we idealize somewhat to forget about the eventual demise of the earth, the series of earth-years starting from a given 0 in which cicadas appear and continuing without end can be viewed as an instantiation of the natural number axioms, with the function ‘s(x)’ being interpreted as ‘the year following year x’.

Consider now a system of objects and relations (mathematical or physical) that instantiates an axiomatically characterised structure. There will of course be a range of truths about that given system. For example, in our set theoretic system instantiating the Peano axioms, when supplemented with definitions of the individual numbers and the ‘less-than’ relation ‘<’), it will be true that the object it calls ‘2’ (ss0, i.e. in this case, $\{\{\emptyset\}\}$ will be a member of the object it calls ‘3’ (sss0, i.e., $\{\{\{\emptyset\}\}\}$), and it will also be true that $2<3$. But only the latter of these is, I claim, true *in virtue of* the axiomatically characterised structure provided by PA $\langle 0, N, s\rangle$. The axiomatic setting helps us to understand this difference. When we supplement the axioms with the appropriate definitions, ‘$2<3$’ is a logical consequence of PA $\langle 0, N, s\rangle$ (and thus true in *all* interpretations of these axioms), whereas ‘$2\in 3$’ is not a logical consequence of the structure-characterising axioms. In general, if structurally characterized axioms are true when interpreted as about a particular system, then we can say that a truth about that system is true *in virtue of* the mathematical structure characterized by those axioms when it is an interpretation of a claim that follows logically from those axioms.

We can now make sense of the notion of a structural explanation to which I wish to draw attention, in cases where the structure in question is characterized by mathematical axioms. Such a structural explanation explains by showing (a) that the system to be explained can be viewed as an instance of a mathematical structure, and (b) showing that the explanandum is true in virtue of that structure, i.e., that it is a consequence of the characterizing axioms and relevant definitions (when suitably interpreted). As such, we can think of the general form of a structural explanation (involving axiomatically characterised mathematical structure) as an explanatory argument as follows:

[Mathematical Premise, MP] Mathematical theorem, modal-structurally characterised (i.e., of the form, ‘necessarily, in any system satisfying <Axioms>, <Theorem>’)[Footnote4].

[Empirical Premises, EP] Empirical claims justifying the claim that <Axioms> are true when interpreted as about the physical system under consideration.

Therefore.

[Explanandum] <Theorem> is true of the system under consideration.

Of course, when it comes to real-life examples of mathematical explanations of physical phenomena, some work may be needed in order to discover this general form in the explanations as provided. We will see below how it might work in some particular cases.

Before we move to examples, however, it is worth noting up front a feature of structural explanations thus characterised, that may ring alarm bells given the history of accounts of explanation in the philosophy of science. By couching the general form of structural explanations as *deductive, explanatory arguments* that invoke a necessitated conditional (in their modal-structurally characterised mathematical premise) this account makes structural explanations look suspiciously close to covering law explanations, in this case a deductive-nomological explanation where a mathematical 'law' takes the place of an empirical one, providing 'nomic expectability' to the conclusion of the argument as explanandum. Despite the well-trodden concerns about the covering law model as a general account of explanation, here I will embrace this similarity. I note some relevant differences: (1) the restriction of our 'law' to a modal-structurally characterised mathematical theorem avoids some difficulties concerning the kinds of generalisations that can be cited as 'laws' in such explanations; (2) the requirement that the empirical premises serve to justify the claim that the axioms applied in the mathematical premises apply to the physical system under consideration avoids some concerns about permitting arguments with irrelevant premises to count as explanations. However, one feature that my account does share with the D–N model is its tolerance for explanatory symmetries: given that (in the famous 'flagpole' case) it is equally a theorem that if the length of the shadow is x, the height of the flagpole is y *and* that, if the height of the flagpole is y, the length of the shadow is x, we can just as well use information about the length of the shadow along with structurally interpreted mathematical results to provide a structural explanation of the height of the flagpole as we can use information about the height of the flagpole to explain the length of the shadow. I do not have the space for a full discussion of this case here, so I will simply note that this is a bullet that I am willing to bite[Footnote5].

Examples

A simple example of a structural explanation is provided if one considers a rather mundane puzzle about the difference between mattress flipping and tyre rotating, discussed by Brian Hayes (2005) in a popular article in *American Scientist* (2005). We are advised to flip/rotate double-sided mattresses periodically in order to ensure even wear. There are four possible ways of fitting a mattress into a standard rectangular bed, and by flipping/rotating mattresses periodically, our aim is to cycle through these four configurations so that over its lifetime the mattress gets equal use in each position. Similarly, it was once considered good practice to move the tyres on a car around, so as to even out wear on the tyres, and it is prudent, if one is doing this, to ensure that the tyres are moved around evenly so that no one tyre spends too long in the same position. There is (what Hayes calls) a simple 'golden rule' for moving tyres: a single operation that one can do each time one moves the tyres to ensure that, after enough applications, each tyre will have occupied each of the four positions it can take exactly once, so that wear is even. If one simply rotates the tyres around the car a single turn at each change (making an arbitrary choice at the start of whether to move clockwise or anticlockwise), one can be confident that, after applying this same operation repeatedly all tyres will have occupied all positions. This means that we do not need to remember how we positioned the tyres on previous occasions. If we simply resolve always to move a single turn clockwise (say), we will ensure even wear without having to keep track of previous positions.

If we think of the three main symmetric operations one can do to 'flip' a mattress (i.e., rotate 180° across each of the three orthogonal axes through its centre point), clearly no single one of these on its own would provide us with a golden rule for mattress turning: if we were always to rotate around its short vertical axis, for example, we would find ourselves always sleeping on the same side of the mattress, with the head and foot flipping each time. Hayes wondered, though, whether there was a single *combination* of these operations which, if one cycled through that combination at each 'flipping', would ensure that, over a period of time, the mattress would take all possible configurations, thus avoiding the problem of remembering how the mattress was configured on previous configurations before choosing which operation to take next time.

A quick internet search of mattress flipping advice suggests that no such solution has been found: the best advice available, Hayes tells us, seems to be to practice seasonal flipping, e.g. flipping across the short horizontal axis for one season and then across the long horizontal axis the next, to cycle through the four possible mattress positions over a year. Why is this? Why isn't there a single combination of moves, which if one repeats that same combination at each flipping would ensure that all positions of the mattress would be taken? And why does the mattress case differ from the superficially similar problem of tyre rotation? The answers can be seen once one realizes that the rotations of mattress form a 4-group. Since groups are closed, any combination of any number rotations is equivalent to a single rotation. And since (as we have already noted) no single rotation provides a golden rule, no combination of these will either. Why are things different in the tyre case? Because, though both are 4-groups, the tyre group is (an instance of) the cyclic group of order 4 (the group of 2-dimensional rotations of the square that we saw earlier), whereas the mattress rotations instantiate the Klein 4-group, which contains no operation which, when repeatedly applied to itself, cycle through all the four operations in the group.[Footnote6] So the fact that the two rotation sets are instances of two different groups explains Hayes's explanandum: why is there a golden rule for tyres but not for mattresses?

Couched in our general terms, we can present this explanation of a contrast as involving two separate structural explanatory arguments, one showing that there is single move which, repeated, will cycle through all of the possible mattress positions, and another showing that there is a single move which, repeated, will cycle through all of the possible tyre positions. If we bundle the definitions of the Klein 4-group and the cyclic group of order 4 into our 'structure characterising' axioms respectively, one argument will take the form:

[MP] Necessarily, if 〈0, a, b, ab〉 is a Klein 4-group, then there is no element in 〈0, a, b, ab〉 whose repeated application will cycle through all the members of the group.

[EP], when '0' is interpreted as no movement, 'a' as flipping across the vertical axis, 'b' as flipping across the short horizontal axis, and 'ab' as flipping across the long horizontal axis, these rotations of a mattress form a Klein 4-group.

Therefore.

[Explanandum]: there is no mattress rotation whose repeated application will cycle through all the possible positions of the mattress.

Similarly, the 'tyres' argument will use the empirical premise that, when '0' is interpreted as no movement, 'a' as moving all tyres one space clockwise, a^2 as moving all tyres two spaces, and a^3 as moving all tyres 3 spaces clockwise (or one space anti-clockwise), these tyre rotations form a cyclic group of order 4, to conclude that there is a golden rule for tyre rotation.

To move from this 'toy' example to some of the examples of mathematical explanations of physical phenomena mentioned already, the famous bridges of Königsberg explanation (as discussed by Pincock, 2007) is relatively easily put in this form, as follows:

[MP] Necessarily, if 〈Nodes, Edges〉 is an instance of a connected graph, then if 〈Nodes, Edges〉 permits a Eulerian walk it must have either zero or two nodes with an odd number of edges.

[EP] When 'Nodes' is interpreted as 'landmasses' and 'edges' is interpreted as 'bridges', Königsberg in 1735 is an instance of a connected graph with four nodes with an odd number of edges.

Therefore.

[Explanandum] Königsberg in 1735 does not permit a Eulerian walk.

Finally, as mentioned before, while Baker's cicada example requires something of an idealization to fit straightforwardly this model (assuming that the sequence of years in which cicadas appear has no end)[Footnote7], having made that idealization we can we can sketch the explanation in rough terms as follows (following Baker in adding a biological premise to fill in the evolutionary constraints):

[MP] Necessarily, if PA 〈0, N, s〉, arithmetic progressions of length n and m that have the same first member overlap minimally when n and m are coprime, and a number m is coprime with all numbers n <2m iff m is prime.

[EP] When '0' is interpreted as some first year in which two broods of periodical magicicada cicadas appear together, 'N' as interpreted as the collection of years including and following that first year, and 's' is interpreted as 'the year after', PA 〈0, N, s〉 hold, and the sequence of years in which a magicicadas with period length m occur form an arithmetic progression of length m.

[Biological premise] It is advantageous for cicadas to choose periods which minimize overlap with periods of other periodical creatures.

Therefore:

[Explanandum] Prime number periods are advantageous for cicadas.

Three questions arise. First of all, do so-called structural explanations deserve to be viewed as genuine explanations? Second, does the use of structure-characterizing mathematical axioms in these explanations commit us to assigning an explanatory role to abstract mathematical objects (or indeed, abstract mathematical structures)? And finally, can all or even most purported examples of mathematical explanations of empirical phenomena be accounted for as structural explanations? My answer to these questions, in brief, are: yes; no; and not quite. I will consider the first two of these questions here, before turning to the third in Sect. 3.

Are Structural 'Explanations' Genuine Explanations?

Structural explanations certainly provide answers to the kind of 'why' questions that we ask in demanding explanations of phenomena. Why do the cicadas have the period length they have? Because the sequence of successive years is an instance of a natural number structure; the sequence of years in which the cicadas appear is an instance of an arithmetic progression within that structure; in any natural number structure, arithmetic progressions with4 prime differences between terms will overlap minimally with other progressions; and non-overlapping periods are advantageous. Why is there no golden rule for mattress flipping when there is one for tyre rotation? Because the mattress operations are an instance of the Klein 4-group; in any Klein 4-group, no one element can be repeatedly applied to itself to cycle through all four operations; and a golden rule would require there to be such a cycle. Perhaps this in itself is enough to present these purported explanations as genuinely explanatory. But for those looking for something more, note that as I have presented these explanations, the explanans in each case involves appeal to a general (structural) law (whose modal status as a logically necessary truth is supported by the fact that the consequent is derivable from the antecedent): "necessarily, in any natural number structure, arithmetic progressions with prime differences between terms will overlap minimally with other progressions"; "necessarily, in any Klein 4-group, no one element can be repeatedly applied to cycle through all four operations". By deriving observed phenomenon from premises that include a modal-structural law (a claim about what must be true in all structures of a given sort), I have already noted that these structural explanations share important features of DN-explanations—they explain by providing what Salmon (1989, p. 57) calls "nomic expectability—the expectability on the basis of lawful connections".

Another reason to think of structural explanations as genuinely explanatory is that they meet the criteria required for 'distinctively mathematical explanations' outlined in Marc Lange's recent defence of non-causal explanations. Lange (2016, pp. 5–6) holds that what he calls 'distinctively mathematical explanations' work.

> by showing how the fact to be explained could not have been otherwise—indeed, was inevitable to a stronger degree than could result from the action of causal powers.

The modal elements of structural explanations as I have characterised them help to show how these explanations establish that, given the structural features of instantiated in the target system, the explanandum of a structural explanation could not have been otherwise. Modality occurs in these explanations at two points. First is in the modal-structural mathematical premise. Here the necessity at work is logical necessity, and our justification for taking the MPs as true is the existence of a derivation of the consequent from the antecedent. The second modal element in these explanations is the fact that they are deductively valid arguments: the inevitability of the explanandum *given the premises* is established through showing that it is a consequence of those premises. While this second modal element justifies the claim that these explanations work by showing how the fact to be explained could not have been otherwise, it is the first modal element that meets Lange's criterion for these explanations to count as 'distinctively mathematical'. By showing that their explananda follow from logically necessary truths about what holds in *any* structure satisfying certain structure-characterising axioms, these explanations display by their form that the inevitability of the explanandum is stronger than causal.

Finally, to the extent that unification is a form of explanation, by displaying physical phenomena as consequences of the mathematical structure instantiated in a physical system, it is easy to see how structural explanations can serve to unify. Structural explanations involve a mathematical premise which is a modal-structurally characterised mathematical theorem, as well as an empirical premise containing information to establish that the physical system under consideration is such as to satisfy the antecedent of the modal-structural theorem in the mathematical premise. Structural explanations of this sort can unify apparently disparate phenomena when it can be shown that the structural explanations of those phenomena appeal to the very same mathematical result.

Ontological Commitments of Structural Explanations

As mentioned above, in the recent debate over platonism and anti-platonism in the philosophy of mathematics, the existence of genuine mathematical explanations of physical phenomena has been held to support mathematical platonism. And my talk above of 'instantiation of a mathematical structure' in a physical system might suggest that structural explanations as I have characterised them are no less committed to a form of platonism, in the form of realism about the mathematical structures instantiated. However, the modal-structural characterisation of the mathematical premises of structural explanations shows that an inference from the existence of mathematical-structural explanations of physical phenomena to platonism is not warranted. In fact, structural explanations as I have characterised them (where a mathematical structure is shown to be instantiated in an empirical system, so that truths about that system can be displayed as holding *in virtue of* that structure) require no specifically mathematical ontology.

Although mathematical theories (such as, in the examples we have considered, number theory, group theory, and graph theory) are used in structural explanations of physical phenomena, we are not required to assume that the axioms of such theories are true of a realm of abstract mathematical objects. Rather, as indicated by the modal-structural formulation of the MPs in our examples, we may simply view the pure mathematical theory that is involved in the explanation as telling us what *would have to be* true, *were* there a system instantiating the structure characterized by the axioms, something that we can discover simply by inquiring into the consequences of the axioms of the theory. Having shown mathematically that any system exhibiting a given structure has a particular feature (e.g., that in any instance of the Klein 4-group, no element can be repeatedly applied to itself to cycle through all four elements of the group), we can transfer this information to the concrete instantiation we have found. Structural explanations of this sort may make essential use of mathematical theories to explain empirical phenomena, but such essential use does not require us to posit the existence of a special realm of mathematical objects about which these theories assert truths, only that such a theory is, when appropriately interpreted, true of the concrete system whose behaviour we are trying to explain.

There are, of course, modal commitments incurred in viewing these explanations as involving claims about what would have to be true, were any system to instantiate the axioms. As I have said, the necessity at hand here is logical necessity (where, 'necessarily, if 〈Axioms〉 then 〈Theorem〉' holds if and only if 〈Theorem〉 is a logical consequence of 〈Axioms〉). So the nominalist who wishes to adopt this account of explanation and hold that the structural laws involved in these explanations are literally true will have to commit to the truth of some logical necessities (or, equivalently, to the truth of some claims about what follows from our axioms). But these modal commitments are no more than are already incurred in the leading fictionalist accounts of mathematics. Hartry Field, for example, is clear in *Science* *without Numbers* (1980) and papers following (including Field (1984), Field (1989) about the requirement to include primitive modal operators in his fictionalist account of mathematics, endorsing the considerations in favour of primitive modality outlined by Georg Kreisel (1967). Likewise I defend the use of such operators in my own version of 'easy road' nominalism in Leng (2007) arguing that attempts to reduce modal claims to truths about set theoretic models fail since it is modal facts themselves that determine whether a set theoretic reduction is adequate. There will of course be those who will worry that the nominalist appeal to modality is problematic—perhaps because we often make use of the mathematical machinery of model theory to discover modal truths, or because they think that modal truths *just are* truths about set theoretic models. However, to the extent that these worries about the modal commitments of fictionalism arise, they arise already independently of the issue of the use of modal truths (about what follows from structure-characterizing axioms) in structural explanations. So this modal element of the mathematical explanations we have considered raises no new problem for mathematical fictionalists.

FROM STRUCTURAL EXPLANATIONS TO STRUCTURAL MODEL EXPLANATIONS

Are most, or even many, mathematical explanations of physical phenomena best understood as structural explanations, explaining by showing that their target system is an instance of a mathematical structure? I have argued that the cicada explanation can be understood as a structural explanation, where the axioms of number theory are interpreted as truths about the system of years consisting of some initial years in which cicadas appear, with the 'successor' relation being the 'the following year' relation. And I have presented a very simple example of a structural explanation involving group theory and a relevant difference between features of the cyclic group of order 4 and the Klein 4-group. In the first of these examples we had to introduce an element of idealization to allow the axioms to be interpreted as truths: we had to assume that the sequence of Earth years continued without end. (This idealization though false, was innocuous enough given that it was only behaviour at a finite initial segment that was needed for the explanation.) The second example required no idealization, but was admittedly rather simple, as is the explanation in the example given of the Königsberg bridges. It is difficult to find many serious examples of genuine structural explanations of this sort in empirical science (though group theory is a powerful tool in chemistry when applied to symmetries of molecules). While for some finite mathematical structures we can find physical instantiations, these structures are often so simple that any empirical phenomena we might try to explain with reference to the structure will be independently obvious already (as, arguably, is the case with the mattress-flipping explanation). And even for finite mathematical structures, these may be more clearly instantiated in idealized models of empirical phenomena rather than directly. Furthermore, most mathematical structures are not finite, and so may not have a physical instantiation (or are at best approximately instantiated as in the case of modelling years using the natural numbers or, as, for example, when we consider localized physical space to be an instantiation of Euclidean geometry). While our scientific theories are mathematical to the core, where complex mathematical structures, and sophisticated, genuinely informative explanations, are involved in these theories it is generally the case that much work needs to be done to fit the mathematics to physical reality. Simple instantiation of a structure is rare. More often a process of modelling must occur in order to bring the phenomena into contact with mathematical theory. Thus, most interesting structural explanations in mathematics will take the form of what, again following Bokulich (2008, p. 147), I will call *structural model explanations*, explanations where a mathematical structure is instantiated not in a physical system but in an idealized model of that system.

Mathematical Explanations as Structural Model Explanations

The most basic form of a structural model explanation explains by hypothesizing a model instantiating the structure of a given mathematical theory, showing that some facts about that model are true in virtue of that structure, and then relating that model to some empirical phenomenon to be explained by means of the model. The 'modelling' relation is as ever a complex one: it may involve resemblance, approximation, or perhaps more formal mappings (isomorphisms, homomorphisms), and will very often involve viewing the phenomenon modelled as related not to the whole structure in the model, but to some smaller substructure embedded in that model. The modelling relation may be described formally by means of a partial structures approach, as developed, e.g., by Bueno, French and Ladyman (see, e.g., French (2000), Bueno et al.(2002)), though we may find that the ultimate tie of model to modelled will be looser than can be formally characterized by such a theory. Indeed, it is likely that the formal framework of partial structures may only apply after a degree of modelling and idealization has already occurred, so as to exhibit relations between various models of increasing abstraction, rather than model and reality itself, which may be tied by fundamentally informal links (a possibility acknowledged by defenders of the partial structures view). Pincock (2012) provides a book-length study of the use of mathematical models to represent physical phenomena. Rather than attempt a full discussion of this complex issue how idealized models represent, though, let us simply assume that where models are aptly chosen, we can draw inferences about the real systems they represent. In such cases, we may ask, how can we transfer explanatory considerations from model to system modelled?

Given an axiomatically characterized mathematical theory, then, we can imagine that that theory is instantiated in some model system. Suppose that this model system has a subsystem that is held to be 'apt'—to be appropriately related by some informal 'representation' relation to an empirical system we wish to investigate (often, this subsystem will be the model system itself). The relation between the subsystem of our mathematical model and the empirical system under investigation may of course be mediated by further models—e.g., we may first need to abstract an idealized system from the empirical system we wish to consider, and then show that this idealized system bears appropriate structural relations to the subsystem of our original mathematical theory. But without going into the complexities of how such a relation may be established (and therefore of how it is the subsystem in the model is held to aptly represent the system modelled), let us simply assume for now that the subsystem of our model that instantiates a mathematical structure is indeed held to be true (enough) to the empirical system under investigation. Suppose now that we show some properties to be true in this subsystem of our mathematical model simply in virtue of the structure it instantiates. And suppose that these properties of the subsystem are seen to correspond (via our loose modelling relation) to empirical properties observed in the empirical system under investigation. Then an explanation of these empirical properties may be that they hold of the empirical system in virtue of its mathematical structure: given that the empirical system is well modelled by a subsystem of the larger mathematical structure, the empirical phenomena observed were to be expected, as (interpreted) consequences of the axioms characterizing that structure.

An example will help us to understand the processes at work in this loose sketch. Given that there is wind at all, at any point in time there will be at least one point on the earth's surface with no wind. Why is this? The 'Hairy Ball Theorem' from topology provides an explanation. In order to give this explanation we must start by preparing the empirical phenomena (wind patterns) for mathematical description. This involves some idealization. First, we forget the fact that the layer of air in the atmosphere above the earth's surface has depth, and that wind movements are different at different depths. Instead we think of a single layer of air on the earth's surface, with no depth. We then need to think about the direction and strength of wind at various points on the earth's surface. Over a large scale, we can measure prevailing wind direction at a point by a weather vane centred on that point—this gives us a horizontal direction of the wind as a 2-dimensional tangent to the earth's surface at that point (e.g., as coming from a North North Easterly direction). We can also measure its speed/strength at that point by means of a spinning anemometer (measuring the number of rotations of spinning cups in a given time period). To each point where measuring instruments are located, our measurements therefore allow us to associate a direction and a magnitude: or, in mathematical terms, a vector (measured using an appropriate measurement scale), where the direction of each vector is always at a tangent to the earth's surface. Extrapolating this to the small scale, we can think of wind direction as being defined at each point on the earth's surface by a tangent vector. Furthermore, we can assume that changes to the direction and magnitude of these vectors as we move across the earth's surface are continuous. Thus we can think of the essential features of the wind as being represented by a continuous function corresponding points on the earth's surface to vectors.

This 'prepared description' (to use Nancy Cartwright's terminology (Cartwright, 1983, p. 15)) of wind behaviour, achieved by a number of idealizations as well as by applying a mathematical measurement scale[Footnote8], enables us to see the phenomenon of wind movement in a wider mathematical perspective. In particular, we have described the wind as a tangential vector field on the surface of the Earth. If we now take that surface to be topologically equivalent to a sphere (ignoring inconvenient tunnels) then we can apply a theorem of topology to the wind system we have described. According to the 'Hairy Ball Theorem' of topology, "there does not exist an everywhere nonzero tangent vector field on the 2-sphere $\mathbf{S}^2$" (Weisstein, web resource). So given that there is any wind at all on the earth's surface (so that the vector field in our model is not everywhere zero), there must be some point on the earth's surface where the value of the tangent vector representing the wind speed and direction is zero (i.e., there is no wind at that point). What happens in the area around a point with zero wind? Well, wind cannot flow in or out of that point as we are hypothesizing that wind speed and direction is zero there, so while there may still be zero wind at adjacent points, when the wind in the vicinity of this zero point is non-zero once more it must initially spiral around the no-wind area, as in a cyclone. To envisage what is going on we may imagine the hairy ball of the theorem's title. What the theorem tells us is that one cannot continuously comb a hairy ball flat—the best we can do is to create a cowlick with a hair sticking straight up in the middle and adjacent hairs circling around. (The flat hairs represent non-zero tangent vectors; the hair that remains sticking up would of course no longer be a tangent vector, so in order for the combing to remain a continuous tangential vector field it must be zeroed.)

We have, therefore, a mathematical explanation of an empirical phenomenon (the inevitable occurrence of certain wind patterns)[Footnote9]. The explanation is a model explanation: we do not apply the mathematics directly to the empirical phenomenon, but first prepare the phenomenon for mathematical description through a process of idealization and abstraction. Our prepared description of wind on the earth's surface as a tangent vector field enables us to apply the resources of topology to this description and so to derive a conclusion about the properties of this vector field. And returning to the original phenomenon modelled, we are able to give a physical interpretation of this conclusion, stating what this should means we should expect about actual wind behaviour. This explains actual wind behaviour structurally, to the extent that it is shown to be a consequence of the mathematical structure of the physical system, so the explanation is a structural model explanation.

Must a structural model explanation to be true in order to explain? In Bokulich's discussion of such explanations, she focuses on structural explanations where the model is a classical system and the phenomena to be explained are quantum phenomena, so that the mathematical structure appealed to in the structural explanation, using the mathematical of classical mechanics rather than quantum mechanics, is not straightforwardly 'true of' the system to be modelled. But whether, in these proposed structural model explanations, the models are close enough in structural terms to the phenomena modelled to be genuinely explanatory is something that need not immediately concern us here. In all model explanations, the issue of how close a model must relate to the phenomenon modelled (and therefore of how 'true' the model is to the phenomenon it models—or, we may say, how 'apt' the model is) in order to be explanatory is a complex and contentious matter, but this is not the sense of truth that matters for the purposes of the metaphysical question of what status should be given to models in model explanations. What we need to ask is, in cases where we do think that the model in a structural model explanation is *close enough* to the system modelled in relevant respects to be genuinely explanatory, must we accept the existence of the model system itself in order to accept this explanation as genuine? What is the metaphysical status of the idealized models appealed to in structural model explanations, and in particular, does commitment to the existence of mathematical explanations as structural model explanations incur an undesirable commitment to abstracta?

There is a strong tradition in discussions of models in the philosophy of science of thinking of models as merely imaginary objects. For example, Peter Godrey-Smith (2009, p. 102) characterises "model-based" science as.

- a style of theoretical work in which an imaginary system is introduced and investigated—an imaginary population, ecology, neural network, stock market, or society. The behavior of the imaginary system is explored, and this is used as the basis for an understanding of more complex real-world systems.

This suggests a picture of model building as analogous with storytelling: although we appear to speak as if the objects in our models really exist, we are actually just telling a story that fleshes out the supposition that there are such things, without commitment to the truth of that supposition. (Similarly, with theorizing in the context of a pure mathematical theory, we can view theorists as working out the consequences of the supposition that the axioms of that theory are true, without any commitment to the actual truth of those axioms.) The literal truth of statements uttered in describing the models used in model explanations may not be required for those explanations to be explanatory if all we are doing by uttering those statements is elaborating on what *would* be the case *were* there objects of the sort described.

I do not have the space here develop a nominalist understanding of ideal models in empirical science. In Leng (2010) I endorse an account of models as representations that builds on Kendall Walton's (1990) account of representation as "prop oriented make-believe", an account that has been developed further in the modelling literature in various ways e.g. by Frigg (2010), Toon (2010) and Salis (2019). To the extent that a fictionalist account of models in science can be defended, the idealised models in 'structural model explanations' do not need to exist in order to be utilised in explaining physical phenomenon.

What I would like to suggest, then, is that whether we really *believe* that there are abstract systems of ideal objects instantiating an axiomatically characterized mathematical structure and appropriately related to a system of physical objects, or merely *pretend* that there are such things, makes no difference to our ability to exploit the 'framing effect' of a structural model explanation in enabling us to see empirical systems as essentially mathematically structured. In mathematical theorizing we discover what would be true of any system instantiating the axioms (including in any subsystem of such a system), and in describing a model (whether real or imagined) that allows us to see a physical system as structurally related to a (real or imagined) subsystem of a system instantiating a mathematical structure, we are able to conclude that certain empirical interpretations of our mathematical results ought to be true of the physical system under discussion simply by virtue of its mathematical structure. The framing effect of seeing wind patterns on the earth's surface as modelled by a tangent vector field no more requires the existence of the mathematical system one imagines than does the framing effect of seeing those same wind patterns as modelled by a hairy ball that one is trying to comb flat requires the real existence of said ball. In either case the appropriateness of the imagined model allows us to frame facts about the system modelled as holding in virtue of its sharing the structure of the model system. And in either case, the explanatory work done by the model is that it shows us that those facts were to be expected given the structure of the situation modelled.

Explanatory Models, or Explanatory Structures?

The introduction of idealized models that represent actual physical phenomenon into our account of explanation may introduce a new wrinkle, however, as it has sometimes been suggested that showing that phenomenon to be explained holds in a structurally similar model of a target system cannot suffice to explain that phenomenon. For example, James R. Brown argues that it is *because* mathematics generally finds application via enabling us to form tractable models of physical phenomena that mathematics cannot play a genuine explanatory role:

- Mathematics hooks on to the world by providing representations in the form of structurally similar models. The fact that it works this way means that it cannot explain physical facts, except in some derivative sense that is far removed from the doctrines of explanation employed in indispensability arguments. (Brown, 2012, p. 8)

In Brown's view, the role that structural models have in our theories is to provide representations of physical systems. These may aid understanding, by being easier to work with than the systems themselves (and allowing us to use familiar descriptive tools). But providing this kind of aide to understanding is a *derivative sense* of explaining, and does not amount to mathematics playing a genuine explanatory role. Despite Brown's own platonism (which he holds on independent grounds), Brown's view of the role of mathematics in explanations is thus in line with those nominalists who hold that all the genuine explanatory work in so-called mathematical 'explanations' of empirical phenomena resides in the nominalistic content that the mathematics helps us to grasp.

By separating out the structural elements of structural model explanations from the model elements, we can see where Brown and others go wrong in this regard. Brown is absolutely right that, to the extent that the role played by mathematics is to provide tractable *models* of empirical phenomena, the ease of understanding that results from having such tractable models is only 'explanatory' in a derivative sense—the models may make the ultimately nominalistic features of the systems easier for us to grasp, but they are not playing an explanatory role in showing us why the phenomena we observe *had* to be true. However, focus on mathematics as providing models diverts attention from the structural features of these models, which is where their explanatory work resides. What makes a structural explanation explanatory is not just that it displays some ultimately nonmathematical content to be true, but rather that it displays that content to be true *in virtue of* the mathematical structure of the empirical system under investigation: it shows it to be a consequence of mathematical axioms that are true under an empirical interpretation. And what makes a structural *model* explanation explanatory is again not (just) that it displays some nonmathematical content to be true, but that it shows it to be true in virtue of the mathematical structure of the situation to be explained, by relating that situation in an appropriate manner to a model that instantiates (or is a subsystem of a system that instantiates) a given mathematical structure. Structural explanations of either sort show why the observed phenomenon had to happen or was to be expected given the mathematical structure of the empirical system under study. The nominalistic content of these explanations, on the other hand, do no such thing: the insight the mathematical structure provides is lost if we simply focus on the content of the true descriptive claim that the empirical situation is such as to make the models used in these explanations appropriate.

The added explanatory work done by representing empirical phenomena as essentially mathematically structured can make sense of a complaint that Otávio Bueno and Mark Colyvan have expressed about the limitations of 'mapping accounts' of the application of mathematics (such as that of Pincock (2004) and suggested in Leng (2005)). Mapping accounts try to explain the applicability of mathematics by noting that our mathematical theories and the physical world to which they apply are related by structural similarity relations, much like a map is related to a city. It is not surprising that we can learn things about a city from studying its map, given the structural similarity relation holding between the map and the city, and it is also not surprising that it is helpful for us in discussing the spatial arrangements of objects in a city in terms of the map rather than the city itself—it provides a useful simplification that can make navigation problems tractable and allow us to ignore the mass of irrelevant detail. But, Bueno and Colyvan (2011) note, it would be odd to think that the map of the city could by itself *explain* facts about the city (unless, perhaps, we discover that the map was the blueprint from which the city was built). More needs to be said in mapping accounts of applications to show how the mathematical theories we claim to be structurally related to the physical world can *explain* features of that world.

The difference between a mathematical theory and a road map in explanatory uses of mathematics is that, while both are models (and hence both are structurally similar to the reality they represent), only the mathematical theory involves a *structural model*, in the sense of representing the physical world as (approximating) an instantiation of (a substructure of) some wider mathematical structure. Rather than simply mirroring the structure of an empirical system (as in any model), a *structural model* represents that system as an instance of a mathematical structure. As such, it enables us to explain features of that system as holding *in virtue of* its mathematical structure, whenever they can be shown to be empirical interpretations of mathematical statements that are derivable from the structure's characterizing axioms. It is not mere similarity that matters in the use of mathematical models to explain, but rather, the realization that this similarity means that the inferential structure of our mathematical theories carries over to the empirical situation modelled, so that truths about the empirical situation can be seen as holding in virtue of its mathematical structure.

CONCLUSION

I have, in this paper, agreed with platonists including Baker and Colyvan that mathematics sometimes plays an indispensable explanatory role in empirical science, with mathematical hypotheses sometimes doing genuine explanatory (or at least, explanation-like) work that is not exhausted by the nominalistic content that those hypotheses enable us to represent. Nevertheless, I have argued, the involvement of mathematical hypotheses in these explanations does not support platonism. Mathematical hypotheses can play this kind of explanatory (or explanation-like) role even if there are no abstract mathematical objects, since the role mathematics plays in such explanations is of showing physical phenomena to be true in virtue of the mathematical structure instantiated (or approximately instantiated) in the physical system under study, rather than by appealing to abstract mathematical objects per se. In structural explanations, we have examples of distinctively mathematical explanations which show their explananda to hold by virtue of logical necessity given the mathematical structure instantiated in the physical system. When structure-characterising axioms are interpreted so as to be true of a particular physical system, we can generate mathematical explanations of physical phenomena that do not appeal to any abstract mathematical objects, but instead only require modal truths about what follows logically from our mathematical assumptions, together with the recognition that the assumptions of our mathematical theories are true when interpreted as about the physical system under examination.

Given, however, the amount of idealization that is generally required in order to apply mathematics to physical systems, most mathematical explanations of empirical phenomena will involve intermediate idealized models, rather than the direct instantiation of mathematical structures in physical systems, where what is directly explained by these structural explanations is features of an idealized model that instantiates our mathematical axioms. The presence of these models as intermediaries may raise concerns that such explanations are committed to abstract mathematical objects, or at the very least, to abstract idealizations of physical objects whose status is arguably as questionable as the abstract mathematical objects that mathematical fictionalists try to avoid. I have suggested that when idealized models are introduced into explanations, mathematical theories are used to provide structural explanations of the features of these models, which can then be used to explain features of the physical systems they model to the extent that the models provide apt representations. And I have proposed an understanding of idealized models in science that views them as a form of 'make-believe' or pretense. In such cases, when we speak of similarities between the physical system and the model, we are indirectly asserting that the physical system is the way *it* would have to be to make the pretense appropriate. So we can speak *as if* there are objects as imagined in our idealized theoretical models in order to represent how things are taken to be with the physical systems with which we are ultimately concerned. The models in structural model explanations need not, then, present any new worry the nominalist who takes it that fictions can be used to represent without the objects of fictions existing.

What this paper has not, of course, established is that *all* explanatory or explanation-like uses of mathematical hypotheses occur in the context of structural model explanations. I have not argued (and indeed I do not hold) that all explanation is structural explanation, and it is at least conceivable that examples could be found where mathematics plays a genuine explanatory role but where the explanation given is not structural. However, the dual role of mathematics identified in this discussion—as providing amenable, abstracted *models* of reality that are easy to work with, and as identifying features of those models that hold in virtue of *structure-characterizing mathematical axioms*, seems to me at least to get at some of the key elements of what it is so special about mathematics that makes it such a useful tool in both describing and explaining empirical phenomena. If there are other features of mathematics in application that have been overlooked, and that can be exploited to show mathematics to be explanatory in other ways, I would certainly be keen to hear of them. But what I hope to have shown in this paper is that there is a gap between showing that mathematics can play an indispensable explanatory (or explanation-like) role and showing that the existence of mathematical objects (or the truth of our mathematical theories) is required for mathematics to play such a role. We should not automatically infer, of the best explanation we have of a phenomena, that it is *true*, but only that it does indeed *explain*. The question then needs to be asked, "How does it explain?", and it is in the details of answering this question that we may hope to uncover the metaphysical commitments of our taking the explanation to be explanatory.

Notes

1. Talk of shared structural features may suggest to some readers a Platonist interpretation in the form of the ante rem structuralism of Shapiro (1997). In Sect. 3.3 I show that this Platonist interpretation can be avoided by adopting a modal structuralist understanding of the notion of mathematical structure (following Hellman, 1989).
2. The labels ‘algebraic’ and ‘assertory’ are due to Geoffrey Hellman (2003), but the debate over how to understand axiomatic theories is older, going back at least to Frege and Hilbert, who corresponded over this matter (with Frege on the assertory and Hilbert on the algebraic side of the debate Frege, 1980).
3. In Leng (2007) I argue that an algebraic approach to mathematical theories is also shared by other contemporary philosophical accounts of mathematics, including fictionalism and full-blooded platonism.
4. The modal structural characterisation follows Hellman (1989).
5. The issue of explanatory symmetries is also raised as a problem for Lange’s ‘explanations by constraint’, which also look like they are best cast as explanatory arguments. In *Because without Cause*, Lange tries to avoid symmetries by arguing that ‘reversed’ versions of his explanations by constraint are ruled out because they appeal to features that are “not understood to be constitutive of the physical arrangement with which [the explanatory why question] is concerned” (Lange, 2016, p. 43). Craver and Povich (2017) find this account wanting (though see Lange, 2018 for a reply). I wish to embrace the potential for empirical symmetries in part because I think the kinds of features that contextual information might determine to be *constitutive* of a physical arrangement when we consider a why question might well be such as to allow perfectly acceptable reversals. Those who hold that true explanations cannot admit of symmetries might wish to resist taking ‘displaying the phenomenon to be nomically expectable’ to be a way of explaining things. My own view, though, is that ‘showing it to be nomically expectable’ should be considered a perfectly good way of explaining a phenomenon, and it is only a prejudice in favour of the causal that prevents us from accepting that perfectly good explanations may sometimes run in more than one direction.
6. The group tables are as follows:
   Klein 4-group

0	a	b	ab
a	0	ab	b
b	ab	0	a
ab	b	a	0

Cyclic group of order 4

0	a	a^2	a^3
a	a^2	a^3	0
a^2	a^3	0	a
a^3	0	a	a^2

This idealization is inessential. It is made for the convenience of using a straightforward instantiation of the Peano axiom structure in the sequence of years in our explanatory argument. Since the theorem we are using will also apply also to apply to finite initial segments of the natural numbers, we could avoid the idealization and talk instead about theorems that hold in finite initial segments of n. I prefer to make the idealization for simplicity in formulating the explanatory argument, since, as I will argue in the next section, introducing idealizations into our mathematical explanations of physical phenomena will often be required anyway, and doing so incurs no additional platonistic debt.

7. Field (1980) explains how the use of mathematics in such measurements can be dispensed with. Without actually following Field in dispensing with this use of mathematics, Field's machinery should convince us that our initial use of real numbers in measurement are merely a means of quantitatively representing qualitative differences between wind strength and direction at various points: there is some nominalistic content to these measurements, even though they are mathematically indexed. Beyond this measurement step, though, I wish to suggest that the subsequent use of a mathematical theory to model wind behaviour so-measured is an essential explanatory use whose value does not solely reside in its representational content.
8. In fact, as Alan Baker (2005) has pointed out, examples such as these are somewhat tenuous, since the phenomenon to be explained is not one that has been independently noticed or even verified: it is more a prediction of the mathematics than a previously noted puzzle crying out for explanation. Nevertheless, since I am presuming for the purposes of this paper that there are some genuine mathematical explanations of empirical phenomena, rather than trying to establish

the existence of genuine examples, I have chosen to stick with this example for its relative simplicity. It provides, at least, an explanation in the sense that, had the phenomenon been noted prior to the mathematical prediction, it would have explained that phenomenon.

REFERENCES

1. Baker, A. (2005). Are there genuine mathematical explanations of physical phenomena? *Mind, 114*, 223–238.
2. Baker, A. (2017). Mathematics and explanatory generality. *Philosophia Mathematica, 25*, 194–209.
3. Bokulich, A. (2008). *Re-examining the quantum-classical relation*. CUP.
4. Brown, J. R. (2012). *Platonism, naturalism, and mathematical knowledge*. Routledge.
5. Bueno, O., French, S., & Ladyman, J. (2002). On representing the relationship between the mathematical and the empirical. *Philosophy of Science, 69*, 452–473.
6. Bueno, O., & Linnebo, Ø. (Eds.). (2009). *New waves in philosophy of mathematics*. Palgrave Macmillan.
7. Cartwright, N. (1983). *How the laws of physics lie*. OUP.
8. Cellucci, C., & Gillies, D. (Eds.). (2005). *Mathematical reasoning and heuristics*. King's College Publications.
9. Craver, C. F., & Povich, M. (2017). The directionality of distinctively mathematical explanations. *Studies in History and Philosophy of Science, 63*, 31–38.
10. Cushing, J. T., & McMullin, E. (Eds.). (1989). *Philosophical consequences of quantum theory: Reflections on Bell's theorem*. University of Notre Dame Press.
11. Daly, C., & Langford, S. (2009). Mathematical explanation and indispensability arguments. *Philosophical Quarterly, 59*(237), 641–658.
12. Dasgupta, S., Dotan, R., & Weslake, B. (Eds.). (2021). *Current controversies in the philosophy of science*. Routledge.
13. Field, H. (1980). *Science without numbers*. Princeton University Press.
14. Field, H. (1984). 'Is mathematical knowledge just logical knowledge? Philosophical review 93: 509–52. *Reprinted With a Postscript in Field, 1989*, 79–124.
15. Field, H. (1989). *Realism, mathematics, and modality*. Blackwell.
16. Frege, G. (1980). Philosophical and mathematical correspondence. In G. Gabriel, H. Hermes, F. Kambartel, C. Thiel, A. Veraart, B. McGuinness, & H. Kaal (Eds.). Oxford: Blackwell Publishers.

17. French, S. (2000). The reasonable effectiveness of mathematics: Partial structures and the applicability of group theory to physics. *Synthese, 136*, 31–56.

18. Frigg, R. (2010). Models and fiction. *Synthese, 172*, 251–268.

19. Godfrey-Smith, P. (2009). Models and fictions in science. *Philosophical Studies, 143*, 101–116.

20. Hayes, B. (2005). Group theory in the bedroom. *Scientific American, 93*(5), 395.

21. Hellman, G. (1989). *Mathematics without numbers: Towards a modal-structural interpretation*. Clarendon Press.

22. Hellman, G. (2003). Does category theory provide a framework for mathematical structuralism? *Philosophia Mathematica, 11*, 129–157.

23. Hughes, R. I. G. (1989). Bell's theorem, ideology, and structural explanation. In Cushing and McMullin, (Eds.), (1989) Philosophical consequences of quantum theory: Reflections on Bell's theorem (pp. 195–207). Notre Dame, IN: University of Notre Dame Press.

24. Kreisel, G. (1967). 'Informal rigour and completeness proofs'. In Lakatos (Ed.) (1967) (pp. 138–171).

25. Lakatos, I. (Ed.). (1967). *Problems in the philosophy of mathematics*. Amsterdam: North Holland.

26. Lange, M. (2016). *Because without cause: Non-causal explanations in science and mathematics*. OUP.

27. Lange, M. (2018). A reply to Craver and Povich on the directionality of distinctively mathematical explanations. *Studies in the History of Philosophy of Science, 67*, 85–88.

28. Leng, M. (2005). Mathematical explanation. In Cellucci and Gillies (2005) (pp. 167–189).

29. Leng, M. (2007). What's there to know? A fictionalist approach to mathematical knowledge. In Leng, Paseau, and Potter (2007), pp. 84–108.

30. Leng, M. (2009). Algebraic approaches to mathematics. In Bueno and Linnebo (2009) (pp. 117–134).

31. Leng, M. (2010). *Mathematics and reality*. OUP.

32. Leng, M. (2012). Taking it easy: A response to Colyvan. *Mind, 121*, 983–995.

33. Leng, M. (2021). Mathematical explanation does not require Mathematical Truth. In Dasgupta et al (2021) (pp. 51–59).

34. Leng, M., Paseau, A., & Potter, M. (Eds.). (2007). *Mathematical knowledge*. Oxford: Oxford University Press.

35. Melia, J. (2000). Weaseling away the indispensability argument. *Mind, 109*, 458–479.

36. Otávio Bueno and Mark Colyvan. (2011). An inferential conception of the applications of mathematics. *Noûs, 45*, 345–374.

37. Pincock, C. (2004). A revealing flaw in Colyvan's indispensability argument. *Philosophy of Science, 71*, 61–79.

38. Pincock, C. (2007). A role for mathematics in the physical sciences. *Noûs, 41*, 253–275.

39. Pincock, C. (2012). *Mathematics and scientific representation*. Oxford: Oxford University Press.

40. Railton, P. (1980). *Explaining explanation: A realist account of scientific explanation and understanding* (Ph.D. Dissertation, Princeton University).

41. Saatsi, J. (2011). The enhanced indispensability argument: Representational versus explanatory role of mathematics in science. *British Journal for the Philosophy of Science, 62*, 143–154.

42. Salis, F. (2019). The new fiction view of models. *British Journal for the Philosophy of Science* (advanced access). https://doi.org/10.1093/bjps/axz015.

43. Salmon, W. (1989). *Four decades of scientific explanation*. University of Minnesota Press.

44. Shapiro, S. (1997). *Philosophy of mathematics: Structure and ontology*. Oxford University Press.

45. Toon, A. (2010). The ontology of theoretical modelling: Models as make-believe. *Synthese, 172*, 301–315.

46. Walton, Kendall L. (1990). *Mimesis as make-believe: On the foundations of the representational arts*. Harvard University Press.

47. Weisstein, E.W. (web resource), 'Hairy Ball Theorem.' In MathWorld-A Wolfram Web Resource. Retrieved Dec 18, 2020 from, https://mathworld.wolfram.com/HairyBallTheorem.html.

Chapter

2

THE REAL AND THE MATHEMATICAL IN QUANTUM MODELING: FROM PRINCIPLES TO MODELS AND FROM MODELS TO PRINCIPLES

Arkady Plotnitsky

Theory and Cultural Studies Program, Purdue University, West Lafayette, IN, United States

The history of mathematical modeling outside physics has been dominated by the use of classical mathematical models, C-models, primarily those of a probabilistic or statistical nature. More recently, however, quantum mathematical models, Q-models, based in the mathematical formalism of quantum theory have become more prominent in psychology, economics, and decision science. The use of Q-models in these fields remains controversial, in part because it is not entirely clear whether Q-models are necessary for dealing with the phenomena in question or whether C-models would still suffice. My aim, however, is not to assess the necessity of Q-models in these fields, but instead to reflect on what the possible applicability of Q-models may tell us about the corresponding phenomena there, vis-à-vis

Citation: (APA): Medová, Plotnitsky, A. (2017). The real and the mathematical in quantum modeling: From principles to models and from models to principles. Frontiers in Physics, 5, 19. (13 pages).

quantum phenomena in physics. In order to do so, I shall first discuss the key reasons for the use of Q-models in physics. In particular, I shall examine the fundamental principles that led to the development of quantum mechanics. Then I shall consider a possible role of similar principles in using Q-models outside physics. Psychology, economics, and decision science borrow already available Q-models from quantum theory, rather than derive them from their own internal principles, while quantum mechanics was derived from such principles, because there was no readily available mathematical model to handle quantum phenomena, although the mathematics ultimately used in quantum did in fact exist then. I shall argue, however, that the principle perspective on mathematical modeling outside physics might help us to understand better the role of Q-models in these fields and possibly to envision new models, conceptually analogous to but mathematically different from those of quantum theory, that may be helpful or even necessary there or in physics itself. I shall, in closing, suggest one possible type of such models, singularized probabilistic models, SP-models, some of which are time-dependent, TDSP-models. The necessity of using such models may change the nature of mathematical modeling in science and, thus, the nature of science, as it happened in the case of Q-models, which not only led to a revolutionary transformation of physics but also opened new possibilities for scientific thinking and mathematical modeling beyond physics.

Keywords: principles, models, probability, statistics, reality, realism

INTRODUCTION

The history of mathematical modeling outside physics has been dominated by classical mathematical models, C-models, based on mathematical models developed in classical physics, especially probabilistic or statistical models, borrowed from classical statistical physics or chaos and complexity theories. More recently, however, models based in the mathematical formalism of quantum theory, Q-models, primarily borrowed from quantum mechanics but occasionally also quantum field theory, became more current outside physics, specifically in psychology, economics, and decision science, the fields (beyond physics) with which I will be primarily concerned here [e.g., 1, 2][1]. My abbreviations follows P. Dirac's distinction between c-numbers (classical numbers) and q-numbers (quantum numbers), because the variables used in Q-models are in fact q-numbers. Quantum mechanics and Q-models are based in the mathematics of Hilbert spaces over *complex* numbers, **C**, with Hilbert-space *operators* used as physical variables in the equations of quantum mechanics, as against functions of real (mathematical) variables, c-numbers, that serve as physical variables in classical physics. The use of Q-models in these fields remains controversial, because it is not entirely clear whether they are necessary for dealing with the phenomena in question or whether C-models would suffice. It is true that debates and sometimes controversies have also accompanied quantum mechanics since its birth in 1925. These debates, initiated by the famous confrontation between N. Bohr and A. Einstein on, in Bohr's phrase, "epistemological problems in atomic physics," used in the title of his account of this confrontation, have never lost their intensity and appear to be interminable [3, v. 2, pp. 32–66]. However, as Bohr's phrase indicates, the reasons for these controversies have been primarily philosophical. The effectiveness of quantum mechanics or higher-level quantum theories, such as quantum field theory, has not been in question: they are among the best-confirmed theories in physics. The situation is different in psychology, economics, and decision science, where it is the *scientific* effectiveness or at least necessity of Q-models that is doubted. My aim here, however, is not to assess this effectiveness or necessity, but instead to reflect on what *the possible applicability* of Q-models may tell us about the corresponding phenomena in these fields vis-à-vis quantum phenomena in physics. In order to do so, I shall first consider the key reasons for the use of Q-models in physics. In particular, I shall examine the fundamental principles that *grounded* and indeed *led* to the development of quantum theory. Then I shall consider a possible role of similar principles *in using* Q-models beyond quantum theory. My emphases are due to the fact that psychology, economics, and decision science *borrow* already available Q-models from quantum theory, rather than derive them from their own fundamental principles, while quantum mechanics and then quantum field theory were derived from such principles. This is not surprising because there was at the time no available mathematical model or (a more general concept, which includes an interpretation of the model used) theory to effectively handle quantum phenomena. The "old quantum theory" of M. Planck, A. Einstein, N. Bohr, and A. Sommerfeld, which ushered in the quantum revolution, became manifestly inadequate by the time W. Heisenberg began his work on quantum mechanics that he discovered in 1925 [4]. For the reasons explained below (mostly a search for a more rigorous derivation of the formalism), the research in quantum foundations is still concerned with deriving quantum theory from such principles, a project in part motivated by the rise of quantum information theory. That does not appear to be a significant concern outside physics where the use of Q-models is motivated primarily by their predictive capacities, which is of course a crucial consideration in physics as well. It may, however, be beneficial to consider the deeper reasons for the possible use of Q-models in these fields, or, in terms of my title, the *real* that gives rise to the *mathematical* of Q-models there. The principle perspective on mathematical modeling beyond physics might help us to do this and possibly to envision new, *post*-quantum, models there or even in physics. I shall, in closing, suggest one possible type of such models, singularized probabilistic models, SP-models, some of which are time-dependent, TDSP-models, and consider their implications for mathematical modeling in science and for our understanding of the nature of science[2].

PHYSICAL PRINCIPLES AND MATHEMATICAL MODELS IN QUANTUM MECHANICS

Theories, Principles, and Models in Fundamental Physics

I would like to begin by outlining the key features of the standard mathematical model of quantum mechanics, more customarily used as a probabilistically or statistically predictive model in view of the difficulties of in maintaining its representational capacities, which continue to be debated:

- *The Hilbert-space formalism over the field of complex numbers*, **C**, an abstract vector space of any dimension, finite or infinite (in quantum mechanics, either finite or countably infinite), possessing the structure of an inner product that allows lengths and angles to be measured, analogously to an n-dimensional Euclidean space (which is a Hilbert space over real numbers **R**);
- *The noncommutativity* of the Hilbert-space operators, also known as “observables,” which are *mathematical* entities associated, in terms of probabilistic or statistical predictions, with *physically* observable quantities;
- *The nonadditive nature of the probabilities involved*: the joint probability of two or more mutually exclusive alternatives in which an event might occur is, in general, not equal to the sum of the probabilities for each alternative, and instead obey the law of the addition of the so-called “quantum amplitudes,” associated

with complex Hilbert-space vectors, for these alternatives (technically, these amplitudes are linked to probability densities);

- *Born's rule* or an analogous rule (such as von Neumann's projection postulate or Lüder's postulate) added to the formalism, which establishes the relation between amplitudes as complex entities and probabilities as real numbers (by using square moduli or, equivalently, the multiplication of these quantities and their complex conjugates) and (3) above[3].

In the development of quantum mechanics, discovered in 1925, these features were not initially assumed, but were derived from certain physical features of quantum phenomena and principles arising from these features. The formalism was only given a properly Hilbert-space form by J. von Neumann, in 1932, in *The Mathematical Foundations of Quantum Mechanics*, a standard text ever since [7][4].

I shall now explain the concepts of theory, principle, and model, as they will be understood here. By a theory, I mean an organized assemblage of concepts, explanations, principles, and models by means of which one is able to relate, in one way or another, to the phenomena or (they are not always the same) objects the theory considers. In defining principles, I follow Einstein's distinction between "constructive" and "principle" theories, two contrasting, although in practice often intermixed, types of theories [8, 9, pp. 35–50]. "Constructive theories" aim "to build up a picture of the more complex phenomena out of the materials of a relatively simple formal scheme from which they start out" [8, p. 228]. Thus, according to Einstein, the kinetic theory of gases, as a constructive theory in classical physics, "seeks to reduce mechanical, thermal, and diffusional processes to movements of molecules—i.e., to build them up out of the hypothesis of molecular motion," described by the laws of classical mechanics [8, p. 228]. By contrast, principle theories "employ the analytic, not the synthetic, method. The elements which form their basis and starting point are not hypothetically constructed but empirically discovered ones, general characteristics of natural processes, principles that give rise to mathematically formulated criteria which the separate processes or the theoretical representations of them have to satisfy" [8, p. 228]. Thus, thermodynamics, a classical principle theory (parallel to the kinetic theory of gases as a constructive theory), "seeks by analytical means to deduce necessary conditions, which separate events have to satisfy, from the universally experienced fact that perpetual motion is impossible" [8, p. 228].

Principles, then, are "empirically discovered, general characteristics of natural processes, ...that give rise to mathematically formulated criteria which the separate processes or the theoretical representations of them have to satisfy." I shall adopt this definition, but with the following qualification, which is likely to have been accepted by Einstein. Principles are not empirically discovered but formulated, constructed, on the basis of empirically established evidence. "The impossibility of perpetual motion" is hardly empirically given; it is as a principle formulated on the basis of such evidence.

Constructive theories are, more or less by definition, realist theories, and conversely, many realist theories are constructive. Realist theories represent, commonly causally, the phenomena or objects they consider and their behavior, in science by mathematical models, assumed to idealize how nature or reality works, in the case of constructive theories at the simpler, or deeper, level of reality constructed by a theory. In other words, a constructive theory offer a representation of the processes underlying and connecting the observable phenomena considered, commonly by understanding the ultimate character of these processes on the model of classical mechanics or classical electrodynamics, as in the kinetic theory of gases, as described above or other forms of classical statistical physics. All such theories assume that the individual behavior of the ultimate constituents of the systems they consider is described by the laws of classical mechanics. A realist theory may represent objects or phenomena it considers in a more direct, if still idealized, manner, as classical mechanics (which deals with individual or sufficiently small systems) or classical electrodynamics do. I shall discuss the concepts of reality and realism, which encompasses that of realist theory, in more detail below. First, however, I shall define a mathematical model.

By a "mathematical model" I refer to a mathematical structure or set of mathematical structures that enables any type of relation to the (observed) phenomena or objects considered. (As I shall only deal with mathematical models here, the term "model" hereafter refers to mathematical models.) All modern, post-Galilean, physical theories are defined by their uses of such models. The requirement of using *mathematical models* may be seen as a principle, the mathematization principle, "the M principle," arguably the single defining principle of all modern physics, from Galileo on. Such models may be realist, representational, as in classical physics, specifically classical mechanics, or predictive, as in classical statistical physics (the models of which are, however, underlain by representational models of classical mechanics), or in quantum mechanics, without assuming realism and causality even in considering elementary individual quantum processes, such as those concerning elementary quantum objects, "elementary particles." This assumption is expressly abandoned or even precluded in non-realist interpretations of quantum phenomena and quantum mechanics, following Bohr and "the spirit of Copenhagen," as Heisenberg called it [10, p. iv][5]. The M principle is upheld in quantum mechanics, but, in non-realist interpretations, in a way different from how it is used in realist theories.

The probabilistic or statistical character of quantum predictions must also be maintained by realist interpretations of these theories or alternative theories (such as Bohmian theories) of quantum phenomena, in conformity with quantum experiments, in which only probabilistic or statistical predictions are possible. The reasons for this is that the repetition of identically prepared quantum experiments in general leads to different outcomes, a difference that cannot be improved beyond a certain limit (defined by Planck's constant, h) by improving the conditions of measurement, which is possible in classical physics. This fact is also manifested in Heisenberg's uncertainty relations, which are statistical in character as well. This situation leads to the quantum probability or (depending on interpretation) quantum statistics principle, the QP/QS principle, arguably the single defining principle in Q-models in physics and beyond, keeping in mind that in psychology, economics, and decision science, we do not have anything corresponding to elementary individual physical processes, involving the ultimate elementary constituents of nature, "elementary particles." Nor do we have anything analogous to h. The probabilities themselves necessary for making correct predictions, in either quantum mechanics or in using Q-models elsewhere, are, thus far, calculated by using the Hilbert-space or mathematically equivalent formalisms and the (non-additive) procedure described above that uses quantum amplitudes and Born's or a similar rule[6].

Realist models are, then, representational models, idealizing the nature of objects or phenomena they consider. The term "realism" will be primarily understood here as referring to the possibility, at least, again, in principle, of such models, and, in the first place, theories allowing for such models. One could define another type of realism, which would refer to theories that presuppose an independent architecture of reality they consider, while allowing that this architecture cannot be represented, either at a given moment in history or perhaps ever, but if so, only due to practical human limitations [9, pp. 11–23]. In the first case, a theory that is strictly predictive may be accepted, but with the hope that a future theory will do better, by being a realist theory of the representational type. Einstein adopted this attitude toward quantum mechanics, which he expected to be eventually replaced by a (representational) realist theory. Even in the second case, the ultimate nature of reality is commonly deemed to be conceivable on realist models of classical physics, possibly adjusting them to accommodate new phenomena. However, this type of realism implies that there is no representational theory or model of the ultimate nature of the phenomena or objects considered. Either type of realism is abandoned or even precluded in quantum mechanics, when interpreted in the spirit of Copenhagen. However, such interpretations do assume the concept of *reality*, by which I refer to what exists or is assumed to exist, without making any claim upon the *character* of this existence, which type of claims defines realist theories. By existence I refer to a capacity to have effects on the world, ultimately, which also assume the existence of the world by virtue of its capacity to have effects upon itself, effects which establish by means of and thus in terms as effects of our interactions with the world. In physics, the primary reality considered is that of nature or matter. It is generally assumed to exist independently of our interaction with it, which also assumes that it has existed when we did not exist and will continue to exist when we will no longer exist. This assumption is also made in non-realist interpretations of quantum mechanics, in the absence of a representation or even (as against the second, non-representational type of realism defined above) conception of the character of this existence. Thus, if *realism* presupposes a representation or at least a conception of reality, this concept of *reality* is that of "reality *without* realism" [9, 11]. The assumption of this concept of reality is a principle, *the RWR principle*. The existence or *reality* of quantum objects, a form of reality beyond representation or even conception, is inferred from effects they have on our world, specifically on experimental technology. It has not been possible, at least thus far, to observe a moving electron or photon, or for that matter even stationary electrons (there are no stationary photons, which only exist in motion before they are absorbed by other forms of matter, such as electrons). It is only possible to observe traces of their interactions with measuring instruments, traces that do not allow us to reconstitute the independent behavior of quantum objects movement, an impossibility reflected in Heisenberg's uncertainty relations. In non-realist, RWR-principle-based, interpretations, quantum mechanics only predicts, in probabilistic or statistical terms (no other predictions are, again, possible on experimental grounds), effects manifested in measuring instruments impacted by quantum objects.

While a principle theory, which, as I explained, need not be constructive in Einstein's sense, could be either realist or non-realist, a constructive theory is by definition realist. Realist or, it follows, constructive theories do involve principles, such as the equivalence principle in general relativity, or the principle of causality, which, to adopt Kant's definition, commonly used ever since, states that, if an event takes place, it has a cause of which it is an effect [12, p. 305, 308][7]. Asymmetrically, however, a principle theory need not involve constructive aspects or be realist. In non-realist, RWR-principle-based, interpretations, quantum mechanics is a principle theory by definition, by virtue of the RWR principle. It is not possible, in such interpretations, to have a constructive theorization of the ultimate entities, quantum objects, which are responsible for the observable quantum phenomena, unless one sees quantum objects as *constructed* as in principle *unconstructible*. According to Bohr, thus formulating the RWR principle, "in quantum mechanics we are not dealing with an arbitrary renunciation of a more detailed analysis of atomic phenomena, but with a recognition that such an analysis is *in principle* excluded," beyond a certain point [3, v. 2, p. 62]. In this interpretation, quantum mechanics divorces itself from the representation of the connections between observed quantum phenomena, which it only relates in terms of predictions, in general probabilistic or statistical in character, thus fulfilling the M principle under the conditions of the RWR principle.

Finally, the present view does not assume a permanent, Platonist, essence to any given principle, which can always be abandoned under the pressure of new experimental findings or new ways of theorizing previously available experimental findings. Indeed, one might argue that the greatest form of creative thinking in science or other theoretical fields is that which lead to the invention of new principles, which implies the transformation of principles, rather than any Platonist permanence to them.

The Physical Principles of the Quantum Theory

The RWR principle and the corresponding interpretation of quantum mechanics emerged only in the 1930s. Heisenberg's discovery of quantum mechanics in 1925 and Bohr's initial interpretation of it, proposed in 1927, were based on the following principles, with Bohr's complementarity principle added in 1927:

- the proto-RWR principle, according to which, "quantum mechanics does not deal with a space–time description of the motion of atomic particles" [3, v. 1, p. 48];

- the principle of discreteness or the QD principle, according to which all observed quantum phenomena are individual and discrete in relation to each other, which is fundamentally different the atomic discreteness of quantum objects themselves;
- the principle of the probabilistic or statistical nature of quantum predictions, the QP/QS principle, even (in contrast to classical statistical physics) in the case of primitive or elementary quantum processes, in which nature also reflects a special, non-additive, nature of quantum probabilities and rules, such as Born's rule, for deriving them, and
- the correspondence principle, which, as initially understood by Bohr, required that the predictions of quantum theory must coincide with those of classical mechanics in the classical limit, but was given by Heisenberg a new and more rigorous form of "the mathematical correspondence principle," which required that the equations of quantum mechanics convert into those of classical mechanics in the classical limit, thus, in accordance with the M principle.

I speak of the proto-RWR principle because Heisenberg saw the project of describing the motion of electrons as unachievable at the time, rather than "*in principle* excluded," as Bohr assumed a decade later [3, v. 2, p. 62]. This was, nevertheless, a radical move on Heisenberg's part, as Bohr was the first to realize: "In contrast to ordinary [classical] mechanics, the new quantum mechanics does not deal with a space–time description of the motion of atomic particles. It operates with manifolds of quantities [matrices] which replace the harmonic oscillating components of the motion and symbolize the possibilities of transitions between stationary states in conformity with the correspondence principle. These quantities satisfy certain relations which take the place of the mechanical equations of motion and the quantization rules [of the old quantum theory]" [3, v. 1, p. 48].

Quantum discreteness was eventually (as part of Bohr's ultimate interpretation) recast by Bohr in terms of his concept of "phenomenon," defined in terms of what is observed in measuring instruments under the impact of quantum objects, in contradistinction to quantum objects themselves, which cannot be observed or represented [3, v. 2, p. 64]. Quantum phenomena are, in Bohr's interpretation, irreducibly discrete in relation to each other, and there is no continuous or any other conceivable process that could be assumed to connect them. Probability has a temporal structure by virtue of its futural and discrete nature: one can only verifiably estimate future discrete events. Such events may, however, be continuously and causally connected, as they are in classical physics, even though we may not be able to track these connections to make exact predictions, as happens in classical statistical mechanics or chaos theory. By contrast, in non-realist, RWR-principle-based, interpretations, the nature of quantum phenomena and events precludes us from causally (or otherwise) connecting them. This means that only probabilistic or statistical predictions are possible, even ideally and in principle, and even in dealing with elementary individual quantum objects, such as those known as "elementary particles," and the processes and events they lead to, objects and processes that cannot be decomposed into a smaller objects and processes. This qualification distinguishes quantum mechanics from classical probabilistic or statistical theories, or of course classical mechanics where such predictions could, at least ideally, be exact in dealing with individual classical objects or a small number of classical objects. In quantum mechanics, in non-realist interpretations, this type of idealization is not possible, a fact reflected in the uncertainty relations. The theory only estimates the probabilities or statistics of the outcomes of discrete future events, on the basis of previous events, and tells us nothing about what happens between events. Nor does it describe the data observed in measuring instruments and hence quantum phenomena. They are described by classical physics, which, however, cannot predict them.

The QP/QS principle was *mathematically expressed* in Heisenberg's scheme by matrices containing the necessary probability amplitudes cum Born's rule. Heisenberg only formulated this rule in the case of electrons' quantum jumps in the hydrogen atom, rather than as universally applicable in quantum mechanics, as Born did. Born's rule is not inherent in the formalism but is added to it—it is *postulated*.

The correspondence principle was central to Heisenberg's derivation of quantum mechanics. In its mathematical form, introduced by Heisenberg, the principle required that both the equations of quantum mechanics, which were formally those of classical mechanics, and the variables used, which were different, convert into those of classical mechanics in the classical limit, a conversion automatic in the case of equations but not variables. (The processes themselves, however, are still quantum even in this limit.) Thus, the principle gave Heisenberg a half of the mathematical architecture he needed.

An important qualification is in order. Heisenberg's derivation of quantum mechanics from principles cannot be considered a strictly rigorous derivation, especially in a mathematical sense. As he noted in *The Physical Principles of the Quantum Theory* (from which title I borrow my title of this section): "The deduction of the fundamental equation of quantum mechanics is not a deduction in the mathematical sense of the word, since the equations to be obtained form themselves the postulates of the theory. Although made highly plausible, their ultimate justification lies in the agreement of their predictions with the experiment" [10, p. 108]. While Heisenberg, again, borrowed the form of equations themselves from classical mechanics by the mathematical correspondence principle, he virtually guessed the variables he needed—one of the most extraordinary guesses in the history of physics. A more rigorous derivation of quantum mechanics from fundamental principles may, thus, be pursued. More recent work in this direction has been in quantum information theory in the case of discrete quantum variables, such as spin, which require finite-dimensional Hilbert spaces, as opposed to infinite-dimensional ones for continuous variables, such as position and momentum (e.g., 13–15)[8]. I shall comment on this work below.

Bohr's interpretation of quantum phenomena and quantum mechanics added a new principle, *the complementarity principle*. It arises from Bohr's *concept of complementarity* and may be defined as requiring: "*(a) a mutual exclusivity of certain phenomena, entities, or conceptions; and yet (b) the possibility of considering each one of them separately at any given point, and (c) the necessity of considering all of them at different moments for a comprehensive account of the totality of phenomena that one must consider in quantum physics*" [9, p. 70].

In Bohr's ultimate interpretation, this concept applies strictly to what is observed in measuring instruments, *quantum phenomena*, and not to *quantum objects*, placed beyond representation or even conception. Complementarity is a reflection of the fact that, in a radical departure from classical physics or relativity, the behavior of quantum objects of the same type, say, electrons, is not governed by the same physical law, especially a representational physical law, in all possible contexts, specifically in complementary contexts. In other words, the behavior of quantum objects has mutually incompatible effects in complementary set-ups, although this mutual incompatibility is, generally, manifested collectively, in multiple identically prepared experiments. On the other hand, the mathematical formalism of quantum mechanics offers correct probabilistic or statistical predictions of quantum phenomena *in all* *contexts*, in non-realist interpretations, under the assumption, that quantum objects and processes are beyond representation or even conception, by the RWR principle.

In some non-realist interpretations, such as the one the present author would favor, following W. Pauli, individual quantum events are not subject even to the *probabilistic* laws of quantum mechanics. This makes these laws collective, *statistical* [9, pp. 173–186; 11]. The QP/QS principle, accordingly, becomes strictly the QS principle. According to Pauli:

As this indeterminacy is an unavoidable element of every initial state of a system that is at all possible according to the [quantum-mechanical] laws of nature, the development of the system can never be determined as was the case in classical mechanics. The theory predicts only the statistics of the results of an experiment, when it is repeated under a given condition. Like the ultimate fact without any cause, the individual outcome of a measurement is, however, in general not comprehended by laws. This must necessarily be the case, if quantum or wave mechanics is interpreted as a rational generalization of classical physics, which take into account the finiteness of the quantum of action [*h*]. The probabilities occurring in the new laws have then to be considered to be primary, which means not deducible from deterministic laws. [19, p. 32]

Thus, in Pauli or the present view, this "beyond the law" includes the probabilistic or, in this view, *statistical* laws of quantum mechanics, laws that, thus, only apply to statistical multiplicities of repeated quantum events. Individual quantum events are not subject to laws, even to the probabilistic or statistical laws of quantum mechanics. Their outcomes cannot, in general, be assigned a probability: they are strictly random[9]. Only the statistics of multiple (identically prepared) experiments could be predicted and repeated, which repeatability appears to have been, thus far, necessary for scientific practice. Whether, however, one interprets quantum mechanics on such statistical lines or on the Bayesian lines, by assigning probability to individual events, we are compelled to rethink the concept of physical law as unavoidably contextual. This is "an entirely new situation as regards the description of physical phenomena that, the notion of *complementarity* aims at characterizing" [20, p. 700].

There are other important features of quantum phenomena, mathematically expressed in the quantum-mechanical formalism, in particular, the so-called "quantum non-locality," which refers to the existence of the statistical correlations between spatially separated quantum events, and "quantum entanglement," which reflects these correlations in the formalism. These features were discovered later and played no role in the initial derivation of quantum mechanics by either Heisenberg or Schrödinger. They do figure significantly in quantum information theory and recent attempts, mentioned above, to derive quantum mechanics from the principles of quantum information. Their analysis would require a treatment beyond my scope[10]. A few key points may, however, be mentioned. First, while quantum entanglement is a clearly defined feature of the formalism, the situation is different in the case of quantum non-locality. Although originating in the experimentally well-confirmed fact that certain spatially separated quantum phenomena or events exhibit statistical correlations (not found in classical physics), quantum non-locality is a complex and much debated issue. The problematic was *in effect* introduced in 1935 in the famous article by Einstein et al. [22]. I qualify because neither EPR's article nor Bohr's equally famous reply to it [20] used the language of correlations or entanglement. The latter term was introduced, in both German [*Verschränkung*] and English, by Schrödinger in his response to EPR's article, known as "the cat-paradox paper," after the paradox found there [23]. The subject remained dormant until the 1960s, when it was rekindled by the Bell and Kochen-Specker theorems, even to the point of nearly defining the current debate concerning quantum foundations. The theoretical and experimental research on the subject during the last decades has been massive and literature concerning it is immense. The term "non-locality" is not uniformly used in referring to quantum correlations, because it may suggest some sort of instantaneous physical connections between distant events, a "spooky action at a distance," as Einstein called it. Such connections are incompatible with relativity, although the principle of *locality*, which prohibits such connections, is independent of relativity. This type of *physical* non-locality, which is found, for example, in Bohmian mechanics, is commonly viewed as undesirable. The absence of realism allows one to avoid physical non-locality, as Bohr argued in his reply to EPR's article, which contended that quantum mechanics is either incomplete or physically nonlocal [20, 22].

FROM MODELS TO PRINCIPLES IN Q-MODELING OUTSIDE PHYSICS

Q-Models, Fundamental Principles, and Reality without Realism Outside Physics

In addressing Q-models in physics in preceding discussion, my main question, arising from the history of quantum theory, was: Given certain fundamental physical principles, established on the basis experimental evidence, in particular the QD and QP/QS principles, and perhaps adopting additional principles, such as the correspondence principle or the RWR (or proto-RWR) principle, what are the mathematical models that would enable us to handle this evidence? In turning now to the Q-models beyond physics, my main question is reverse: Assuming that mathematical Q-models apply in psychology, economics, and decision science, which features and which fundamental principles are behind such models, and how they accord with the fundamental principles of quantum mechanics? There are two sets of principles I have in mind. The first contains the principles that led to the emergence of quantum mechanics; and the second the principles of quantum information theory, which are, however, in accord with most principles of the first set. I shall be primarily concerned with this first set (apart from the correspondence principle, unique to quantum theory), but will also comment on the second[11].

But why is this question important in the first place? As noted from the outset, if there are phenomena outside physics that appear to require Q-models, one need, unlike at the time of the introduction of quantum mechanics, not invent such models at this point. One can borrow them, "ready-made," from quantum theory, which is what happed in the case of Q-modeling outside physics. Nevertheless, establishing, now inferentially, fundamental principles behind Q-models might allow us to make important conclusions about the nature of the phenomena handled by these models. To put it in stronger terms, finding the fundamental principles behind a given model, even if this model is already available, is important because otherwise we don't have a rigorous theory or a rigorous model, which is true even if a constructive theory is available, but is all the more important if it is not. Otherwise, we don't really know what our models are models *of*, especially, again, in the absence of a constructive theory and realism, which absence is likely if Q-models apply and is my main interest here.

These considerations are also relevant in pursuing projects of more rigorous derivation of quantum mechanics from principles in physics, for example on lines of quantum information theory, even though the theory itself is already established. Part of the reason is, again, that doing so can give us a deeper understanding of quantum phenomena and quantum theory. More, however, is at stake. The main value of such projects lies in solving outstanding problems of fundamental physics, as in quantum field theory (which still has unresolved problems, its extraordinary successes notwithstanding) or quantum gravity, which has no model as yet [24, 25]. The same argument applies to Q-modeling beyond physics. The future of mathematical modeling there is at stake as well.

Before addressing the relationships between fundamental principles and Q-models in psychology, economics, and decision science, it may be helpful to summarize the non-realist, the RWR-principle-based, interpretation of quantum phenomena and quantum mechanics outlined in Section Physical Principles and Mathematical Models in Quantum Mechanics. While quantum objects are assumed to exist, the character of this existence or reality is, by the RWR principle, assumed to be beyond representation and even conception. As such, this reality is different from the reality of quantum phenomena, which are defined by what is observed in measuring instruments under the impact of quantum objects and, thus, can be represented. There are no mathematically expressed *physical laws* corresponding to the behavior of quantum objects. There are, however, *mathematical laws* that, expressing the QP/QS principle, enable correct probabilistic or statistical predictions of the outcomes of quantum experiments, manifested in measuring instruments, in all contexts. In addition, there are two interpretations of these mathematical laws. The first is probabilistic, along Bayesian lines, in which case these laws are seen as allowing one to assign probabilities to the outcomes of individual quantum events in accordance with one or the other law of the available set of laws, specifically those applicable in complementary situations. The second is statistical, when no such probabilities could be assigned because the outcomes of individual quantum experiments are not comprehended even by these laws and are seen as random, while these laws are assumed to predict the statistics of multiple identically prepared experiments in the corresponding contexts.

It is clear, however, that this conceptual architecture, in either the Bayesian or statistical interpretation, cannot apply unaltered in considering, along non-realist lines, human phenomena found in psychology, economics, or decision science and the possible Q-models there. This is because, while there are individual objects or, the case may be, (human) subjects and processes to consider, there are no elementary objects of the type found in quantum physics. There is nothing analogous to elementary particles, such as electrons or photons, and there is rarely a completely random individual behavior. When one deals in these fields with large multiplicities one can, either in using C- or Q-models, average the individual behavior and statistically disregard the differences in this behavior, differences defined by psychological or other human and social factors, in which case one could apply either a Bayesian or statistical interpretation of the Q-model used. While, however, this averaging is sometimes possible in psychology, economics, and decision science, there are often significant obstacles in using it. Each sequence of events considered in such situations is singular, unique. Accordingly, if a Q-model applies in *a given class* of such cases, it would have to be interpreted on Bayesian lines, if one can establish such a class. If not, then, as discussed below, another type of models may be possible, the singularized probabilistic (SP) models, some of which are time-dependent (TDSP). Each such model is unique to the individual situation considered, rather than applicable to a class of individual situations; and this uniqueness may pose difficulties for scientific use of such models.

The QP/QS Principle and the Complementarity Principle

Beginning with Tversky and Kahneman's work in the 1970–80's [e.g., 26], it has been primarily the presence of probabilistic data akin to those encountered in quantum physics that suggested using Q-models in cognitive psychology, decision science, and economics [e.g., 1, 2][12]. Economic behavior may also involve psychological factors of the type analyzed by Tversky and Kahneman. (Kahneman was eventually awarded a Nobel Prize in economics.) The recourse to Q-models is motivated by the fact that one could not effectively use the classical (additive) rules but could use the quantum-mechanical-like (non-additive) rules for predicting the probabilities of the outcomes of certain psychological experiments, such as those involving responses to certain specific questions, asked sequentially. These responses were found to be statistically dependent on the order in which they were asked, which, again, in parallel with quantum mechanics, suggested that a non-commutative model and, in combination with the non-additive rules for calculating the probabilities involved, a Q-model could be more effective[13]. To clarify this parallel, in quantum mechanics, simultaneously measuring, or simultaneously asking questions concerning, two or more complementary variables, such as the position and the momentum of a given quantum object, are mutually exclusive or incompatible. Correlatively, changing the order of measuring (of asking the question concerning) the position and then the momentum of a quantum object, in general, changes the outcomes and hence our predictions concerning them. This circumstance is reflected, *experimentally*, in the uncertainty relations, and *mathematically*, in the non-commutativity of the multiplication of the corresponding Hilbert-space operators in the formalism, and *epistemologically*, in the complementarity of these two measurements. One can, analogously, consider psychologically incompatible and, thus, complementary questions in psychology and attempt to handle the corresponding events statistically by a Q-model [e.g., 1, pp. 259–260]. The situation involves further complexities in and outside quantum physics, which I put aside here. I would like, however, to mention R. Spekkens's article, which introduced "a toy theory," based on the following principle, linked to complementarity: "the number of questions about the physical state of a system that are answered must always be equal to the number that are unanswered in a state of maximal knowledge. Many quantum phenomena are found to have analogs within this toy theory." Many but not all! For the theory expressly fails to reproduce some among the crucial features of quantum theory, specifically and intriguingly some of those related to correlations and entanglement, such as "violations of Bell inequalities and the existence of a Kochen-Specker theorem" [27, p. 032110]. This failure reminds us that models based on the existence of incompatible questions, in and outside physics, may mathematically differ from quantum mechanics.

Q-models are, then, used to predict probabilities and correlations found in such experiments, without being expressly concerned with the principles characterizing the situations considered, but only assuming certain mathematical principles inherent in the quantum-mechanical formalism. Some among the principles of the first kind are, nevertheless, implicitly at work, specifically the QP/QS principle or the principle of incompatibility, in effect complementarity[14]. Whether these Q-models are required or C-models, models derived from the mathematics of classical physics, suffice remains, again, an open question, although it is difficult to assume that C-models could provide the non-additive probabilities necessary in such cases. A model alternative to that of quantum mechanics, possibly also free of quantum amplitudes and dealing directly with probabilities, is, in principle, possible even, as noted earlier, in quantum physics, but such a model is unlikely to be akin to those of classical physics. Thus, while they are both realist and causal, Bohmian models are mathematically different from those of classical physics. It may also be possible to construct a realist and causal mathematical model that would represent a deeper level of reality and that would have quantum mechanics as its limit, and then extend this model beyond physics [e.g., 30].

In any event, one can see the QP/QS principle, in part in conjunction with complementarity, as the main principle behind the use of Q-models beyond physics, accompanied, as in quantum mechanics, by the specific (non-additive) calculus of probability. Indeed, the QP/QS principle, along with the QD principle, was the starting principle for Heisenberg. The role of complementarity, only implicit initially by virtue of the non-commutative nature of Heisenberg's scheme, became apparent shortly thereafter, helped by Heisenberg's discovery of the uncertainty relations in 1927. It became clear that non-commutativity, the uncertainty relations, and complementarity were correlative, representing, respectively, the mathematical, physical, and epistemological aspects of the quantum-mechanical situation, defined by quantum discreteness (the QD principle). As noted earlier, quantum discreteness was eventually rethought by Bohr in terms of quantum phenomena, defined by what is observed in measuring instruments impacted by quantum objects, as opposed to the nature of quantum objects and processes, which are beyond conception and, hence, cannot be thought of as either discrete or continuous.

The psychological, economic, and decision-making phenomena treated by means of Q-models do not exhibit this type of irreducible discreteness or individuality. The processes that connect these phenomena are more akin to processes considered in classical physics, especially in chaos or complexity theory, again, often providing mathematical models, C-models, used in these fields. Now, assuming the defining role of, jointly, the QP/QS principle and the complementarity principle in considering these phenomena, could some form of the QD principle, correlative to the QP/QS principle in quantum mechanics, find its place in considering or even in order to derive Q-models in these fields? And if so, or in the first place, would the RWR principle, or a proto-RWR principle of the type used by Heisenberg, also be applicable? There are reasons to believe that such might be the case.

The RWR and QD Principles

Bohr thought that, along with the complementarity principle, the RWR principle might apply in biology and psychology. In considering biology, he argued as follows:

The existence of life must be considered as an elementary fact that cannot be explained, but must be taken as a starting point in biology, in a similar way as the quantum of action, which appears as *an irrational element* from the point of view of the classical mechanical physics, taken together with the existence of elementary particles, forms the foundation of atomic physics. The asserted impossibility of a physical or chemical explanation of the function peculiar to life would in this sense be analogous to the insufficiency of the mechanical analysis for the understanding of the stability of atoms. [31, p. 458; emphasis added]

The ultimate character of biological processes may, thus, be beyond representation or even conception, in accord with the RWR principle. Once the theory suspends accounting for the connections between the phenomena considered, these phenomena are unavoidably discrete, leading to the QD principle, and our predictions concerning them are unavoidably probabilistic, leading to the QP/QS principle. Our predictions concerning them are likely to follow a (non-additive) probability calculus of the type used in quantum probability, and thus are likely to require a Q-model. This is because, by the RWR or proto-RWR principle, it would be difficult or even impossible to treat the processes connecting the phenomena considered as either continuous or causal. Bohr's appeal to "an irrational element" is noteworthy, and I shall comment on it below. It is important that, as Bohr clearly implies here, this approach is possible even if the nature of biological processes is not physically quantum in the sense of being able to have physically *quantum effects*. (The ultimate constitution of all matter is quantum, but this constitution does not manifest itself apart from quantum experiments.) If they were quantum, such processes would be unrepresentable or inconceivable in Bohr's interpretation. At stake here, however, are *parallel*, rather than physically connected, situations that may require using the same type of mathematical models, Q-models, without possible connections between the systems defining these situations[15].

A recent article by Haven and Khrennikov provides an instructive example for possible roles of both the RWR and QD principle in market economics in their Q-modeling of market phenomena involving arbitrage as analogous to quantum tunneling [33]. The term "quantum tunneling" refers to a quantum object's capacity to "tunnel" through an energy barrier that it would not be able to surmount if it behaved classically. It is a quantum phenomenon *par excellence*. The quantum process itself behind any given case of quantum tunneling cannot be observed. One only ascertains that a particle can be found beyond the barrier, which is to say, that the corresponding measurement will register an impact of this particle on the measuring instrument beyond the barrier. Thus, in accord with the general situation that obtains in quantum mechanics, one deals with two discrete phenomena, connected by probabilistic or (in which case, we need multiple trials) statistical predictions concerning the second event on the basis of the first. “Arbitrage” is the practice of taking advantage of a price difference between two or more markets: striking a combination of matching deals that capitalize on the imbalance, the profit being the difference between the market prices. An arbitrage is a transaction that involves no negative cash flow at any probabilistic or temporal state and a positive cash flow in at least one state; in simple terms, it is the possibility, *ideally*, of a risk-free profit at zero cost. In practice, there are always risks in arbitrage, sometimes minor (such as fluctuation of prices decreasing profit margins) and sometimes major (such as devaluation of a currency or derivative). In most ideal models, an arbitrage involves taking advantage of differences in price of a *single* asset or *identical* cash-flows.

Now, if arbitrage can be modeled analogously to quantum tunneling in physics, one might expect features analogous to those found in quantum tunneling, which dramatically exhibits the character of quantum phenomena. Haven and Khrennikov are primarily concerned with the use of Q-models in predicting the probabilities involved, by QP/QS principle (accompanied by the non-additive calculus of probabilities), rather than with the QD and the RWR, or proto-RWR, principles. They do, however, offer some considerations concerning discreteness:

> We believe that the equivalent of quantum discreteness in this paper corresponds to the idea that each act of arbitrage is a discrete event corresponding to the detection of a quantum system after it passed …the barrier. In reality arbitrage opportunities do not occur on a continuous time scale. They appear at discrete time spots and often experience very short lives. We would like to argue that it is the tunneling effect which is closely associated to the occurrence of arbitrage. …We also mentioned the wave function in the discussion above, and quantum discreteness is narrowly linked with quantum probabilities. [33, p. 4095]

This view at least allows for an interpretation of the phenomenon of arbitrage in terms of the QD and the RWR principles, even if it does not require it. Haven and Khrennikov, while, again, allowing for the applicability of the QD principle, do not appear to subscribe to the RWR principle, or even to the proto-RWR principle[16]. In effect, however, they follow the proto-

RWR principle, insofar as they are not concerned with representing *how* arbitrage actually occurs, any more than Heisenberg was concerned with representing the behavior of the electron in the hydrogen atom in deriving his formalism. They are only concerned with predicting the probabilities or statistics of future events of arbitrage.

Thus, situations governed the QD, QP/QS, and RWR (or proto-RWR) principles are possible in economics, psychology, and decision science, and just as in quantum mechanics, they may allow for either a statistical or Bayesian view of the Q-model used. When finite-dimensional Q-models (dealing with discrete variables, such a spin) are used, as they often are in these fields, one can also consider the application of the principles of quantum information theory. While I cannot address the subject in detail, the operational framework, used in this field, merits a brief detour. This framework allows one to arrive at Q-models in a more rigorous and first-principle-like way, by using the rules governing the structure of operational devices, "circuits," via recent work on monoidal categories and linear logic [13–15, 34].

According to Chiribella et al.: "The operational-probabilistic framework combines the operational language of circuits with the toolbox of probability theory: on the one hand experiments are described by circuits resulting from the connection of physical devices, on the other hand each device in the circuit can have classical outcomes and the theory provides the probability distribution of outcomes when the devices are connected to form closed circuits (that is, circuits that start with a preparation and end with a measurement)" [13, p. 3]. A circuit is an arrangement of measuring instruments capable of quantum measurements and predictions, which are, again, probabilistic or statistical, and sometimes, as in the EPR type of experiments, are correlated, which gives a circuit a very specific architecture, corresponding only to quantum but not classical experiments. A realist representation of a circuit is possible because a circuit is described by classical physics, even though it interacts with quantum objects, and thus has a quantum stratum, enabling this interaction. Hence, the information obtained by means of a circuit is physically classical, too, but the architecture and mode of transmission of this information is quantum: they cannot be generated by a classical process.

As discussed earlier, Heisenberg found the formalism of quantum mechanics by adopting, in addition to the QD, QP/QS, and proto-RWR principles, the mathematical correspondence principle and, by the latter principle, using the equations of classical mechanics, while changing the variables in these equations. This principle was not exactly the first principle. In particular, it depended on formally adopting the equations of classical mechanics, while one might prefer these equations to be a consequence of fundamental quantum principles. Heisenberg's variables were new, which was his great discovery. But they were new more of a guess, a logical guess, fitting the probabilities of transitions between the energy levels of the electron in the hydrogen atom he worked with. In the operational framework, one derives finite-dimensional quantum theory in a more first-principle-like way, in particular, independently of classical mechanics (which does not exist for discrete variables, such as spin). This derivation is made possible by applying the rules that define the operational language of circuits, as the language of monoidal categories and linear logic, and thus giving a mathematical structure to operational circuits themselves and thus, in effect, to measuring instruments [13, p. 4, 33]. These rules are *more empirical*, but they are *not completely empirical* (which no rules may ever be), because circuits *are given* a mathematical structure, from which the mathematical architecture of the theory emerges[17]. The resulting formalism is equivalent to the standard Hilbert-space formalism. As in Heisenberg, one only deals with "mathematical representations" providing the probabilities or statistics of the outcomes of discrete quantum experiments, in accord with the QD and QP/QS principles, without providing a representation of quantum processes themselves, in accord with the RWR principle.

In the areas of social science, which concerns human subjects, establishing the mathematical architecture for such "circuits" is a formidable task. However, given important recent work along the lines of category theory beyond physics [e.g., 35], this approach may prove to be viable in enabling a principle approach in Q-modeling outside physics[18].

Q-Theories as Rational Theories of the Irrational

As indicated earlier, while the main reasons for using Q-models in psychology, economics, and decision science are due to the quantum-like nature or calculus of the probabilities associated with predicting certain phenomena, the underlying dynamics of the cognitive or psychological processes leading to each such phenomenon individually might, in principle, be causal or partially causal. This dynamics might also not be causal, especially given the quantum (non-additive) character of the probabilities involved. If it is causal or partially causal, then, unlike quantum processes, *in non-realist interpretations*, an analysis of these psychological processes may be possible, rather than "*in principle* excluded" [3, v. 2, p. 62]. This is because one might expect psychological, social, or economic reasons shaping these situations, and one of the tasks of analyzing them to explain these reasons, an imperative that is hard to avoid, as is clearly apparent in Tversky and Kahneman's articles [26, 37] or in Pothos and Buseymeyer's survey [1].

Psychological, social, or economic research using Q-models may renounce this task, especially in statistical analysis, thus in effect assuming a form of proto-RWR principle, akin to that used by Heisenberg. Even in this case, however, the question would still arise to what degree the QP/QS, QD, and (strictly) RWR principles, or the principles of quantum information theory, could apply in these fields, in particular in considering individual situations. As explained earlier, in quantum mechanics, in non-realist interpretations, the latter could either be treated on Bayesian lines or, in statistical interpretations, assumed to be random, which assumption would, again, be difficult in the fields in question at the moment. Some considerations of discreteness are unavoidable because, as noted, probability has an irreducibly futural and discrete character by dealing with estimates concerning discrete future events.

It is a more complex question whether one can renounce, as one does in quantum mechanics, in non-realist interpretations, considering or even assuming the existence of continuous processes connecting these events. I would surmise that such may be the case and that our brains may work, at least sometimes, in accordance with the QD, the QP/QS, and the RWR principles. This means they would not be relying on and calculating hidden causality connecting events but would instead functions by relying on the quantum-like workings of probabilities and correlations. This type of brain functioning would define what may be called a Bayesian Q-brain, which would require the corresponding Bayesian models. Importantly, however, this kind of Bayesian brain is fundamentally different from rational Bayesian agents, associated with the term Bayesian in cognitive psychology. Indeed, Q-models there are in part advanced in these fields against this concept of human agency. A Bayesian Q-brain need not always function "rationally," at least, not in accordance with any single concept of rationality. A corresponding Bayesian Q-model, if possible, would allow one to predict the outcomes of decisions governed by the brain processes of the individual subjects involved without having, even conjecturally, a full access to these processes, by the RWR principle. Nor do those who make these decisions have this access: these processes are unconscious, and, if one assumes the RWR principle, this part of the unconscious is not causal or "rational" (in its own way), as S. Freud, for example, saw it [38]. Freud's thinking on this point was, however, ultimately more complex, even if against his own grain.

It is instructive to return, in this context, to Bohr's invocation of "an irrational element," in the passage cited above and repeated elsewhere in his writings. The idea and even the language of irrationality have often been seen as problematic by Bohr's critics and even by some of his advocates. I would argue this assessment to be a result of misunderstanding Bohr's meaning. This "irrationality" is not any "irrationality" of quantum mechanics, which Bohr saw as a rational theory, a "rational quantum mechanics," and argued for its rational character throughout his writing (e.g., 3, v. 1, p. 48; 3, v. 2, p. 63). However, he did see it as a rational theory of something—the nature of quantum objects and processes—that is inaccessible to rational thinking, or at least to a rational representation. If, as he says, "the quantum of action [*h*], which appears as *an irrational element* from the point of view of the classical mechanical physics," it only means that cannot be rationally incorporated into the latter [31, p. 458].

Tversky and Kahneman's and related arguments are, too, sometimes seen as related to "irrational" elements in decision-making. This decision-making replaces purportedly "rational" Bayesian agents with at least partially "irrational" Bayesian agents. The "rational" Bayesian agents, as explained above, use probabilistic reasoning subject to updating their estimates on the basis of new information (which defines the Bayesian approach to probability). The irrationality of "irrational" Bayesian agents may be divided into three main, sometimes overlapping, types. The first type is in effect a form of rationality. This rationality is, however, different from rationality presumed to be dominant in the class of situations considered, say, the rationality of maximizing one's monetary benefits. In addition, this alternative rationality may be unconscious. The second type of irrationality refers to something that could be explained. However, it defies explaining it as anything assumed to be rational, say, as a form of rational behavior, beforehand. This irrationality may, upon further analysis, reveal itself to be the irrationality of the first type, but it may also be an alternative form of rationality[19]. Finally, the third type of irrationality is that invoked by Bohr: a realist theory cannot incorporate it in its handling of the corresponding phenomena, while a non-realist Q-model or theory can make it part of its probabilistically predictive scheme without explaining it. In this way, QD, QP (or, if averaging is possible QS), and RWR principles can be brought together in this domain.

There is yet another possibility, which leads to a different type of models or theories, conforming to the QD, QP (but not QS), and RWR principles. I shall call such models or theories singularized probabilistic (SP) models or theories, keeping in mind their non-realist, RWR-principle-based, character. Realist SP models are possible, but I shall not be concerned with them. SP-models may also be time-dependent (TDSP). Such models can only be briefly sketched here in conceptual and somewhat abstract terms, but their possibility is intriguing. SP- or TDSP-models need not be mathematically related to Q-models, but they might be, given the shared principles in which they are based.

Singularized Probabilistic (SP) Theories and Models

Let us recall that, as reflected in the complementarity principle, in quantum mechanics there is no single, uniform physical law applicable to quantum behavior in all contexts, while the same mathematical formalism or model can be used in all contexts. Depending on whether an interpretation is statistical or (Bayesian) probabilistic, the individual quantum behavior is either assumed to be random or to be subject to the probabilistic law, the application of which is defined by the context. By contrast, in the case an SP-model or theory, the following situation obtains. While, as in quantum physics, there is no single uniform physics law, realist or not, each individual behavior obeys its own singular *law*, defined by its own mathematical model, rather than conforms to one or another contextual probabilistic or statistical law, from a (determinable) set of such laws determined by the theory, using a single mathematical model. Under the RWR principle, assumed here for SP-models, such a model still does not represent the reality of the ultimate processes considered, which makes the absence of not only determinism but also causality automatic, just as in quantum mechanics under the RWR principle. One cannot, however, any longer adopt a statistical view, which assumed multiplicities of events that could be averaged (in quantum mechanics, contextually). In each case, only a Bayesian view of the corresponding (unique) model is possible. Such individual laws and accompanying mathematical models may also be changing in time, a change observed each time a new observation occurs. If so, the corresponding model or theory becomes time dependent, TDSP.

The concept of an SP and especially a TDSP model or theory is a radical idea, to my knowledge, rarely, if ever, entertained, at least in science[20]. Indeed, it is not clear whether such theories and, especially, the mathematical models defined by them are scientifically viable, particularly if the corresponding mathematical laws are assumed to be changing in time, possibly on small scales. For an effective scientific practice to be possible, one might need regularities beyond those found in each singular situation, for which a mathematical model, unique to it, would be introduced, say, in order to predict the outcome of events. Such changes of laws and models could, in principle, be governed mathematically, have an overall mathematical model. Thus, one could have a set of models mathematically parameterized so as to allow one to use them for different individual situations and to adjust them to make effective predictions in all of these situations. If not, then each case would require its own mathematical model. Would mathematical-experimental sciences, as they are practiced now, still be possible, then?

Furthermore, there might, in a given domain, be individual cases the character of which will defeat our attempt to treat them by mathematical means. Indeed, this is already so in the case individual quantum processes if one adopts a statistical view, according to which each individual process is random, beyond the law. Now, however, there would not be statistical regularities, of the type found in quantum physics, applicable to multiplicities of repeatable cases (handled, moreover, by the same model, even if contextually), because there would be no repeatable cases in any meaningful sense. There would be neither statistical averaging, nor individual mathematical probabilistic treatment. This situation may be more familiar in literature, which is concerned with the particular or the singular, for example, with a unique life history of a novel's protagonist. One also encounters this singularity or uniqueness in life itself. Such histories resist and even preclude statistical averaging, again, allowed by, otherwise equally unique, histories (which cannot be thought of as classical trajectories of motion) of individual quantum objects, as well as mathematical handling. But they may become, at least outside physics, perhaps especially, in psychology (which often deals with the same human conditions as literature), part of science, a science that will combine science and non-science, or at least mathematical, both of the more standard or the SP/TDSP type, and nonmathematical modeling. Indeed, as just indicated, the SD/TDSP-modeling already poses complexities for scientific practice. Could this situation also emerge in physics, for example, in dealing with quantum gravity? This is not inconceivable. If it does, it will not end mathematical modeling in physics or, again, beyond, or the mathematical-experimental character of modern science, which has defined it beginning with Galileo. It might, however, change both, just as it happened in the case of quantum theory, which not only led to a revolutionary transformation—physical, mathematical, and philosophical—of physics itself but also opened new possibilities for scientific thinking and mathematical modeling beyond physics.

CONFLICT OF INTEREST STATEMENT

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ACKNOWLEDGMENTS

I would like to thank Mauro G. D'Ariano, Emmanuel Haven, Gregg Jaeger, and Andrei Khrennikov for helpful discussions concerning the subjects considered in this article. I am grateful to both readers of the article for their constructive criticisms, especially one of these readers, who made helpful specific suggestions for revisions and who directed my attention to Robert Spekkens's article "Evidence for the epistemic view of quantum states: A toy theory."

FOOTNOTES

1. ^I shall only discuss the standard quantum mechanics or quantum field theory, bypassing alternative theories of quantum phenomena, such as Bohmian theories, which are sometimes used in mathematical modeling outside physics, but which would require a separate consideration. By "quantum phenomena" I refer to those physical phenomena in considering which Planck's constant, *h*, must be taken into account, and by "quantum objects" (thus different from quantum phenomena) to those entities in nature that are responsible for the appearance of quantum phenomena, manifested in measuring instruments involved in quantum experiments or in certain natural phenomena.

2. ^The discussion to follow in part builds on two previous articles [5, 6], but only in part: overall the present argument is different, especially (but not exclusively) by virtue of considering SP-models.

3. ^I bypass more technical definitions, found in standard texts and reference sources.

4. ^There are alternative formalisms, such as those in terms of C*-algebras

or more recently category theory, thus far, all mathematically equivalent to the Hilbert-space formalism.

5. ^The designation "the spirit of Copenhagen" is preferable to a more common "the Copenhagen interpretation," because there is no single Copenhagen interpretation.

6. ^That does not mean that an alternative way of doing so, for example, by bypassing amplitudes or by using some an alternative formalism (not mathematically equivalent to the standard one) is impossible.

7. ^Causality is, thus, an ontological category, characterizing the nature of reality. It proceeds by connecting a cause (an event, phenomenon, a state of a system, or force) to an effect, while the *principle* of causality connects an event to a cause. Determinism is assumed here to be an epistemological category. It designates our ability to predict the state of a system (ideally) exactly at any moment of time once we know its state at a given moment of time. In classical mechanics (which deals with a small number of objects), causality and determinism coincide. Once a classical system is large, one can no longer predict its causal behavior exactly. In other words, a system may be causal without our theory of its behavior being deterministic, as is the case, for example, in classical statistical physics or chaos theory. Causal influences are generally, although not always, assumed to propagate from past or present towards future. Relativity theory further precludes the propagation of physical influences faster than the speed of light in a vacuum, *c*. Principle theories do not require causality, which is, again, difficult to assume in quantum physics without, however, violating relativity or more generally the principle of locality, which requires that all physical influences are local (still under the assumption that they cannot, locally, propagate faster than *c*).

8. ^Among the key earlier approaches are [16], Fuchs's work, which "mutated" to the program of quantum Bayesianism or QBism [17], and [18].

9. ^Randomness may be defined by this impossibility. This concept of randomness is not ontological, because one cannot ascertain the *reality* of this randomness, but *epistemological*. It is ultimately a matter of assumption or belief, practically justified in a given interpretation.

10. ^I have discussed the subject, also in relation to complementarity, in Plotnitsky (9, pp. 136–54). These connections also bring in a related (EPR-correlation) concept, "contextuality." This concept plays a significant role in Q-modeling beyond physics [1, pp. 363–5, 21].

11. ^I have discussed the role of principles of quantum information theory beyond physics in Plotnitsky [6].

12. ^I also refer to these works for more detailed discussions of the ways in which Q-models are used in these fields.

13. ^As noted earlier, this does not mean that such probabilities could not be predicted by means of alternative models even in quantum physics.

14. ^Complementarity has received some attention outside physics, beginning with Bohr’s own (tentative) suggestions. Inspired by Bohr and others did propose using the concept in philosophy, biology, and psychology. See Plotnitsky [28, pp. 158–66] and [29].

15. ^There are several recent arguments for such connections, most prominent of which is arguably that by Penrose [32] and developed in several subsequent studies. The model itself that Penrose has in mind is, thus far, only mathematically conjectured, following certain approaches to quantum gravity.

16. ^As indicated earlier, elsewhere Khrennikov argued for a classical-like model at the ultimate level of the constitution of nature in physics [30].

17. ^See also Plotnitsky [9, pp. 248–58] and Hardy [15].

18. ^See also a recent approach to representing sensation-perception dynamics in terms of quantum-like mental instruments, which are akin to “circuits,” in Khrennikov [36].

19. ^Some might still see, as Freud did, this “irrationality” as a form of unconscious “rationality.” Once again, however, Freud, against his own grain, could not ultimately avoid giving the unconscious a stratum that is beyond representation, if not conception.

20. ^Something akin to this possibility has been suggested in physics in Ungar and Smolin [39], but in a different context and based it on a very different set of principles than those adopted here, most especially because, as against the present argument, they assume realism and causality.

REFERENCES

1. Pothos E, Busemeyer JR. Can quantum probability provide a new direction for cognitive modeling? *Behav Brain Sci.* (2013) **36**:255. doi: 10.1017/S0140525X12001525
2. Haven E, Khrennikov A. *Quantum Social Science.* Cambridge: Cambridge University Press (2013).
3. Bohr N. *The Philosophical Writings of Niels Bohr*, Vol. 3. Woodbridge, CT: Ox Bow Press (1987).
4. Heisenberg W. Quantum-theoretical re-interpretation of kinematical and mechanical relations. In: Van der Waerden BL, editor, *Sources of Quantum Mechanics.* New York, NY: Dover (1925/1968). p. 261–77.
5. Plotnitsky A. The visualizable, the representable, and the inconceivable: realist and non-realist mathematical models in physics and beyond. *Philos Trans R Soc A* (2016) **374**:20150101. doi: 10.1098/rsta.2015.0101
6. Plotnitsky A. Quantum principles and mathematical models in physics and beyond. In: Haven E, Khrennikov A, editors. *The Palgrave Book of Quantum Models in Social Science.* London: Palgrave-MacMillan (2016). p. 335–57.
7. Von Neumann J. *Mathematical Foundations of Quantum Mechanics*, Transl. by Beyer, RT. Princeton, NJ: Princeton University Press (1932/1983).
8. Einstein A. What is the Theory of Relativity. The London Times, 28 November, 1919. In *Einstein, A. Ideas and Opinions*. New York, NY: Bonanza Books (1919/1954).
9. Plotnitsky A. *The Principles of Quantum Theory, from Planck's Quanta to the Higgs Boson: The Nature of Quantum Reality and the Spirit of Copenhagen*. New York, NY: Springer/Nature (2016).
10. Heisenberg W. *The Physical Principles of the Quantum Theory*. Transl. by Eckhart K, Hoyt FC, New York, NY: Dover (1930:1949).
11. Plotnitsky A, Khrennikov A. Reality without realism: on the ontological and epistemological architecture of quantum mechanics. *arXiv:1502.06310* (2015). doi: 10.1007/s10701-015-9942-1
12. Kant I. *Critique of Pure Reason*, Transl. by Guyer P, Wood, AW. Cambridge: Cambridge University Press (1997).

13. Chiribella G, D'Ariano GM, Perinotti P. Informational derivation of quantum theory. *Phys Rev A* (2011) **84**:012311. doi: 10.1103/PhysRevA.84.012311

14. D'Ariano GM, Chiribella G, Perinotti P. *Quantum Theory from First Principles: An Informational Approach.* Cambridge: Cambridge University Press (2017).

15. Hardy L. Foliable operational structures for general probabilistic theory. In: Halvorson H, editor. *Deep Beauty: Understanding the Quantum World through Mathematical Innovation*. Cambridge: Cambridge University Press (2011). p. 409–42. doi: 10.1017/CBO9780511976971.013

16. Zeilinger A. A foundational principle for quantum mechanics. *Found. Phys.* (1999) **29**:631–43. doi: 10.1023/A:1018820410908

17. Fuchs CA, Mermin ND, Schack R. An introduction to QBism with an application to the locality of quantum mechanics. *Am J Phys.* (2014) **82**:749. doi: 10.1119/1.4874855

18. Hardy L. Quantum mechanics from five reasonable axioms. *arXiv:quant-ph/0101012* (2001).

19. Pauli W. Writings on physics and philosophy. Berlin: Springer (1994).

20. Bohr N. Can quantum-mechanical description of physical reality be considered complete? *Phys Rev.* (1935) **48**:696.

21. Dzhafarov E, Jordan SR, Zhang R, Cervantes V. (editors). *Reality, Contextuaity, and Probability in Quantumtheory and Beyond.* Singapore: World Scientific (2016). p. 93–138.

22. Einstein A, Podolsky B, Rosen N. Can quantum-mechanical description of physical reality beconsidered complete? In: Wheeler JA, Zurek WH, editors, *Quantum Theory and Measurement*, Princeton, NJ: Princeton University Press (1935/1983). p. 138–41; 152–67.

23. Schrödinger E. The present situation in quantum mechanics. In Wheeler JA and Zurek, WH editors. *Quantum Theory and Measurement*, Princeton, NJ: Princeton University Press (1935/1983), 152–67.

24. Hardy L. Towards quantum gravity: a framework for probabilistic theories with non-fixed causal structure. *J Phys.* (2007) **A40**:3081–99. doi: 10.1088/1751-8113/40/12/S12

25. D'Ariano GM, Perinotti P. Derivation of the Dirac equation from principles of information processing. *Phys Rev A* (2014) **90**:062106. doi: 10.1103/PhysRevA.90.062106

26. Tversky A, Kahneman D. Availability: a heuristic for judging frequency and probability. *Cogn Psychol.* (1973) **5**:207. doi: 10.1016/0010-0285(73)90033-9
27. Spekkens R. Evidence for the epistemic view of quantum states: a toy theory. *Phys Rev A* (2007) **75**:032110. doi: 10.1103/PhysRevA.75.032110
28. Plotnitsky A. *Niels Bohr and Complementarity: An Introduction*. New York, NY: Springer (2012).
29. Wang Z, Busemeyer J. Reintroducing the concept of complementarity into psychology. *Front Psychol.* (2015) **6**:1822. doi: 10.3389/fpsyg.2015.01822
30. Khrennikov A. Quantum probabilities and violation of CHSH-inequality from classical random signals and threshold type detection scheme. *Progr. Theor. Phys.* (2012) **128**:31. doi: 10.1143/PTP.128.31
31. Bohr N. Life and light. *Nature* (1931) **131**:458.
32. Penrose R. *The Emperor's New Mind.* Oxford: Oxford University Press (1995).
33. Haven E, Khrennikov A. Quantum-like tunnelling and levels of arbitrage. *Int J Theor Phys.* (2013). **52**:4083. doi: 10.1007/s10773-013-1722-0
34. Coecke B. Quantum picturalism. *Contemp Phys.* (2009) **51**:59–83. doi: 10.1080/00107510903257624
35. Abramsky S, Brandenburger A. The sheaf-theoretic structure of non-locality and contextuality. *arXiv:1102.0264 [quant-ph]* (2011). doi: 10.1088/1367-2630/13/11/113036
36. Khrennikov A. Quantum-like modeling of cognition. *Front Phys.* (2015) **22**:77. doi: 10.3389/fphy.2015.00077
37. Tversky A, Kahneman D. Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychol. Rev.* (1983) **90**:293. doi: 10.1037/0033-295X.90.4.293
38. Freud S. *General Psychological Theory: Papers on Metapsychology*. New York, NY: Touchstone (2008).
39. Ungar RM, Smolin L. *The Singular Universe and the Reality of Time: A Proposal in Natural Philosophy.* Cambridge: Cambridge University Press (2014).

Chapter

3

THE NATURE AND MATHEMATICAL BASIS FOR MATERIAL STABILITY IN THE CHEMICAL AND BIOLOGICAL WORLDS

Robert Pascal[1] and Addy Pross[2]

[1]Institut des Biomolécules Max Mousseron, UMR5247 CNRS–Universités Montpellier 1 & Montpellier 2, CC17006, Place E. BataillonCC17006, Place E. Bataillon, Montpellier F-34095, France.
[2]Department of Chemistry, Ben-Gurion University of the Negev, Be'er Sheva 84105, Israel.

ABSTRACT

The conceptual divide separating the physical and biological sciences continues to challenge modern science. In this perspective it is proposed that the two sciences can be directly connected through the fundamental concept of stability. Physicochemical stability is shown to have a logical, rather than an empirical basis, and able to manifest itself in two distinct and often contrary ways, one thermodynamic, reflecting energetic considerations, and

Citation: (APA): Pascal, R., & Pross, A. (2014). The nature and mathematical basis for material stability in the chemical and biological worlds. Journal of Systems Chemistry, 5(1), 1-8. (8 pages).

the other kinetic, reflecting time/persistence considerations. Each stability kind is shown to rest on a particular mathematical truism. Thermodynamic stability, the energetic expression, has a probabilistic/statistical basis due to Boltzmann, and leads to the Second Law of Thermodynamics. Dynamic kinetic stability (DKS), the time/persistence expression, is attributed to the stability associated with persistent replicating systems, and derives from the mathematics of exponential growth. The existence of two distinct stability kinds, each mathematically-based, leads to two distinct organizational forms of matter, animate and inanimate. That understanding offers insight into the reasons for the observation of just those two organizational forms, their different material characteristics, and provides a logical basis for understanding the nature of chemical and biological transformations, both within, and between, the two forms.

Keywords: Thermodynamic stability, Dynamic kinetic stability, Origin of life, Autocatalysis, Abiogenesis

REPORT

Introduction

The question of why matter exists in two starkly distinct material categories – living and non-living – has puzzled mankind for millennia. Our understanding of the living world was, of course, revolutionized through Darwin’s landmark ideas of natural selection and common descent [1], and that understanding has been both deepened and extended by the dramatic advances in molecular biology over the past 60 years. Yet, despite those insights into the workings of life’s molecular machinery, the perennial and more general question – how living and non-living relate to one another - continues to elude us. Why does matter exist in two distinctly different organizational forms? Why not just one (or three)? Is there some basis in the laws of nature which would make the existence of two distinct forms expected, even inevitable? And with regard to those living forms, what would the material prerequisites for a generalized living system be? What is it about living systems that makes their properties so different to those of non-living ones? And, finally, of the two organizational forms, which is naturally preferred, and why? Such questions are not merely theoretical. Being able to understand the relationship between those two material forms would be a prerequisite for understanding how, at least in principle, one would go about transforming one form to the other. Needless to say, it is trivially easy to convert animate to inanimate, yet how tantalizingly difficult to proceed in the opposite direction. And the problem does not lie with a particular technical difficulty in one particular step along the way. The problem is much deeper. Despite those 60 years of mechanistic advances in molecular biology, the essence of the living state continues to elude us. As Kauffman [2] put it recently: "…we know many of the parts and many of the processes. But what makes a cell alive is still not clear to us. The center is still mysterious".

Then there is the perennial origin of life problem, fascinating in its own right [3–10]. How and why did this distinct organizational form of matter emerge in the first place? There is a broad scientific consensus that life on earth emerged from inanimate beginnings, and at first sight the origin of life question might appear unrelated to the other life questions, but that impression is false. While the *historic* path linking inanimate to animate will likely remain buried in the mists of time [4, 5], the question how life *could have* emerged from inanimate matter is intimately connected to the means by which one would go about synthesizing a living system. The two questions constitute two sides of the one coin; solve one and you've taken a major step toward solving the other. By what process and based on what physicochemical principles was it possible for matter to be transformed from the relatively well-understood inanimate state into that extraordinarily complex and thermodynamically unstable animate state. Certainly from a purely thermodynamic perspective such a transition would seem to be spectacularly improbable [11–13].

But the confusion surrounding the living state goes further. Consider the extraordinary characteristics of living systems, which seem to defy simple chemical explanation. Whereas chemistry is readily able to explain the characteristics of inanimate materials – why ice is hard, why metals conduct electricity, why helium is chemically inert and a gas at room temperature, and so on, living matter's strikingly different properties do not lend themselves to that kind of chemical approach [5]. Take the concept of function, for example, one that permeates all of biology. There is an entire area of biology, functional biology, which continually asks 'what is its function, how does it work' type questions, as the purposeful (teleonomic) character of living systems is empirically irrefutable. But how is it at all possible for any biological system, ultimately just chemical in its composition, able to express the characteristic we denote by the term function, which of necessity also signifies purpose? In the inanimate world, in the world of 'regular' chemistry there is no function, no purpose. That, after all, was the essence of the scientific revolution of the 17th century. Teleology was banished from the scientific discourse [14]. How then are biological systems, entirely material in nature, able to manifest function? In the context of the origin of life question, the issue can be rephrased as: how could biological function have emerged from an inanimate world devoid of function? How could biological systems have acquired properties seemingly inconsistent with nature's objective character?

And, finally, to the heart of the problem – the nature of biological organization, as it is within that special kind of organization that the essence of the animate - inanimate distinction presumably needs to be sought. How can biological organization as a phenomenon, characterized by inordinate dynamic (homeostatic) complexity and quite distinct to the static complexity of the inanimate world, be understood, an issue glossed over in the neo-Darwinian view? In response to these probing questions, directed toward clarifying the nature of the chemistry–biology connection, modern biology has taken a defensive posture and battened down the hatches. The unstated but implicit message in contemporary biology appears to be: yes, there are innumerable apparent contradictions when biology is directly confronted with physics and chemistry [2, 11–14]. However, since the physical sciences have not provided biology with the appropriate conceptual and methodological tools for resolving these contradictions, biology can avoid these awkward questions by fencing itself off from the physical sciences. The result: biology of the 20th century has been overtaken by an 'autonomy of biology' philosophy, one openly endorsed by Ernst Mayr [15], one of the leading evolutionary biologists of the 20th century, whereby biology is treated as a disparate science governed by a separate philosophy to the one underpinning the physical sciences. There are two kinds of matter, inanimate and animate, the physical sciences deal with the former, the biological sciences deal with the latter, and that's that! Thus in the neo-Darwinian perspective, biology's essence resides in the genome and the information coded therein, and from this vantage point, questions of origins – how did genomic information come about, how does information emerge from non-information – are conveniently brushed aside. But if, as is now widely believed, on planet Earth some 3.5 to 4 billion years ago chemistry *did* become biology [3–10], then the two subjects must in some sense be one, making it clear that the *historical* merging that took place in the distant past must be accompanied by a corresponding *conceptual* merging. The dissonance that continues to radiate from the glaring contradictions inherent in the biological and physical world views gives no respite.

In several recent papers the authors, together and separately, have attempted to address these questions, to help bridge the chemistry-biology gap, through the characterization of a unique stability kind in nature, termed *dynamic kinetic stability (DKS)*[4, 5, 12, 13, 16–19]. This stability kind, quite distinct to traditional thermodynamic stability, applies to systems able to maintain a presence over time through a *process of self-replication.* Thus replicative stability, whether chemical or biological, is able to lead to a distinct and separate organizational state - a *kinetic state of matter*, thereby offering a physicochemical framework for relating biological systems to replicative chemical ones. Through that approach several of the puzzling issues regarding the relationship of animate to inanimate appear resolvable - the continuity and underlying unity of chemical and biological evolution [4, 5, 13, 16, 17], its physicochemical characterization [12, 13], the source of life's functional nature [20], its extraordinary and distinct kind of complexity [20], its metabolic (energy-consuming) character [12, 13, 21], to mention central ones.

In this paper we wish to refine and extend the argument regarding the nature of stability in the physicochemical world by pointing out that the concept of stability can be logically defined, and that the two stability kinds that govern physicochemical processes in the inanimate and animate worlds - thermodynamic stability and DKS respectively, are not arbitrary and empirically derived, but have a mathematical basis. Through an understanding of that basis for the two respective material forms, insight is offered into *why* there are two organizational material forms in nature, why the animate state once formed is inherently preferred over the inanimate state, and a clearer understanding as to why the origin of life question (meaning that initial transformation of inanimate to animate) is continuing to prove so intractable.

DISCUSSION

The Nature of Stability

Let us begin by considering what the term 'stability' actually means within a physicochemical context. Our starting point is the observation that matter is not immutable, that the material world is undergoing continual change. That statement is, of course, empirically self-evident. Wherever one looks in the world, one can discern change, both physical and chemical. Significantly, however the *direction* of change can be summed up by the qualitative statement: *all physicochemical systems tend from less stable to more stable forms.* This general statement, not normally discussed (though alluded to by Dawkins [22]) may be thought of as axiomatic. It is inherent in the definition of the term 'stable' – unchanging, persistent over time. The statement is axiomatic in the sense that it is tautological to state that changing systems change, whereas unchanging ones do not. But within that tautology lies hidden a deeper truth. It is implicit that if matter does tend to undergo change, over time that change will necessarily be in the direction *from systems more susceptible to change* (i.e., less stable/persistent forms), *toward systems that are less susceptible to change* (i.e., more stable/persistent forms). Indeed, even if at some point the system were to change in the reverse direction, namely, from a relatively unstable form to a form that is even *less* stable, then, by definition, that change would be transitory, as the system would change yet again (by definition), until reaching a more stable form, one less susceptible to change (in the present context change is understood as one that is spontaneous, without the work or action of an external agent). Thus the direction of change is implicit in the very definition of stability. Stability is logically rather than empirically defined.

Note, however, that the above discussion has to an extent switched the concept of stability, normally associated with a system's *energy* to one that focuses on the system's *persistence, i.e.,* its stability over time. The question then arises: is the stability of a system manifest through its energetic properties, or by its unchanging character over time, regardless of energetic considerations? As we will now discuss, stability in its energetic sense necessarily leads to stability in its time (persistence) sense, but not all systems that are stable in a time (persistent) sense, are necessarily stable in an energetic sense.

Existence of Two Stability Kinds

The concept of stability as part of our consideration of the physicochemical world is of course fundamental and well-established, but the focus tends to be on just one kind of stability – thermodynamic stability, a stability kind associated with a system's energy Accordingly, the general 'less stable to more stable' rule described earlier is expressed by the Second Law of Thermodynamics, a law which formalizes the stability concept by providing a means for its quantification. And being the rule that specifies the direction of all irreversible processes, there is no doubting the Second Law's status as one of the fundamental tenets of physics and chemistry, one that operates at both macroscopic and cosmological levels. Of course the Second Law is not merely an empirical law, even though it was initially formulated as such by Clausius and Kelvin, but, as Boltzmann pointed out over a century ago, there is a mathematical logic, a mathematical underpinning to the law with the concept of entropy as its centerpiece [23]. Thus the most stable macrostate of a system (in energy terms) is the one described by the largest number of contributing microstates and the Second Law formulation of *'less stable to more stable'* can be restated more insightfully as *'less probable to more probable'*. Indeed, it is that inherent mathematical/statistical logic that elevates the broader concept of stability from one that is merely qualitative to one that is quantitative, thereby giving the law its almost hallowed status as one that is supremely incontestable. Importantly, a system that is stable in this energetic sense will also be stable in a time (persistence) sense. A system that has reached its lowest energy state, the equilibrium state, remains unchanged over time; *energetic stability invariably leads to time stability (persistence).*

Though energetic stability necessarily leads to time stability, the reverse does not necessarily apply. A system may well be stable in a time sense (persistent) without being stable in an energy sense. The familiar concept of *kinetic stability* characterizes that other stability kind, as exemplified by a hydrogen and oxygen gas mixture. Such a mixture is *highly unstable* in an energetic sense (a spark or catalyst will immediately result in water formation), but can be *highly stable* in a time sense – a mixture of the two gases may well persist over long periods of time.

But, as noted earlier, within the biological world as well as parts of the chemical world, an *alternative* kinetic stability kind exists and governs the nature of transformations within that world - DKS, a stability kind associated solely with the replicative world, and distinct to the more familiar static kinetic stability mentioned above. Indeed it is that concept of DKS that can help explain both replicative chemical, as well as biological, phenomena. Accordingly, replicating systems, though unstable in thermodynamic terms, are able to persist over time through continuing self-replication, and so are stable in kinetic terms. They are stable, not because they *do not* react, but because they *do* – to make more of themselves – thereby opening a door to a distinctly different organizational form of matter [4, 5, 13, 16, 18].

But what is the basis for this other stability kind? Is it just empirical, or is there some underlying imperative that enables it to circumvent the probabilistic drive of the Second Law? Does the DKS concept also express some underlying, but alternative mathematical logic, to the probabilistic one? The answer to this last question appears to be *yes*, DKS is also governed by a mathematically-based directive – the enormous kinetic power associated with systems able to undergo exponential growth due to the kinetic character of some (though not all) autocatalytic systems [24, 25]. The central role of autocatalysis in the emergence of life has long been recognized and has been described within different theoretical models [26–28]. And, indeed, it is the kinetic power associated with autocatalysis which initiates the beginning of divergence from a thermodynamically-directed world by the establishment of what is effectively a parallel kinetic world in which systems are found to be dynamic, energy consuming, far-from-equilibrium, and necessarily open to material and energy resources [12, 13]. Let us describe how this comes about.

Once a DKS state does emerge, it turns out that its key reactivity characteristic, its potential ability to evolve, is also governed by that same mathematical directive. Due to the action of the Second Law, a stable DKS system will over time necessarily undergo variation leading to competition between the variants for resources. It has been recognized since Lotka [29] that autocatalytic systems can exhibit a range of complex kinetic behaviors [30–32], but it was Lifson [33] who explicitly pointed out that two competing autocatalysts that exhibit exponential growth and feed off common resources cannot coexist. Solution of the relevant rate equations leads to an unambiguous result – the more stable replicator (in the time/persistent sense) drives the less stable one into extinction.

Of course, for the above evolutionary mechanism to be operative, the DKS system must be inherently evolvable [34, 35]. Thus an autocatalytic network of reactions, such as the one involved in the formose reaction [36], would not satisfy this condition, as it lacks any possibility of evolving toward a state of increased DKS. Similarly, a system involving the autocatalytic production of fatty acids leading to vesicle division [37] would also be unable to satisfy this condition. But once evolvability is present within the system, such as is naturally found in template-based biopolymeric replicating systems, the DKS formulation opens up a mechanism for the stabilization of inherently less stable replicating entities. In fact the drive toward greater DKS can be expected to favor those systems whose evolvability is greater, so that initially weak evolvability will itself likely evolve into stronger evolvability, giving rise to an open-ended evolutionary path [38][a]. The key point is, however, that within the above mentioned constraints, the DKS selection rule - *from DK less stable to DK more stable* – can also be seen to rest on a mathematical truism, the mathematics of exponential growth. The DKS selection rule thereby becomes an additional sub-set of the global selection rule, 'less stable to more stable', discussed earlier.

Given the different mathematical foundation for each of the two stability kinds, it should come as no surprise that the evolutionary process for each of the two material kinds follows different kinetic patterns. For an isolated system (exchanging neither matter nor energy with the environment) the system is directed toward the lowest energy state – the equilibrium state, where entropy is maximal, and the drift toward that state tends to be monotonic. For a persistent DK system, however, governed as it is by divergent and intrinsically non-linear autocatalytic processes, the drift toward its stationary state can result in periodic or even chaotic behavior [29, 39–41]. And being divergent, the system is not directed toward one specific state, but, rather, any number of feasible evolutionary pathways are possible. Moreover, the DKS formulation suggests that, in contrast to an isolated thermodynamic system, where the maximal (energetic) stability of the equilibrium state is achievable, in biological systems maximal stability (in the time/persistent sense) is unachievable. There can be no formal stability maximum in DKS systems given the almost infinite possibilities of variability that the divergent and open system description offers and the nature of stability in its time/persistence facet.

To summarize, whereas thermodynamic stability (for isolated systems) involves a *probabilistic reordering of the existing*, a drive toward entropically measured randomness, and is defined in energy terms, DKS is governed by the *kinetic power of exponential growth* acting on particular replicative systems and is manifest through its persistence over time. It is that kinetic power which both establishes the DK state and then drives it so as to channel that kinetic power most effectively, i.e., to exploit energetic and material resources most efficiently. Thus the empirical observation of an evolutionary process toward enhanced stability within the replicative world (whether replicative chemical or biological) *also* has its roots in a mathematical truism. Indeed DKS may be thought of as a Malthusian stability, in recognition of the contribution of Malthus to the appreciation of the consequences of exponential growth on replicating populations [42], and its subsequent influence on Darwin's formulation of the concept of natural selection. The result – a continually expanding replicative network able to penetrate and exploit most any ecological niche, whether deep under the sea, high in the earth's atmosphere, in polar ice-caps or tropical forests, above ground or miles below the earth's surface - *two mathematically-based stability kinds leading to two distinct material forms.*

Relationship between the Two Stability Kinds

We have attempted to explain why there are just two material categories in nature, as well as the basis of those two categories, so let us now apply that insight to address aspects of the relationship that links those two material kinds. The fact that there are two stability kinds, each underpinned by its particular mathematical logic, means that the corresponding material forms can be expected to exhibit very different characteristics. And indeed they do. Whereas the properties of non-living things are largely explicable in well-established physical and chemical terms, the world of living things has proven resistant to similar characterization. We return to the issue of teleonomy, the term popularised by Monod specifically to describe the behavior of biological systems [14]. All living systems appear to have an agenda, to be goal-directed, as evident in their actions - building a nest, raising young, fighting off predators, and so on. But how can living things, ultimately nothing more than a form of material organization, act in a goal-directed fashion? How does life's unequivocal teleonomic character cohabit with the essence of the modern scientific revolution – nature's objective character?

It turns out that cohabitation need not be contradictory, that nature can be both objective *and* goal-directed. Once it is recognized that change is written into nature's laws, and that nature *is* goal-directed in a way that is *logically prescribed* - toward systems of greater stability, then the existence of two worlds, one living, one non-living, becomes explicable. There are *two* distinct kinds of stabilities in nature, so nature's goal directedness reflects that duality - in the non-living world nature follows the thermodynamic directive (termed by Mayr *teleomatic*[15]), the probabilistic drive toward uniformity, toward so-called heat death, whereas for persistent replicating systems, nature's drive is toward replicative stability (DKS), with its teleological undertones (though consistent with the requirements of the Second Law). Thus nature's goal-directedness with respect to persistent replicating systems, though teleologically tinged, can now be understood *as* *manifesting an aspect of its objective character* – the fundamental drive of *all* material systems toward ever greater stability.

Once the issue of goal-directedness in the biological world is resolved, two of biology's seemingly incompatible bedfellows – stability and complexity – can be harmoniously wedded, and this can be brought about through the mediating concept of function. As one of us has recently described, in the replicating world stability and function are directly related - greater replicative stability is induced through enhanced replicative function [20]. But, as is readily verifiable, there is also a logical connection between function and complexity. Function, of whatever kind, biological or technological, is almost invariably enhanced through complexity. Indeed, to paraphrase Carl Sagan's famous aphorism, one could say: *extraordinary function requires extraordinary complexity,* thereby offering insight into the connection between life's extraordinary functionality and its staggering complexity. But from these two relationships it then follows that (replicative) stability and complexity are also linked – greater complexity is necessary for greater stability. The physical-biological relationality can be summed up by the triad: *stability – function – complexity*, all interconnected and interrelating [20].

As a final point, let us now address a purely material aspect, the issue of material transfer between the two worlds, as evidenced on this planet. First, why was inanimate transformed into animate matter in the first place, i.e., why did life emerge. Second, it is obvious that once life was established on our planet, there has been a continual transfer of matter between the two material forms; living things die and their material form reverts to inanimate, while in the reverse direction, inanimate matter is drawn into the web of life, and thereby transformed into animate matter. But which process is dominant, and why? What can one say about the rates of material transformation in the two directions starting from that moment when earliest life was able to emerge?

The fact that life presumably started off in some limited physical location and expanded rapidly to occupy just about every conceivable planetary niche capable of sustaining life states unambiguously that once a stable and evolvable DKS system emerged on earth the rate of inanimate to animate transformation exceeded the reverse process, i.e., the rate of animate degradation. The fact that this difference is fundamental, not merely incidental, is confirmed by a recent estimate of the ongoing rate of growth of the earth's biomass, ca. 10^{17} g C/year [43]. That rate of growth turns out to be a significant percentage of the earth's total estimated biomass, ca. 10^{18} g C [44]. In other words from the moment that life on earth emerged, there is every indication that for much of the time animate formation exceeded animate degradation. In fact, one might even see in man's attempt to physically explore the universe beyond our planet as an expression of animate matter's tendency to expand into all available niches, to continue life's relentless drive to expand wherever possible.

The reason for the clear imbalance in the rate of animate formation compared its decay, can now be pointed out. Based on the stability kinds involved, the conversion of animate to inanimate – death – is expected to be *slower* than the process of animate formation leading to life. Animate to inanimate is governed by the Second Law, by the more muted directive, the one concerned with material reorganization based on probabilistic considerations, while inanimate to animate is autocatalytic and driven by the kinetic power of exponential growth. Thus once a stable DK system emerges, i.e., once a network of far-from-equilibrium metabolic reactions that is holistically replicative is firmly established, the on-going drive toward greater DKS wins out, and the Second Law directive is circumvented and marginalized. In fact an energy-gathering metabolic capability must be an intrinsic component of the DKS system [12, 13] for the Second Law requirement to be satisfied. Or put another way, once the necessary conditions for life's emergence are met and life is established, the *kinetic* drive toward more life, and more efficient life, overshadows the *thermodynamic* directive toward death, though of course that continuing transformation is conditional on a continuing source of energy. And the environmental consequences of that kinetic imbalance is dramatic and clear to see – life is (effectively) everywhere. The cosmological implications of these simple ideas need further consideration, but the preliminary conclusion seems to be that, provided a continual source of energy is available (most likely fed by nuclear processes in suns), matter will preferentially be driven from inanimate to animate, from non-replicative to replicative, that life will invariably prevail over non-life.

Notwithstanding the above comments, it should also be made clear that the dominance of animate formation over its degradation should not be seen as smoothly monotonic, but rather one that can itself be highly contingent, as is evident in the evolutionary process itself. It is generally believed that in the long evolutionary process toward ever more effective replicating networks there may have been periods of regression as a result of drastic ecological and/or climatic changes, such as the emergence of oxygen as a significant component of the earth's atmosphere [45]. Such an event could likely have led to the destruction of anaerobic life forms that populated the early planet. But the underlying long-term trend is unmistakable – the exponential driving force of living processes overwhelms the mathematically weaker Second Law directive.

The Question of Life's Contingency

The above discussion on life's emergence, and its explosive (and continuing) expansion since its emergence, leads us to the problematic issue of *life's contingency*. In fact it is the issue of contingency which remains the central unresolved dilemma in our attempt to place animate systems squarely within a comprehensive material framework. In order to connect between the inanimate and animate worlds, it is presumed that the life process would have begun with the *contingent* emergence of a persistent and evolvable DK system, even though the likelihood of such a system emerging spontaneously currently remains unknown. So how contingent is life? What materials and reaction conditions would facilitate the emergence of a suitable DKS system? We do not know and we are still far from being able to answer these questions. That is the prime reason we are unable to specify how likely it is for life to exist elsewhere in the universe. But we may be at a turning point. Through recent advances in systems chemistry [46, 47], the path to enlightenment now seems more clearly marked, with preliminary results, both experimental [48] and theoretical [49], suggesting that replicative networks are under certain circumstances able to emerge spontaneously. Thus the immediate goal: the synthesis of stable DK systems so as to enhance our understanding of how DKS systems can be generated, and how readily they can be maintained. The DKS state is a chemically intricate and dynamic entity so its synthesis cannot be assumed to be a trivial one. Theory now needs to give way to experiment, very much in line with Richard Feynman's aphorism: "What I cannot create, I do not understand". And with regard that most intriguing of questions: how likely is it that life exists elsewhere in the universe, paradoxically, it could well be that through experiments conducted on earth, that we may finally reveal the likelihood of life existing elsewhere in the universe. In any case, a prevailing perception that protolife might be created through incorporating some replicating entity and its building blocks within a vesicle-like structure, seems unlikely to be productive, as several of the prerequisites of the DK state would be absent. The evolutionary process by which life was able to undergo such extraordinary complexification can only be understood in the context of exponential replicating systems.

CONCLUSION

This perspective has attempted to demonstrate that through an appreciation of nature's axiomatic drive toward stable/persistent forms, the underlying connection between the two organizational forms of matter – animate and inanimate - can be understood. In simplest terms, nature is able to express its spontaneous drive toward ever increasing stability, not in *one*, but in *two* fundamentally different ways, one based on energetic considerations – thermodynamic stability; the other based on time/persistence considerations – dynamic kinetic stability (DKS), with each leading to a particular manifestation of material organization. That basic reality means that material organization and reactivity take place in two seemingly parallel, yet intersecting, worlds. One hundred years after Ludwig Boltzmann laid down the statistical basis for the Second Law, and two hundred years after Thomas Malthus pointed out the profound consequences of exponential growth on living populations, it is now possible to see that within those two fundamental mathematical truths can be found not just the basis for a dual material world, but also the basis for change both within, and between, those two worlds. Life, in its stupendous diversity and extraordinary complexity, is just the inevitable consequence of mathematical law (exponential growth) operating on very particular replicating chemical systems. The answer to Schrödinger's 'what is life' question may finally be within reach.

ENDNOTES

[a] The ways in which DKS behavior and evolvability could emerge in a far-from-equilibrium system are certainly diverse and the possibility that both of them can appear at the same time cannot be excluded. See ref. [50].

AUTHORS' CONTRIBUTIONS

This paper is the result of extensive discussions and correspondence over time between the two authors. AP wrote an initial draft and RP modified and extended the draft. The final paper reflects the views of both authors. Both authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

The authors thank Prof. Jan Engberts for helpful comments on an early draft and the COST Action CM0703 “Systems chemistry” for providing a fruitful context for scientific exchanges during the realization of this work.

REFERENCES

1. Darwin C: *On the origin of species*. Cambridge, MA: Harvard University Press; 1859.
2. Kauffman SA: *Investigations*. Oxford: Oxford University Press; 2000.
3. Ruiz-Mirazo K, Briones C, de la Escosura A: Prebiotic Systems Chemistry: New Perspectives for the Origins of Life. *Chem Rev* 2014,114(1):285–366. doi:10.1021/cr2004844 10.1021/cr2004844
4. Pross A, Pascal R: The origin of life: what we know, what we can know, what we will never know. *Open Biol* 2013, 3: 120190. doi:10.1098/rsob.120190 10.1098/rsob.120190
5. Pross A: *What is life? How chemistry becomes biology*. Oxford: Oxford University Press; 2012.
6. Forterre P, Gribaldo S: The origin of modern terrestrial life. *HFSP J* 2007, 1: 156–168. doi:10.2976/1.2759103 10.2976/1.2759103
7. Shapiro R: Small molecule interactions were central to the origin of life. *Q Rev Biol* 2006, 81: 105–125. doi:10.1086/506024 10.1086/506024
8. Luisi PL: *The emergence of life: from chemical origins to synthetic biology*. Cambridge, UK: Cambridge University Press; 2006.
9. Popa R: *Between necessity and probability: searching for the definition and origin of life*. Berlin: Springer; 2004.
10. Fry I: *The emergence of life on Earth*. Piscataway, NJ: Rutgers University Press; 2000.
11. Schrödinger E: *What is life? The physical aspects of the living cell*. Cambridge, UK: Cambridge University Press; 1944.
12. Pascal R: Suitable energetic conditions for dynamic chemical complexity and the living state. *J Syst Chem* 2012, 3: 3. doi:10.1186/1759–2208–3-3 10.1186/1759-2208-3-3
13. Pascal R, Pross A, Sutherland JD: Towards an evolutionary theory of the origin of life based on kinetics and thermodynamics. *Open Biol* 2013, 3: 130156. http://dx.doi.org/10.1098/rsob.130156 10.1098/rsob.130156
14. Monod J: *Chance and necessity*. New York: Random; 1972.
15. Mayr E: *Toward a new philosophy of biology*. Cambridge: Harvard University Press; 1988.
16. Pross A: Toward a general theory of evolution: Extending Darwinian theory to inanimate matter. *J Syst Chem* 2011, 2: 1. 10.1186/1759-

2208-2-1

17. Pross A: Seeking the Chemical Roots of Darwinism: Bridging between Chemistry and Biology. *Chem Eur J* 2009, 15: 8374–8381. 10.1002/chem.200900805
18. Pross A: Stability in chemistry and biology: Life as a kinetic state of matter. *Pure Appl Chem* 2005, 77: 905–1921.
19. Pross A, Khodorkovsky V: Extending the concept of kinetic stability: toward a paradigm for life. *J Phys Org Chem* 2004, 17: 312–316. 10.1002/poc.729
20. Pross A: The evolutionary origin of biological function and complexity. *J Mol Evol* 2013, 76: 185–191. doi:10.1007/s00239–013–9556–1 10.1007/s00239-013-9556-1
21. Wagner N, Pross A, Tannenbaum E: Selective advantage of metabolic over non-metabolic replicators: a kinetic analysis. *Biosystems* 2009, 99: 126–129.
22. Dawkins R: *The selfish gene*. Oxford: Oxford University Press; 1989.
23. Sethna J: *Statistical mechanics*. Oxford: Oxford University Press; 2006:78.
24. Bissette AJ, Fletcher SP: Mechanisms of Autocatalysis. *Angew Chem Int Ed* 2013, 52: 12800–12826. doi:10.1002/anie.201303822 10.1002/anie.201303822
25. Plasson R, Brandenburg A, Jullien L, Bersini H: Autocatalysis: At the root of self-replication. *Artificial life* 2011, 17: 219–236. 10.1162/artl_a_00033
26. Eigen M: Selforganisation of matter and the evolution of biological macromolecules. *Naturwissenschaften* 1971, 58: 465–523. 10.1007/BF00623322
27. Kauffman SA: Cellular homeostasis, epigenesis and replication in randomly aggregated macromolecular systems. *J Cybernetics* 1971, 1: 71–96. 10.1080/01969727108545830
28. Gánti T: *The principles of life*. Oxford: Oxford University Press; 2003.
29. Lotka AJ: Contribution to the theory of periodic reaction. *J Phys Chem* 1910, 14: 271–4.
30. Eigen M, Schuster P: The hypercycle. A principle of natural self-organization. Part A: Emergence of the Hypercycle. *Naturwissenschaften* 1977, 64: 541–565. doi:10.1007/BF00450633 10.1007/BF00450633

31. von Kiedrowski G: A self-replicating hexadeoxynucleotide. *Angew Chem Int Ed* 1986, 25: 932–935.

32. Szathmáry E, Gladkih I: Sub-exponential growth and coexistence of non-enzymatically replicating Templates. *J Theor Biol* 1989, 138: 55–58. 10.1016/S0022-5193(89)80177-8

33. Lifson S: On the crucial stages in the origin of animate matter. *J Mol Evol* 1997, 44: 1–8. 10.1007/PL00006115

34. Vasas V, Szathmáry E, Santos M: Lack of evolvability in self-sustaining autocatalytic networks: A constraint on the metabolism-first path to the origin of life. *Proc Natl Acad Sci USA* 2010, 107: 1470–1475. 10.1073/pnas.0912628107

35. Vasas V, Fernando C, Santos M, Kauffman S, Szathmáry E: Evolution before genes. *Biol Direct* 2012, 7: 1. doi:10.1186/1745–6150–7-1 10.1186/1745-6150-7-1

36. Breslow R: On the mechanism of the formose reaction. *Tetrahedron Lett* 1959,1(21):22–26. 10.1016/S0040-4039(01)99487-0

37. Schrum JP, Zhu TF, Szostak JW: The Origins of Cellular Life. *Cold Spring Harb Perspect Biol* 2010, 2: a002212. doi:10.1101/cshperspect.a002212

38. Mavelli F, Ruiz-Mirazo K: Theoretical conditions for the stationary reproduction of model protocells. *Integr Biol* 2013, 5: 324–341. 10.1039/c2ib20222k

39. Goldbeter A: *Biochemical oscillations and cellular rhythms*. Cambridge, UK: Cambridge University Press; 1996.

40. Tyson JJ: *Biochemical oscillations: in Computational cell biology.* Edited by: Fall CP, Marland ES, Wagner JM, Tyson JJ. New York: Springer-Verlag; 2002.

41. McKane AJ, Nagy JD, Newman TJ, Stefanini MO: Amplified biochemical oscillations in cellular systems. *J Stat Phys* 2007, 128: 165–191. 10.1007/s10955-006-9221-9

42. Malthus T: *An essay on the principle of population.* London: Printed for Johnson J, in St. Paul's Church-Yard; 1798.http://www.esp.org/books/malthus/population/malthus.pdf

43. Field CB, Behrenfeld MJ, Randerson JT, Falkowski P: Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science* 1998, 281: 237–240. doi: 10.1126/science.281.5374.237

44. Kallmeyer J, Pockalny R, Adhikari RR, Smith DC, D'Hondt S: Global distribution of microbial abundance and biomass in subseafloor sediment. *Proc Natl Acad Sci USA* 2012, 109: 162131–16216. doi: 10.1073/pnas.1203849109

45. Sessions AL, Doughty DM, Welander PV, Summons RE, Newman DK: The continuing puzzle of the great oxidation event. *Current Biol* 2009, 19: R567-R574. 10.1016/j.cub.2009.05.054

46. Stankiewicz J, Eckardt LH: Chembiogenesis 2005 and systems chemistry workshop. *Angew Chem Int Ed* 2006, 45: 342. 10.1002/anie.200504139

47. Ludlow RF, Otto S: Systems chemistry. *Chem Soc Rev* 2008, 37: 101–108. 10.1039/b611921m

48. Vaidya N, Manapat ML, Chen IA, Xulvi-Brunet R, Hayden EJ, Lehman N: Spontaneous network formation among cooperative RNA replicators. *Nature* 2012, 491: 72–77. 10.1038/nature11549

49. Hordijk W, Steel M: Detecting autocatalytic, self-sustaining sets in chemical reaction systems. *J Theor Biol* 2004,227(4):451–461. 10.1016/j.jtbi.2003.11.020

50. Wu M, Higgs PG: Origin of Self-Replicating Biopolymers: autocatalytic Feedback Can Jump-Start the RNA World. *J Mol Evol* 2009, 69: 541–554. 10.1007/s00239-009-9276-8

Chapter

A RICCATI-BERNOULLI SUB-ODE METHOD FOR NONLINEAR PARTIAL DIFFERENTIAL EQUATIONS AND ITS APPLICATION

4

Xiao-Feng Yang[1], Zi-Chen Deng[2,3], and Yi Wei[1]

[1]Department of Applied Mathematics, Northwestern Polytechnical University, Xi'an, 710072, P.R. China.

[2]School of Mechanics, Civil Engineering and Architecture, Northwestern Polytechnical University, Xi'an, 710072, P.R. China.

[3]State Key Laboratory of Structural Analysis of Industrial Equipment, Dalian University of Technology, Dalian, 116023, P.R. China.

ABSTRACT

The Riccati-Bernoulli sub-ODE method is firstly proposed to construct exact traveling wave solutions, solitary wave solutions, and peaked wave solutions for nonlinear partial differential equations. A Bäcklund transformation of the Riccati-Bernoulli equation is given. By using a traveling wave transformation and the Riccati-Bernoulli equation, nonlinear partial differential equations

Citation: (APA): Yang, X. F., Deng, Z. C., & Wei, Y. (2015). A Riccati-Bernoulli sub-ODE method for nonlinear partial differential equations and its application. Advances in Difference equations, 2015(1), 1-17. (17 pages).

can be converted into a set of algebraic equations. Exact solutions of nonlinear partial differential equations can be obtained by solving a set of algebraic equations. By applying the Riccati-Bernoulli sub-ODE method to the Eckhaus equation, the nonlinear fractional Klein-Gordon equation, the generalized Ostrovsky equation, and the generalized Zakharov-Kuznetsov-Burgers equation, traveling solutions, solitary wave solutions, and peaked wave solutions are obtained directly. Applying a Bäcklund transformation of the Riccati-Bernoulli equation, an infinite sequence of solutions of the above equations is obtained. The proposed method provides a powerful and simple mathematical tool for solving some nonlinear partial differential equations in mathematical physics.

MSC: 35Q55, 35Q80, 35G25

Keywords: Riccati-Bernoulli sub-ODE method, Bäcklund transformation, traveling wave solution, solitary wave solution, peaked wave solution

INTRODUCTION

Nonlinear partial differential equations (NLPDEs) are known to describe a wide variety of phenomena not only in physics, but also in biology, chemistry, and several other fields. The investigation of traveling wave solutions for NLPDEs plays an important role in the study of nonlinear physical phenomena. In recent years, many powerful methods were used to construct traveling wave solutions of NLPDEs. For example, the inverse scattering method [1], the Bäcklund and Darboux transformation method [2], the homotopy perturbation method [3], the first integral method [4–6], the $\left(\frac{G'}{G}\right)$-expansion method [7–9], the sub-equation method [10, 11], Hirota's method [12], the homogeneous balance method [13–15], the variational iteration method [16, 17], the tanh-sech method [18], the Jacobi elliptic function method [19], the modified simple equation method [20–23], the $\exp(-\Phi(\xi))$-expansion method [24], the alternative functional variable method [25], and so on.

Many well-known NLPDEs can be handled by those traditional methods. However, there is no unified method which can be used to deal with all types of NLPDEs. Moreover, we always encounter the fractional NLPDEs, the NLPDEs which have nonlinear terms of any order or peaked wave solutions. It is significant to construct traveling wave solutions of NLPDEs by a uniform method. Based on those problems, the Riccati-Bernoulli sub-ODE method is firstly presented.

In this paper, the Riccati-Bernoulli sub-ODE method is proposed to construct traveling wave solutions, solitary wave solutions, and peaked wave solutions of NLPDEs. By using a traveling wave transformation and the Riccati-Bernoulli equation, NLPDEs can be converted into a set of algebraic equations. Exact solutions of NLPDEs can be obtained by solving the set of algebraic equations. The Eckhaus equation, the nonlinear fractional Klein-Gordon equation, the generalized Ostrovsky equation, and the generalized Zakharov-Kuznetsov-Burgers (ZK-Burgers) equation are chosen to illustrate the validity of the Riccati-Bernoulli sub-ODE method. A Bäcklund transformation of the Riccati-Bernoulli equation is given. If we get a solution of NLPDEs, we can search for a new infinite sequence of solutions of the NLPDEs by using a Bäcklund transformation.

The remainder of this paper is organized as follows: the Riccati-Bernoulli sub-ODE method is described in Section 2. In Section 3, a Bäcklund transformation of the Riccati-Bernoulli equation is given. In Sections 4-7, we apply the Riccati-Bernoulli sub-ODE method to the Eckhaus equation, the nonlinear fractional Klein-Gordon equation, the generalized Ostrovsky equation, and the generalized ZK-Burgers equation, respectively. In Section 8, our results are compared with the first integral method, the $\left(\frac{G'}{G}\right)$-expansion method, and physical explanations of the obtained solutions are discussed. In Section 9, some conclusions and directions for future work are given.

Description of the Riccati-Bernoulli sub-ODE method

Let there be given a NLPDE, say, in two variables,

$$P(u, u_t, u_x, u_{xx}, u_{xt}, \ldots) = 0, \tag{1}$$

where *P* is in general a polynomial function of its arguments, the subscripts denote the partial derivatives. The Riccati-Bernoulli sub-ODE method consists of three steps.

Step 1. Combining the independent variables *x* and *t* into one variable

$$\xi = k(x + Vt), \tag{2}$$

with

$$u(x, t) = u(\xi), \tag{3}$$

where the localized wave solution u(ξ) travels with speed *V*, by using Eqs. (2) and (3), one can transform Eq. (1) to an ODE

$$P(u, u', u'', u''', \ldots) = 0, \tag{4}$$

where u′ denotes $\frac{du}{d\xi}$.

Step 2. Suppose that the solution of Eq. (4) is the solution of the Riccati-Bernoulli equation

$$u' = au^{2-m} + bu + cu^{m}, \tag{5}$$

where *a*, *b*, *c*, and *m* are constants to be determined later.

From Eq. (5) and by directly calculating, we get

$$u'' = ab(3-m)u^{2-m} + a^2(2-m)u^{3-2m} + mc^2u^{2m-1} + bc(m+1)u^{m} + \left(2ac + b^2\right)u, \tag{6}$$

$$\begin{aligned} u''' = \Big(&ab(3-m)(2-m)u^{1-m} + a^2(2-m)(3-2m)u^{2-2m} \\ &+ m(2m-1)c^2u^{2m-2} + bcm(m+1)u^{m-1} + \left(2ac + b^2\right)\Big)u', \\ \ldots \end{aligned} \tag{7}$$

Remark

When ac ≠ 0 and m = 0, Eq. (5) is a Riccati equation. When a ≠ 0, c = 0, and m ≠ 1, Eq. (5) is a Bernoulli equation. Obviously, the Riccati equation and Bernoulli equation are special cases of Eq. (5). Because Eq. (5) is firstly proposed, we call Eq. (5) the Riccati-Bernoulli equation in order to avoid introducing new terminology.

Equation (5) has solutions as follows:

Case 1. When m = 1, the solution of Eq. (5) is

$$u(\xi) = Ce^{(a+b+c)\xi}. \tag{8}$$

Case 2. When m ≠ 1, b = 0, and c = 0, the solution of Eq. (5) is

$$u(\xi) = \left(a(m-1)(\xi + C)\right)^{\frac{1}{m-1}}. \tag{9}$$

Case 3. When m ≠ 1, b ≠ 0, and c = 0, the solution of Eq. (5) is

$$u(\xi) = \left(-\frac{a}{b} + Ce^{b(m-1)\xi}\right)^{\frac{1}{m-1}}. \tag{10}$$

Case 4. When m ≠ 1, a ≠ 0, and $b^2 - 4ac < 0$, the solutions of Eq. (5) are

$$u(\xi) = \left(-\frac{b}{2a} + \frac{\sqrt{4ac-b^2}}{2a}\tan\left(\frac{(1-m)\sqrt{4ac-b^2}}{2}(\xi + C)\right)\right)^{\frac{1}{1-m}} \tag{11}$$

and

$$u(\xi) = \left(-\frac{b}{2a} - \frac{\sqrt{4ac-b^2}}{2a}\cot\left(\frac{(1-m)\sqrt{4ac-b^2}}{2}(\xi + C)\right)\right)^{\frac{1}{1-m}}. \tag{12}$$

Case 5. When m ≠ 1, a ≠ 0, and $b^2 - 4ac > 0$, the solutions of Eq. (5) are

$$u(\xi) = \left(-\frac{b}{2a} - \frac{\sqrt{b^2-4ac}}{2a}\coth\left(\frac{(1-m)\sqrt{b^2-4ac}}{2}(\xi + C)\right)\right)^{\frac{1}{1-m}} \tag{13}$$

and

$$u(\xi) = \left(-\frac{b}{2a} - \frac{\sqrt{b^2-4ac}}{2a}\tanh\left(\frac{(1-m)\sqrt{b^2-4ac}}{2}(\xi + C)\right)\right)^{\frac{1}{1-m}}. \tag{14}$$

Case 6. When m ≠ 1, a ≠ 0, and $b^2 - 4ac = 0$, the solution of Eq. (5) is

$$u(\xi) = \left(\frac{1}{a(m-1)(\xi + C)} - \frac{b}{2a}\right)^{\frac{1}{1-m}}, \tag{15}$$

where *C* is an arbitrary constant.

Step 3. Substituting the derivatives of *u* into Eq. (4) yields an algebraic equation of *u*. Noticing the symmetry of the right-hand item of Eq. (5) and setting the highest power exponents of *u* to equivalence in Eq. (4), *m* can be determined. Comparing the coefficients of u^i yields a set of algebraic equations for *a*, *b*, *c*, and *V*. Solving the set of algebraic equations and substituting *m*, *a*, *b*, *c*, *V*, and $\xi = k(x + Vt)$ into Eq. (8)-(15), we can get traveling wave solutions of Eq. (1).

In the subsequent section, we will give a Bäcklund transformation of the Riccati-Bernoulli equation and some applications to illustrate the validity of the Riccati-Bernoulli sub-ODE method.

BÄCKLUND TRANSFORMATION OF THE RICCATI-BERNOULLI EQUATION

When $u_{n-1}(\xi)$ and $u_n(\xi)$ ($u_n(\xi) = u_n(u_{n-1}(\xi))$) are the solutions of Eq. (5), we get

$$\frac{du_n(\xi)}{d\xi} = \frac{du_n(\xi)}{du_{n-1}(\xi)}\frac{du_{n-1}(\xi)}{d\xi} = \frac{du_n(\xi)}{du_{n-1}(\xi)}\left(au_{n-1}^{2-m} + bu_{n-1} + cu_{n-1}^{m}\right),$$

namely

$$\frac{du_n(\xi)}{au_n^{2-m} + bu_n + cu_n^{m}} = \frac{du_{n-1}(\xi)}{au_{n-1}^{2-m} + bu_{n-1} + cu_{n-1}^{m}}.$$

Integrating above equation once with respect to ξ and simplifying it, we get

$$u_n(\xi) = \left(\frac{-cA_1 + aA_2(u_{n-1}(\xi))^{1-m}}{bA_1 + aA_2 + aA_1(u_{n-1}(\xi))^{1-m}}\right)^{\frac{1}{1-m}}, \tag{16}$$

where A_1 and A_2 are arbitrary constants.

Equation (16) is a Bäcklund transformation of Eq. (5). If we get a solution of Eq. (5), we can search for new infinite sequence of solutions of Eq. (5) by using Eq. (16). Then an infinite sequence of solutions of Eq. (1) is obtained.

APPLICATION TO THE ECKHAUS EQUATION

The Eckhaus equation reads

$$i\psi_t + \psi_{xx} + 2\left(|\psi|^2\right)_x\psi + |\psi|^4\psi = 0, \tag{17}$$

where $\psi = \psi(x, t)$ is a complex-valued function of two real variables x, t.

The Eckhaus equation was found [26] as an asymptotic multiscale reduction of certain classes of nonlinear Schrödinger type equations. A lot of the properties of the Eckhaus equation were obtained [27]. The Eckhaus equation can be linearized by making some transformations of dependent variables [28]. An exact traveling wave solution of the Eckhaus equation was obtained by the $(\frac{G'}{G})$-expansion method [8] and the first integral method [5].

In this section, new type of exact traveling wave solutions of the Eckhaus equation are obtained by using the Riccati-Bernoulli sub-ODE method.

Using the traveling wave transformation

$$\psi(x,t) = u(\xi)e^{i(\alpha x+\beta t)}, \tag{18}$$

Eq. (17) is reduced to

$$k^2u'' - (\alpha^2+\beta)u + 4ku^2u' + u^5 = 0, \tag{19}$$

where

$$\xi = k(x-2\alpha t), \tag{20}$$

and *k*, *α*, *β* are real constants to be determined later.

Suppose that the solution of Eq. (19) is the solution of Eq. (5). Substituting Eqs. (5) and (6) into Eq. (19), we get

$$\begin{aligned} &k^2\left(ab(3-m)u^{2-m} + a^2(2-m)u^{3-2m} + mc^2u^{2m-1} + bc(m+1)u^m\right.\\ &\quad\left. + \left(2ac+b^2\right)u\right) - \left(\alpha^2+\beta\right)u + 4ku^2\left(au^{2-m} + bu + cu^m\right) + u^5 = 0. \end{aligned} \tag{21}$$

Setting m = −1 and c = 0, Eq. (21) becomes

$$\left(k^2b^2 - \left(\alpha^2+\beta\right)\right)u + \left(4k^2ab + 4kb\right)u^3 + \left(3k^2a^2 + 4ka + 1\right)u^5 = 0. \tag{22}$$

Setting each coefficient of u^j (j = 1,3,5) to zero, we get

$$k^2b^2 - \left(\alpha^2+\beta\right) = 0, \tag{23a}$$

$$4k^2ab + 4kb = 0, \tag{23b}$$

$$3k^2a^2 + 4ka + 1 = 0. \tag{23c}$$

Notice that k ≠ 0, otherwise we can only get trivial solution.

Case A. If b = 0, from Eqs. (23a)-(23c) and (5), we get

$$u(x,t) = \pm\frac{1}{\sqrt{-2ka(x-2\alpha t)+C}}, \tag{24a}$$

$$\alpha^2 + \beta = 0, \tag{24b}$$

$$ka = -1\left(ka = -\frac{1}{3}\right), \tag{24c}$$

where C is an arbitrary real constant.

Case A-1. When ka = −1, we get exact traveling wave solutions of Eq. (17),

$$\psi_1(x,t) = \pm\frac{1}{\sqrt{2(x-2\alpha t)+C}}e^{i(\alpha x-\alpha^2 t)}, \tag{25}$$

where C and α are arbitrary real constants.

Case A-2. When $ka = -\frac{1}{3}$, we get exact traveling wave solutions of Eq. (17),

$$\psi_2(x,t) = \pm\frac{1}{\sqrt{\frac{2}{3}(x-2\alpha t)+C}}e^{i(\alpha x-\alpha^2 t)}, \tag{26}$$

where C and α are arbitrary real constants.

Case B. If b ≠ 0, from Eqs. (23a)-(23c), we get

$$b = \pm\frac{\sqrt{\alpha^2+\beta}}{k}, \tag{27a}$$

$$a = -\frac{1}{k}. \tag{27b}$$

Case B-1. When $b = -\frac{\sqrt{\alpha^2+\beta}}{k}$ and $a = -\frac{1}{k}$, from Eqs. (10) and (18), we get an exact traveling wave solution of Eq. (17),

$$\psi_3(x,t) = \left(-\frac{1}{\sqrt{\alpha^2+\beta}} + Ce^{2\sqrt{\alpha^2+\beta}(x-2\alpha t)}\right)^{-\frac{1}{2}} e^{i(\alpha x+\beta t)}, \tag{28}$$

where C, α, and β are arbitrary real constants.

Especially, if we choose $C = C_1 = \frac{1}{\sqrt{\alpha^2+\beta}}$, Eq. (28) becomes

$$\psi_4(x,t) = \left(\frac{\sqrt{\alpha^2+\beta}}{2}\left(-1+\coth\left(\sqrt{\alpha^2+\beta}(x-2\alpha t)\right)\right)\right)^{\frac{1}{2}} e^{i(\alpha x+\beta t)}, \tag{29}$$

where α and β are arbitrary real constants.

If we choose $C = C_2 = -\frac{1}{\sqrt{\alpha^2+\beta}}$, Eq. (28) becomes

$$\psi_5(x,t) = \left(\frac{\sqrt{\alpha^2+\beta}}{2}\left(-1+\tanh\left(\sqrt{\alpha^2+\beta}(x-2\alpha t)\right)\right)\right)^{\frac{1}{2}} e^{i(\alpha x+\beta t)}, \tag{30}$$

where α and β are arbitrary real constants.

Case B-2. When $b = \frac{\sqrt{\alpha^2+\beta}}{k}$ and $a = -\frac{1}{k}$, from Eqs. (10) and (18), we get an exact traveling wave solution of Eq. (17),

$$\psi_6(x,t) = \left(\frac{1}{\sqrt{\alpha^2+\beta}} + Ce^{-2\sqrt{\alpha^2+\beta}(x-2\alpha t)}\right)^{-\frac{1}{2}} e^{i(\alpha x+\beta t)}, \tag{31}$$

where C, α, and β are arbitrary real constants.

Especially, if we choose $C = C_3 = \frac{1}{\sqrt{\alpha^2+\beta}}$, Eq. (31) becomes

$$\psi_7(x,t) = \left(\frac{\sqrt{\alpha^2+\beta}}{2}\left(1+\tanh\left(\sqrt{\alpha^2+\beta}(x-2\alpha t)\right)\right)\right)^{\frac{1}{2}} e^{i(\alpha x+\beta t)}, \tag{32}$$

where α and β are arbitrary real constants.

If we choose $C = C_4 = -\frac{1}{\sqrt{\alpha^2+\beta}}$, Eq. (31) becomes

$$\psi_8(x,t) = \left(\frac{\sqrt{\alpha^2+\beta}}{2}\left(1+\coth\left(\sqrt{\alpha^2+\beta}(x-2\alpha t)\right)\right)\right)^{\frac{1}{2}} e^{i(\alpha x+\beta t)}, \tag{33}$$

where α and β are arbitrary real constants.

Applying Eq. (16) to $\psi_j(x, t)$ ($j = 1, 2, \ldots, 8$), we can get an infinite sequence of solutions of Eq. (17). For example, by applying Eq. (16) to Eq. (32), we get a new solution of Eq. (17),

$$\psi_7^*(x,t) = \left(\frac{A_2\sqrt{\alpha^2+\beta}(1+\tanh(\sqrt{\alpha^2+\beta}(x-2\alpha t)))}{2A_2 + A_1\sqrt{\alpha^2+\beta}(-1+\tanh(\sqrt{\alpha^2+\beta}(x-2\alpha t)))}\right)^{\frac{1}{2}} e^{i(\alpha x+\beta t)}, \tag{34}$$

where A_1, A_2, α, and β are arbitrary real constants.

APPLICATION TO THE NONLINEAR FRACTIONAL KLEIN-GORDON EQUATION

The nonlinear fractional Klein-Gordon equation [23] reads

$$\frac{\partial^{2\alpha} u(x,t)}{\partial t^{2\alpha}} = \frac{\partial^2 u(x,t)}{\partial x^2} + \beta u(x,t) + \gamma u^3(x,t), \quad t > 0, 0 < \alpha \leq 1, \tag{35}$$

where *β* and *γ* are known constants.

As is well known, linear and nonlinear Klein-Gordon equations model many problems in classical and quantum mechanics, solitons and condensed matter physics. For example, the nonlinear sine Klein-Gordon equation models a Josephson junction, the motion of rigid pendula attached to a stretched wire, and dislocations in crystals [17, 29–31]. A non-local version of these equations are properly described by the fractional version of them. Exact traveling wave solutions of the nonlinear fractional Klein-Gordon equation were obtained by the homotopy perturbation method [29] and the first integral method [6].

In this section, exact traveling wave solutions of the nonlinear fractional Klein-Gordon equation are obtained by using the Riccati-Bernoulli sub-ODE method.

Using the transformation

$$u(x,t) = u(\xi), \tag{36}$$

with

$$\xi = lx - \frac{\lambda t^\alpha}{\Gamma(1+\alpha)}, \tag{37}$$

where *l* and *λ* are constants to be determined later, Eq. (35) becomes

$$u'' - \frac{\beta}{\lambda^2 - l^2} u - \frac{\gamma}{\lambda^2 - l^2} u^3 = 0. \tag{38}$$

Suppose that the solution of Eq. (38) is the solution of Eq. (5). Substituting Eq. (6) into Eq. (38), we get

$$ab(3-m)u^{2-m}+a^2(2-m)u^{3-2m}+mc^2u^{2m-1}$$
$$+bc(m+1)u^m+\left(2ac+b^2\right)u-\frac{\beta}{\lambda^2-l^2}u-\frac{\gamma}{\lambda^2-l^2}u^3=0. \tag{39}$$

Setting m = 0, Eq. (39) is reduced to

$$3abu^2+2a^2u^3+bc+\left(2ac+b^2\right)u-\frac{\beta}{\lambda^2-l^2}u-\frac{\gamma}{\lambda^2-l^2}u^3=0. \tag{40}$$

Setting each coefficient of u^i (i = 0, 1, 2, 3) to zero, we get

$$bc=0, \tag{41a}$$

$$2ac+b^2-\frac{\beta}{\lambda^2-l^2}=0, \tag{41b}$$

$$3ab=0, \tag{41c}$$

$$2a^2-\frac{\gamma}{\lambda^2-l^2}=0. \tag{41d}$$

Solving Eqs. (41a)-(41d), we get

$$b=0, \tag{42a}$$

$$ac=\frac{\beta}{2(\lambda^2-l^2)}, \tag{42b}$$

$$a=\pm\sqrt{\frac{\gamma}{2(\lambda^2-l^2)}}. \tag{42c}$$

Case A. When $\frac{\beta}{\lambda^2-l^2}>0,$ substituting Eqs. (42a)-(42c) and (37) into Eqs. (11) and (12), we get exact traveling wave solutions of Eq. (35),

$$u_{1,2}(x,t)=\pm\sqrt{\frac{\beta}{\gamma}}\tan\left(\sqrt{\frac{\beta}{2(\lambda^2-l^2)}}\left(lx-\frac{\lambda t^\alpha}{\Gamma(1+\alpha)}\right)+C\right), \tag{43a}$$

and

$$u_{3,4}(x,t)=\pm\sqrt{\frac{\beta}{\gamma}}\cot\left(\sqrt{\frac{\beta}{2(\lambda^2-l^2)}}\left(lx-\frac{\lambda t^\alpha}{\Gamma(1+\alpha)}\right)+C\right), \tag{43b}$$

where C, l, and λ are arbitrary constants.

Case B. When $\frac{\beta}{\lambda^2-l^2}<0,$ substituting Eqs. (42a)-(42c) and (37) into Eqs. (13) and (14), we get exact traveling wave solutions of Eq. (35),

$$u_{5,6}(x,t) = \pm\sqrt{-\frac{\beta}{\gamma}}\tanh\left(\sqrt{-\frac{\beta}{2(\lambda^2 - l^2)}}\left(lx - \frac{\lambda t^\alpha}{\Gamma(1+\alpha)}\right) + C\right), \tag{44a}$$

and

$$u_{7,8}(x,t) = \pm\sqrt{-\frac{\beta}{\gamma}}\coth\left(\sqrt{-\frac{\beta}{2(\lambda^2 - l^2)}}\left(lx - \frac{\lambda t^\alpha}{\Gamma(1+\alpha)}\right) + C\right), \tag{44b}$$

where *C*, *l*, and *λ* are arbitrary constants.

Applying Eq. (16) to $u_j(x, t)$ (j = 1, 2, … , 8), we can get an infinite sequence of solutions of Eq. (35). For example, by applying Eq. (16) to $u_j(x, t)$ (j = 1, 2, … , 8) once, we get new solutions of Eq. (35),

$$u^*_{1,2}(x,t) = \frac{-\frac{\beta}{\gamma} \pm A_3\sqrt{\frac{\beta}{\gamma}}\tan(\sqrt{\frac{\beta}{2(\lambda^2-l^2)}}(lx - \frac{\lambda t^\alpha}{\Gamma(1+\alpha)}) + C)}{A_3 \pm \sqrt{\frac{\beta}{\gamma}}\tan(\sqrt{\frac{\beta}{2(\lambda^2-l^2)}}(lx - \frac{\lambda t^\alpha}{\Gamma(1+\alpha)}) + C)},$$

$$u^*_{3,4}(x,t) = \frac{-\frac{\beta}{\gamma} \pm A_3\sqrt{\frac{\beta}{\gamma}}\cot(\sqrt{\frac{\beta}{2(\lambda^2-l^2)}}(lx - \frac{\lambda t^\alpha}{\Gamma(1+\alpha)}) + C)}{A_3 \pm \sqrt{\frac{\beta}{\gamma}}\cot(\sqrt{\frac{\beta}{2(\lambda^2-l^2)}}(lx - \frac{\lambda t^\alpha}{\Gamma(1+\alpha)}) + C)},$$

$$u^*_{5,6}(x,t) = \frac{-\frac{\beta}{\gamma} \pm A_3\sqrt{-\frac{\beta}{\gamma}}\tanh(\sqrt{-\frac{\beta}{2(\lambda^2-l^2)}}(lx - \frac{\lambda t^\alpha}{\Gamma(1+\alpha)}) + C)}{A_3 \pm \sqrt{-\frac{\beta}{\gamma}}\tanh(\sqrt{-\frac{\beta}{2(\lambda^2-l^2)}}(lx - \frac{\lambda t^\alpha}{\Gamma(1+\alpha)}) + C)},$$

$$u^*_{7,8}(x,t) = \frac{-\frac{\beta}{\gamma} \pm A_3\sqrt{-\frac{\beta}{\gamma}}\coth(\sqrt{-\frac{\beta}{2(\lambda^2-l^2)}}(lx - \frac{\lambda t^\alpha}{\Gamma(1+\alpha)}) + C)}{A_3 \pm \sqrt{-\frac{\beta}{\gamma}}\coth(\sqrt{-\frac{\beta}{2(\lambda^2-l^2)}}(lx - \frac{\lambda t^\alpha}{\Gamma(1+\alpha)}) + C)},$$

where A_1, A_2, *C*, *l*, and *λ* are arbitrary real constants.

APPLICATION TO THE GENERALIZED OSTROVSKY EQUATION

The generalized Ostrovsky equation reads

$$\left(u_t + 3uu_x - \frac{\beta}{4}u_{xxx}\right)_x - \frac{\varepsilon^2}{2}\left(u + \delta u^2\right) = 0, \tag{45}$$

where *β*, *ε*, and *δ* are known constants.

The generalized Ostrovsky equation is a model for the weakly nonlinear surface and internal waves in a rotating ocean. Exact peaked wave solutions were obtained by the undetermined coefficient method [32].

In this section, exact peaked wave solutions of the generalized Ostrovsky equation are obtained by using the Riccati-Bernoulli sub-ODE method.

Using the transformation

$$u(x,t) = u(\xi), \tag{46}$$

with

$$\xi = k(x + Vt), \tag{47}$$

where *k* and *V* are the wave number and wave speed, respectively, Eq. (45) becomes

$$k^2 Vu'' + 3k^2(u')^2 + 3k^2 uu'' - \frac{\beta}{4}k^4 u'''' - \frac{\varepsilon^2}{2}(u + \delta u^2) = 0. \tag{48}$$

Suppose that the solution of Eq. (48) is the solution of Eq. (5). From Eqs. (5) and (6), we get

$$\begin{aligned} u'''' = {} & \big(ab(3-m)(2-m)(1-m)u^{-m} + a^2(2-m)(3-2m)(2-2m) \\ & \times u^{1-2m} + m(2m-1)(2m-2)c^2u^{2m-3} + bcm(m+1)(m-1)u^{m-2}\big) \\ & \times (u')^2 + \big(ab(3-m)(2-m)u^{1-m} + a^2(2-m)(3-2m)u^{2-2m} \\ & + m(2m-1)c^2u^{2m-2} + bcm(m+1)u^{m-1} + (2ac+b^2)\big)u''. \end{aligned} \tag{49}$$

Substituting Eqs. (5), (6), and (49) into Eq. (48), we get

$$\begin{aligned} & (k^2V + 3k^2u)\big(ab(3-m)u^{2-m} + a^2(2-m)u^{3-2m} \\ & \quad + mc^2u^{2m-1} + bc(m+1)u^m + (2ac+b^2)u\big) + \frac{\beta}{4}k^4\Sigma = 0, \end{aligned} \tag{50}$$

where

$$\begin{aligned} \Sigma = {} & (au^{2-m} + bu + cu^m)^2\big(ab(3-m)(2-m)(1-m)u^{-m} + a^2(2-m)(3-2m) \\ & \times (2-2m)u^{1-2m} + m(2m-1)(2m-2)c^2u^{2m-3} + bcm(m+1)(m-1)u^{m-2}\big) \\ & + \big(ab(3-m)u^{2-m} + a^2(2-m)u^{3-2m} + mc^2u^{2m-1} + bc(m+1)u^m + (2ac+b^2)u\big) \\ & \times \big(ab(3-m)(2-m)u^{1-m} + a^2(2-m)(3-2m)u^{2-2m} + m(2m-1)c^2u^{2m-2} \\ & + bcm(m+1)u^{m-1} + (2ac+b^2)\big). \end{aligned}$$

Setting m = 2 and c = 0, Eq. (50) is reduced to

$$\left(k^2Vab + 3k^2a^2 - \frac{\beta}{4}k^4ab^3\right) + \left(k^2Vb^2 - \frac{\beta}{4}k^4b^4 + 9k^2ab - \frac{1}{2}\varepsilon^2\right)u$$
$$+ \left(6k^2b^2 - \frac{1}{2}\varepsilon^2\delta\right)u^2 = 0.$$

(51)

Setting each coefficient of u^j (j = 0, 1, 2) to zero, we get

$$6k^2b^2 - \frac{1}{2}\varepsilon^2\delta = 0,$$

(52a)

$$k^2Vb^2 - \frac{\beta}{4}k^4b^4 + 9k^2ab - \frac{1}{2}\varepsilon^2 = 0,$$

(52b)

$$k^2Vab + 3k^2a^2 - \frac{\beta}{4}k^4ab^3 = 0.$$

(52c)

Solving Eqs. (52a)-(52c), we get

$$-\frac{a}{b} = -\frac{1}{\delta},$$

(53a)

$$kb = \pm\sqrt{\frac{\delta\varepsilon^2}{12}},$$

(53b)

$$V = \frac{\beta\delta^2\varepsilon^2 - 144}{48\delta}.$$

(53c)

Substituting Eqs. (53a)-(53c) and (47) into Eq. (10), we get exact peaked wave solutions of Eq. (45),

$$u_{1,2}(x,t) = -\frac{1}{\delta} + Ce^{\pm\sqrt{\frac{\delta\varepsilon^2}{12}}|x+(\frac{\beta\delta^2\varepsilon^2-144}{48\delta})t|},$$

(54)

where *C* is an arbitrary constant.

Similar to Sections 4 and 5, by using a Bäcklund transformation, we can get an infinite sequence of solutions of the generalized Ostrovsky equation. It being a similar process, we omit it.

APPLICATION TO THE GENERALIZED ZK-BURGERS EQUATION

The generalized ZK-Burgers equation [33] reads

$$u_t + \alpha u^{\lambda} u_x + \beta u_{xxx} + \gamma\left(u_{yy} + u_{zz}\right) + \sigma u_{xx} = 0, \tag{55}$$

where α, β, γ, σ, and λ are known constants.

The generalized ZK-Burgers equation retains the strong nonlinear aspects of the governing equation in many practical transport problems such as nonlinear waves in a medium with low-frequency pumping or absorption, transport and dispersion of pollutants in rivers, and sediment transport. Wang *et al.* obtained a solitary wave of the generalized ZK-Burgers equation with a positive fractional power term by using the HB method and with the aid of sub-ODEs [33].

In this section, exact traveling wave solutions of the generalized ZK-Burgers equation are obtained by using the Riccati-Bernoulli sub-ODE method.

Using the transformation

$$u(x, y, z, t) = u(\xi), \tag{56}$$

with

$$\xi = k(x + ly + nz + Vt), \tag{57}$$

where k, l, n, and V are constants to be determined later, Eq. (55) becomes

$$kVu' + k\alpha u^{\lambda} u' + k^3 \beta u''' + k^3 \gamma\left(l^2 + n^2\right) u''' + k^2 \sigma u'' = 0. \tag{58}$$

Suppose that the solution of Eq. (55) is the solution of Eq. (5). Noticing $u' \neq 0$ and $k \neq 0$, otherwise we can only get trivial solution. Substituting Eqs. (5), (6), and (7) into Eq. (58), we get

$$\begin{aligned}
&\rho k^2 \Big(ab(3-m)(2-m)u^{1-m} + a^2(2-m)(3-2m)u^{2-2m} \\
&\quad + m(2m-1)c^2 u^{2m-2} + bcm(m+1)u^{m-1} + \left(2ac + b^2\right)\Big) \\
&\quad + \sigma k\left(a(2-m)u^{1-m} + mcu^{m-1} + b\right) + V + \alpha u^{\lambda} = 0,
\end{aligned} \tag{59}$$

where

$$\rho = \beta + \gamma\left(l^2 + n^2\right). \tag{60}$$

Setting $m = 1 - \frac{\lambda}{2}$ and c = 0, Eq. (59) is reduced to

$$\left(V + \rho k^2 b^2 + \sigma k b\right) + \left(\frac{(\lambda+2)(\lambda+4)\rho k^2 ab}{4} + \frac{(\lambda+2)\sigma ka}{2}\right)u^{\frac{\lambda}{2}} + \left(\alpha + \frac{(\lambda+2)(\lambda+1)\rho k^2 a^2}{2}\right)u^{\lambda} = 0. \tag{61}$$

Setting each coefficient of $u^j\ (j = 0, \frac{\lambda}{2}, \lambda)$ to zero, we get

$$V + \rho k^2 b^2 + \sigma k b = 0, \tag{62a}$$

$$\frac{(\lambda+2)ka}{2}\left(\frac{(\lambda+4)\rho kb}{2} + \sigma\right) = 0, \tag{62b}$$

$$\alpha + \frac{(\lambda+2)(\lambda+1)\rho k^2 a^2}{2} = 0. \tag{62c}$$

Solving Eqs. (62a)-(62c), we get

$$b = \frac{-2\sigma}{k\rho(\lambda+4)}, \tag{63a}$$

$$a = \pm\frac{1}{k}\sqrt{\frac{-2\alpha}{\rho(\lambda^2+3\lambda+2)}}, \tag{63b}$$

$$V = \frac{2\sigma^2(\lambda+2)}{\rho(\lambda+4)^2}. \tag{63c}$$

Substituting Eqs. (63a)-(63c) and (57) into Eq. (10), we get exact traveling wave solutions of Eq. (55),

$$u_{1,2}(x,t) = \left(\pm\frac{\rho(\lambda+4)}{\sigma}\sqrt{\frac{-\alpha}{2\rho(\lambda^2+3\lambda+2)}} + Ce^{\frac{\lambda\sigma}{\rho(\lambda+4)}\left(x+ly+nz+\left(\frac{2\sigma^2(\lambda+2)}{\rho(\lambda+4)^2}\right)t\right)}\right)^{\frac{-2}{\lambda}}, \tag{64}$$

where C, l, and n are arbitrary constants.

Equation (64) is new type of traveling wave solution of the generalized ZK-Burgers equation. Especially, if we choose $C = C_1 = \frac{\rho(\lambda+4)}{\sigma}\sqrt{\frac{-\alpha}{2\rho(\lambda^2+3\lambda+2)}}$, we get the solitary wave solutions of Eq. (55),

$$u_3(x,t) = \left(\frac{1}{2C_1}\left(1 - \tanh\frac{\eta}{2}\right)\right)^{\frac{2}{\lambda}}, \tag{65}$$

$$u_4(x,t) = \left(\frac{1}{2C_1}\left(-1 + \coth\frac{\eta}{2}\right)\right)^{\frac{2}{\lambda}}, \tag{66}$$

where *l*, *n* are arbitrary constants and

$$\eta = \frac{\lambda\sigma}{\rho(\lambda+4)}\left(x + ly + nz + \left(\frac{2\sigma^2(\lambda+2)}{\rho(\lambda+4)^2}\right)t\right). \tag{67}$$

If we choose $C = C_2 = -\frac{\rho(\lambda+4)}{2\sigma}\sqrt{\frac{-2\alpha}{\rho(\lambda^2+3\lambda+2)}}$, we get the solitary wave solutions of Eq. (55),

$$u_5(x,t) = \left(\frac{1}{2C_2}\left(-1 + \coth\frac{\eta}{2}\right)\right)^{\frac{2}{\lambda}}, \tag{68}$$

$$u_6(x,t) = \left(\frac{1}{2C_2}\left(1 - \tanh\frac{\eta}{2}\right)\right)^{\frac{2}{\lambda}}, \tag{69}$$

where *l* and *n* are arbitrary constants.

Similar to Sections 4 and 5, by using a Bäcklund transformation, we can get an infinite sequence of solutions of the generalized ZK-Burgers equation. It being a similar process, we omit it.

COMPARISONS AND EXPLANATIONS OF THE SOLUTIONS

In this section, the physical interpretation of the results of Sections 4-7 are given, respectively. We will compare the Riccati-Bernoulli sub-ODE method with the $(\frac{G'}{G})$-expansion method, the first integral method, and so on. Some of our obtained exact solutions are in the figures represented with the aid of Maple software.

- **The Eckhaus equation**: Applying the Riccati-Bernoulli sub-ODE method, Eqs. (25), (26), (28), (31), and (34) are new types of exact traveling wave solutions of the Eckhaus equation. Equations (29), (30), (32), and (33), which are expressed by the hyperbolic functions, are a kind of kink-type envelope solitary solutions. They could not be obtained by the method presented in Ref. [27]. Equation (26), which is expressed by the rational functions, could not be obtained by the $(\frac{G'}{G})$-expansion method [8] and the first integral method [5].

- **The nonlinear fractional Klein-Gordon equation**: Applying the Riccati-Bernoulli sub-ODE method and comparing our results with Golmankhaneh's results [29], it is easy to find that $u_j(x, t)$ $(j = 1, \ldots, 8)$ are new and identical to results by the first integral method [6]. $u_j(x, t)$ $(j = 1, 2, 3, 4)$, which are expressed by the trigonometric functions, are periodic wave solutions. $u_j(x, t)$ $(j = 5, 6, 7, 8)$, which are expressed by the hyperbolic functions, are a kind of kink-type envelope solitary solutions. The shape of $u = u_1(x, t)$ is represented in Figure 1 with $\alpha = \frac{1}{90}, \beta = 1, \gamma = 1, \lambda = \frac{9}{5}, C = 0$ and $l = \frac{3}{2}$ within the interval $-5 \leq x \leq 5$ and $0 \leq t \leq \frac{1}{2}$. The shape of $u = u_5(x, t)$ is represented in Figure 2 with $\alpha = \frac{1}{5}, \beta = -1, \gamma = 1, \lambda = 2, C = 0,$ and $l = 1$ within the interval $-6 \leq x \leq 6$ and $0 \leq t \leq 6$.

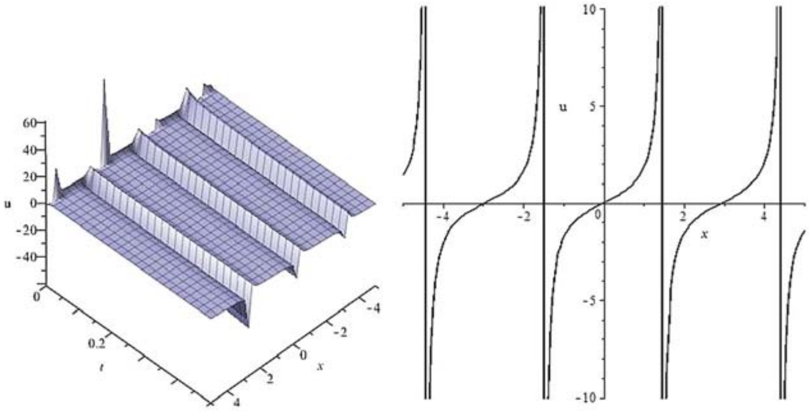


Figure 1. Graph of solution $u = u_1(x, t)$ of the nonlinear fractional Klein-Gordon equation for $\alpha = \frac{1}{90}$, $\beta = 1$, $\gamma = 1$, $\lambda = \frac{9}{5}$, $C = 0$, and $l = \frac{3}{2}$. The left figure shows the 3-D plot and the right figure shows the 2-D plot for $t = 0$.

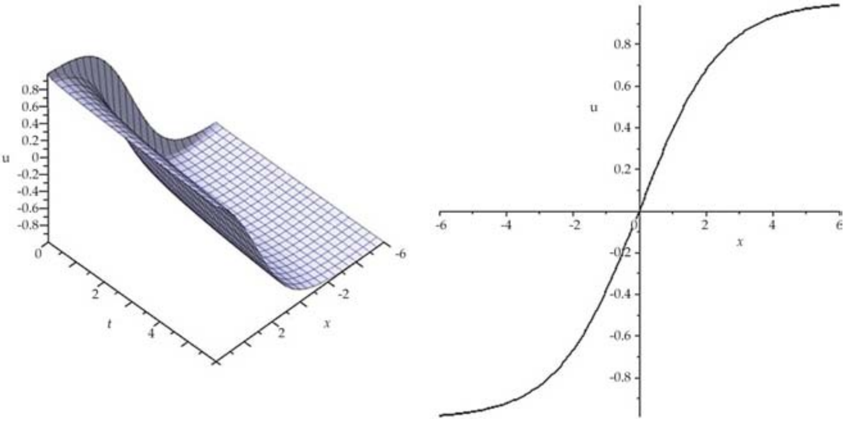


Figure 2. Graph of solution $u = u_5(x, t)$ of the nonlinear fractional Klein-Gordon equation for $\alpha = \frac{1}{5}$, $\beta = -1$, $\gamma = 1$, $\lambda = 2$, $C = 0$, and $l = 1$. The left figure shows the 3-D plot and the right figure shows the 2-D plot for $t = 0$.

- **The generalized Ostrovsky equation**: Applying the Riccati-Bernoulli sub-ODE method, it is easy to find that our results are identical to results presented in Ref. [32]. $u = u_{1,2}(x, t)$ are peaked wave solutions of the generalized Ostrovsky equation. The shape of $u = u_1(x, t)$ is represented in Figure 3 with $\delta = 6$, $\beta = 6$, $\varepsilon = 1$, $\lambda = 2$, and $C = \frac{1}{10}$ within the interval $-5 \leq x, t \leq 5$.

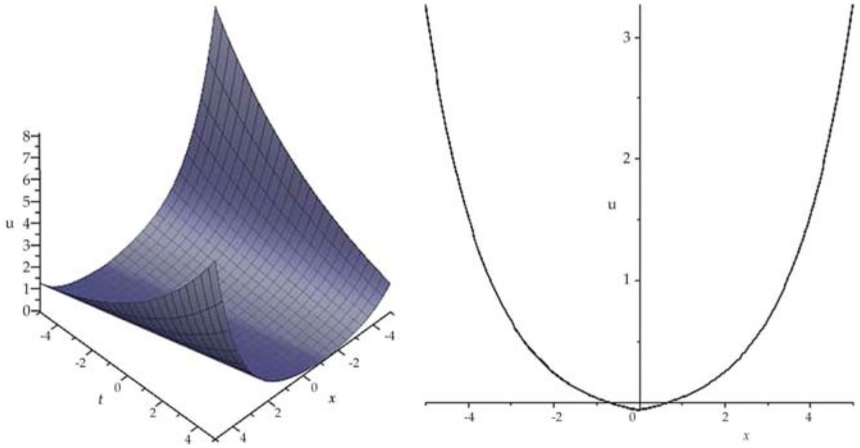


Figure 3. Graph of solution $u = u_5(x, t)$ of the generalized Ostrovsky equation for $\delta = 6$, $\beta = 6$, $\varepsilon = 1$, $\lambda = 2$, and $C = \frac{1}{10}$. The left figure shows the 3-D plot and the right figure shows the 2-D plot for $t = 0$.

- **The generalized ZK-Burgers equation**: By applying the Riccati-Bernoulli sub-ODE method to the generalized ZK-Burgers equation, we find that if λ is a positive fraction, our results degenerate to the results of Ref. [33]. Moreover, we enlarge the value range of parameters λ of the generalized ZK-Burgers equation so that the parameter λ can be an arbitrary constant ($\lambda \neq -1, -2, -4$). $u_j(x, t)$ ($j = 1, \ldots, 6$) are exact traveling wave solutions of the generalized ZK-Burgers equation. $u_j(x, t)$ ($j = 3, 4, 5, 6$), which are expressed by the hyperbolic functions, are a kind of kink-type envelope solitary solutions. The shape of $u = u_1(x, t)$ is represented in Figure 4 with $\alpha = \beta = \gamma = l = n = y = z = 1$, $\lambda = -\sqrt{2}$ and $\sigma = 2$ within the interval $-5 \leq x, t \leq 5$.

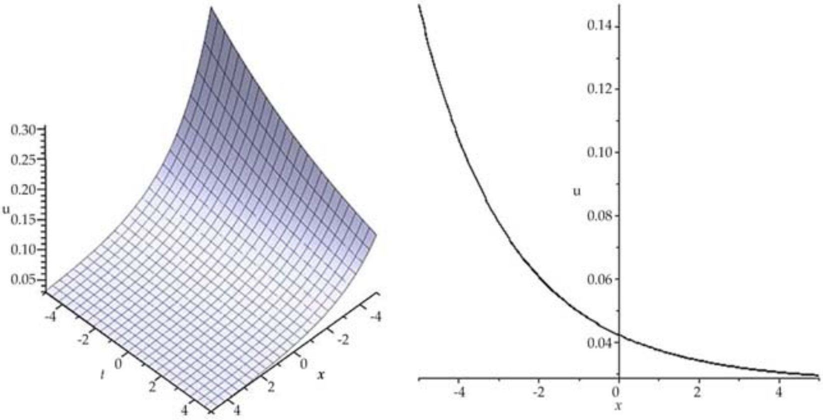


Figure 4. Graph of solution $u = u_\varsigma(x, t)$ of the generalized ZK-Burgers equation for $\alpha = \beta = \gamma = l = n = y = z = 1$, $\lambda = -\sqrt{2}$, and $\sigma = 2$. The left figure shows the 3-D plot and the right figure shows the 2-D plot for $t = 3$.

Moreover, by using a Bäcklund transformation, we can get an infinite sequence of solutions of these NLPDEs which cannot be obtained by the $(\frac{G'}{G})$-expansion method and the first integral method. The graphical demonstrations of some obtained solutions are shown in Figures 1-4.

CONCLUSIONS

The Riccati-Bernoulli sub-ODE method is successfully used to establish exact traveling wave solutions, solitary wave solutions and peaked wave solutions of NLPDEs. A Bäcklund transformation of the Riccati-Bernoulli equation is given. By applying a Bäcklund transformation of the Riccati-Bernoulli equation to the NLPDEs, an infinite sequence of solutions of the NLPDEs is obtained. The Eckhaus equation, the nonlinear fractional Klein-Gordon equation, the generalized Ostrovsky equation, and the generalized ZK-Burgers equation are chosen to illustrate the validity of the Riccati-Bernoulli sub-ODE method. Many well-known NLPDEs can be handled by this method. The performance of this method is found to be simple and efficient. The availability of computer systems like Maple facilitates the tedious algebraic calculations. The Riccati-Bernoulli sub-ODE method is also a standard and computable method, which allows us to perform complicated and tedious algebraic calculations.

It is well known that it is difficult to propose an uniform analytical method for all types of the NLPDEs, and the Riccati-Bernoulli sub-ODE method is no exception. Similar to the first integral method, the $(\frac{G'}{G})$-expansion method and the homogeneous balance method, the Riccati-Bernoulli sub-ODE method is used to obtain exact solutions of the form of Eq. (1). Constructing more powerful sub-ODE and Bäcklund transformations is future work and aims to search for exact solutions of NLPDEs

ACKNOWLEDGEMENTS

The research is supported by the National Natural Science Foundation of China (11372252, 11372253 and 11432010) and the Fundamental Research Funds for the Central Universities (3102014JCQ01035).

REFERENCES

1. Ablowitz, MJ, Clarkson, PA: Solitons, Nonlinear Evolution Equations and Inverse Scattering. Cambridge University Press, New York (1991)
2. Rogers, C, Schief, WK: Bäcklund and Darboux Transformation Geometry and Modern Applications in Solitons Theory. Cambridge University Press, Cambridge (2002)
3. He, JH: An approximate solution technique depending on an artificial parameter: a special example. Commun. Nonlinear Sci. Numer. Simul. 3, 92-97 (1998)
4. Feng, ZS: The first-integral method to study the Burgers-Korteweg-de Vries equation. J. Phys. A, Math. Gen. 35, 343-349 (2002)
5. Taghizadeh, N, Mirzazadeh, M, Filiz, T: The first-integral method applied to the Eckhaus equation. Appl. Math. Lett. 25, 798-802 (2012)
6. Lu, B: The first integral method for some time fractional differential equations. J. Math. Anal. Appl. 395, 684-693 (2012)
7. Wang, ML, Li, XZ, Zhang, JL: The $(\frac{G'}{G})$-expansion method and travelling wave solutions of nonlinear evolution equations in mathematical physics. Phys. Lett. A 372, 417-423 (2008)
8. Zhang, H: New application of the $(\frac{G'}{G})$-expansion method. Commun. Nonlinear Sci. Numer. Simul. 14, 3220-3225 (2009)
9. Khan, K, Akbar, MA: Study of analytical method to seek for exact solutions of variant Boussinesq equations. SpringerPlus 3, 324-340 (2014)
10. Xu, SL, Liang, JC: Exact soliton solutions to a generalized nonlinear Schrödinger equation. Commun. Theor. Phys. 53, 159-165 (2010)
11. Wang, ML: Applications of *F*-expansion to periodic wave solutions for a new Hamiltonian amplitude equation. Chaos Solitons Fractals 24, 1257-1268 (2005)
12. Hirota, R: Exact solution of the Korteweg-de Vries equation for multiple collisions of solitons. Phys. Rev. Lett. 27, 1192-1194 (1971)
13. Wang, ML: Solitary wave solutions for variant Boussinesq equations. Phys. Lett. A 199, 169-172 (1995)
14. Wang, ML, Zhou, YB: Application of a homogeneous balance method to exact solutions of nonlinear equations in mathematical physics. Phys. Lett. A 216, 67-75 (1996)

15. Bai, CL: Extended homogeneous balance method and Lax pairs, Bäcklund transformation. Commun. Theor. Phys. 37, 645-648 (2002)

16. He, JH: An new approach to nonlinear partial differential equations. Commun. Nonlinear Sci. Numer. Simul. 2, 230-235 (1997)

17. Yusufoglu, E: The variational iteration method for studying the Klein-Gordon equation. Appl. Math. Lett. 21, 669-674 (2008)

18. Wazwaz, AM: The tanh method: exact solutions of the sine-Gordon and the sinh-Gordon equations. Appl. Math. Comput. 167, 1196-1210 (2005)

19. Yan, ZL: Abunbant families of Jacobi elliptic function solutions of the-dimensional integrable Davey-Stewartson-type equation via a new method. Chaos Solitons Fractals 18, 299-309 (2003)

20. Khan, K, Akbar, MA, Rayhanul Islam, SM: Exact solutions for (1 + 1)-dimensional nonlinear dispersive modified Benjamin-Bona-Mahony equation and coupled Klein-Gordon equations. SpringerPlus 3, 724-731 (2014)

21. Khan, K, Akbar, MA: Solitary wave solutions of some coupled nonlinear evolution equations. J. Sci. Res. 6, 273-284 (2014)

22. Khan, K, Akbar, MA: Traveling wave solutions of the (2 + 1)-dimensional Zoomeron equation and the Burgers equations via the MSE method and the Exp-function method. Ain Shams Eng. J. 5, 247-256 (2014)

23. Ahmed, MT, Khan, K, Akbar, MA: Study of nonlinear evolution equations to construct traveling wave solutions via modified simple equation method. Phys. Rev. Res. Int. 3, 490-503 (2013)

24. Khan, K, Akbar, MA: The $\exp(-\Phi(\xi))$-expansion method for finding travelling wave solutions of Vakhnenko-Parkes equation. Int. J. Dyn. Syst. Differ. Equ. 5, 72-83 (2014)

25. Zerarka, A, Ouamane, S, Attaf, A: Construction of exact solutions to a family of wave equations by the functional variable method. Waves Random Complex Media 21, 44-56 (2011)

26. Calogero, F, Eckhaus, W: Nonlinear evolution equations, rescalings, model PDEs and their integrability: I. Inverse Probl. 3, 229-262 (1987)

27. Calogero, F, Lillo, SD: The Eckhaus PDE $i\psi_t + \psi_{xx} + 2(|\psi|^2)_x\psi + |\psi|^4\psi = 0$. Inverse Probl. 4, 633-682 (1987)

28. Calogero, F: The evolution partial differential equation $u_t = u_{xxx} + 3(u_{xx}u^2 + 3u_x^2 u) + 3u_x u^4$. J. Math. Phys. 28, 538-555 (1987)

29. Golmankhaneh, AK, Baleanu, D: On nonlinear fractional Klein-Gordon equation. Signal Process. 91, 446-451 (2011)

30. El-Sayed, S: The decomposition method for studying the Klein-Gordon equation. Chaos Solitons Fractals 18, 1025-1030 (2003)

31. Odibat, Z, Momani, SA: Numerical solution of sine-Gordon equation by variational iteration method. Phys. Lett. A 370, 437-440 (2007)

32. Lu, Y: A simple method for solving nonlinear wave equations for their peaked soliton solutions and its applications. Acta Phys. Sin. 58, 7452-7456 (2009)

33. Wang, ML, Li, LX, Li, EQ: Exact solitary wave solutions of nonlinear evolutions with a positive fractional power term. Commun. Theor. Phys. 61, 7-14 (2014)

Chapter

5

MATHEMATICAL MODELLING OF MANTLE CONVECTION AT A HIGH RAYLEIGH NUMBER WITH VARIABLE VISCOSITY AND VISCOUS DISSIPATION

Sumaiya B. Islam, Suraiya A. Shefa, and Tania S. Khaleque

Department of Applied Mathematics, University of Dhaka, Dhaka 1000, Bangladesh

ABSTRACT

In this paper, the classical Rayleigh–Bénard convection model is considered and solved numerically for extremely large viscosity variations (i.e., up to 10^{30}) across the mantle at a high Rayleigh number. The Arrhenius form of viscosity is defined as a cut-off viscosity function. The effects of viscosity variation and viscous dissipation on convection with temperature-dependent viscosity and also temperature- and pressure-dependent viscosity are shown through the figures of temperature profiles and streamline contours. The values of Nusselt number and root mean square velocity indicate that the

Citation: Citation (APA): Islam, S. B., Shefa, S. A., & Khaleque, T. S. (2022). Mathematical modelling of mantle convection at a high Rayleigh number with variable viscosity and viscous dissipation. Journal of the Egyptian Mathematical Society, 30(1), 1-17. (17 pages).

convection becomes significantly weak as viscosity variation and viscous dissipation are increased at a fixed pressure dependence parameter.

Keywords: Mantle convection, Variable viscosity, Viscous dissipation, Rayleigh–Bénard convection, Viscosity variation

Mathematics Subject Classifcation: 76D05, 80A19, 8610

INTRODUCTION

Convection in mantle is responsible for most of the physical and chemical phenomena happening on the surface and in the interior of the Earth, and it is caused by the heat transfer from the interior to the Earth's surface. Even though there are some debates, it is quite well established that convection in the mantle is the driving mechanism for plate tectonics, seafloor spreading, volcanic eruptions, earthquakes, etc. [1]. However, the mechanism of mantle convection is still an unsolved mystery since the rheology of mantle rocks is extremely complicated [2,3,4]. Temperature, pressure, stress, radiogenic elements, creep, and many other factors influence the mantle's behavior on a large scale. One of its significant but complex characteristics is its viscosity, which is dependent mainly on temperature, pressure, and stress [5]. In earlier studies of mantle convection, scientists assumed constant viscosity (e.g. [6, 7]) but later, among many others Moresi and Solomatov [8, 9], studied the temperature-dependent viscosity case numerically and concluded that the formation of an immobile lithosphere on terrestrial planets like Mars and Venus seems to be a natural result of temperature-dependent viscosity. However, studies with purely temperature-dependent viscosity cannot portray the true convection pattern of the Earth's mantle. As a result, convection with temperature and pressure-dependent viscosity is becoming more important, and some notable works in this area have recently been published [10,11,12,13,14]. Christensen [10] showed that additional pressure dependence of viscosity strongly influences the flow regimes. In a 2D axi-symmetrical model, Shahraki and Schmeling [15] examined the simultaneous effect of pressure and temperature-dependent rheology on convection and geoid above the plumes, and Fowler et al. [16] studied the asymptotic structure of mantle convection at high viscosity contrast.

According to King et al. [17], when pressure increases through the mantle, there is a corresponding increase in density due to self-compression. In a vigorously convecting mantle, the rate at which viscous dissipation, which is the irreversible process that changes other forces into heat, is non-negligible and contributes to the heat energy of the fluid, resulting in adiabatic temperature and density gradients that reduce the vigour of convection. Conrad and Hager [18] proposed that the viscous dissipation and resisting force to plate motion may have significant effects on convection and the thermal evolution history of the Earth's mantle. Leng and Zhong [19] concluded that the dissipation occurring in a subduction zone is 10–20% of the total dissipation for cases with only temperature-dependent viscosity, whereas Morgan et al. [20] declared that when slabs subduct, about 86% of the gravitational energy for the whole mantle flow is mostly transformed into heat by viscous dissipation. According to Balachandar et al. [21], numerical simulations of 3D convection with temperature-dependent viscosity and viscous heating at realistic Rayleigh numbers for Earth's mantle reveal that, in the strongly time-dependent regime, very intense localized heating takes place along the top portion of descending cold sheets and also at locations where the ascending plume heads impinge at the surface. They also found that the horizontally averaged viscous dissipation is concentrated at the top of the convecting layer and has a magnitude comparable to that of radioactive heating. King et al. [17] worked on a benchmark for 2-D Cartesian compressible convection in the Earth's mantle where they used steady-state constant and temperature-dependent viscosity cases as well as time-dependent constant viscosity cases. In their work, the Rayleigh numbers are near 10^6 and dissipation numbers are between 0 and 2, and they conclude that the most unstable wavelengths of compressible convection are smaller than those of incompressible convection. As the research on mantle convection is growing, the importance of studying viscous dissipation is also increasing since it was suggested that the bending of long and highly viscous plates at subduction zones dissipates most of the energy that drives mantle convection [22]. Some notable recent works on numerical studies of convection and effects of variable viscosity and viscous dissipation have been done by Ushachew et al. [23], Megahed [24], Ferdows et al. [25], Ahmed et al. [26], Fetecau et al. [27].

Although mantle convection is a 3D problem, many 2D codes have been developed to gain an understanding of the fundamental mechanism and to minimize the computational cost and complexity. As the Earth's mantle has been affected by many complexities, its basic understanding has been constructed through research on simple Rayleigh–Bénard convection [2]. Over the years, the Rayleigh–Bénard convection has become a benchmark problem in computational geophysics as a paradigm for convection in the Earth's mantle. Although Rayleigh–Benard convection with viscosity variation is a well-known topic for mantle convection, very high viscosity variation (up to 10^{30}) for mantle convection is not widely covered. To the best of our knowledge, mantle convection with strongly variable viscosity, which is temperature dependent and also both temperature and pressure dependent with the inclusion of viscous dissipation, has not been studied so far. The governing equation in two-dimensional form ensures the conservation of mass, momentum, and energy and the thermodynamic equation of state. In this study, incompressible mantle convection will be considered where the mantle viscosity depends strongly on both temperature and pressure, and viscous dissipation is also considered. The convection will be investigated at a high Rayleigh number with high viscosity variations across the mantle.

In "Methods" section the full governing equations for mantle convection and the appropriate boundary conditions for classical Rayleigh–Bénard convection in a 2D square cell are described. The equations are non-dimensionalized and the dimensionless parameters are identified. Though the variable viscosity is defined in an Arrhenius form, a modified form of viscosity is used to improve the efficiency of numerical computation. The computational method for simulation is also described, and the code is verified using some benchmark values. Then the governing model is solved numerically in a unit aspect-ratio cell for extremely large viscosity variations, and steady solutions for temperature and streamlines are obtained. The numerical and graphical results of the computation are described in "Result and discussion" section. Finally, in "Conclusion" section some concluding remarks on the results are given.

METHODS

Governing Equations

A classical Rayleigh–Bénard convection in a two-dimensional unit aspect ratio cell with a free slip boundary condition is taken into account. The temperature difference is fixed between the horizontal boundaries. The convective cell is assumed to be a section of a periodic structure in the associated infinite horizontal layer. When adopting Cartesian coordinates (x, z) with horizontal x-axis and vertical z-axis, the Boussinesq approximation is assumed, which suggests that density variation is barely vital within the buoyancy term of the momentum equation, so that mass conservation takes the shape of the incompressibility condition [16]. The inertia terms within the Navier–Stokes equations (taking the limit of an infinite Prandtl number) are neglected as well. According to Solomatov [28], the integral viscous dissipation within the layer is often balanced by the integral mechanical work done by thermal convection per unit time, and if the viscosity contrast is large, dissipation in the cold boundary layer becomes comparable with the dissipation in the internal region. Thus, in order to balance the energy equation, the extended Boussinesq approximation is used. Here, "extended Boussinesq approximation" means that apart from the driving buoyancy forces, the fluid is treated as being incompressible all over. The non-Boussinesq effects of the adiabatic gradient and frictional heating are introduced into the energy equation [29]. The governing equations ensure the conservation of mass, momentum, and energy. This also ensures a suitable thermodynamic equation of state. The Navier–Stokes equations, which describe the motion in component forms, are [30]

$$\begin{aligned}
&\frac{\partial u}{\partial x}+\frac{\partial w}{\partial z}=0,\\
&\qquad\frac{\partial P}{\partial x}=\frac{\partial \tau_1}{\partial x}+\frac{\partial \tau_3}{\partial z},\\
&\qquad\frac{\partial P}{\partial z}=\frac{\partial \tau_3}{\partial x}-\frac{\partial \tau_1}{\partial z}-\rho g,\\
&\qquad\tau_1=2\eta\frac{\partial u}{\partial x},\\
&\qquad\tau_3=\eta\left(\frac{\partial u}{\partial z}+\frac{\partial w}{\partial x}\right),\\
&\qquad\rho=\rho_0[1-\alpha(T-T_b)],
\end{aligned}\tag{1}$$

The energy equation is

$$\frac{\partial T}{\partial t}+\boldsymbol{u}.\nabla T-\frac{\alpha T}{\rho C_{\mathrm{p}}}\left(\frac{\partial P}{\partial t}+\boldsymbol{u}.\nabla P\right)=\kappa\left(\nabla^2 T\right)+\frac{\tau^2}{2\eta\rho C_{\mathrm{p}}}.\tag{2}$$

Here, *P* is the pressure, τ is the deviatoric stress tensor, *t* is time, ρ is the density, u = (u, 0, w) is the fluid velocity, where *u* and *w* are velocity components in the *x*- and *z*-directions, *g* is the assumed constant gravitational acceleration acting downwards (the variation of *g* across the mantle is quite small that it is taken as constant), τ_1 and τ3τ3 are the longitudinal and shear components of the deviatoric stress tensor, respectively, ηη is the viscosity, T_b is the basal temperature, ρ_0 is the basal density, κ is the thermal diffusivity, *T* is the absolute temperature, C_p is the specific heat at constant pressure, and α is the thermal expansion coefficient.

The deviatoric stress tensor, τ can be expressed as

$$\boldsymbol{\tau} = \tau_1^2 + \tau_3^2$$

where τ_1 and τ_3 are the longitudinal and shear components of the deviatoric stress tensor, respectively.

The Arrhenius form of viscosity function is

$$\eta = \frac{1}{2A\left(\tau_1^2 + \tau_3^2\right)^{\frac{n-1}{2}}} \exp\left[\frac{E + pV}{RT}\right], \tag{3}$$

where *A* is the rate factor, *n* is the flow index, *E* is the activation energy, *V* is the activation volume, and *R* is the universal gas constant [5].

A unit aspect-ratio cell with a free-slip boundary condition is considered. The temperatures at the bottom and top boundaries are taken as constant, and thermal insulation is assumed on the side walls. The boundary conditions are

$$\begin{aligned} &w = 0, \quad \tau_3 = 0, \quad = T_b \quad \text{on} \quad z = 0, \\ &w = 0, \quad \tau_3 = 0, \quad T = T_s \quad \text{on} \quad z = d, \\ &u = 0, \quad \tau_3 = 0, \quad \frac{\partial T}{\partial x} = 0 \quad \text{on} \quad x = 0, d. \end{aligned} \tag{4}$$

where *d* is the depth of the convection cell, T_b and T_s are the basal and top temperatures, respectively (Fig. 1).

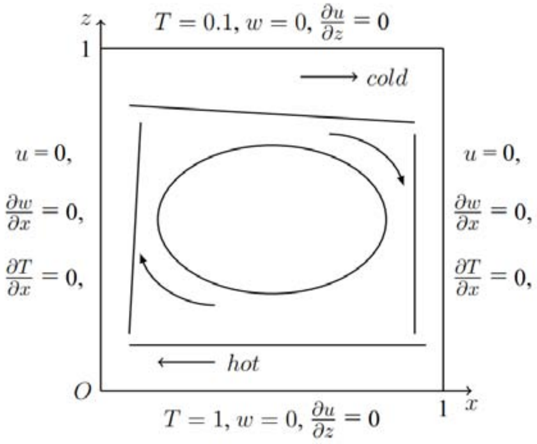

$T = 0.1, w = 0, \frac{\partial u}{\partial z} = 0$
z
1
cold
$u = 0,$
$\frac{\partial w}{\partial x} = 0,$
$\frac{\partial T}{\partial x} = 0,$
$u = 0,$
$\frac{\partial w}{\partial x} = 0,$
$\frac{\partial T}{\partial x} = 0,$
hot
O
1
x
$T = 1, w = 0, \frac{\partial u}{\partial z} = 0$

Figure 1. Schematic diagram of a basally heated non-dimensional unit aspect-ratio cell in mantle.

Throughout this work, Newtonian rheology is considered with n = 1 in the viscosity relation and internal heating is neglected. To see the effects of variable viscosity (both temperature-dependent and temperature-and pressure-dependent viscosity) and viscous dissipation on convection, these assumptions are made to make the model less complicated.

Non-dimensionalization and Simplification

In order to non-dimensionalize the model, the variables are set as [7, 30]

$$\boldsymbol{u} = \frac{\kappa}{d}\boldsymbol{u}^{*}, \quad (x, z) = d(x^{*}, z^{*}), \quad \boldsymbol{\tau} = \frac{\eta_0 \kappa}{d^2}\boldsymbol{\tau}^{*}, \quad \eta = \frac{e^{(1+\mu)/\varepsilon}}{2A}\eta^{*} = \eta_0 \eta^{*},$$
$$P = \rho_0 g d(1 - z^{*}) + \frac{\eta_0 \kappa}{d^2}P^{*}, \quad \rho = \rho_0 \rho^{*}, \quad t = \frac{d^2}{\kappa}t^{*}, \quad T = T_{\mathrm{b}}T^{*} \tag{5}$$

Using these in equations from (1) to (3) and dropping the asterisk decorations, the dimensionless equations becomes

$$\begin{aligned}
&\frac{\partial u}{\partial x} + \frac{\partial w}{\partial z} = 0, \\
&\frac{\partial P}{\partial x} = \frac{\partial \tau_1}{\partial x} + \frac{\partial \tau_3}{\partial z}, \\
&\frac{\partial P}{\partial z} = \frac{\partial \tau_3}{\partial x} - \frac{\partial \tau_1}{\partial z} - \mathrm{Ra}(1 - T), \\
&\tau_1 = 2\eta\frac{\partial u}{\partial x}, \\
&\tau_3 = \eta\left(\frac{\partial u}{\partial z} + \frac{\partial w}{\partial x}\right),
\end{aligned} \tag{6}$$

$$\frac{\partial T}{\partial t} + \boldsymbol{u} \cdot \nabla T - DT\frac{\bar{B}}{\mathrm{Ra}}\frac{\partial P}{\partial t} + DTw - DT\frac{\bar{B}}{\mathrm{Ra}}u \cdot \nabla P = \nabla^2 T + \frac{D}{\mathrm{Ra}}\frac{\tau^2}{2\eta}, \tag{7}$$

while the dimensionless version of constitutive relation (3) reads

$$\eta = \exp\left[\frac{(1 - T)(1 + \mu) - \mu z + \mu\bar{B}p/\mathrm{Ra}}{\varepsilon T}\right], \tag{8}$$

in which the dimensionless parameters are,

$$\text{Dissipation number}, D = \frac{\alpha g d}{C_\mathrm{p}},$$

$$\text{Viscous temperature parameter}, \varepsilon = \frac{R T_\mathrm{b}}{E},$$

$$\text{Viscous pressure number}, \mu = \frac{\rho_0 g \mathrm{d} V}{E},$$

$$\text{Boussinesq number}, \bar{B} = \alpha T_\mathrm{b},$$

$$\text{Rayleigh number}, \mathrm{Ra} = \frac{\alpha \rho_0 g T_\mathrm{b} d^3}{\eta_0 \kappa}. \tag{9}$$

Since this model was developed for the mantle, the typical values of the parameters are given in Table 1, and it is found that for $\mathrm{Ra} >> 1, \bar{B}/\mathrm{Ra}$ can be easily ignored. Therefore, the dimensionless energy equation (7) becomes

$$\frac{\partial T}{\partial t} + \boldsymbol{u} \cdot \nabla T + DTw = \nabla^2 T + \frac{D}{\mathrm{Ra}} \frac{\tau^2}{2\eta}, \tag{10}$$

and viscosity relation (8) becomes

$$\eta = \exp\left[\frac{1 - T + \mu(1 - z - T)}{\varepsilon T}\right]. \tag{11}$$

This Eq. (11) is known as full form of Arrhenius viscosity function.

Table 1. Typical parameter values for mantle convection with variable viscosity

Parameter	**Symbol**	**Value**
Mantle depth	d	3×10^6 m
Thermal expansion coefficient	α	2×10^{-5} K^{-1}
Reference density	ρ_0	4×10^3 kg m^{-3}
Gravitational acceleration	g	10 m s^{-2}
Temperature at the top of the mantle	T_s	300 K
Temperature at the base of the mantle	T_b	3000 K
Temperature difference	ΔT	2700 K
Thermal conductivity	k	4 W m^{-1}K^{-1}
Specific heat at constant pressure	C_p	10^3 J kg^{-1} K^{-1}
Activation energy	E	300 – 525 kJ mol^{-1}

Activation volume	V	6×10^{-6} m^3 mol^{-1}
Gas law constant	R	8.31 J mol^{-1} K^{-1}
Viscous rate constant	A	10^5 MPa^{-1} s^{-1}
Thermal diffusivity	κ	1×10^{-6} m^2 s^{-1}
Rayleigh number	Ra	$10^7 - 10^9$
Viscous temperature parameter	ε	0.042 − 0.083
Viscous pressure number	μ	1.2 − 2.4
Boussinesq number	$\bar{B}$	0.06
Dimensionless surface temperature	θ_0	0.1
Dissipation number	D	0.6

The dimensionless boundary conditions (4) become

$$\begin{aligned}
&w = 0, \quad \tau_3 = 0, \quad T = 1 \quad \text{on} \quad z = 0,\\
&w = 0, \quad \tau_3 = 0, \quad T = \frac{T_s}{T_b} = \theta_0 \quad \text{on} \quad z = 1,\\
&u = 0, \quad \tau_3 = 0, \quad \frac{\partial T}{\partial x} = 0 \quad \text{on} \quad x = 0, 1.
\end{aligned} \qquad (12)$$

The dimensionless model consists of governing Eqs. (6), (10), viscosity relation (11) and boundary conditions (12).

Low Temperature Cut-Off Viscosity

To investigate the convection with extremely high viscosity contrasts in the mantle layer, a low temperature cut-off viscosity function is used. This cut-off viscosity relation helps reduce the computational stiffness while retaining the sensitivity of the viscosity to the changes in temperature and pressure across the mantle. It is a well-established fact that in strongly temperature-dependent viscous convection, most of the viscosity variation occurs in a stagnant lid in which the velocity is essentially zero. Based on this fact, the sub-lid convection field is calculated accurately (but not the stress field) by cutting off the dimensionless viscosity at a sufficiently high value that the lid thickness, which essentially only depends on the interaction of the lid temperature with the underlying convection flow, is unaffected.

The low temperature cut-off viscosity function has the following form

$$\eta = \begin{cases} \exp[Q/\varepsilon] & \frac{Q}{\varepsilon} \le \log 10^r, \\ 10^r & \frac{Q}{\varepsilon} > \log 10^r, \end{cases} \qquad (13)$$

where

$$Q = \frac{1 - T + \mu(1 - z - T)}{T}, \tag{14}$$

and the cut-off viscosity value 10r is to be chosen appropriately; in numerical experiments, it is chosen r = 6. Similar type of Arrhenius law with an imposed cut-off viscosity was applied by Huang et al. [31], Huang and Zhong [32], King [33] and Khaleque et al. [13]. A comparison between full-form viscosity function and cut-off viscosity function is shown in "Comparison with benchmark values and validation" section.

Three useful diagnostic quantities which will be used to characterize are viscosity contrast, Nusselt number and root mean square velocity respectively.

The viscosity contrast Δη is the ratio between the surface and basal values of the viscosity, defined as

$$\Delta\eta = \exp\left(\frac{1 - \theta_0 - \mu\theta_0}{\varepsilon\theta_0}\right),$$

where $\theta_0 = \frac{T_s}{T_b}$.

The Nusselt number Nu is the ratio of the average surface heat flow from the convective solution to the heat flow due to conduction. It is calculated in the present case of a square cell by the dimensionless relation

$$\mathrm{Nu} = -\frac{1}{(1 - \theta_0)} \int_0^1 \frac{\partial T}{\partial z}(x, 1)\mathrm{d}x.$$

Nu is equal to unity for conduction and exceeds unity as soon as convection starts.

The vigour of the circulating flow is characterised by the non-dimensional RMS (root mean square) velocity. Here RMS velocity is defined by

$$V_{\mathrm{rms}} = \left[\int_0^1 \int_0^1 (u^2 + w^2)\mathrm{d}x\mathrm{d}z\right]^{1/2},$$

where *u* is the horizontal component of velocity and *w* is the vertical component of velocity.

Computational Method

In order to solve the dimensionless governing Eqs. (6), (10), (11) with boundary conditions (12) a finite element method based PDE solver *'COMSOL Multiphysics 5.3'* is used. The modules for creeping flow, heat transfer in fluids, and Poisson's equation are chosen based on the physics of the model. Free triangular meshing with some refinement near the boundaries of 200 × 200 and *COMSOL*'s "extra fine" setting results in a complete mesh of a total of 18,000 elements. As the basis functions or shape functions, Lagrangian P2–P1 elements for creeping flow are selected, which means the shape functions for the velocity field and pressure are Lagrangian quadratic polynomials and Lagrangian linear polynomials, respectively. Similarly, Lagrangian quadratic elements for both temperature in the heat equation and the stream function in Poisson's equation are chosen. For Lagrange elements, the values of all the variables at the nodes are called degrees of freedom (dof) and in this case, our specific discretization finally produces 153,816 degrees of freedom (N_{dof}). The following convergence criterion is applied for all cases:

$$\left(\frac{1}{N_{\text{dof}}}\sum_{i=1}^{N_{\text{dof}}}|E_i|^2\right)^{\frac{1}{2}} < \varepsilon \tag{15}$$

where E_i is the estimated error and $\varepsilon = 10^{-6}$. Further details of the method can be found in Zimmerman [34].

Comparison with Benchmark Values and Validation

The values of Nusselt number Nu and root mean square velocity V_{rms} are compared with the benchmark values from Blankenbach et al. [35][a] and Koglin Jr et al. [36][b] in Table 2 for constant viscosity case. Their values were computed for *Ra* up to 10^6 and 10^7 respectively. From Table 2, it is evident that the agreement is within a very good range.

Table 2. Comparison of computed Nusselt number Nu and RMS velocity V_{rms} with benchmark values from Blankenbach et al. [35][a] and Koglin Jr et al. [36][b]

Ra	**Nu**		**V_{rms}**	
	This work	**Benchmark**	**This work**	**Benchmark**
10^4	4.884409	4.884409[a]	42.864973	42.864947[a]
10^5	10.534113	10.534095[a]	193.215527	193.21454[a]
10^6	21.972563	21.972465[a]	834.004359	833.98977[a]
10^7	45.638611	45.62[b]	3633.932754	–

Table 3. Comparison of Nusselt number, Nu of full-form viscosity function (11) and cut-off viscosity function (13) for $\mu = 0.0$ and $\mu = 0.5$ at Ra = 10^7 and $\theta_0 = 0.1$

$\Delta\eta$	**Full form η**		**Cut-off η**	
	$\mu = 0.0$	$\mu = 0.5$	$\mu = 0.0$	$\mu = 0.5$
10^{10}	6.76217	8.06845	6.76800	8.06845
10^{15}	5.35744	6.98327	5.36157	6.98490
10^{20}	4.44652	6.29296	4.45036	6.29310
10^{25}	3.79703	5.79253	3.80090	5.79442
10^{30}	3.25274	5.39134	3.25696	5.39304

Then the computation is done with variable viscosity with a high viscosity contrast across the mantle layer. The values of Nusselt number Nu that are compared in Table 3 are found using the full form viscosity function (11) and the cut-off viscosity function (13) for $\mu = 0.5$ and $\mu = 0.0$. It should be noted that $\mu = 0.0$ indicates temperature-dependent viscosity, whereas $\mu \neq 0$ implies that viscosity depends on both temperature and pressure. From Table 3 it can be seen that the values of Nusselt number, Nu with full form viscosity function and the values of Nusselt number, Nu with cut-off viscosity function are very close, which validates the use of the cut-off viscosity function for numerical computation.

RESULT AND DISCUSSION

After validating the model, the governing Eqs. (6), (10) and (13) with boundary conditions (12) are solved. Throughout the computation, the constants $\theta_0 = 0.1$ and Ra = 10^7 are used, and the values of the Nusselt number, Nu, and root mean square velocity, V_{rms} for different dissipation numbers, *D*, pressure dependent parameter μ, and temperature dependent parameter ε are calculated. By varying μ and ε, different viscosity contrast is obtained across the mantle layer. The numerical computations with D = 0.3 and D = 0.6 at Ra = 10^7 when μ = 0.0, μ = 0.5 and μ = 1.0 are performed, and the calculated Nusselt number and the RMS velocity values for high viscosity contrasts from 10^{10} to 10^{30} are shown in Tables 4 and 5.

Table 4. Nusselt number Nu computed for μ = 0.0, μ = 0.5, μ = 1.0 with different viscous dissipation number *D* at Ra = 10^7 and $\theta_0 = 0.1$

Δη	**Nusselt number Nu**								
	μ = 0.0			**μ = 0.5**			**μ = 1.0**		
	D = 0.0	**D = 0.3**	**D = 0.6**	**D = 0.0**	**D = 0.3**	**D = 0.6**	**D = 0.0**	**D = 0.3**	**D = 0.6**
10^{10}	6.76800	3.83912	2.22500	8.06845	4.46772	2.45665	9.35884	5.22977	2.76453
10^{15}	5.36157	2.90745	1.73657	6.98490	3.62823	1.93308	8.20184	4.58926	2.25912
10^{20}	4.45036	2.34275	1.48655	6.29310	3.09495	1.63913	6.85055	4.18656	1.93551
10^{25}	3.80090	2.00320	1.34157	5.79442	2.71655	1.45780	5.43699	3.88971	1.71162
10^{30}	3.25696	1.77292	1.24780	5.39304	2.40512	1.33803	4.79113	3.64133	1.54933

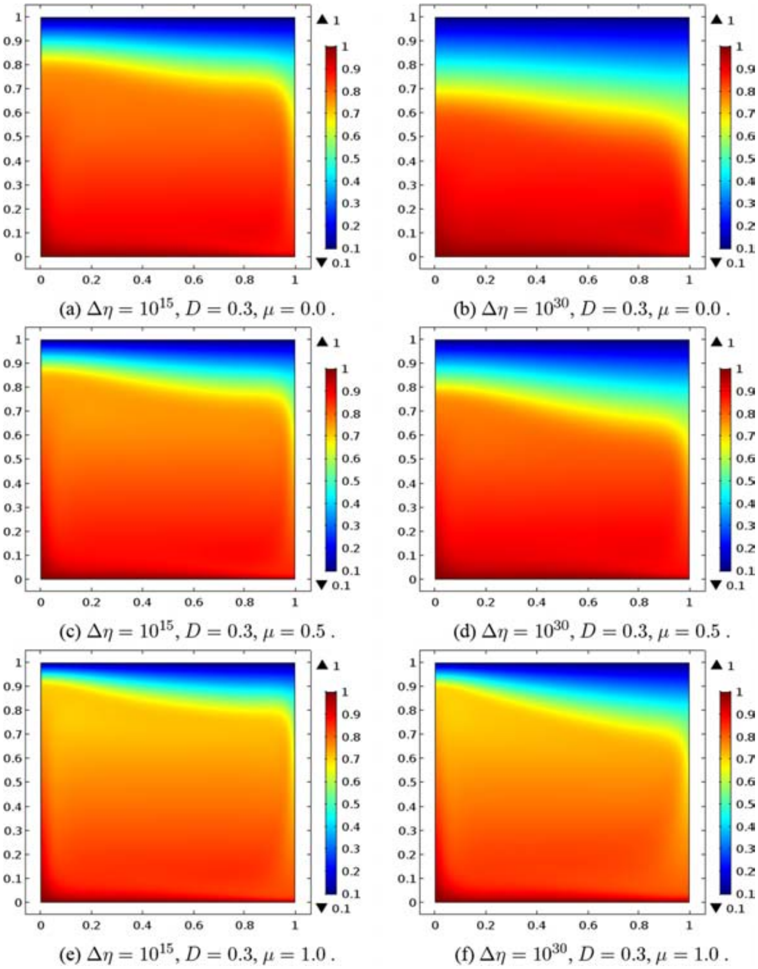
Table 5. RMS velocity V_{rms} computed for μ = 0.0, μ = 0.5, μ = 1.0 with different viscous dissipation number *D* at Ra=10^7 and $\theta_0 = 0.1$

Δη	**RMS velocity V_{rms}**								
	μ = 0.0			**μ = 0.5**			**μ = 1.0**		
	D = 0.0	D = 0.3	D = 0.6	D = 0.0	D = 0.3	D = 0.6	D = 0.0	D = 0.3	D = 0.6
10^{10}	753.149	450.644	244.368	1000.932	580.307	292.445	1189.156	722.471	350.050
10^{15}	594.920	314.292	161.594	894.191	468.522	204.168	949.421	640.405	273.675
10^{20}	483.396	222.950	115.209	806.197	388.104	149.557	585.505	562.901	214.198
10^{25}	398.149	173.658	85.850	721.782	324.978	113.462	307.379	481.827	169.948
10^{30}	308.093	139.281	65.259	634.734	259.877	87.966	273.349	398.414	136.766

Tables 4 and 5 show that for each fixed value of μ and *D*, Nu and V_{rms} decrease as the viscosity contrast increases (i.e., the temperature dependence parameter decreases) across the mantle. It confirms that at the higher viscosity variation, convection becomes weaker, which can also be seen clearly in the thermal distribution Figs. 2 and 3. Nu and V_{rms} values also decrease as *D* increases for every particular value of μ.

It is also observed that at a specific viscosity contrast as the pressure dependence parameter μ is increased, both Nu and V_{rms} values increase for a fixed dissipation number D = 0.3 and D = 0.6. The reason behind this is that even though μ is increased, ε is actually decreased to maintain the fixed viscosity contrast. However, for D = 0.0, the trend is not that smooth at higher viscosity variations. Comparing the V_{rms} values between D = 0.0 and D = 0.3 at μ = 1.0 it can be seen that at high viscosity contrasts, the V_{rms} values for D = 0.3 are larger than those for D = 0.0 which are unlike the other values.

The thermal distribution and stream function contours for μ = 0.0, μ=0.5 and μ = 1.0 are presented in Figs. 2, 3 and 4.



(a) $\Delta\eta = 10^{15}, D = 0.3, \mu = 0.0$.

(b) $\Delta\eta = 10^{30}, D = 0.3, \mu = 0.0$.

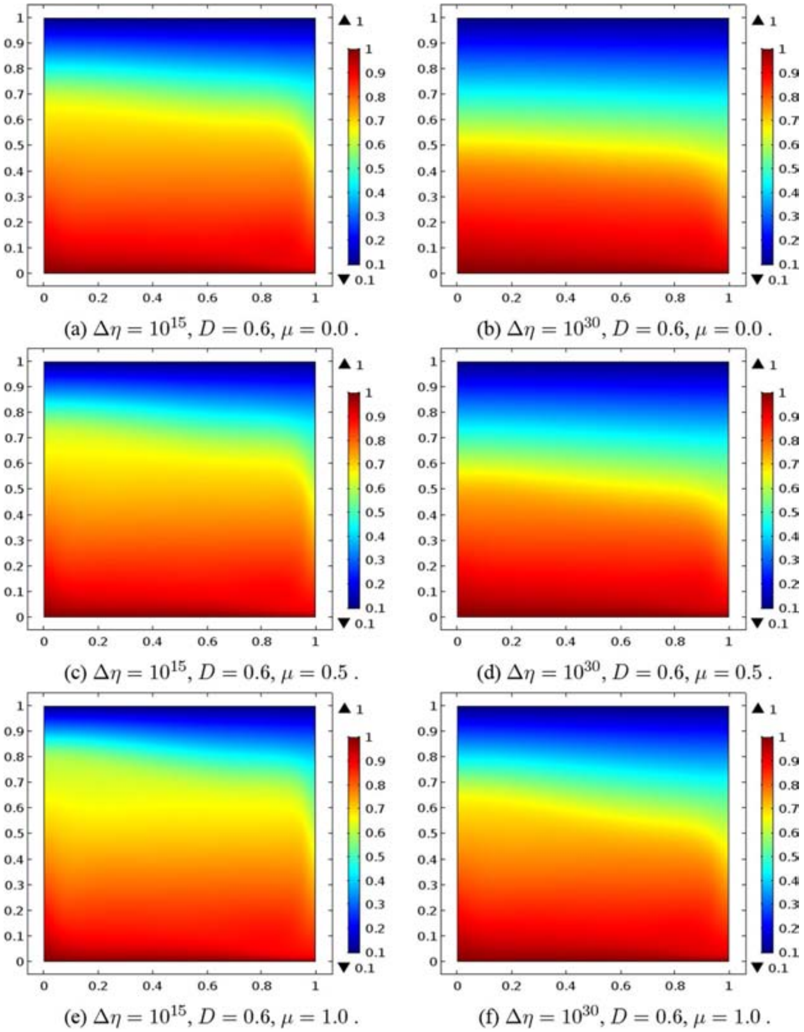
(c) $\Delta\eta = 10^{15}, D = 0.3, \mu = 0.5$.

(d) $\Delta\eta = 10^{30}, D = 0.3, \mu = 0.5$.

(e) $\Delta\eta = 10^{15}, D = 0.3, \mu = 1.0$.

(f) $\Delta\eta = 10^{30}, D = 0.3, \mu = 1.0$.

Figure 2. Thermal distributions of a convection at different viscosity variations and at different pressure numbers for a fixed viscous dissipation number D = 0.3 with $\theta_0 = 0.1$ and Ra = 10^7.



(a) $\Delta\eta = 10^{15}, D = 0.6, \mu = 0.0$.

(b) $\Delta\eta = 10^{30}, D = 0.6, \mu = 0.0$.

(c) $\Delta\eta = 10^{15}, D = 0.6, \mu = 0.5$.

(d) $\Delta\eta = 10^{30}, D = 0.6, \mu = 0.5$.

(e) $\Delta\eta = 10^{15}, D = 0.6, \mu = 1.0$.

(f) $\Delta\eta = 10^{30}, D = 0.6, \mu = 1.0$.

Figure 3. Thermal distributions of a convection at different viscosity variations and at different pressure numbers for a fixed viscous dissipation number D = 0.6 with $\theta_0 = 0.1$ and Ra = 10^7.

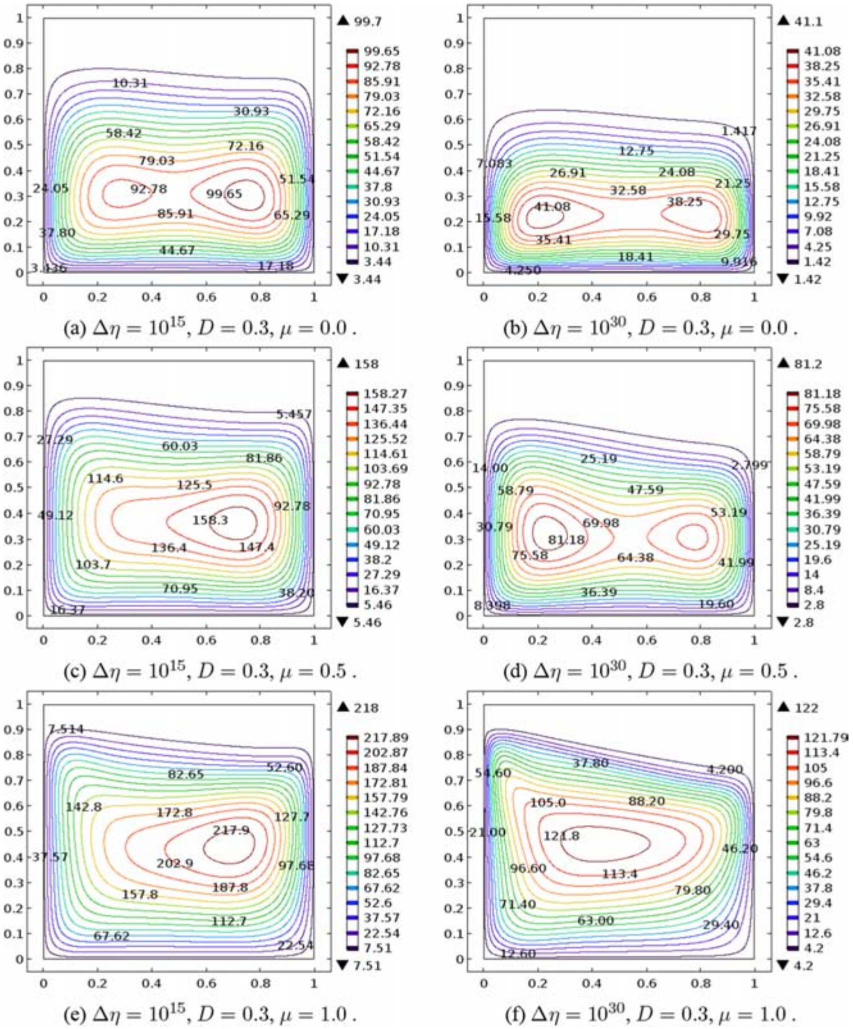
In Figs. 2 and 3 the thermal distribution of the unit aspect ratio convection cell for the values of D = 0.3 and D = 0.6 respectively are presented for different viscosity contrasts. In panel 2a, b and 3a, b, the viscosity depends only on temperature (i.e. μ = 0.0) and in panel 2c, f and 3c, f, the viscosity depends on both temperature and pressure (i.e. μ ≠ 0.0). At each plot of the temperature profile, the blue region corresponds to the cooler temperature whereas the red region corresponds to the high temperature.

For μ = 0.0, μ = 0.5 and μ = 1.0, from Figs. 2 and 3 we see that as viscosity contrast Δη increases the thickness of the cold thermal boundary layer at the top of the cell. At the lower mantle, which is near the core of the Earth, the boundary is hot as the temperature is very high and this temperature continues to increase as the viscosity contrast gets larger. The interior temperature decreases significantly as the pressure dependence parameter is included. The convection cell is quite different when viscosity is both temperature and pressure dependent rather than only temperature dependent. Compared to μ = 0.5 the significance of pressure can be seen clearly for μ = 1.0 from both Figs. 2 and 3.

The stream function contours where stream function Ψ(x, z) defined as

$$u = -\Psi_z, \qquad w = \Psi_x, \tag{16}$$

are presented in Fig. 4 for D = 0.3. As the streamlines represent fluid flow, the absence of a streamline confirms that fluid in that region is immobile. In other words, this immobile region represents the stagnant lid. With increasing viscosity contrast and viscous dissipation, the changes in the convection pattern are very clear. It is observed that the cold thermal boundary layer thickness increases with viscosity contrast. But for a fixed dissipation number, the cold thermal boundary thickness is reduced with the inclusion of the pressure-dependent parameter μ. Clearly, the lid thickness decreases as the pressure dependence parameter is increased at a fixed viscosity variation. However, the lid thickness increases when viscosity variation is increased at a fixed pressure dependence parameter μ and dissipation number *D*. The Tables 4 and 5 clearly indicate that the heat transfer rate and the root mean square velocity decrease, and Figs. 2, 3 and 4 show that the immobile lid thickness increases as the viscosity contrast at a fixed pressure dependent parameter is increased. The decrease in Nu and V_{rms} values, as well as the increase in the thickness of the cold thermal boundary layer, imply that the convection becomes significantly weaker.



(a) $\Delta\eta = 10^{15}, D = 0.3, \mu = 0.0$.

(b) $\Delta\eta = 10^{30}, D = 0.3, \mu = 0.0$.

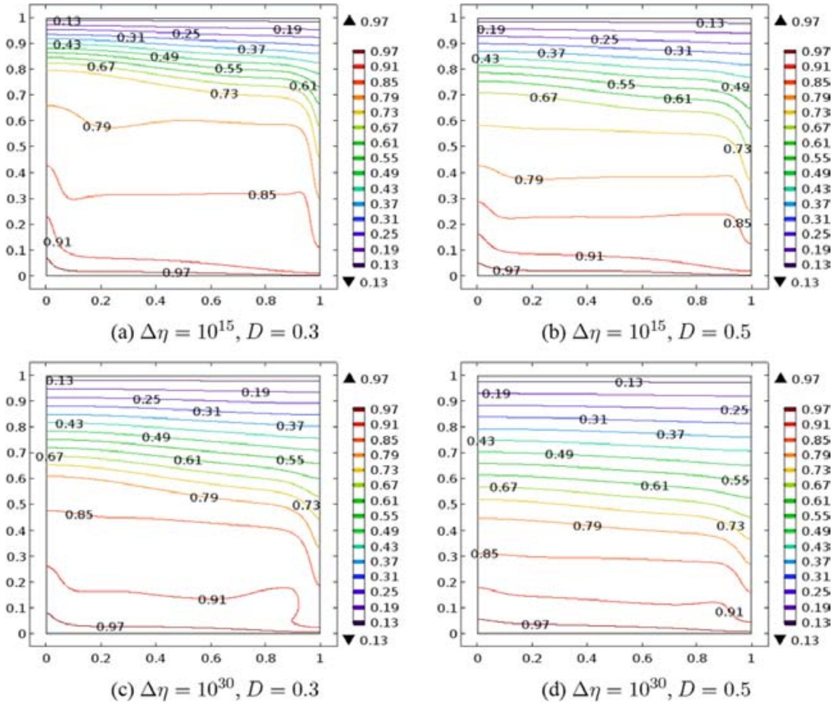
(c) $\Delta\eta = 10^{15}, D = 0.3, \mu = 0.5$.

(d) $\Delta\eta = 10^{30}, D = 0.3, \mu = 0.5$.

(e) $\Delta\eta = 10^{15}, D = 0.3, \mu = 1.0$.

(f) $\Delta\eta = 10^{30}, D = 0.3, \mu = 1.0$.

Figure 4. Stream function contours of a convection at different viscosity variations and at different pressure numbers for a fixed viscous dissipation number D = 0.3 with $\theta_0 = 0.1$ and Ra=10^7.



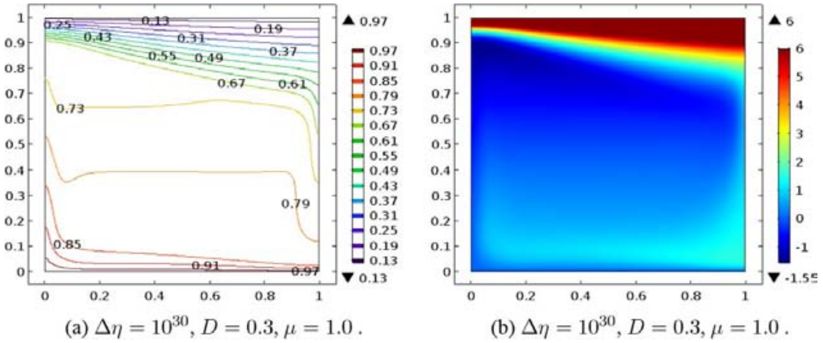
(a) $\Delta\eta = 10^{15}$, $D = 0.3$

(b) $\Delta\eta = 10^{15}$, $D = 0.5$

(c) $\Delta\eta = 10^{30}$, $D = 0.3$

(d) $\Delta\eta = 10^{30}$, $D = 0.5$

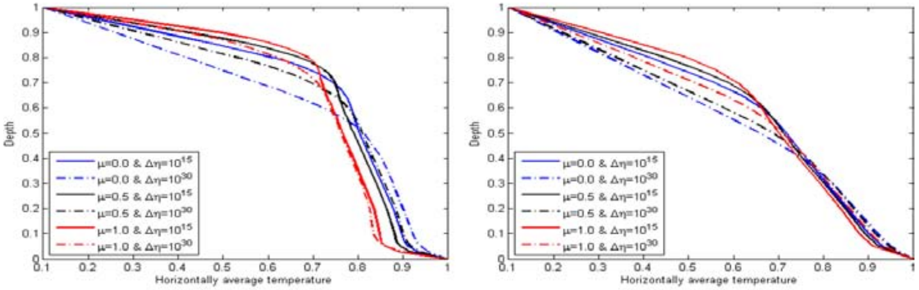
Figure 5. Isothermal contours of a temperature dependent viscosity convection at different viscosity variations and viscous dissipation number with $\theta_0 = 0.1$ and Ra = 10^7.



(a) $\Delta\eta = 10^{30}$, $D = 0.3$, $\mu = 1.0$.

(b) $\Delta\eta = 10^{30}$, $D = 0.3$, $\mu = 1.0$.

Figure 6. a Isothermal contour and **b** distribution of $\log_{10}\eta$ for $\mu = 1.0$ at $\Delta\eta=10^{30}$ and viscous dissipation D = 0.3 with $\theta_0 = 0.1$ and Ra = 10^7.



(a) $D = 0.3$.

(b) $D = 0.6$.

Figure 7. Horizontally average temperature vs depth profiles at viscosity contrasts $\Delta\eta = 10^{15}$ and $\Delta\eta = 10^{30}$ for convection with $\mu = 0.0$, $\mu = 0.5$ and $\mu = 1.0$ at $\theta_0 = 0.1$ and $Ra = 10^7$ with viscous dissipation D = 0.3 and D = 0.6.

A visualization of the isothermal contours in Fig. 5 shows that the hot thermal boundary layer is very thin compared to the cold thermal boundary layer. This figure represents the isothermal contours of a convection cell with temperature dependent viscosity at different viscosity contrast (i.e $\Delta\eta = 10^{15}$ and $\Delta\eta = 10^{30}$) when viscous dissipation numbers are D = 0.3 and D = 0.5. There might not be any significant difference in the convection pattern (i.e., isothermal contours), but the contours are not similar. They are clearly affected by different viscous dissipation numbers at different viscosity contrasts.

Isothermal contours (Fig. 6a) and viscosity distribution (Fig. 6b) for $\mu = 1.0$ at $\Delta\eta = 10^{30}$ and viscous dissipation D = 0.3 are shown in Fig. 6. The viscosity variation from top to bottom is shown in Fig. 6b, and the resulting color ranges from the lowest value (blue) to 10^6 (brown). Clearly, the cut-off viscosity function simply ignores the high value of the lid viscosity and considers it as a constant there. Figure 6b shows a low viscosity region in the upper mantle and a relatively high viscosity region in the lower mantle just above the bottom boundary layer. This implies that the interior is not isoviscous.

Horizontally average temperature vs depth profiles for viscous dissipation of D = 0.3 and D = 0.6 are presented in Fig. 7. These figures show how the horizontally averaged temperature varies with depth at different viscous dissipation numbers and at different viscosity variations. It also shows how it changes for temperature-dependent viscosity and temperature- and pressure-dependent viscosity. The rapid change in temperature near the cold upper boundary and the hot lower boundary explains the strong temperature gradients in those regions. The plots also indicate that the core of the mantle, i.e. the interior, is not isothermal for both the temperature dependent viscosity case and the temperature and pressure dependent viscosity case. The interior of the convection cell undergoes a larger jump in temperature when dissipation effect is stronger (D = 0.6). The figures show that the interior temperature increases with the increase of viscosity contrast across the mantle layer for μ = 0.0 and μ = 0.5 at D = 0.3 and D = 0.6. Similar situation occurs for μ= 1.0 at D = 0.6 but when D = 0.3, temperature decreases at high viscosity contrast (i.e. at $\Delta\eta = 10^{30}$).

CONCLUSION

The study of a basally heated convection model with a strongly temperature and pressure dependent viscous fluid relative to the Earth's mantle in the presence of viscous dissipation has been the principal aim of this work. The classical Rayleigh–Bénard convection model was solved using a low temperature cut-off viscosity function to avoid the stiffness of computation. It was aimed to pursue viscosity that is dependent only on temperature and simultaneously dependent on both temperature and pressure, and a comparison is presented through figures and tables.

According to Jarvis and Mckenzie [37], the dissipation number is between 0.25 and 0.8, whereas Leng and Zhong [19] estimate *D* to be 0.5 to 0.7. Ricard [38] found that its value is about 1.0 near the surface, and decreases to about 0.2 near the CMB. From Table 1, D ≈ 0.6 has been found. Thus, the effect of various viscous dissipation numbers for mantle like convection with $Ra=10^7$ is checked. The different values of viscous dissipation number show the changes in heat transfer rate Nu and root mean square velocity V_{rms}. It is shown that the fluid is not isothermal and isoviscous in the presence of viscous dissipation in both cases when viscosity is temperature-dependent and temperature-pressure-dependent. The viscosity distribution at high viscosity contrast for μ = 1.0 also showed that the fluid is not isoviscous.

Analysis of the results can predict that if the dissipation number is increased, the lid thickness will increase more and the convection rate will decrease notably. But it is also clear that the inclusion of viscous dissipation does not affect the convection pattern in any drastic way. The convection becomes weaker as viscosity contrast becomes larger and the viscous dissipation number is increased. However, the variation in Nu, V_{rms} increase as μ goes from 0 to 0.5, but the trend is different when μ goes from 0.5 to 1.0. Thus, strong pressure dependence in viscosity affects the convection in a different way. For a temperature-dependent viscosity case and a temperature and pressure-dependent viscosity case, the horizontally averaged temperature increases with viscosity contrast in the interior, but the trend is opposite in the top boundary layer, i.e., the stagnant lid. In this study we investigated convection with high viscosity contrast, because for the typical parameter values, it is estimated that the viscosity contrast for the Earth's mantle is 10^{50} or more. Without extreme parameter values, it is quite impossible to obtain a proper asymptotic structure of mantle convection for the Earth and other planets. Thus, it is believed that this study will have a significant impact on the study of thermal convection in the Earth's mantle and other planets where viscosity is strongly variable and the variation of the order of magnitude is very large.

REFERENCES

1. Runcorn, S.: Mechanism of plate tectonics: mantle convection currents, plumes, gravity sliding or expansion? Tectonophysics 63, 297–307 (1980)
2. Bercovici, D.: Treatise on Geophysics, Mantle Dynamics, vol. 7. Elsevier, Amsterdam (2010)
3. Karato, S.: Rheology of the Earth's mantle: a historical review. Gondwana Res. 18, 17–45 (2010)
4. Rose, I.R., Korenaga, J.: Mantle rheology and the scaling of bending dissipation in plate tectonics. J. Geophys. Res. Solid Earth 116 (2011). https://doi.org/10.1029/2010JB008004
5. Schubert, G., Turcotte, D.L., Olson, P.: Mantle Convection in the Earth and Planets. Cambridge University Press, Cambridge (2001)
6. Turcotte, D.L., Oxburgh, E.R.: Finite amplitude convective cells and continental drift. J. Fluid Mech. 28, 29–42 (1967). https://doi.org/10.1017/S0022112067001880
7. Jarvis, G.T., Peltier, W.R.: Mantle convection as a boundary layer phenomenon. Geophys. J. Int. 68, 389–427 (1982). https://doi.org/10.1111/j.1365-246X.1982.tb04907.x
8. Moresi, L.N., Solomatov, V.: Numerical investigation of 2d convection with extremely large viscosity variations. Phys. Fluids 7, 2154–2162 (1995). https://doi.org/10.1063/1.868465
9. Solomatov, V., Moresi, L.N.: Three regimes of mantle convection with non-Newtonian viscosity and stagnant lid convection on the terrestrial planets. Geophys. Res. Lett. 24, 1907–1910 (1997). https://doi.org/10.1029/97GL01682
10. Christensen, U.: Convection with pressure- and temperature-dependent non-Newtonian rheology. Geophys. J. Int. 77, 343–384 (1984). https://doi.org/10.1111/j.1365-246X.1984.tb01939.x
11. Doin, M.P., Fleitout, L., Christensen, U.: Mantle convection and stability of depleted and undepleted continental lithosphere. J. Geophys. Res. Solid Earth 77, 2771–2787 (1997)
12. Dumoulin, C., Doin, M.P., Fleitout, L.: Heat transport in stagnant lid convection with temperature- and pressure-dependent Newtonian or non-Newtonian rheology. J. Geophys. Res. Solid Earth 104, 12759–12777 (1999)

13. Khaleque, T., Fowler, A., Howell, P., Vynnycky, M.: Numerical studies of thermal convection with temperature- and pressure-dependent viscosity at extreme viscosity contrasts. Phys. Fluids 27, 076603 (2015)
14. Maurice, M., Tosi, N., Samuel, H., Plesa, A.C., Hüttig, C., Breuer, D.: Onset of solid-state mantle convection and mixing during magma ocean solidification. J. Geophys. Res. Planets 122, 577–598 (2017). https://doi.org/10.1002/2016JE005250
15. Shahraki, M., Schmeling, H.: Plume-induced geoid anomalies from 2D axi-symmetric temperature- and pressure-dependent mantle convection models. J. Geodyn. 50–60, 193–206 (2012)
16. Fowler, A.C., Howell, P.D., Khaleque, T.S.: Convection of a fluid with strongly temperature and pressure dependent viscosity. Geophys. Astrophys. Fluid Dyn. 110, 130–165 (2016). https://doi.org/10.1080/03091929.2016.1146264
17. King, S.D., Lee, C., Van Keken, P.E., Leng, W., Zhong, S., Tan, E., Tosi, N., Kameyama, M.C.: A community benchmark for 2-D Cartesian compressible convection in the Earth's mantle. Geophys. J. Int. 180, 73–87 (2010)
18. Conrad, C.P., Hager, B.H.: The thermal evolution of an earth with strong subduction zones. Geophys. Res. Lett. 26, 3041–3044 (1999)
19. Leng, W., Zhong, S.: Constraints on viscous dissipation of plate bending from compressible mantle convection. Earth Planet Sci. Lett. 297, 154–164 (2010)
20. Morgan, J.P., Rüpke, L.H., White, W.M.: The current energetics of earth's interior: agravitational energy perspective. Front. Earth Sci. 4, 46 (2016)
21. Balachandar, S., Yuen, D., Reuteler, D., Lauer, G.: Viscous dissipation in three-dimensional convection with temperature-dependent viscosity. Science 267, 1150–1153 (1995)
22. Conrad, C., Hager, B.: Effects of plate bending and fault strength at subduction zones on plate dynamics. J. Geophys. Res. 104, 17551–17571 (1999)
23. Ushachew, E.G., Sharma, M.K., Makinde, O.D.: Numerical study of MHD heat convection of nanofluid in an open enclosure with internal heated objects and sinusoidal heated bottom. Comput. Therm. Sci. 13(5), 1–16 (2021)

24. Megahed, A.M.: Williamson fluid flow due to a nonlinearly stretching sheet with viscous dissipation and thermal radiation. J. Egypt Math. Soc. 27, 12 (2019). https://doi.org/10.1186/s42787-019-0016-y
25. Ferdows, M., Murtaza, M.G., Shamshuddin, M.: Effect of internal heat generation on free convective power-law variable temperature past a vertical plate considering exponential variable viscosity and thermal conductivity. J. Egypt Math. Soc. 27, 56 (2019). https://doi.org/10.1186/s42787-019-0062-5
26. Ahmed, Z., Nadeem, S., Saleem, S., Ellahi, R.: Numerical study of unsteady flow and heat transfer CNT-based MHD nanofluid with variable viscosity over a permeable shrinking surface. Int. J. Numer. Method H 29, 4607–4623 (2019)
27. Fetecau, C., Vieru, D., Abbas, T., Ellahi, R.: Analytical solutions of upper convected Maxwell fluid with exponential dependence of viscosity under the influence of pressure. Mathematics 9(4), 334 (2021)
28. Solomatov, V.S.: Scaling of temperature and stress dependent viscosity convection. Phys. Fluids 7, 266–274 (1995). https://doi.org/10.1063/1.868624
29. Christensen, U.R., Yuen, D.A.: Layered convection induced by phase transitions. J. Geophys. Res. Solid Earth 90, 10291–10300 (1985)
30. Fowler, A.: Mathematical Geoscience, vol. 36. Springer, Berlin (2011)
31. Huang, J., Zhong, S., van Hunen, J.: Controls on sublithospheric small-scale convection. J. Geophys. Res. Solid Earth 108 (2003). https://doi.org/10.1029/2003JB002456
32. Huang, J., Zhong, S.: Sublithospheric small-scale convection and its implications for the residual topography at old ocean basins and the plate model. J. Geophys. Res. Solid Earth. 110 (2005). https://doi.org/10.1029/2004JB003153
33. King, S.D.: On topography and geoid from 2-d stagnant lid convection calculations. Geochem. Geophys. Geosyst. 10 (2009) https://doi.org/10.1029/2008GC002250
34. Zimmerman, W.B.: Multiphysics Modeling with Finite Element Methods. World Scientific Publishing Company, Singapore (2006)
35. Blankenbach, B., Busse, F., Christensen, U., Cserepes, L., Gunkel, D., Hansen, U., Harder, H., Jarvis, G., Koch, M., Marquart, G., et al.: A benchmark comparison for mantle convection codes. Geophys. J. Int. 98, 23–38 (1989). https://doi.org/10.1111/j.1365-246X.1989.tb05511.x

36. Koglin Jr, D.E., Ghias, S.R., King, S.D., Jarvis, G.T., Lowman, J.P.: Mantle convection with reversing mobile plates: a benchmark study. Geochem. Geophys. Geosyst. 6 (2005). https://doi.org/10.1029/2005GC000924
37. Jarvis, G.T., Mckenzie, D.P.: Convection in a compressible fluid with infinite Prandtl number. J. Fluid Mech. 96, 515–583 (1980)
38. Ricard, Y.: Physics of mantle convection. In: Bercovici, D. (ed.) Treatise on Geophysics. 7, 31–87 (2009)

Chapter

6

EXTENDING THE PERSISTENT PRIMARY VARIABLE ALGORITHM TO SIMULATE NON-ISOTHERMAL TWO-PHASE TWO-COMPONENT FLOW WITH PHASE CHANGE PHENOMENA

Yonghui Huang[1,2], Olaf Kolditz[1,2], and Haibing Shao[1,3]

[1]Helmholtz Centre for Environmental Research - UFZ, Permoserstr. 15, 04318 Leipzig, Germany.
[2]Technical University of Dresden, Helmholtz-Strane 10, 01062 Dresden, Germany.
[3]Freiberg University of Mining and Technology, Gustav-Zeuner-Strasse 1, 09596 Freiberg, Germany.

ABSTRACT

In high-enthalpy geothermal reservoirs and many other geo-technical applications, coupled non-isothermal multiphase flow is considered to be

Citation: (APA): Huang, Y., Kolditz, O., & Shao, H. (2015). Extending the persistent primary variable algorithm to simulate non-isothermal two-phase two-component flow with phase change phenomena. Geothermal Energy, 3(1), 1-23. (23 pages).

the underlying governing process that controls the system behavior. Under the high temperature and high pressure environment, the phase change phenomena such as evaporation and condensation have a great impact on the heat distribution, as well as the pattern of fluid flow. In this work, we have extended the persistent primary variable algorithm proposed by (Marchand et al. Comput Geosci 17(2):431–442) to the non-isothermal conditions. The extended method has been implemented into the OpenGeoSys code, which allows the numerical simulation of multiphase flow processes with phase change phenomena. This new feature has been verified by two benchmark cases. The first one simulates the isothermal migration of H_2 through the bentonite formation in a waste repository. The second one models the non-isothermal multiphase flow of heat-pipe problem. The OpenGeoSys simulation results have been successfully verified by closely fitting results from other codes and also against analytical solution.

Keywords: Non-isothermal multiphase flow, Geothermal reservoir modeling, Phase change, OpenGeoSys

BACKGROUND

In deep geothermal reservoirs, surface water seepages through fractures in the rock and moves downwards. At a certain depth, under the high temperature and pressure condition, water vaporizes from liquid to gas phase. Driven by the density difference, the gas steam then migrates upwards. Along with its path, it will condensate back into the liquid form and release its energy in the form of latent heat. Often, this multiphase flow process with phase transition controls the heat convection in deep geothermal reservoirs. Besides, such multiphase flow and heat transport are considered to be the underlying processes in a wide variety of applications, such as in geological waste repositories, soil vapor extraction of Non-Aqueous Phase Liquid (NAPL) contaminants (Forsyth and Shao 1991), and CO_2 capture and storage (Park et al. 2011; Singh et al. 2012). Throughout the process, different phase zones may exist under different temperature and pressure conditions. At lower temperatures, water flows in the form of liquid. With the rise of temperature, gas and liquid phases may co-exist. At higher temperature, water is then mainly transported in the form of gas/vapor. Since the physical behaviors of these phase zones are different, they are mathematically described by different governing equations. When simulating the geothermal convection with phase change phenomena, this imposes challenges to the numerical models. To numerically model the above phase change behavior, there exist several different algorithms so far. The most popular one is the so-called primary variable switching method proposed by Wu and Forsyth (2001). In Wu's method, the primary variables are switched according to different phase states. For instance, in the two phase region, liquid phase pressure and saturation are commonly chosen as the primary variables; whereas in the single gas or liquid phase region, the saturation of the missing phase will be substituted by the concentration or mass fraction of one light component. This approach has already been adopted by the multiphase simulation code such as TOUGH (Pruess 2008) and MUFTE (Class et al. 2002). Nevertheless, the governing equations deduced from the varying primary variables are intrinsically non-differentiable and often lead to numerical difficulties. To handle this, Abadpour and Panfilov (2009) proposed the negative saturation method, in which saturation values less than zero and bigger than one are used to store extra information of the phase transition. Salimi et al. (2012) later extended this method to the non-isothermal condition, and also taking into account the diffusion and capillary forces. By their efforts, the primary variable switching has been successfully avoided. Recently, Panfilov and Panfilova (2014) has further extended the negative saturation method to the three-component three-phase scenario. As the negative saturation value does not have a physical meaning, further extension of this approach to general multi-phase multi-component system would be difficult. For deep geothermal reservoirs, it requires the primary variables of the governing equation to be persistent throughout the entire spatial and temporal domain of the model. Following this idea, Neumann et al. (2013) chose the pressure of non-wetting phase and capillary pressure as primary variables. The two variables are continuous over different material layers, which make it possible to deal with heterogeneous material properties. The drawback of Neumann's approach is that it can only handle the disappearance of the non-wetting phase, not its appearance. As a supplement, Marchand et al. (2013) suggested to use mean pressure and molar fraction of the light component as primary variables. This allows both of the primary variables to be constructed independently of the phase status and allows the appearance and disappearance of any of the two phases. Furthermore, this algorithm could be easy to be extended to multi-phases (≥3) multi-components (≥3) system.

In this work, as the first step of building a multi-component multi-phase reactive transport model for geothermal reservoir simulation, we extend Marchand's component-based multi-phase flow approach (Marchand et al. 2013) to the non-isothermal condition. The extended governing equations (‘Governing equations’ section), together with the Equation of State (EOS) (‘Constitutive laws’ section), were solved by nested Newton iterations (‘Numerical solution of the global equation system’ section). This extended model has been implemented into the OpenGeoSys software. To verify the numerical code, two benchmark cases were presented here. The first one simulates the migration of H_2 gas produced in a waste repository (‘Benchmark I: isothermal injection of H_2 gas’ section). The second benchmark simulates the classical heat-pipe problem, where a thermal convection process gradually develops itself and eventually reaches equilibrium (‘Benchmark II: heat pipe problem’ section). The numerical results produced by OpenGeoSys were verified against analytical solution and also against results from other numerical codes (Marchand et al. 2012). Furthermore, details of numerical techniques regarding how to solve the non-linear EOS system were discussed (‘Numerical solution of EOS’ section). In the end, general ideas regarding how to include chemical reactions into the current form of governing equations are introduced.

METHOD

Governing Equations

Following Hassanizadeh and Gray (1980), we write instead the mass balance equations of each chemical component by summing up their quantities over every phase. According to Gibbs Phase Rule (Landau and Lifshitz 1980), a simplest multiphase system can be established with two phases and two components. Considering a system with water and hydrogen as constitutive components (with superscript *h* and *w*), they distribute in liquid and gas phase, with the subscript α∈*L*,*G*. The component-based mass balance equations can be formulated as

$$\Phi\frac{\partial(S_L\rho_L^w + S_G\rho_G^w)}{\partial t} + \nabla(\rho_L^w v_L + \rho_G^w v_G) + \nabla(j_L^w + j_G^w) = F^w \tag{1}$$

$$\Phi\frac{\partial(S_L\rho_L^h + S_G\rho_G^h)}{\partial t} + \nabla(\rho_L^h v_L + \rho_G^h v_G) + \nabla(j_L^h + j_G^h) = F^h, \tag{2}$$

where S_L and S_G indicate the saturation in each phase. ρ_α^i $(i \in \{h, w\}, \alpha \in \{L, G\})$ represents the mass density of i-component in α phase. Φ refers to the porosity. F^h and F^w are the source and sink terms. The Darcy velocity v_L and v_G for each fluid phase are regulated by the general Darcy Law

$$v_L = -\frac{KK_{rL}}{\mu_L}(\nabla P_L - \rho_L \mathbf{g}) \tag{3}$$

$$v_G = -\frac{KK_{rG}}{\mu_G}(\nabla P_G - \rho_G \mathbf{g}). \tag{4}$$

Here, K is the intrinsic permeability, and $\boldsymbol{g}$ refers to the vector for gravitational force. The terms j_L^w, j_L^h, j_G^w, and j_G^h represent the diffusive mass fluxes of each component in different phases, which are given by Fick's Law as

$$j_\alpha^{(i)} = -\Phi S_\alpha \rho_\alpha D_\alpha^{(i)} \nabla C_\alpha^{(i)}. \tag{5}$$

Here $D_\alpha^{(i)}$ is the diffusion coefficient, and $C_\alpha^{(i)}$ the mass fraction. When the non-isothermal condition is considered, a heat balance equation is added, with the assumption that gas and liquid phases have reached local thermal equilibrium and share the same temperature.

$$\begin{aligned} &\frac{\Phi\partial[(1-S_G)\rho_L u_L + S_G \rho_G u_G]}{\partial t} + \frac{(1-\Phi)\partial(\rho_S c_S T)}{\partial t} \\ &\quad + \nabla[\rho_G h_G v_G] + \nabla[\rho_L h_L v_L] - \nabla(\lambda_T \nabla T) \\ &\quad = Q_T + \Delta h_{vap}\left(\Phi \frac{\partial(\rho_L S_L)}{\partial t} - \nabla(\rho_L v_L)\right) \end{aligned} \tag{6}$$

In the above equation, the phase density ρ_G, ρ_L, the specific internal energy in different phase u_L, u_G and specific enthalpy in different phase h_L and h_G are all temperature and pressure dependent. While ρ_S and c_S are the density and specific heat capacity of the soil grain, λ_T refers to the heat conductivity, and Q_T is source term, $\Delta h_{vap}(\Phi \frac{\partial(\rho_L S_L)}{\partial t} - \nabla(\rho_L v_L))$ represents the latent heat term according to (Gawin et al. 1995). Generally, the specific enthalpy in Eq. 6 can be described as follows

$$h_\alpha = c_{p\alpha} T. \tag{7}$$

Here $c_{p\alpha}$ is the specific heat capacity of phase α at given pressure. At the same time, relationship between internal energy and enthalpy can be described as

$$h_\alpha = u_\alpha + P_\alpha V_\alpha, \tag{8}$$

where P_α and V_α are the pressures and volumes of phase α. Since we consider the liquid phase is incompressible, its volume change can be ignored, i.e. h=u.

Non-isothermal Persistent Primary Variable Approach

Here in this work, we follow the idea of Marchand et al. (2013), where the 'Persistent Primary Variable' concept were adopted. A new choice of primary variables consists of:

- *P* [Pa] is the weighted mean pressure of gas and liquid phase, with each phase volume as the weighting factor. It depends mainly on the liquid saturation *S*.

  $$P = \gamma(S)P_G + (1 - \gamma(S))P_L \quad (9)$$

  Here $\gamma(S)$ stands for a monotonic function of saturation S, with $\gamma(S)\in[0,1], \gamma(0)=0, \gamma(1)=1$. In Benchmark I ('Benchmark I: isothermal injection of H_2 gas' section), we choose

  $$\gamma(S) = 0$$

  In Benchmark II ('Benchmark II: heat pipe problem' section), we choose

  $$\gamma(S) = S^2$$

  When one phase disappears, its volume converges to zero, making the *P* value equal to the pressure of the remaining phase. If we assume the local capillary equilibrium, the gas and liquid phase pressure can both be derived based on the capillary pressure P_c, that is also a function of saturation *S*.

  $$P_L = P - \gamma(S)P_c(S) \quad (10)$$

  $$P_G = P + (1 - \gamma(S))P_c(S) \quad (11)$$

- *X* [-] refers to the total molar fraction of the light component in both fluid phases. Similar to the mean pressure *P*, it is also a continuous function throughout the phase transition zones. We formulate it as

  $$X = \frac{SN_G X_G^h + (1 - S)N_L X_L^h}{SN_G + (1 - S)N_L} \quad (12)$$

  In a hydrogen-water system, X_L^h and X_G^h refer to the molar fraction of the hydrogen in the two phases, and N_L and N_G are the respective molar densities [mol m^{-3}].

  Based on the choice of new primary variables, the mass conservation Eqs. 1 and 2 can be transformed to the molar mass conservation. The governing equations of the two-phase two-component system are then written as

$$\frac{\Phi\partial((SN_G+(1-S)N_L)X^{(i)})}{\partial t} + \nabla\left(N_L X_L^{(i)} v_L + N_G X_G^{(i)} v_G\right) + \nabla\left(N_L S_L W_L^{(i)} + N_G S_G W_G^{(i)}\right) = F^{(i)} \tag{13}$$

with $i \in (h,w)$ and the flow velocity v regulated by the generalized Darcy's law, referred to Eqs. 3 and 4.

The molar diffusive flux can be calculated following Fick's law

$$W_\alpha^i = -D_\alpha^{(i)} \Phi \nabla X_\alpha^{(i)}. \tag{14}$$

- T [K] refers to the Temperature. If we consider the temperature T as the third primary variable, the energy balance equation can then be included.

$$\begin{aligned} &\frac{\Phi\partial\left[(1-S_G)N_L\left(\sum X_L^{(i)}M^{(i)}\right)u_L + S_G N_G\left(\sum X_G^{(i)}M^{(i)}\right)u_G\right]}{\partial t} \\ &\quad + \frac{(1-\Phi)\partial(\rho_S c_S T)}{\partial t} - \nabla\left[N_G\left(\sum X_G^{(i)}M^{(i)}\right)h_G v_G\right] \\ &\quad - \nabla\left[N_L\left(\sum X_L^{(i)}M^{(i)}\right)h_L v_L\right] - \nabla(\lambda_T \nabla T) = Q_T \end{aligned} \tag{15}$$

The non-isothermal system can thus be simulated by the solution of combined Eqs. 13 and 15, with *P*, *X*, and *T* as primary variables. Once these three primary variables are determined, the other physical quantities are then constrained by them and can be obtained by the solution of EOS system. These secondary variables were listed in Table 1. Compared to the primary variable switching (Wu and Forsyth 2001) and the negative saturation (Abadpour and Panfilov 2009) approach, the choice of *P* and *X* as primary variables fully covers all three possible phase states, i.e., the single-phase gas, two-phase, and single-phase liquid regions. It also allows the appearance or disappearance of any of the two phases. Instead of switching the primary variable, the non-linearity of phase change behavior was removed from the global partial differential equations and was embedded into the solution of EOS.

Table 1. List of secondary variables and their dependency on the primary variables

Parameters	**Symbol**	**Unit**
Gas phase saturation	$S(P,X)$	[-]
Molar density of phase α	$N_{\alpha}(P,X,T)$	[mol m^{-3}]
Molar fraction of component i in phase α	$X_{\alpha}^{(i)}(P,X,T)$	[-]
Capillary pressure	$P_{c}(S)$	[Pa]
Relative permeability of phase α	$K_{r\alpha}(S)$	[-]
Specific internal energy of phase α	$u_{\alpha}(P,X,T)$	[J mol^{-1}]
Specific enthalpy of phase α	$H_{\alpha}(P,X,T)$	[J mol^{-1}]
Heat conduction coefficient	$\lambda_{pm}(P,X,S,T)$	[W m^{-1} K^{-1}]

Closure Relationships

Mathematically, the solution for any linear system of equations is unique if and only if the rank of the equation system equals the number of unknowns. In this work, the combined mass conservation of Eqs. 1, 2, and the energy balance Eq. 6 must be determined by three primary variables. Other variables are dependent on them and considered to be secondary. Such nonlinear dependencies form the necessary closure relationships.

Constitutive Laws

Dalton's Law

Dalton's Law regulates that the total pressure of a gas phase is equal to the sum of partial pressures of its constitutive non-reacting chemical component. In our case, a gas phase with two components, i.e., water and hydrogen is considered. Then the gas phase pressure P_G writes as

$$P_G = P_G^h + P_G^w. \qquad (16)$$

Ideal Gas Law

In our model, the ideal gas law is assumed, where the response of gas phase pressure and volume to temperature is regulated as

$$P_G = \frac{nRT}{V}, \tag{17}$$

where R is the Universal Gas Constant (8.314 J mol^{-1} K^{-1}), V is the volume of the gas and n stands for the mole number gas. Reorganizing the above equation gives the molar density of gas phase N_G

$$N_G = \frac{n}{V} = \frac{P_G}{RT}. \tag{18}$$

Combining Dalton's Law of Eq. 16, we have

$$N_G^h = \frac{P_G^h}{RT}, N_G^w = \frac{P_G^w}{RT}. \tag{19}$$

Furthermore, the molar fraction of component i can be obtained by normalizing its partial pressure with the total gas phase pressure,

$$X_G^i = \frac{P_G^i}{P_G}. \tag{20}$$

Incompressible Fluid

Unlike the gas phase, the liquid phase in our model is considered to be incompressible, i.e., the density of the fluid is linearly dependent on the molar amount of the constitutive chemical component. By assuming standard water molar density $N_L^{std} = \frac{\rho_w^{std}}{M^w}$, with ρ_w^{std} refers to the standard water mass density (1000 kg m^{-3} in our model), the in-compressibility of the liquid phase writes as

$$N_L = \frac{N_L^{std}}{1 - X_L^h}. \tag{21}$$

Henry' Law

We assume that the distribution of light component (hydrogen in our case) can be regulated by the Henry's coefficient $H_W^h(T)$, which is a temperature-dependent parameter.

$$P_G^h H_W^h(T) = N_L X_L^h \tag{22}$$

Raoult's Law

For the heavy component (water), we apply Raoult's Law that the partial pressure of the water component in the gas phase changes linearly with its molar fraction in the liquid.

$$P_G^w = X_L^w P_{Gvapor}^w(T) \tag{23}$$

Here X_L^w is the molar fraction of the water component in the liquid phase. $P_{Gvapor}^w(T)$ is the vapor pressure of pure water, and it is a temperature-dependent function in non-isothermal scenarios.

EOS for Isothermal Systems

Based on the constitutive laws discussed in the 'Constitutive laws' section, we have:

$$\frac{X_L^h N_L^{std}}{X_L^w H_W^h(T)} + X_L^w P_{Gvapor}^w(T) = P_G \tag{24}$$

$$P_G X_G^h = \frac{N_L^{std}}{H_W^h(T)} \frac{X_L^h}{X_L^w} \tag{25}$$

According to Eqs. 24 and 25, X_L^h and X_G^h can be calculated explicitly, under the condition:

$$G(T) = \frac{H_W^h(T) P_{Gvapor}^w(T)}{N_L^{std}} < \frac{1}{4} \tag{26}$$

which is obviously satisfied in water-air and water-hydrogen system, i.e., under the condition that the temperature T is 25 $°C$, with $H_W^h(T) = 7.8 \times 10^{-6}[\text{mol m}^{-3}\text{ Pa}^{-1}], P_{Gvapor}^w(T) = 3173.07[\text{Pa}],$ then we could have $G(T) = 4.54 \times 10^{-7} \ll \frac{1}{4}$. Here, if we only consider isothermal condition, the temperature is assumed to be fixed with T_0. In summary, X_L^h and X_G^h could be expressed as:

$$X_L^h = X_m(P_L, S, T_0) = \frac{N_L^{std} + (P_L + P_c)H_W^h(T_0)}{2H_W^h(T_0)P_{Gvapor}^w(T_0)} + \frac{\left(\sqrt{(N_L^{std} + (P_L + P_c)H_W^h(T_0))^2 - 4(P_L + P_c)H_W^h(T_0)N_L^{std}P_{Gvapor}^w(T_0)}\right)}{2H_W^h(T_0)P_{Gvapor}^w(T_0)} \tag{27}$$

$$X_G^h = X_M(P_G, S, T_0) = \frac{X_L^h N_L^{std}}{H_W^h(T_0)P_G(1 - X_L^h)} \tag{28}$$

Where S is the saturation of light component, and P_c represents the capillary pressure. The above equations are the most general way of calculating the distribution of molar fraction. In Benchmark I ('Benchmark I: isothermal injection of H_2 gas' section), we follow Marchand's idea (Marchand and Knabner 2014), by assuming there is no water vaporization and the gas phase contains only hydrogen, which indicate $P_G \equiv P_G^h$ and $X_G^h \equiv 1$. Therefore Eqs. 27 and 28 could be reformulated as:

$$X_L^h = X_m(P_L, S, T_0) = \frac{(P_L + P_c)H_W^h(T_0)}{(P_L + P_c)H_W^h(T_0) + N_L^{std}} \tag{29}$$

$$X_G^h \equiv 1 \tag{30}$$

Here, for simplification purpose, if we combined with Eqs. 10 and 11, X_L^h and X_G^h could be expressed as functions of mean pressure P and gas phase saturation S, and the above formulation can be transformed to

$$X_L^h = X_m(P_L(P, S(P, X)), S(P, X), T_0) = X_m(P, S, T_0) \tag{31}$$

$$X_G^h = X_M(P_G(P, S(P, X)), S(P, X), T_0) = X_M(P, S, T_0). \tag{32}$$

Assuming the local thermal equilibrium of the multi-phase system is reached, then the Equations of State (EOS) are formulated accordingly based on the three different phase states.

- **In two phase region**: Molar fraction of hydrogen (X_L^h and X_G^h) and molar density in each phase (N_G and N_L) are all secondary variables that are dependent on the change of pressure and saturation. They can be determined by solving the following non-linear system.

$$X_L^h = X_m(P, S, T_0) \tag{33}$$

$$X_G^h = X_M(P, S, T_0) \tag{34}$$

$$N_G = \frac{P_G(P, S)}{RT_0} \tag{35}$$

$$N_L = \frac{N_L^{std}}{1 - X_L^h} \tag{36}$$

$$\frac{SN_G(X - X_G^h) + (1 - S)N_L(X - X_L^h)}{SN_G + (1 - S)N_L} = 0 \tag{37}$$

- **In the single liquid phase region**: In a single liquid phase scenario, the gas phase does not exist, i.e., the gas phase saturation S always equals to zero. Meanwhile, the molar fraction of light component in the gas phase X_G^h can be any value, as it will be multiplied with the zero saturation (see Eqs. 13 to 14) and vanish in the governing equation. This also applies to the gas phase molar density N_G, whereas the two parameters can be arbitrarily given, and have no physical impact. So to determine the EOS, we only need to solve for the liquid phase molar fraction and density.

$$X_L^h = X \tag{38}$$

$$N_L = \frac{N_L^{std}}{1 - X} \tag{39}$$

- **In the single gas phase region**: Similarly, in a single gas phase scenario, the liquid phase does not exist, i.e., the gas phase saturation S always equals to 1, whereas the liquid phase saturation remains zero. Meanwhile, the molar fraction of light component in the liquid phase X_L^h can be any value, as it will be multiplied with the zero liquid phase saturation (see Eqs. 13 to 14) and vanish in the governing equation. This also applies to the liquid phase molar density N_G, whereas the two parameters can be arbitrarily given, and have no physical meaning. So to determine the EOS, we only need to solve for the gas phase molar fraction and density.

$$X_G^h = X$$

$$N_G = \frac{P}{RT_0}$$

EOS for Non-Isothermal Systems

As the energy balance of Eq. 6 has to be taken into account under the non-isothermal condition, all the secondary variables not only are dependent on the pressure *P* but also rely on the temperature *T*. Except for the parameters mentioned above, several other physical properties are also regulated by the T/P dependency. Furthermore, in a non-isothermal transport, high non-linearity of the model exists in the complex variational relationships between secondary variables and primary variables. Therefore, how to set up an EOS system for each fluid is a big challenge for the non-isothermal multi-phase modeling. In the literature, (Class et al. 2002; Olivella and Gens 2000; Peng and Robinson 1976; Singh et al. 2013a, and Singh et al. 2013b) have given detailed procedures of solving EOS to predicting the gas and liquid thermodynamic and their transport properties. Here in our model, we follow the idea by Kolditz and De Jonge (2004). Detailed procedure regarding how to calculate the EOS system is discussed in the following.

Vapor Pressure

As we discussed in the 'Constitutive laws' section, vapor pressure is a key parameter for determining the molar fractions of different components in each phase. The equilibrium restriction on vapor pressure of pure water is given by Clausius-Clapeyron equation (Çengel and Boles 1994).

$$P^w_{Gvapor}(T) = P_0 \exp\left[\left(\frac{1}{T_0} - \frac{1}{T}\right)\frac{h^w_G M^w}{R}\right] \tag{40}$$

where h^w_G is enthalpy of vaporization, M^w is molar mass of water. P_0 represents the vapor pressure of pure water at the specific Temperature T_0. In our model, we choose T_0=373K,P_0=101,325Pa. An alternative method is using the Antoine equation, written as

$$log_{10}(P^w_{Gvapor}(T)) = A - \frac{B}{C+T} \tag{41}$$

0with A, B, and C as the empirical parameters. Details regarding this formulation can be found in Class et al. (2002).

Specific Enthalpy

Specific enthalpy h_{α} [J mol^{-1}] is the enthalpy per unit mass. According to Eq. 6, we need to know the specific enthalpy of a certain phase. In particular, since component-based mass balance is considered, we calculate the phase enthalpy as the sum of mole (mass) specific enthalpy of each component in this phase. Here we assume that the energy of mixing is ignored. For instance, the water-air system applied in the second benchmark is formulated as

$$h_G = h_G^{air} X_G^{air} + h_G^{w_{vap}} X_G^{w_{vap}} \tag{42}$$

$$h_L = h_L^{air} X_L^{air} + h_L^{w_{liq}} X_L^{w_{liq}} \tag{43}$$

Here h_G^{air} is the specific enthalpy of air in gas phase, $h_G^{w_{vap}}$ is specific enthalpy of vapor water in gas phase, h_L^{air} represents the specific enthalpy of the air dissolved in the liquid phase, while $h_L^{w_{liq}}$ donates the specific enthalpy of the liquid water in liquid phase. While $X_G^{air}, X_G^{w_{vap}}, X_L^{air}$ and $X_L^{w_{liq}}$ represent molar fraction [-] of each component (air and water) in the corresponding phase (gas and liquid).

Henry Coefficient

We assume Henry's Law is valid under the non-isothermal condition. Therefore Henry coefficient is a secondary variable. In the water-air system, it can be defined as (Kolditz and De Jonge 2004)

$$H_W^h(T) = (0.8942 + 1.47 \exp(-0.04394T)) \times 10^{-10} \tag{44}$$

with T the temperature value in °C.

Heat Conductivity

Since the local thermal equilibrium is assumed, the heat conductivity λ_{pm} [W m^{-1} K^{-1}] of the fluid-containing porous media is averaged from the heat conductivities of the fluid phases and the solid matrix. Thus, it is a function of saturation only.

$$\lambda_{pm} = \lambda_{pm}^{S_L=S_G=0} + \sqrt{S_L}(\lambda_{pm}^{S_L=1} - \lambda_{pm}^{S_L=0}) + \sqrt{S_G}(\lambda_{pm}^{S_G=1} - \lambda_{pm}^{S_G=0}) \tag{45}$$

Fugacity

When the thermal equilibrium is reached, the chemical potentials of component *i* in gas and liquid phase equal with each other. This equilibrium relationship can be formulated as the equation of chemical potential ν

$$\nu_G^{(i)}(P_G, X_L^{(i)}, X_G^{(i)}, T) = \nu_L^{(i)}\left(P_L, X_L^{(i)}, X_G^{(i)}, T\right)$$

In our model, the fugacity was applied instead of chemical potential. The above relationship is then transformed to the equivalence of component fugacities, where

$$f_G^{(i)} = f_L^{(i)}$$

holds for each component *i* in each phase. In order to compute the fugacity of a component in a particular phase, the following formulation is used

$$f_\alpha^{(i)} = P_\alpha X_\alpha^{(i)} \phi_\alpha^{(i)} \tag{46}$$

where $\phi_\alpha^{(i)}$ is the respective fugacity coefficient of component.

NUMERICAL SCHEME

Numerical Solution of EOS

Physical Constraints of EOS

Since the pore space should be fully occupied by either or both the gas and liquid phases, the sum of phase saturation should equal to one. By definition, the saturation for each phase should be no less than zero and no larger than one. This constraint is summarized as

$$\sum_\alpha S_\alpha = 1 \ \wedge \ S_\alpha > 0 \ \ (\alpha \in G, L) \tag{47}$$

Similarly, the sum of the molar fraction for all components in a single phase should also be in unity, and this second constraint can be formulated as

$$\sum_i X_G^{(i)} = 1 \ \wedge \ \sum_i X_L^{(i)} = 1 \ \text{ with } X_G^{(i)} > 0 \ \wedge \ X_L^{(i)} > 0 \ \ (i \in h, w) \tag{48}$$

Combining these constraints, we have

$$S = 0 \wedge X_L^h \leq X_m(P,0,T) \tag{49}$$

$$0 \leq S \leq 1 \wedge X_m(P,S,T) - X_L^h = 0, X_G^h - X_M(P,S,T) = 0 \tag{50}$$

$$S = 1 \wedge X_G^h \geq X_M(P,1,T) \tag{51}$$

For the Eqs. 49 to 51, they contain both equality and inequality relationships, which impose challenges for the numerical solution. In order to solve it numerically, we introduce a minimum function (Kanzow 2004; Kräutle 2011), to transform the inequalities. It is defined as

$$\mathit{\Psi}(a,b) := min\{a,b\} \tag{52}$$

Combined with Eqs. 49 to 51, they can be transformed to

$$\Psi(S, X_m(P,S,T) - X_L^h) = 0 \tag{53}$$

$$\Psi(1-S, X_G^h - X_M(P,S,T)) = 0 \tag{54}$$

$$\frac{SN_G(X - X_G^h) + (1-S)N_L(X - X_L^h)}{SN_G + (1-S)N_L} = 0 \tag{55}$$

Then Eqs. 53 to 55 formulates the EOS system, which needs to be solved on each mesh node of the model domain.

Numerical Scheme of Solving EOS

For the EOS, the primary variables *P* and *X* are input parameters and act as the external constraint. The saturation *S*, gas and liquid phase molar fraction of the light component X_G^h and X_L^h are then the unknowns to be solved. Once they have been determined, other secondary variables can be derived from them. When saturation is less than zero or bigger than one, the second argument of the minimization function in Eq. 53 will be chosen. Then it effectively prevents the saturation value from moving into unphysical value. This transformation will result in a local Jacobian matrix that might be singular. Therefore, a pivoting action has to be performed before the Jacobian matrix is decomposed to calculating the Newton step. An alternative approach to handle this singularity is to treat the EOS system as a nonlinear optimization problem with the inequality constraints. Our tests showed that the optimization algorithms such as Trust-Region method are very robust in solving such a local problem, but the calculation time will be considerably longer, compared to the Newton-based iteration method.

NUMERICAL SOLUTION OF THE GLOBAL EQUATION SYSTEM

In this work, we solve the global governing equation Eqs. 13 to 15 with all the closure relationships simultaneously satisfied. To handle the non-linearities, a nested Newton scheme was implemented (see the flow chart in Fig. 1). All the derivatives in the EOS system Eqs. 53 to 55 are computed exactly and the local Jacobian matrix is constructed in an analytical way, while the global Jacobian matrix is numerically evaluated based on the finite difference method. For the global equations, the time was discretized with the backward Euler scheme, and the spatial discretization was performed with the Galerkin Finite Element method. In each global Newton iteration, the updated global variables *P*, *X*, and *T* from the previous iteration were passed to the EOS system, and acted as constraints to solve for secondary variables. The solution of Eqs. 53 to 55 was performed one after the other on each mesh node of the model domain.

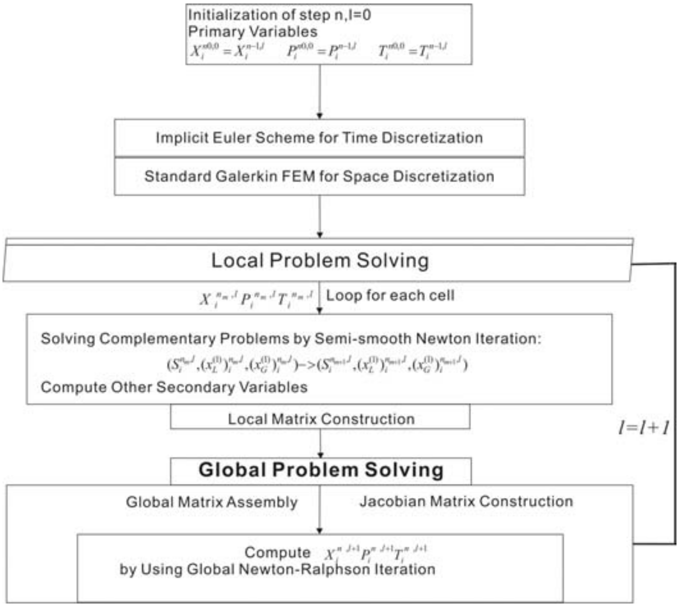

Initialization of step n,l=0
Primary Variables
$X_i^{n0,0}=X_i^{n-1,l}$ $P_i^{n0,0}=P_i^{n-1,l}$ $T_i^{n0,0}=T_i^{n-1,l}$
Implicit Euler Scheme for Time Discretization
Standard Galerkin FEM for Space Discretization
Local Problem Solving
$X_i^{n,l}P_i^{n,l}T_i^{n,l}$
Loop for each cell
Solving Complementary Problems by Semi-smooth Newton Iteration:
$(S_i^{n,l},(x_L^{(l)})_i^{n,l},(x_G^{(l)})_i^{n,l})->(S_i^{n+1,l},(x_L^{(l)})_i^{n+1,l},(x_G^{(l)})_i^{n+1,l})$
Compute Other Secondary Variables
Local Matrix Construction
$l=l+1$
Global Problem Solving
Global Matrix Assembly
Jacobian Matrix Construction
Compute $X_i^{n,l+1}P_i^{n,l+1}T_i^{n,l+1}$
by Using Global Newton-Ralphson Iteration

Figure 1. Scheme of the algorithm for global equation system.

For Newton iterations, the following convergence criteria was applied.

$$\left\| Residual(Step(k)) \right\|_2 \leq \epsilon \tag{56}$$

where $\|\|_2$ denotes the Euclidean norm. A tolerance value $\varepsilon=1\times10^{-14}$ were adopted for the EOS and 1×10^{-9} for the global Newton iterations.

HANDLING UNPHYSICAL VALUES DURING THE GLOBAL ITERATION

In the 'Numerical scheme of solving EOS' section, we have discussed the procedure of handling physical constraint of the EOS system. However, during the global iterations, if the initial value of X is small enough, it may happen that $X\leq0$ can appear. Since the negative value of X would cause failures of further iteration, it is necessary to force the non-negativity constraint on X. To achieve this, a widely used method is extending the definition of the physical variables such as N_G, N_L for $X<0$, as was done in (Marchand et al. 2013), (Marchand and Knabner (2014), and (Abadpour and Panfilov 2009). In our implementation, we chose an alternative and more straightforward method, which is adding a damping factor in each global Newton iteration when updating the unknown vector. The damping factor δ are chosen as follows,

$$\frac{1}{\delta} = \max\{1, 2 * \frac{\Delta P(j)}{P(j)}, 2 * \frac{\Delta X(j)}{X(j)}, 2 * \frac{\Delta T(j)}{T(j)}\} \tag{57}$$

where $P(j)$, $X(j)$ and $T(j)$ denote pressure/molar fraction/temperature at node j.

RESULTS AND DISCUSSIONS

In our work, the model verification was carried out in two separate cases, one under isothermal and the other under non-isothermal conditions. In the first case, a simple benchmark case was proposed by GNR MoMaS (Bourgeat et al. 2009). We simulated the same H_2 injection process with the extended OpenGeoSys code (Kolditz et al. 2012), and compared our results against those from other code (Marchand and Knabner 2014). For the non-isothermal case, there exists no analytical solution, which explicitly involves the phase transition phenomenon. Therefore, we compared our simulation result of the classical heat pipe problem to the semi-analytical solution from Udell and Fitch (1985). This semi-analytical solution was developed for the steady state condition without the consideration of phase change phenomena. Despite of this discrepancy, the OpenGeoSys code delivered very close profile as by the analytical approach.

Benchmark I: Isothermal Injection of H_2 Gas

The background of this benchmark is the production of hydrogen gas due to the corrosion of the metallic container in the nuclear waste repository. Numerical model is built to illustrate such gas appearance phenomenon. The model domain is a two-dimensional horizontal column representing the bentonite backfill in the repository tunnel, with hydrogen gas injected on the left boundary. This benchmark was proposed in the GNR MoMaS project by French National Radioactive Waste Management Agency. Several research groups has made contributions to test the benchmark and provided their reference solutions (Ben Gharbia and Jaffré 2014; Bourgeat et al. 2009; Marchand and Knabner 2014; Neumann et al. 2013). Here we adopted the results proposed in Marchand's paper Marchand and Knabner 2014 for comparison.

Physical Scenario

Here a 2D rectangular domain Ω=[0,200]×[−10,10] m (see Fig. 2) was considered with an impervious boundary at Γ_{imp}=[0,200]×[−10,10] m, an inflow boundary at Γ_{in}={0}×[−10,10] m, and an outflow boundary at Γ_{out}={200}×[−10,10] m. The domain was initially saturated with water, hydrogen gas was injected on the left-hand-side boundary within a certain time span ([0,5×10^4century]). After that the hydrogen injection stopped and no flux came into the system. The right-hand-side boundary is kept open throughout the simulation. The initial condition and boundary conditions were summarized as

- $X(t=0) = 10^{-5}$ and $P_L(t=0) = P_L^{out} = 10^6$ [Pa] on Ω.

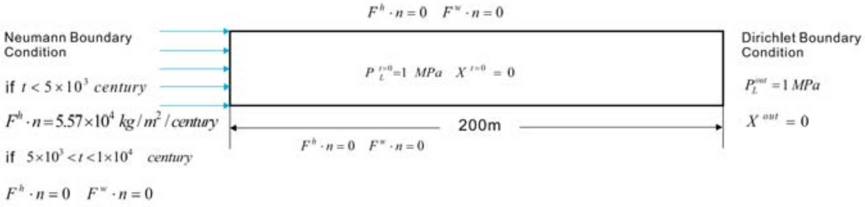

$F^h \cdot n = 0$ $F^w \cdot n = 0$
Neumann Boundary Condition
if $t < 5\times10^3$ century
$F^h \cdot n = 5.57\times10^4$ kg/m^2/century
if $5\times10^3 < t < 1\times10^4$ century
$F^h \cdot n = 0$ $F^w \cdot n = 0$
$P_L^{t=0}$=1 MPa $X^{t=0} = 0$
200m
$F^h \cdot n = 0$ $F^w \cdot n = 0$
Dirichlet Boundary Condition
P_L^{out} =1 MPa
$X^{out} = 0$

Figure 2. Geometry and boundary condition for the H_2 injection benchmark.

- $q^{w} \cdot \nu = q^{h} \cdot \nu = 0$ on Γ_{imp}.
- $q^w \cdot \nu = 0, q^h \cdot \nu = Q_d^h = 0.2785$ [mol century^{-1}m^{-2}] on Γ_{in}
- $X = 0$ and $P_l = P_L^{out} = 10^6$ [Pa] on Γ_{out}.

Model Parameters and Numerical Settings

The capillary pressure P_c and relative permeability functions are given by the van-Genuchten model (Van Genuchten 1980).

$$P_c = P_r\left(S_{le}^{-\frac{1}{m}} - 1\right)^{\frac{1}{n}}$$

$$K_{r_L} = \sqrt{S_{le}}\left(1 - \left(1 - S_{le}^{\frac{1}{m}}\right)^m\right)^2$$

$$K_{rG} = \sqrt{1 - S_{le}}\left(1 - S_{le}^{\frac{1}{m}}\right)^{2m}$$

where $m = 1 - \frac{1}{n}$, P_r and n are van-Genuchten model parameters and the effective saturation S_{le} is given by

$$S_{le} = \frac{1 - S_g - S_{lr}}{1 - S_{lr} - S_{gr}} \qquad (58)$$

here S_{lr} and S_{gr} indicate the residual saturation in liquid and gas phases, respectively. Values of parameters applied in this model are summarized in Table 2.

Table 2. Fluid and porous medium properties applied in the H_2 migration benchmark

Parameters	**Symbol**	**Value**	**Unit**
Intrinsic permeability	K	5×10^{-20}	[m^2]
Porosity	Φ	0.15	[-]
Residual saturation of liquid phase	S_{lr}	0.4	[-]
Residual saturation of gas phase	S_{gr}	0	[-]
Viscosity of liquid	μ_l	10^{-3}	[Pa·s]
Viscosity of gas	μ_g	9×10^{-6}	[Pa·s]
van Genuchten parameter	P_r	2×10^{6}	[Pa]
van Genuchten parameter	n	1.49	[-]

We created a 2D triangular mesh here with 963 nodes and 1758 elements. The mesh element size varies between 1m and 5m. A fixed time step size of

1 century is applied. The entire simulated time from 0 to 10^4 centuries were simulated. The entire execution time is around 3.241×10^4s.

Results and Analysis

The results of this benchmark are depicted in Fig. 3. The evolution of gas phase saturation and the gas/liquid phase pressure at the inflow boundary Γ_{in} over the entire time span are shown. In additional, we compare results from our model against those given in Marchand's paper (Marchand and Knabner 2014). In Fig. 3, solid lines are our simulation results while the symbols are the results from Marchand et al. It can be seen that a good agreement has been achieved. Furthermore, the evolution profile of the gas phase saturation S_g, the liquid phase pressure P_L, and the total molar fraction of hydrogen X are plotted at different time (t=150,1×10^3,5×10^3,6×10^3 centuries) in Fig. 4 a −c, respectively.

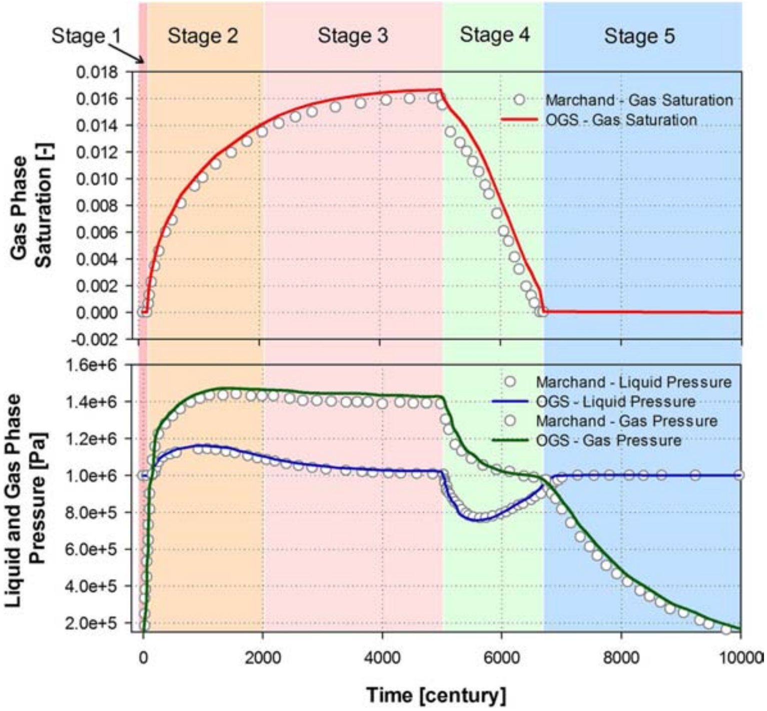


Figure 3. Evolution of pressure and saturation over time.

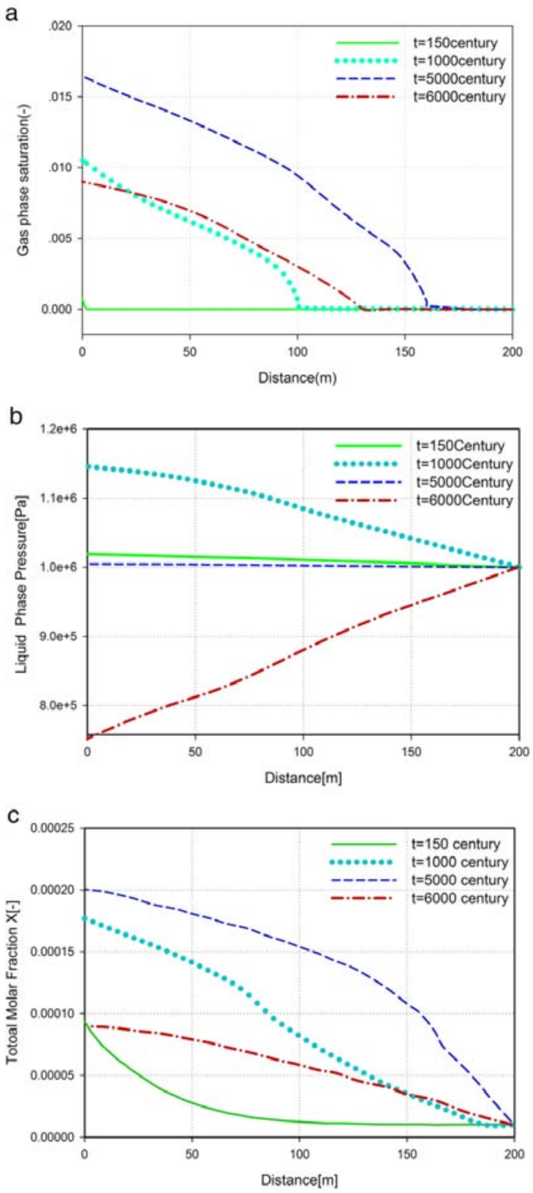


Figure 4. Evolution of (**a**) gas phase saturation, (**b**) liquid phase pressure, and (**c**) total hydrogen molar fraction over the whole domain at different time.

By observing the simulated saturation and pressure profile, the complete physical process of H_2 injection can be categorized into five subsequent stages.

- **The dissolution stage**: After the injection of hydrogen at the inflow boundary, the gas first dissolved in the water. This was reflected by the increasing concentration of hydrogen in Fig. 4 c. Meanwhile, the phase pressure did not vary much and was kept almost constant (see Fig. 4 b).
- **Capillary stage**: Given a constant temperature, the maximal soluble amount of H_2 in the water liquid is a function of pressure. In this MoMaS benchmark case, our simulation showed that this threshold value was about 1×10^{-3} mol H_2 per mol of water at a pressure of 1×10^{6} [Pa]. Once this pressure was reached, the gas will emerge and formed a continuous phase. As shown in Fig. 4 a, at approximately 150 centuries, the first phase transition happens. Beyond this point, the gas and liquid phase pressure quickly increase, while hydrogen gas is transported towards the right boundary driven by the pressure and concentration gradient. In the meantime, the location of this phase transition point also slowly shifted towards the middle of the domain.
- **Gas migration stage**: The hydrogen injection process continued until the 5000th century. Although the gas saturation continues to increase, pressures in both phases begin to decline due to the existence of the liquid phase gradient. Eventually, the whole system will reach steady state with no liquid phase gradient.
- **Recovery stage**: After hydrogen injection was stopped at the 5000th century, the water came back from the outflow boundary towards the left, which was driven by the capillary effect to occupy the space left by the disappearing gas phase. During this stage, the gas phase saturation begins to decline, and both phase pressures drop even below the initial pressure. The whole process will not stop until the gas phase completely disappeared.
- **Equilibrium stage**: After the complete disappearance of the gas phase, the saturation comes to zero again, and the whole system will reach steady state, with pressure and saturation values same as the ones given in the initial condition.

Benchmark II: Heat Pipe Problem

To verify our model under the non-isothermal condition, we adopted the heat pipe problem proposed by Udell and Fitch (1985). They have provided a semi-analytical solution for a non-isothermal water-gas system in porous media, where heat convection, heat conduction as well as capillary forces were considered. A heater installed on the right-hand-side of the domain generated constant flux of heat, and it was then transferred through the porous media by conduction, as well as the enthalpy transport of the fluids. The semi-analytical solution was developed for the steady state condition, and the liquid phase flowed in the opposite direction to the gas phase. If gravity was neglected, the system can be simplified to a system of six ordinary differential equations (ODE), the solution of which was then be obtained in the form of semi-analytical solution. Detailed derivation procedure is available in (Helmig 1997), and the parameters used in our comparison are listed in Table 3. Interested readers may also refer to the supplementary material regarding how this solution was deducted.

Table 3. Parameters applied in the heat pipe problem

Parameters name	**Symbol**	**Value**	**Unit**
Permeability	K	10^{-12}	$[m^{2}]$
Porosity	Φ	0.4	[-]
Residual liquid phase saturation	S_{lr}	0.4	[-]
Heat conductivity of fully saturated porous medium	$\lambda_{pm}^{S_w=1}$	1.13	[W m^{-1} K^{-1}]
Heat conductivity of dry porous medium	$\lambda_{pm}^{S_w=0}$	0.582	[W m^{-1} K^{-1}]
Heat capacity of the soil grains	c_{s}	700	[J kg^{-1} K^{-1}]
Density of the soil grain	ρ_{s}	2600	[kg m^{-3}]
Density of the water	ρ_{w}	1000	[kg m^{-3}]
Density of the air	ρ	0.08	[kg m^{-3}]
Dynamic viscosity of water	μ_{w}	2.938×10^{-4}	[Pa·s]
Dynamic viscosity of air	μ_{g}^{a}	2.08×10^{-5}	[Pa·s]
Dynamic viscosity of steam	μ_{g}^{w}	1.20×10^{-5}	[Pa·s]
Diffusion coefficient of air	D_{g}^{a}	2.6×10^{-5}	[m^{2} s^{-1}]
van Genuchten parameter	P_{r}	1×10^{4}	[Pa]
van Genuchten parameter	n	5	[-]

Physical Scenario

As shown in Fig. 5, the heat pipe was represented by a 2D horizontal column (2.25 m in length and 0.2 m in diameter) of porous media, which was partially saturated with a liquid phase saturation value of 0.7 at the beginning. A constant heat flux (Q_T=100 [W m^{-2}]) was imposed on the right-hand-side boundary Γ_{in}, representing the continuously operating heating element. At the left-hand-side boundary Γ_{out}, Dirichlet boundary conditions were imposed for Temperature T=70 °C, liquid phase pressure P_G=1×10^5 [Pa], effective liquid phase saturation S_{le}=1, and air molar fraction in the gas phase $X_G^a = 0.71$. Detailed initial and boundary condition are summarized as follows.

- $P(t{=}0)=1\times10^5$ [Pa], $S_L(t{=}0)=0.7$, $T(t{=}0)=70$ [°C] on the entire domain.

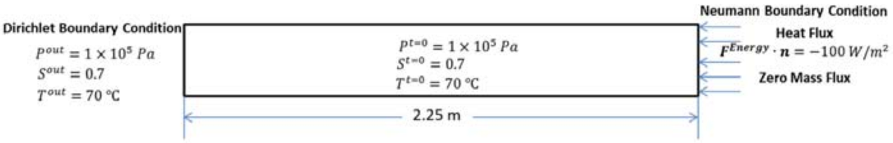

Dirichlet Boundary Condition
$P^{out} = 1 \times 10^5\ Pa$
$S^{out} = 0.7$
$T^{out} = 70\ °C$
$P^{t=0} = 1 \times 10^5\ Pa$
$S^{t=0} = 0.7$
$T^{t=0} = 70\ °C$
2.25 m
Neumann Boundary Condition
Heat Flux
$F^{Energy} \cdot \boldsymbol{n} = -100\ W/m^2$
Zero Mass Flux

Figure 5. Geometry of the heat pipe problem.

$q^w \cdot \nu = q^h \cdot \nu = 0$ on Γ_{imp}.

$q^w \cdot \nu = q^h \cdot \nu = 0$, $q^T \cdot \nu = Q_T$ on Γ_{in}.

P=1×10^5 [Pa], S_L=0.7, T=70 [°C] on Γ_{out}.

Model Parameters and Numerical Settings

For the capillary pressure −saturation relationship, van Genuchten model was applied. The parameters used in the van Genuchten model are listed in Table 3. The water −air relative permeability relationships were described by the Fatt and Klikoffv formulations (Fatt and Klikoff Jr 1959).

$$K_{rG} = (1 - S_{le})^3 \tag{59}$$

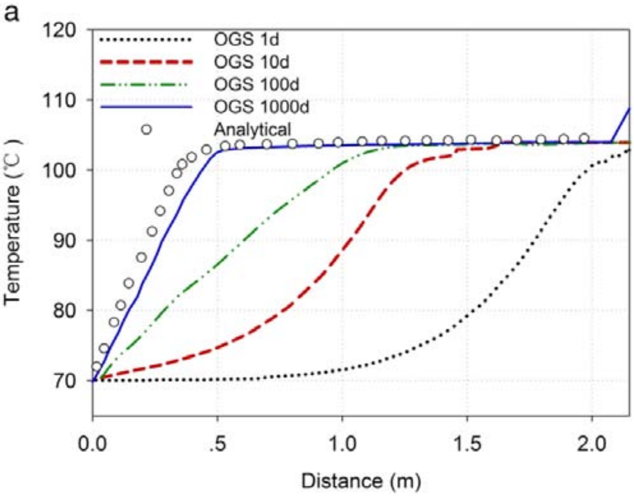
$$K_{rL} = S_{le}^3 \tag{60}$$

where S_{le} is the effective liquid phase saturation, referred to Eq. 58.

We created a 2D triangular mesh here with 206 nodes and 326 elements. The averaged mesh element size is around 6m. A fixed size time stepping scheme has been adopted, with a constant time step size of 0.01 day. The entire simulated time from 0 to 10^4 day were simulated.

Results

The results of our simulation were plotted along the central horizontal profile over the model domain at y = 0.1 m, and compared against semi-analytical solution. Temperature and saturation profiles at day 1, 10, 100, 1000 are depicted in Fig. 6 a, b respectively. As the heat flux was imposed on the right-hand-side boundary, the temperature kept rising there. After 1 day, the boundary temperature already exceeded 100 °C, and the water in the soil started to boil. Together with the appearance of steam, water saturation on the right-hand-side began to decrease. After 10 days, the boiling point has almost moved to the middle of the column. Meanwhile, the steam front kept boiling and shifted to the left-hand-side, whereas liquid water was drawn back to the right. After about 1000 days, the system reached a quasi-steady state, where the single phase gas, two phase and single phase liquid regions co-exist and can be distinguished. A pure gas phase region can be observed on the right and liquid phase region dominates the left side.



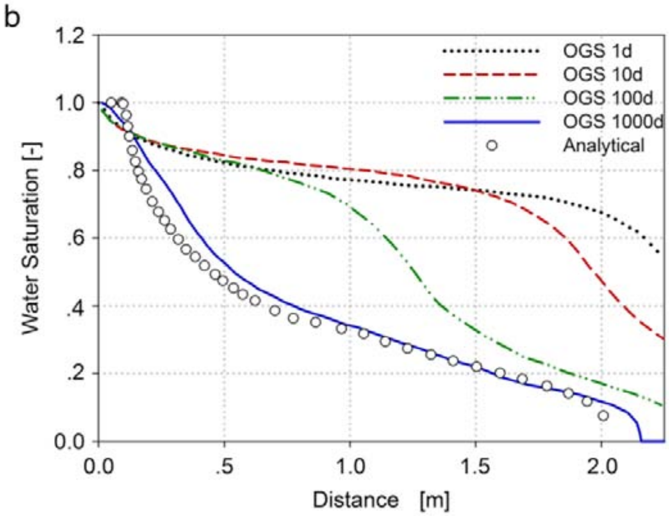


Figure 6. Evolution of (**a**) temperature and (**b**) liquid phase saturation over the whole domain at different time.

Discussion

Analysis of the Differences in Benchmark II

From Fig. 6 a, b, some differences can still be observed in comparison to the semi-analytical solution. Our hypothesis is this difference originates from the capillary pressure −saturation relationship adopted in our numerical implementation. In the original formulation of Udell and Fitch (1985), the Leverett model was applied to produce the semi-analytical solution. It is assumed that the liquid and gas are immiscible and thus there is no gas component dissolved in the liquid phase, and vice versa. In our work, we cannot precisely follow the same assumption, since the dissolution of chemical component in both phases is a requirement for the calculation of phase equilibrium. When considering phase change, we need to allow the saturation *S* to drop below the residual saturation, so that the evaporation as well as the condensation process can occur. In the traditional van Genuchten model, infinite value of capillary pressure may occur in the lower residual saturation region. Therefore we have made regularization that allows water saturation to fall below the residual saturation, as demonstrated in Fig. 7. Every time the capillary pressure needs to be evaluated, an if-else judgment is performed.

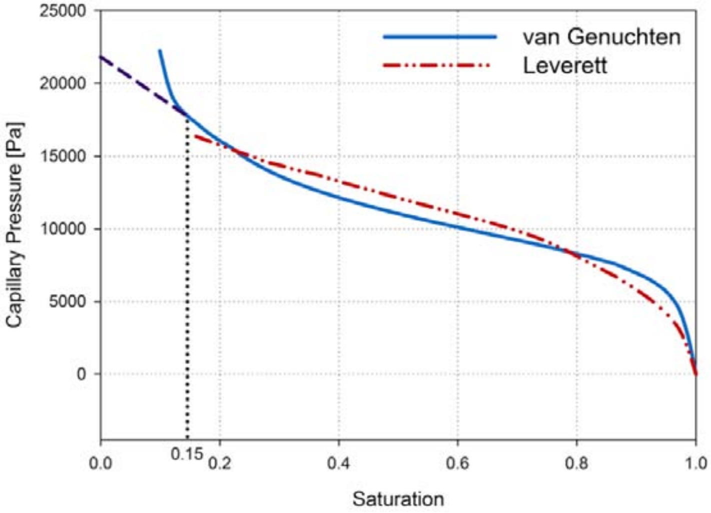


Figure 7. The regularization of the van Genuchten model.

```
if $S_{lr} < S < 1$ then
    $\bar{P}_c(S) = P_c(S)$
else
    if $0 < S < S_{lr}$ then
        $\bar{P}_c(S) = P_c(S_{lr}) - P_c'(S_{lr})(S - S_{lr})$
    end
end
```

Here $\bar{P}_c(S)$ indicates the modified van Genuchtem model, and $P_c'(S_{lr})$ represents the slope of Pc-S curve at the point of residual water phase saturation. The above modified van Genuchten model approximates the same behavior as the original Leverett one in majority part of the saturation region (see Fig. 7), yet still allowing the phase change behavior. However, it is not exactly same as the one in the semi-analytical solution. This is considered to be the reason why the quasi steady-state profile by our numerical model (Fig. 6) deviates from the analytical one.

Continuity of the Global System and Convergence of the Iteration

In this work, we have only considered the homogeneous medium, where the primary variables of P and X are always continuous over the entire domain. For some primary variables, their derivatives in the governing Eqs. (13) − (15) are discontinuous at locations where the phase transition happens, i.e., $X=X_m(P,S=0,T)$ and $X=X_M(P,S=1,T)$. For instance, $\frac{\partial S}{\partial X}$ and $\frac{\partial S}{\partial P}$ might produce singularities at $S=0$ and $S=1$, and they can cause trouble on the conditioning of the global Jacobian matrix. In our simulation, a damped Newton iterations with line search has been adopted (see the 'Handling unphysical values during the global iteration' section). We observed that such derivative terms will result in an increased number of global Newton iterations, and the linear iteration number to solve the Newton step as well. It does not alter the convergence of the Newton scheme, as long as the function is Lipschitz continuous.

We are aware of the fact that this issue may be more difficult to handle for the heterogeneous media, where the primary variable P and X could not be directly applied any more because of the non-continuity over the heterogeneous interface (Park et al. 2011). In that case, choosing the primary variables which are continuous over any interface of the medium is a better option. Based on the analysis by Ern and Mozolevski (2012), if we assume Henry's law is valid, concentration, or in another word, the molar or mass fraction of the hydrogen in the liquid phase ρ_L^h (X_L^h), gas/liquid phase pressure P_G / P_L, as well as the capillary pressure are all continuous over the interface. Therefore, they are the potential choices of primary variable which can be applied in the heterogeneous media (see (Angelini et al. 2011); (Neumann et al. 2013), and (Bourgeat et al. 2013)). We are currently investigating these options and will report on the results in subsequent work.

Conclusions

In this work, based on the persistent primary variable algorithm proposed by Marchand et al. (2013), we extended the isothermal multi-phase flow formulation to the non-isothermal condition. The extended governing equation is based on the mass balance of each chemical component and is nonlinearly coupled with the non-isothermal EOS. The numerical scheme has been implemented into the open source code OpenGeoSys. The verification of our model were carried out in two benchmark cases.

- For the GNR MoMaS (Bourgeat et al. 2009) benchmark (‘Benchmark I: isothermal injection of H_2 gas’ section), the extended model is capable of simulating the migration of H_2 gas including its dissolution in aqueous phase. The simulated results fitted well with those from other codes (Marchand et al. 2013; Marchand and Knabner 2014).
- For the non-isothermal benchmark, we simulated the heat pipe problem and verified our result against the semi-analytical solution (‘Benchmark II: heat pipe problem’ section). Furthermore, our numerical model extended the original heat pipe problem to include the phase change behavior.

Currently, we are working on the incorporation of equilibrium reactions, such as the mineral dissolution and precipitation, into the EOS system. As our global mass-balance equations are already component based, one governing equation can be written for each basis component. Pressure, temperature, and molar fraction of the chemical components can be chosen as primary variables. Inside the EOS problem, the amount of secondary chemical components can be calculated based on the result of basis, which can further lead to the phase properties as density and viscosity. The full extension of including temperature-dependent reactive transport system will be the topic of a separate work in the near future.

NOMENCLATURE

Greek symbols		
ε	Tolerance value for Newton iteration.	[-]
λ_T	Heat Conductivity.	[$W m^{-1} K^{-1}$]
μ_α	Viscosity in α phase.	[Pa · s]
ν_α^i	Chemical potential of i-component in α phase.	[Pa]
Φ	Porosity.	[-]
ϕ_α^i	fugacity coefficient of i-component in α phase.	[-]
ρ_α^i	Mass density of i-component in α phase.	[$K g m^{-3}$]
Operators		
$\wedge$	Logical *“ a n d “*	

$\Vert\Vert_2$	Euclidean norm	
$\Psi(a,b)$	Minimum function	
Roman symbols		
$\boldsymbol{g}$	Vector for gravitational force.	$[m\,s^{-2}]$
$c_{p\,\alpha}$	Specific heat capacity in phase α at given pressure.	$[J\,K\,g^{-1}\,K^{-1}]$
c_S	Specific heat capacity of soil grain.	$[J\,K\,g^{-1}\,K^{-1}]$
D_α^i	Diffusion coefficient of i-component in phase α.	$[m^2\,s^{-1}]$
F^i	Mass source/sink term for i-component.	$[K\,g\,m^{-3}\,s^{-1}]$
f_α^i	Fugacity of i-component in α phase.	[Pa]
H_W^h	Henry coefficient.	$[m\,o\,l\,P\,a^{-1}\,m^{-3}]$
h_α	Specific enthalpy.	$[J\,K\,g^{-1}]$
j_α^i	Diffusive mass flux of i-component in α phase.	$[m\,o\,l\,m^{-2}\,s^{-1}]$
K	Intrinsic Permeability.	$[m^2]$
N_α	Molar density in α phase.	$[m\,o\,l\,m^{-3}]$
P_α	Pressure in α phase.	[Pa]
P_{Gvapor}^w	Vapor pressure of pure water.	$[Pa]$
Pc	Capillary pressure.	[Pa]
Q_T	Heat source/sink term.	$[W\,s^{-2}]$
R	Universal Gas Constant.	$[J\,m\,o\,l^{-1}\,K^{-1}]$
$S_{\alpha r}$	Residual saturation in α phase.	[-]
S_α	Saturation in α phase.	[-]
S_{le}	Effective saturation.	[-]
T	Temperature.	[K]
u_α	Specific internal energy.	$[J\,K\,g^{-1}]$
V_α	Volume in α phase.	$[m^3]$
v_α	Darcy velocity in α phase.	$[m\,s^{-1}]$
X	Total molar fraction of light component in two phases.	[-]
X_α^i	Molar Fraction of i-component in α phase.	[-]

AUTHORS' CONTRIBUTIONS

YH implemented the extended method into the OpenGeoSys software, produced simulation results of the two benchmarks, and also drafted this manuscript. HS designed the numerical algorithm of solving the coupled PDEs, and also contributed to the code implementation. OK coordinated the development of OpenGeoSys, and contributed to the manuscript writing. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

We would thank to Dr. Norihito Watanabe for his thoughtful scientific suggestions and comments on this paper. This work has been funded by the Helmholtz Association through the program POF III-R41 'Geothermal Energy Systems'. The first author would also like to acknowledge the financial support from Chinese Scholarship Council (CSC).

REFERENCES

1. Abadpour, A, Panfilov M (2009) Method of negative saturations for modeling two-phase compositional flow with oversaturated zones. Transp Porous Media 79(2): 197–214.
2. Angelini, O, Chavant C, Chénier E, Eymard R, Granet S (2011) Finite volume approximation of a diffusion–dissolution model and application to nuclear waste storage. Math Comput Simul 81(10): 2001–2017.
3. Ben Gharbia, I, Jaffré J (2014) Gas phase appearance and disappearance as a problem with complementarity constraints. Math Comput Simul 99: 28–36.
4. Bourgeat, A, Jurak M, Smaï F (2009) Two-phase, partially miscible flow and transport modeling in porous media; application to gas migration in a nuclear waste repository. Comput Geosci 13(1): 29–42.
5. Bourgeat, A, Jurak M, Smaï F (2013) On persistent primary variables for numerical modeling of gas migration in a nuclear waste repository. Comput Geosci 17(2): 287–305.
6. Çengel, YA, Boles MA (1994) Thermodynamics: an engineering approach. Property Tables, Figures and Charts to Accompany. McGraw-Hill Ryerson, Limited, Singapore. https://books.google.de/books?id=u2-SAAAACAAJ.
7. Class, H, Helmig R, Bastian P (2002) Numerical simulation of non-isothermal multiphase multicomponent processes in porous media: 1. an efficient solution technique. Adv Water Resour 25(5): 533–550.
8. Ern, A, Mozolevski I (2012) Discontinuous galerkin method for two-component liquid–gas porous media flows. Comput Geosci 16(3): 677–690.
9. Fatt, I, Klikoff Jr WA (1959) Effect of fractional wettability on multiphase flow through porous media. Trans., AIME (Am. Inst. Min. Metall. Eng.),216: 426–432.
10. Forsyth, P, Shao B (1991) Numerical simulation of gas venting for NAPL site remediation. Adv Water Resour 14(6): 354–367.
11. Gawin, D, Baggio P, Schrefler BA (1995) Coupled heat, water and gas flow in deformable porous media. Int J Numer Methods Fluids 20(8-9): 969–987. doi:10.1002/fld.1650200817.
12. Hassanizadeh, M, Gray WG (1980) General conservation equations for multi-phase systems: 3. constitutive theory for porous media flow. Adv Water Resour 3(1): 25–40.

13. Helmig, R (1997) Multiphase flow and transport processes in the subsurface: a contribution to the modeling of hydrosystems. Springer, Berlin.
14. Kanzow, C (2004) Inexact semismooth newton methods for large-scale complementarity problems. Optimization Methods Softw 19(3-4): 309–325.
15. Kolditz, O, De Jonge J (2004) Non-isothermal two-phase flow in low-permeable porous media. Comput Mech 33(5): 345–364.
16. Kolditz, O, Bauer S, Bilke L, Böttcher N, Delfs JO, Fischer T, Görke UJ, Kalbacher T, Kosakowski G, McDermott CI, Park CH, Radu F, Rink K, Shao H, Shao HB, Sun F, Sun YY, Singh AK, Taron J, Walther M, Wang W, Watanabe N, Wu Y, Xie M, Xu W, Zehner B (2012) Opengeosys: an open-source initiative for numerical simulation of thermo-hydro-mechanical/chemical (THM/C) processes in porous media. Environ Earth Sci 67(2): 589–599. doi:10.1007/s12665-012-1546-x.
17. Kräutle, S (2011) The semismooth newton method for multicomponent reactive transport with minerals. Adv Water Resour 34(1): 137–151.
18. Landau, L, Lifshitz E (1980) Statistical physics, part i. Course Theoretical Phys 5: 468.
19. Marchand, E, Müller T, Knabner P (2012) Fully coupled generalised hybrid-mixed finite element approximation of two-phase two-component flow in porous media. part ii: numerical scheme and numerical results. Comput Geosci 16(3): 691–708.
20. Marchand, E, Müller T, Knabner P (2013) Fully coupled generalized hybrid-mixed finite element approximation of two-phase two-component flow in porous media. part i: formulation and properties of the mathematical model. Comput Geosci 17(2): 431–442.
21. Marchand, E, Knabner P (2014) Results of the momas benchmark for gas phase appearance and disappearance using generalized mhfe. Adv Water Resour 73: 74–96.
22. Neumann, R, Bastian P, Ippisch O (2013) Modeling and simulation of two-phase two-component flow with disappearing nonwetting phase. Comput Geosci 17(1): 139–149.
23. Olivella, S, Gens A (2000) Vapour transport in low permeability unsaturated soils with capillary effects. Transp Porous Media 40(2): 219–241.

24. Park, CH, Taron J, Görke UJ, Singh AK, Kolditz O (2011) The fluidal interface is where the action is in $C O _2$ sequestration and storage: Hydro-mechanical analysis of mechanical failure. Energy Procedia 4: 3691–3698.
25. Panfilov, M, Panfilova I (2014) Method of negative saturations for flow with variable number of phases in porous media: extension to three-phase multi-component case. Comput Geosci: 1–15.
26. Park, CH, Böttcher N, Wang W, Kolditz O (2011) Are upwind techniques in multi-phase flow models necessary?. J Comput Phys 230(22): 8304–8312.
27. Pruess, K (2008) On production behavior of enhanced geothermal systems with $C O _2$ as working fluid. Energy Convers Manag 49(6): 1446–1454.
28. Peng, DY, Robinson DB (1976) A new two-constant equation of state. Ind Eng Chem Fundam 15(1): 59–64.
29. Salimi, H, Wolf KH, Bruining J (2012) Negative saturation approach for non-isothermal compositional two-phase flow simulations. Transp Porous Media 91(2): 391–422.
30. Singh, A, Baumann G, Henninges J, Görke UJ, Kolditz O (2012) 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(2): 497–509.
31. Singh, A, Delfs JO, Böttcher N, Taron J, Wang W, Görke UJ, Kolditz O (2013a) A benchmark study on non-isothermal compositional fluid flow. Energy Procedia 37: 3901–3910.
32. Singh, A, Delfs JO, Shao H, Kolditz O (2013b) Characterization of co2 leakage into the freshwater body In: EGU General Assembly Conference Abstracts, 11474.
33. Udell, K, Fitch J (1985) Heat and mass transfer in capillary porous media considering evaporation, condensation, and non-condensible gas effects In: 23rd ASME/AIChE National Heat Transfer Conference, Denver, CO, 103–110.
34. Van Genuchten, MT (1980) A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. Soil Sci Soc Am J 44(5): 892–898.

35. Wu, YS, Forsyth PA (2001) On the selection of primary variables in numerical formulation for modeling multiphase flow in porous media. J Contam Hydrol 48(3): 277–304.

Chapter

MODELLING AND DYNAMIC CHARACTERISTICS FOR A NON-METAL PRESSURIZED RESERVOIR WITH VARIABLE VOLUME

7

Pei Wang[1]**, Jing Yao**[1,2,3]**, Baidong Feng**[1]**, Mandi Li**[1]**, and Dingyu Wang**[1]

[1]School of Mechanical Engineering, Yanshan University, Qinhuangdao 066004, China.
[2]Advanced Manufacturing Forming Technology and Equipment, Qinhuangdao 066004, China.
[3]Hebei Provincial Key Laboratory of Heavy Fluid Power Transmission and Control, Yanshan University, Qinhuangdao 066004, China.

ABSTRACT

With the increasing demand to reduce emissions and save energy, hydraulic reservoirs require new architecture to optimize their weight, space, and volume. Conventional open reservoirs are large, heavy, and easily polluted, and threaten the operation of hydraulic systems. A closed reservoir provides the advantages of small volume and light weight, compared to open

Citation: (APA): Wang, P., Yao, J., Feng, B., Li, M., & Wang, D. (2022). Modelling and Dynamic Characteristics for a Non-metal Pressurized Reservoir with Variable Volume. Chinese Journal of Mechanical Engineering, 35(1), 1-15. (15 pages).

reservoirs. In this study, a non-metallic pressure reservoir with variable volume is designed and manufactured for closed-circuit hydraulic systems. The reservoir housing is made of rubber, and the Mooney-Rivlin model is chosen based on the rubber strain properties. The FEA simulation for the reservoir is performed using ANSYS Workbench to obtain the structural stiffness. The major contribution is the establishment of mathematical models for this reservoir, including the volume equation changing with height, flow equation, and force balance equation, to explore the output characteristics of this reservoir. Based on these results, simulation models were built to analyze the output characteristics of the reservoir. Moreover, the test rig of a conventional hydraulic system was transformed into a closed-circuit asymmetric hydraulic system for the reservoir, and preliminary verification experiments were conducted on it. The results demonstrate that the designed reservoir can absorb and discharge oil and supercharge pump inlet to benefit system operation. The changes in the volume and pressure with displacements under different volume ratios and frequencies were obtained, which verified the accuracy of the mathematical models. Owing to its lightweight design and small volume, the reservoir can replace conventional open reservoirs, and this lays a foundation for future theoretical research on this reservoir.

Keywords: Hydraulic reservoir, Variable volume, Pressure reservoir, Non-metal, Lightweight

INTRODUCTION

Hydraulic transmission is crucial in aerospace, heavy machinery, engineering machinery, and other industries [1,2,3,4,5]. Hydraulic systems often rely on large and heavy reservoirs in mobile hydraulic machinery. It is one of the components with the highest potential for weight reduction [6]. The lighter weight of hydraulic reservoirs can improve the power-to-weight ratio of mobile hydraulic machinery and reduce power consumption, achieving energy conservation and emission reduction. The purpose of this study is to effectively reduce the mass and volume of hydraulic reservoirs while meeting the strict requirements of hydraulic systems.

Open and closed reservoirs are utilized in the hydraulic systems. Although open reservoirs are widely utilized, they are not applied in high-altitude environments and are restricted to mobile hydraulic machinery. To address this challenge, closed reservoirs have installed a pressure-driven device that can relatively stabilize the oil pressure at the pump inlet and enable the hydraulic system to work at a high altitude or in harsh environment [7,8,9,10]. Compared with open reservoirs, closed reservoirs have a smaller volume and mass, and are more widely utilized in mobile hydraulic machinery. There is an urgency to design a new type of closed reservoir with small volume and light weight, to meet the lightweight and pressure requirements of the hydraulic system.

There are several types of closed reservoirs, and the difference between them is mainly reflected in pressure-driven methods. The pressure-driven devices of a closed reservoir are mainly divided into hydraulic, spring, and pneumatic drives. Parker's closed metal reservoir is designed to connect the system pressure to the reservoir drive interface. In addition, system pressure can be converted into a stable low-pressure output; however, it occupies a large space and has a large mass. Spring-driven [11, 12] closed reservoirs rely on a spring force to exert pressure on the oil, but their structure is relatively complex and difficult to process and manufacture. Another type of closed reservoir utilizing a pneumatic drive [13,14,15,16,17,18,19] is protected by a metal housing, but the container inside is coated in rubber skin. Gas is filled between the metal housing and rubber skin to establish the driving pressure, but it is sensitive to temperature, which affects the normal operation of the hydraulic system, such as the contact booster hydraulic [20] and airbag isolated booster hydraulic tanks [21]. There are also closed reservoirs with special structural forms, such as a vacuum reservoir with variable capacity, following the movement of actuators [22]. The corrugated elastic lining and housing form a closed capacity cavity, and the bellows produce telescopic action in the opposite direction, altering the volume [23]. Its disadvantage is that the pressure of the output oil is nonlinear and unsuitable for hydraulic systems that require stable pressure. As described above, the complex structure, large volume, large mass, and nonlinear output pressure of the existing closed reservoir are still not conducive for reducing the weight and space of the hydraulic system.

Generally, conventional hydraulic reservoirs are made of metal materials with simple structures, but larger volumes and weights. Non-metal reservoirs have garnered significant attention owing to their lightweight development. A few companies have researched and developed reservoirs using special materials for various functions. ARGO-HYTOS [24] developed an injection-plastic reservoir with a filtering function. Fuel Safe [25] developed a material for designing a collapsible aerial fuel transport container. It is constructed with multiple layers of ballistic nylon cords and a rugged rubberized polymer coating, but not suitable for hydraulic systems. Turtle-Pac [26] manufactured aircraft tanks using unique fabrics and technologies. It is lightweight, convenient to fold, compact, and is tested at 5 PSI during quality control tests. In addition, Smart Reservoir [27], a company in Canada, produced another type of reservoir with the features of lighter mass, smaller volume, and linear output. It has already been applied in various fields, but the effect of rubber on reservoir performance is unclear. Currently, most of the research on lightweight non-metal hydraulic reservoirs are abroad, but seldom research has been conducted on its characteristics and material.

Therefore, considering the particularity of rubber material, this research takes the variable volume and pressure reservoir (VVPR) as the research object, focuses on the establishment of mathematical models for dynamic characteristics, and explores the interaction between the system parameter and VVPR by simulation and experiments. This paper is divided into four parts: ① Composition and principle of the VVPR. ⍰ Strain properties and structural stiffness of reservoir rubber. ③ Modeling and simulation of the VVPR using AMESIM and MATLAB joint simulation methods. ④ Performance analysis of the VVPR in closed-circuit hydraulic system on test platform. It is important to investigate this closed reservoir to replace open reservoirs in numerous applications, which provides a theoretical basis for further research.

RESERVOIR DESCRIPTION

Working Principle

The VVPR is sealed, airless, and slightly pressurized, with small volume, light weight, low pollution, and portability. It comprises upper and lower covers, rings, connecting rods, rubber housing, springs, pillars, and other components, as illustrated in Figure 1.

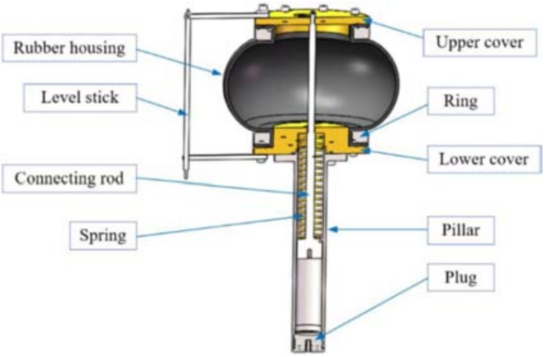


Figure 1. Composition of the reservoir (VVPR).

The connecting rod is fixed to the upper cover, and the rod and spring in the pillar move with reservoir motion. The spring is always in a compressed state with a downward force on the connecting rod. Through the interaction of the spring and internal pressure of the oil, the reservoir can achieve the functions of absorbing and discharging oil.

The VVPR designed in this study plays the role of differential volume compensation in closed-circuit asymmetric hydraulic systems by its own variable volume. In the working cycle of the reservoir, expansion and contraction occur as actuator movement changes. The process of oil absorption and discharge with the expansion and compression of the rubber housing can be achieved using the spring force. As illustrated in Figures 2 and 3, the working principle of the VVPR in a closed-circuit valve-controlled asymmetric hydraulic system is described as follows.

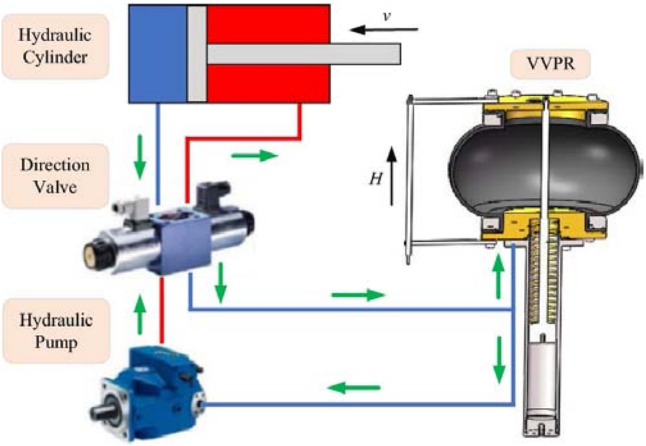


Figure 2. The oil absorption condition of the reservoir.

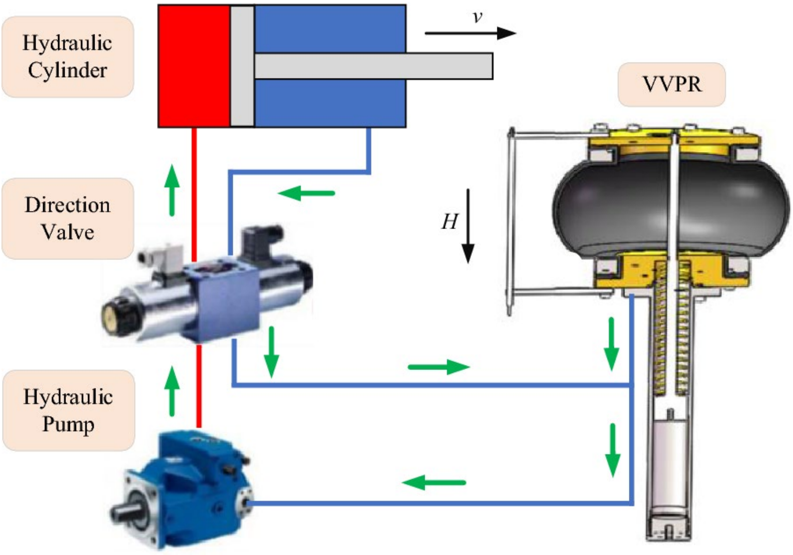


Figure 3. The oil discharging condition of the reservoir.

A closed-circuit valve-controlled asymmetric hydraulic system comprises a hydraulic pump, single-rod cylinder, direction valve, and VVPR. The oil absorption conditions of the reservoir are illustrated in Figure 2. In the retraction process of the cylinder, the oil from the outlet of the hydraulic pump enters the rod chamber of the hydraulic cylinder, and the oil in the rodless chamber enters the suction port of the hydraulic pump through the directional valve. However, owing to the different volumes of oil in the two chambers of the hydraulic cylinder, part of the oil in the rodless chamber enters the reservoir.

The oil discharge conditions of the VVPR are illustrated in Figure 3. During the extension of the cylinder, the oil from the outlet of the hydraulic pump enters the rodless chamber of the hydraulic cylinder, and the oil in the rod chamber enters the suction port of the hydraulic pump through the directional valve. The oil in the VVPR is replenished into the inlet of the pump.

A VVPR was preliminarily designed and manufactured to investigate the reservoir characteristics. The main parameters of the VVPR are as follows: ① Its working pressure is under 0.06 MPa, working volume is 5

L, and structure volume is 9.63 L. ② It is 580 mm in height and 400 mm in width. ③ Its mass is only 13.2 kg, owing to the smaller working cavity made of rubber material.

Rubber Strain Properties and its Constitutive Equation

The rubber housing material utilized in the VVPR is a hyperplastic material that may affect the VVPR dynamic characteristics. Therefore, the physical properties of these materials should be described based on their elasticity and deformation.

To determine the correct mathematical model (constitutive equation) to describe rubber physical properties in the reservoir and define the system dynamic properties, uniaxial tension tests with six rubber samples were conducted at 20 °C (environment temperature), as illustrated in Figure 4. There are several mathematical models to describe the physical properties of hyperplastic materials, including the Mooney-Rivlin model, Ogden, and Yeoh models, but not all of them are suitable for specific hyperplastic materials [28]. The rubber stress and strain data from the uniaxial tension test results on the test pieces were utilized for comparison with the FEA simulation results from ANSYS Workbench for different rubber models, and the results are illustrated in Figure 5.

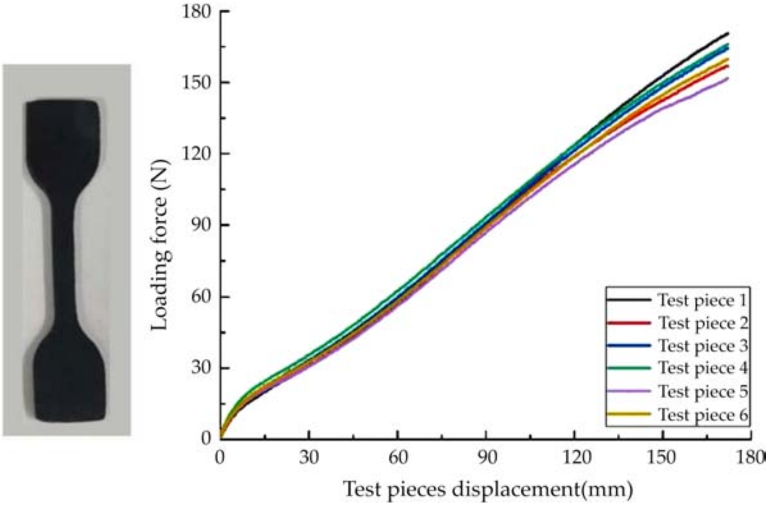


Figure 4. Uniaxial tension test results at 20 ℃.

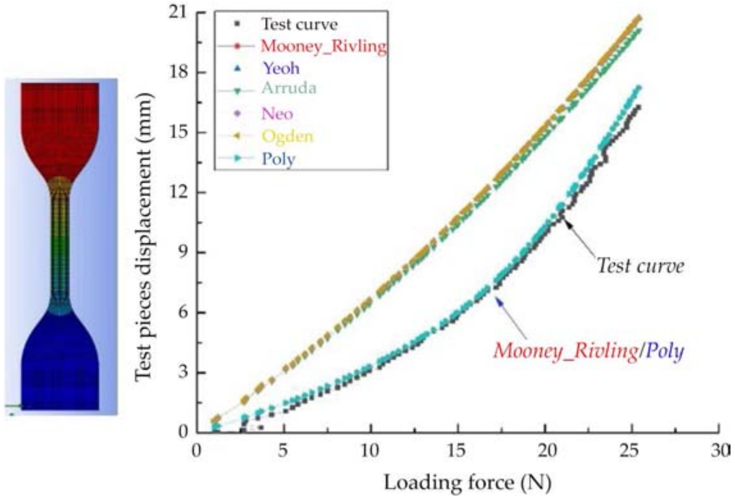


Figure 5. The test pieces displacement curve of FEA simulation results and test result.

It can be observed that there are constitutive models that fit the test data well, i.e., Mooney-Rivlin, and Poly, which can describe the basic rubber hyperplastic properties. In this study, the Mooney-Rivlin model was selected for the next simulation step, and its parameters are presented in Table 1.

Table 1. Parameters of rubber Mooney-Rivlin model

Rubber parameter	**Value(MPa)**
Material constant C10	− 2.4201
Material constant C01	4.2156
Material constant C20	− 0.0035296
Material constant C11	0.027367
Material constant C02	0.81943
Incompressibility parameter	0

Rubber Structural Stiffness of VVPR

Since the hyperplastic material is elastic, the stiffness of the structure on the upper cover should be provided. Once the constitutive model of rubber and the structure of the reservoir are determined, the rubber structural stiffness of the VVPR can also be obtained. Therefore, FEA simulation on ANSYS Workbench was performed by constantly changing the force *F* on the connecting rob and simulating the displacement of the upper cover x. The rubber structural stiffness K_1 of VVPR can be calculated with Eq. (1).

$$K_1 = \frac{\Delta F}{\Delta x}, \qquad (1)$$

The rubber structural stiffness of the VVPR when a force acts on the upper cover is illustrated in Figure 6. It can be observed that as the reservoir height increases, the stiffness also increases.

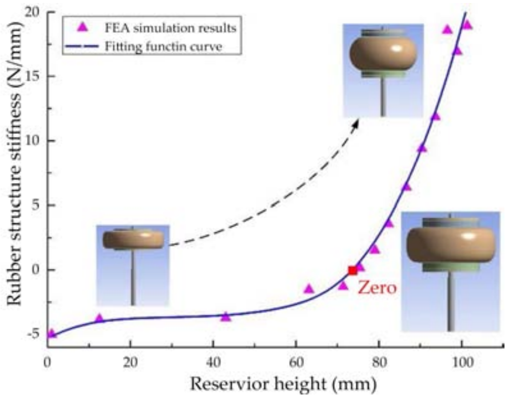


Figure 6. Rubber structural stiffness.

According to simulation results, the fitting function of K_1 has also been obtained using a quartic polynomial curve function with a corresponding fitting precision of R-square = 0.982 and RMSE = 0.1389. The stiffness fitting functions and parameters are presented in Table 2.

Table 2. The stiffness fitting function and parameters

Parameters	**Value**
K_1(N/mm)	$f(x)$=$a_1x^4+a_2x^3+a_3x^2+a_4x+a_5$
a_1(N/mm^5)	2.401×10^{-7}
a_2(N/mm^4)	4.105×10^{-5}
a_3(N/mm^3)	− 0.006157
a_4(N/mm^2)	0.2066
a_5(N/mm)	− 5.315

MODELLING AND SIMULATION

The VVPR continuously absorbs and discharges oil in its working cycle with the movement of the hydraulic cylinder, thereby affecting the changes in its volume and pressure. Volume and pressure are important parameters in the performance of the VVPR; thus, mathematical models must be built for further simulation analysis of the changes in volume and pressure during the working cycle.

Force Balance Equation

The VVPR can be equivalently treated as a single-degree-freedom system with spring-mass-damping, which can be simplified to the model illustrated in Figure 7.

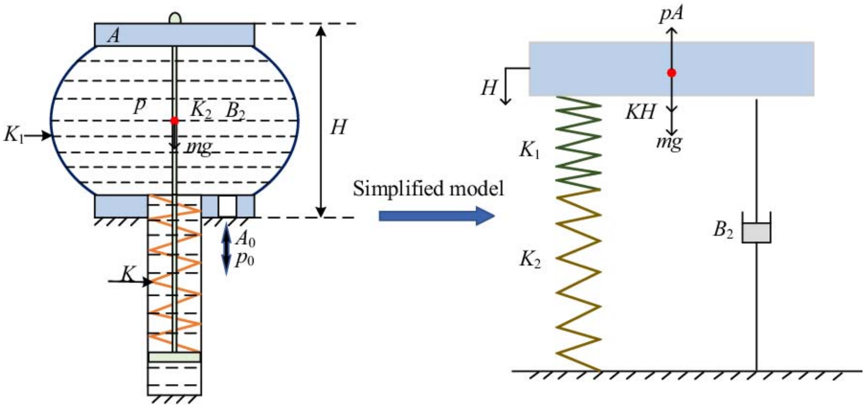


Figure 7. The dynamic model of the VVPR.

In Figure 7, H is the height of the VVPR; K, K_1, and K_2 are the stiffness of the spring, rubber structure, and oil, respectively; m is the mass of the moving parts of the VVPR, A is the area of the upper cover, p is the pressure inside the VVPR, and B_2 is the movement damping of the VVPR.

The oil stillness can be obtained by Eq. (2):

$$\left\{\begin{array}{l} \Delta V = \dfrac{V}{\beta_e}\Delta p, \\ \Delta V = A\Delta H, \quad \Rightarrow K_2 = \dfrac{\beta_e A^2}{V}. \\ K_2 = \dfrac{\Delta p A}{\Delta H}, \end{array}\right. \tag{2}$$

The damping coefficient B_2 can be expressed as Eq. (3):

$$B_2 = 8\pi\mu H, \tag{3}$$

where the $\mu = \mu_0 \cdot e^{\alpha \cdot p}$, $\mu_0 = 0.35$ Pa·s, and $\alpha = (0.015 \sim 0.35)\mathrm{MPa}^{-1}$[29, 30].

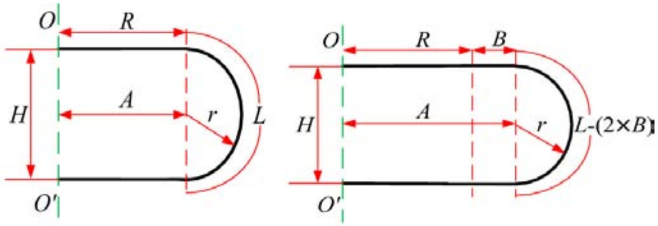
When the VVPR operates by oil absorption and discharge, the upper cover exerts a force on the oil-generating pressure p. According to the dynamic model, this study considers the upper cover as the research object, and the force balance equation is established by considering the inertial force, viscous damping force, spring force, and internal pressure on the upper cover. Hence, the force balance equation can be expressed as Eq. (4):

$$Ap = m\frac{\mathrm{d}^2 H}{\mathrm{d}t^2} + B_2\frac{\mathrm{d}H}{\mathrm{d}t} + \left(K - \frac{K_1 \cdot K_2}{K_1 + K_2}\right)H + mg. \tag{4}$$

The first term on the right side of Eq. (4) is the inertial force, second term is the viscous damping force, and final term is the elastic force exerted by the spring, rubber, and oil.

Volume Equation

As the VVPR moves, the volume changes with the height. Volume calculation is crucial for obtaining the change in height; its dimensions in the vertical plane are illustrated in Figure 8.



(a) Uncompressed volume (b) Compressed volume

Figure 8. The dimension diagram of the rubber housing.

In Figure 8, r is the radius, L is the arc length, R is the radius of the upper cover, A is the distance from the central axis to the arc center, and B is the line segment after arc deformation. The relationships between these variables can be expressed as Eq. (5).

$$\begin{cases} r = \dfrac{H}{2}, \\ B = \dfrac{L - \pi r}{2}, \\ A = B + R. \end{cases} \tag{5}$$

Hence, the volume equation can be expressed as Eq. (6).

$$V(H) = \pi H \left\{ \frac{1}{6} H^2 + R^2 + \frac{1}{4} R \pi H + \left(\frac{L}{2} - \frac{\pi H}{4} \right) \left(\frac{L}{2} + 2R \right) \right\}. \tag{6}$$

Furthermore, the relationship between the volume V and height H was drawn based on the Newton iteration method in MATLAB, as illustrated in Figure 9.

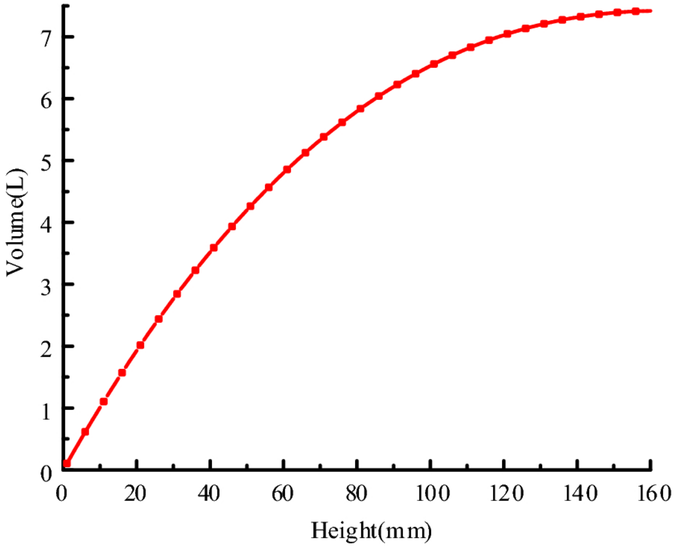


Figure 9. The simulation curve of volume change with height.

Flow Continuity Equation

Some assumptions are made to establish the flow continuity equation of the VVPR: ① pressure loss and pipeline dynamic characteristics of components other than the VVPR are not considered; ② the elastic modulus of oil and oil temperature are constant; and ③ the reservoir leakage is a laminar flow. Hence, the flow continuity equation is expressed as Eq. (7):

$$q = \frac{\mathrm{d}V}{\mathrm{d}t} + C_{ep}p + \frac{V}{\beta_e}\frac{\mathrm{d}p}{\mathrm{d}t}. \tag{7}$$

The first term on the right side of Eq. (7) represents the volume change of the VVPR during operation, second term represents the flow rate change caused by leakage, and third term represents the flow rate change due to compression.

Simulation Analysis

To explore the performance of the VVPR in the system and effect of the VVPR on the system, joint models of the VVPR in MATLAB and a closed-circuit hydraulic system in AMESIM were built for their interaction with different working parameters. The main simulation parameters are presented in Table 3.

Table 3. Simulation parameters

Name	Parameter	Value
Cylinder	Diameter of rod *d*(mm)	90
	Diameter of piston *D*(mm)	110
	Stroke of cylinder *S*(mm)	800
Reservoir	Stiffness of spring *K*(N/mm)	13.63
	Mass *m*(kg)	6
	Damping B_2(N· s/mm)	0.01
	Elastic modulus β_e(MPa)	700

The initial volume was set as 3.7 L, and the ratio between the actual working volume Δ*V* and maximum working volume *V* is defined as the volume ratio, expressed as Eq. (8):

$$\alpha = \frac{\Delta V}{V}. \tag{8}$$

The relationship between the actual working volume Δ*V* of the VVPR and hydraulic cylinder displacement *y* is expressed as Eq. (9):

$$\frac{\pi d^2}{4}y = \Delta V. \tag{9}$$

In this study, the hydraulic cylinder is given a sine displacement reference to form the volume difference between extension and retraction. The corresponding values of the volume ratio, stroke, and amplitude are presented in Table 4.

Table 4. The relationship among the volume ratio, stroke, and amplitude

Volume ratio	Stroke(mm)	Amplitude(mm)
0.05	39.30	19.65
0.1	78.60	39.30
0.2	157.20	78.60
0.3	235.80	117.90
0.4	314.40	157.20
0.5	393.00	196.50
0.6	471.60	235.80
0.7	550.20	275.10

The displacement of the hydraulic cylinder was set as a sine curve with amplitudes of 117.90 mm at 0.016 Hz, to analyze the output characteristics of the VVPR. As can be observed from Figures 10 and 11, the height and volume of the VVPR gradually increases when oil is absorbed. At half a cycle, the flow rate into the VVPR is 0, and both the volume and height of the VVPR reach their maximum values, but the pressure of the VVPR does not. During the oil discharge, the volume and height of the VVPR gradually decreases, while the pressure of the VVPR increases.

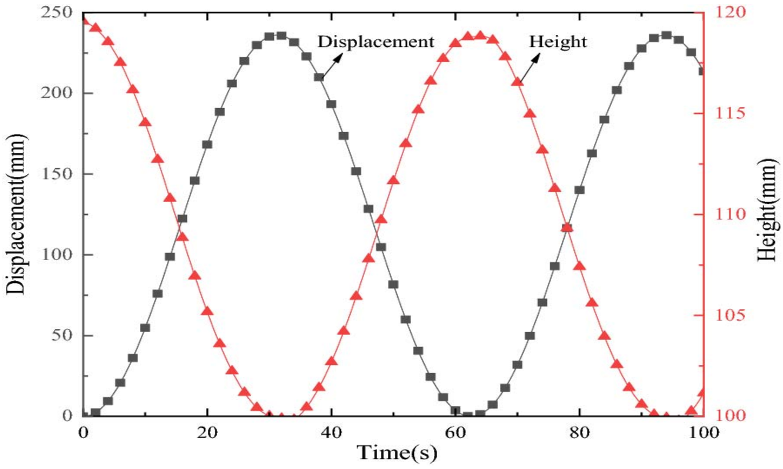


Figure 10. The height changes with displacement.

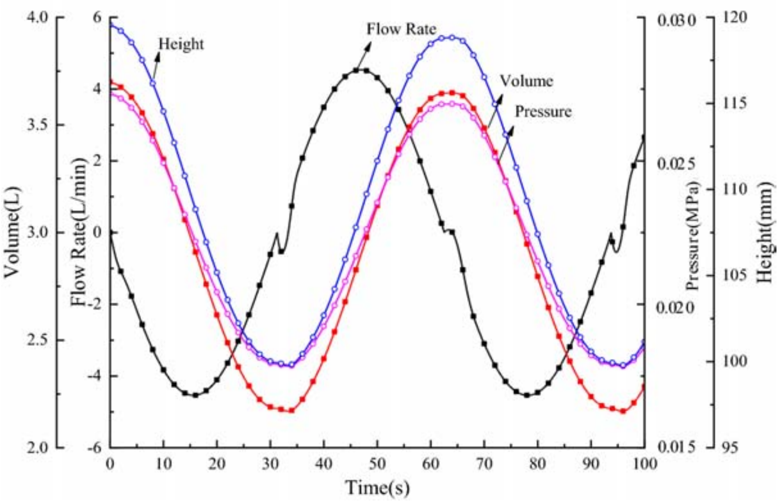


Figure 11. The changes in the main parameters of the VVPR.

Through the simulation analysis, it is determined that the volume and height of the VVPR change with the flow rate described earlier, but the change in the pressure of the VVPR has a certain delay. This may be caused by the rubber material characteristics and damping force.

Performance Analysis at Different Volume Ratios

In this section, the VVPR performance with different volume ratios is analyzed by altering the amplitudes of the sinusoidal displacements. The cylinder displacement is selected as a sine curve at 0.016 Hz, and the volume ratios of the reservoir are set as 0.05, 0.1, 0.2, 0.3 0.4, 0.5, 0.6 and 0.7, changed by the different strokes.

Figures 12, 13, 14 and 15 illustrate that the frequencies and amplitudes of the pressure and volume change with displacement, thereby displaying a positive relationship. This illustrates that the reservoir achieved the function of flow inlet and outlet with the extension and retraction of the cylinder. The cylinder displacement tracked the reference well, proving that the introduction of the VVPR had no effect on the output characteristics of the system.

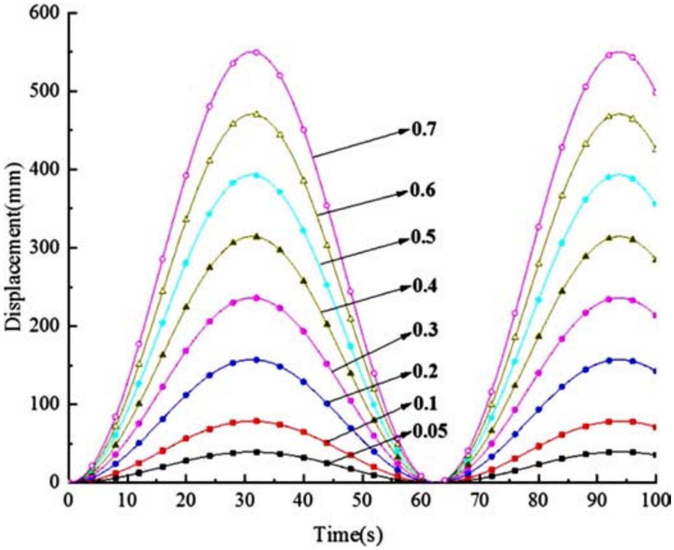


Figure 12. Displacement curve of hydraulic cylinder.

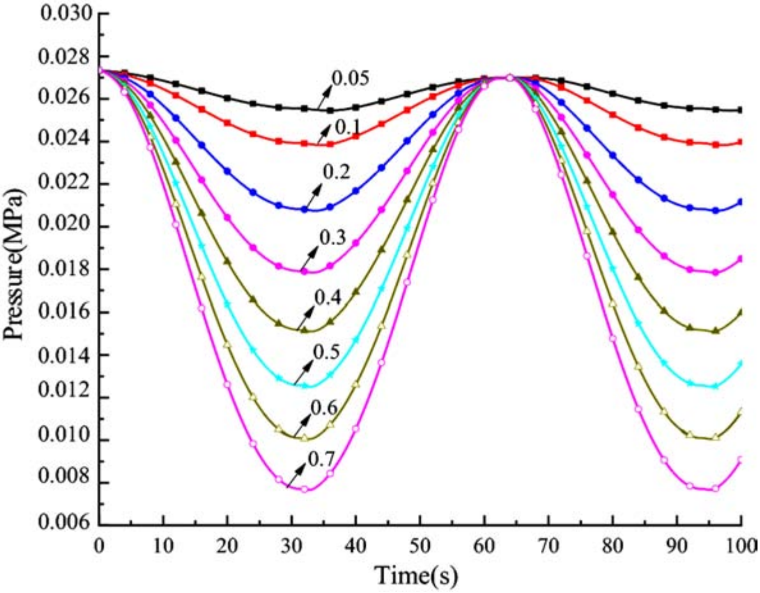


Figure 13. Changes in pressure at different ratios.

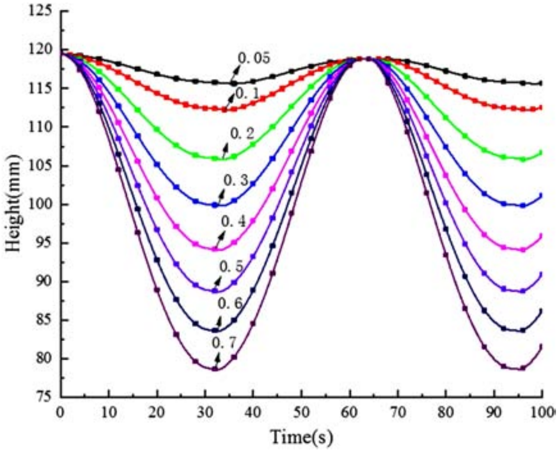


Figure 14. Changes in height at different ratios.

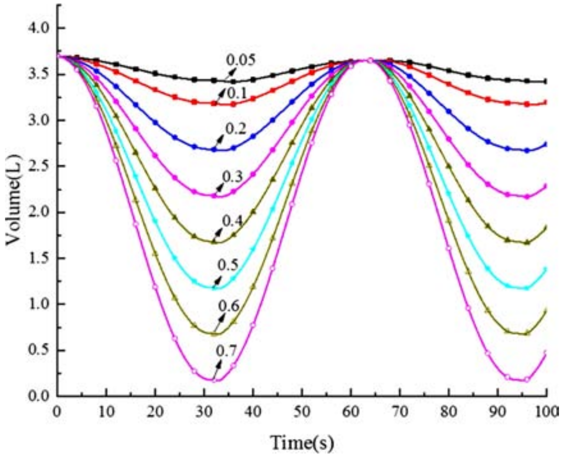


Figure 15. Changes in volume at different ratios.

Performance Analysis at Different Frequencies

In this section, we further explored the effect of frequency on the output characteristics of the reservoir. The volume ratio was set to 0.3, and the working frequencies were selected as 0.1 Hz, 0.5 Hz, and 1 Hz to simulate the closed-circuit hydraulic system.

It can be observed from Figures 16, 17, 18 and 19 that the frequencies of the pressure, volume, and height are the same as the displacement. When the cylinder was extended, the reservoir provided the system with oil. The pressure, volume, and height decrease with increasing displacement. In contrast, the system provided oil to the reservoir when the cylinder was retracted. The pressure, volume, and height increased with decreasing displacement. This indicated that the reservoir achieved the function of flow inlet and outlet with the extension and retraction of the cylinder. The cylinder displacement tracked the reference well, proving that the introduction of the VVPR had approximately no effect on the system characteristics.

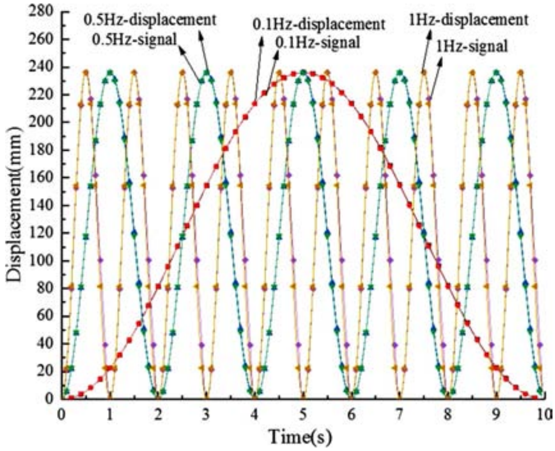


Figure 16. Displacement curves of hydraulic cylinder.

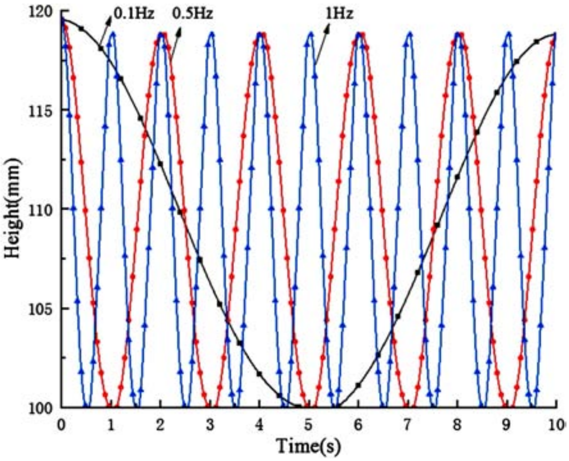


Figure 17. Changes in height at different frequencies.

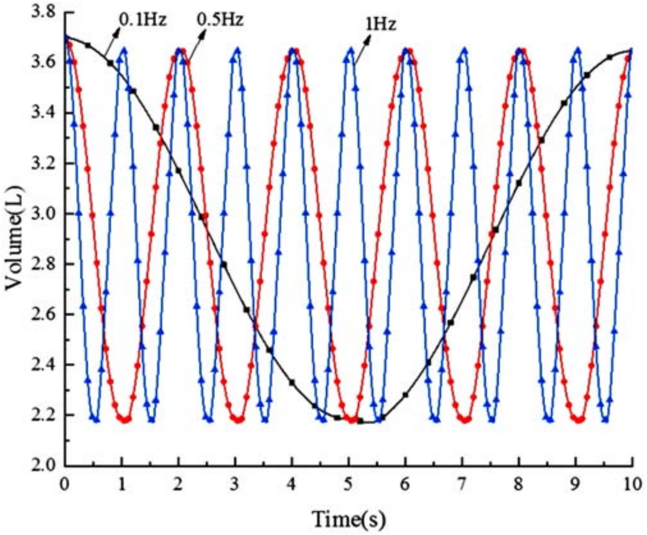


Figure 18. Changes in volume at different frequencies.

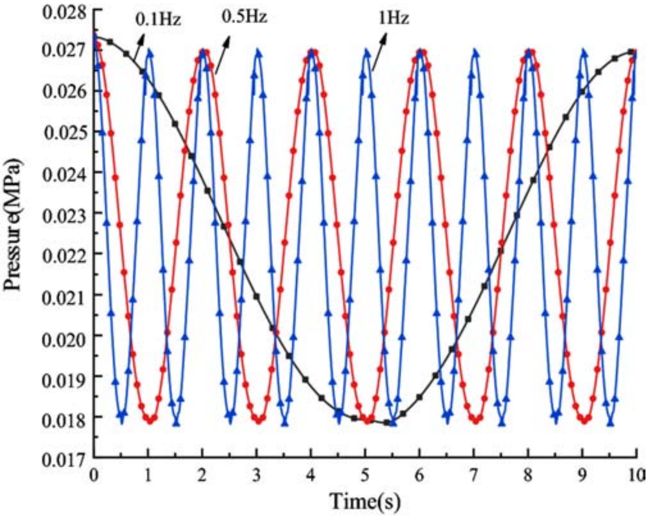


Figure 19. Changes in pressure at different frequencies.

EXPERIMENTAL RESULTS AND DISCUSSION

The test rig was transformed based on a conventional open hydraulic system to verify the accuracy of the mathematical models and simulation analysis, as illustrated in Figure 20. The test platform can achieve an open-or closed-circuit hydraulic system by the on/off of the shut-off valve (2.2) and union tee (3). Before the VVPR worked normally, it needed to replenish oil by opening 2.2 and 2.3 in retraction to discharge the internal air and set an initial volume (3.7 L in this study) by the open-circuit hydraulic system. The VVPR can then work normally to compensate for the volume difference of the cylinder in the closed-circuit hydraulic system.

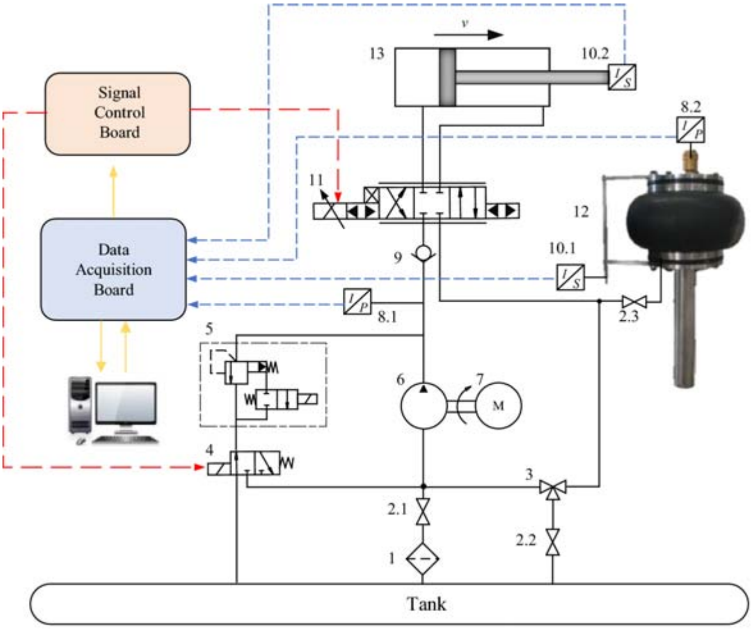


Figure 20. Principle of VVPR dynamic test. 1. Filter; 2. Shut-off valves; 3. Union tee; 4. Direction valve; 5. Unloading relief valve; 6. Hydraulic pump; 7. Electrical motor; 8. Pressure sensors; 9. One-way valve; 10. Displacement sensors; 11. Proportional reverse valve; 12. VVPR; 13 Hydraulic cylinder.

The parameters of the test rig and main parameters between the conventional open reservoir and designed closed reservoir are presented in Table 5. In this study, volume and mass were dramatically reduced by approximately 98% and 94%, respectively.

Table 5. Main parameters of the test platform

Name	**Parameter**	**Value**
System	Flow rate(L/min)	20
Cylinder	Diameter of rod(mm)	90
	Diameter of piston(mm)	110
	Stroke of cylinder(mm)	475
	Difference in volume(L)	3
Open Reservoir	Length(mm)	1100
	Width(mm)	700
	Height(mm)	760
	Volume(L)	590
	Mass with oil(kg)	383.6
Closed Reservoir	Stiffness of spring(N/mm)	13.63
	Total height (mm)	580
	Width(mm)	400
	Structural volume (L)	9.63
	Mass with oil(kg)	21.28

Repeatability Analysis

In this section, static tests are conducted to explore the pressure and volume repeatability.

It can be observed from Figures 21 and 22 that the rubber housing exhibits a hysteresis phenomenon in the rising and falling processes, and the pressure curve is particularly obvious. However, the test volume-height curve is basically coincident, and the error of the pressure repeatability is within 0.002 MPa. Both the pressure and volume characteristics have high coincidence and repeatability in the three tests, which proves the volume and pressure performance and stability of the rubber material.

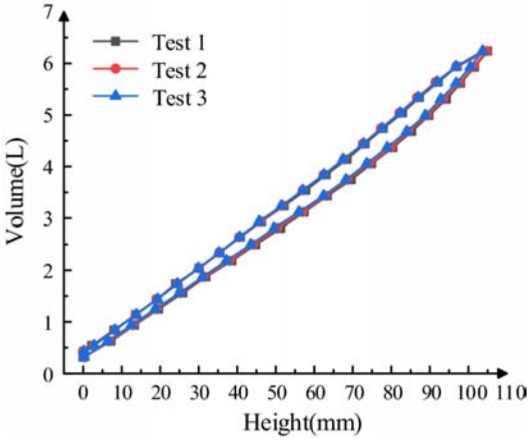


Figure 21. Volume repeatability.

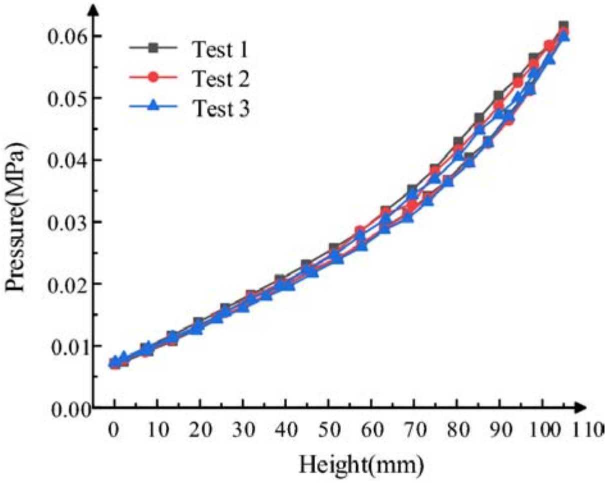


Figure 22. Pressure repeatability.

Experiment Analysis

To explore the relationship between the changes in pressure, height, and volume of the VVPR and the flow rate of the VVPR, the volume ratio was set to 0.3 with an initial value of 3.7 L. A sine reference with a 118 mm amplitude of 0.016 Hz was selected to test the performance of the VVPR. The relationships between the height, pressure, and volume of the VVPR were obtained. The key parameter changes in the working cycle are illustrated in Figures 22 and 23, respectively.

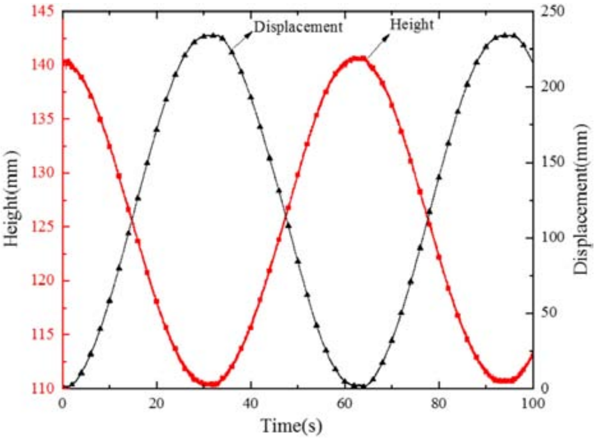


Figure 23. Change in height of the VVPR as displacement of the cylinder.

It can be observed from Figures 23 and 24 that the height of the VVPR can follow the displacement of the cylinder. When the cylinder was extended and retracted, the height decreased and increased at the same frequency, with approximately no hysteresis. Meanwhile, the changes in pressure and volume had the same frequency and tendency as the height. The experimental results indicated a similar tendency for these parameters as the simulation results, and it was demonstrated that the mathematical models and simulation results were corrected. Furthermore, the performance was analyzed at different working volumes and frequencies.

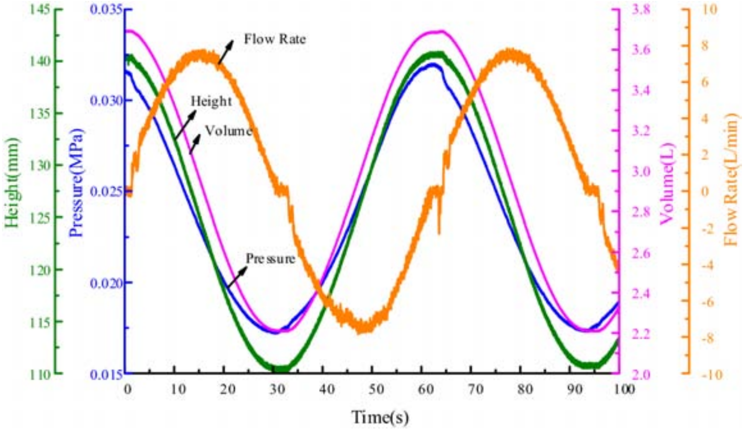


Figure 24. The changes in the main parameters of the VVPR.

Performance Analysis at Different Volume Ratios

To verify the processes of parameter changes in the VVPR under different working volume ratios, the strokes of the cylinder were set to different values in the process of extension and retraction, to achieve different working volume ratios (0.05, 0.1, 0.2, 0.3, 0.4, 0.5 and 0.6) at 0.016 Hz. The corresponding strokes were calculated using Eq. (9). The test results are as follows.

It can be observed from Figures 25, 26, 27 and 28 that the change tendency in pressure, height, and volume of the VVPR under different working volume ratios are the same; i.e., the performance of the VVPR under different working volume ratios does not change, only the variation range of these parameters changes. This is consistent with the earlier simulation findings, suggesting that the performance of the VVPR is stable at different volume ratios.

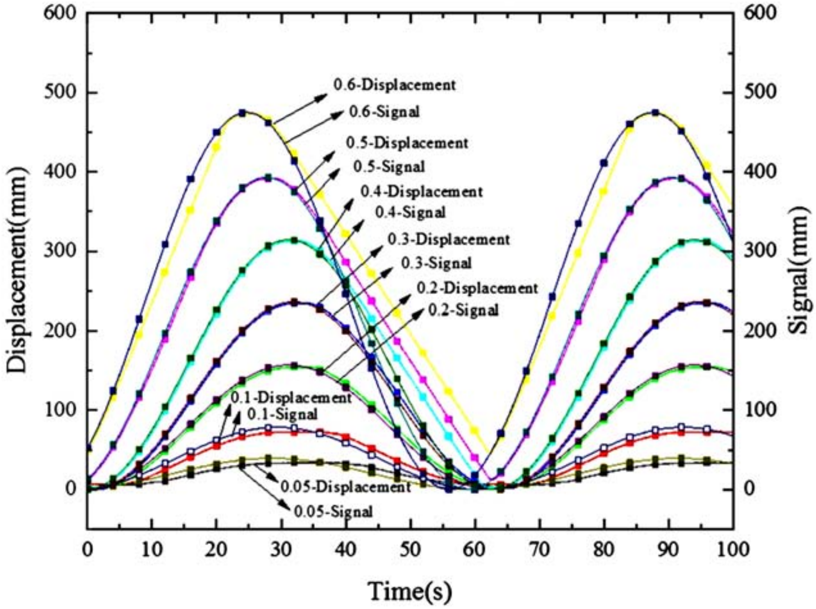


Figure 25. Hydraulic cylinder displacement curve.

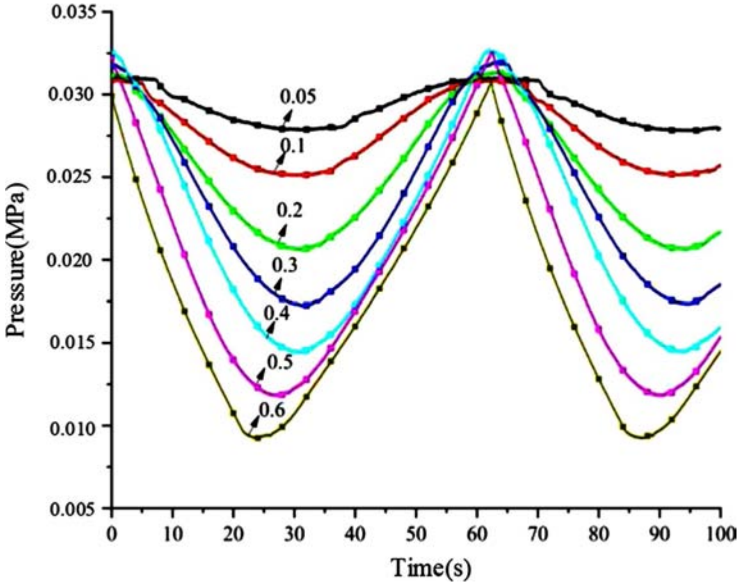


Figure 26. Changes in pressure at different volume ratios.

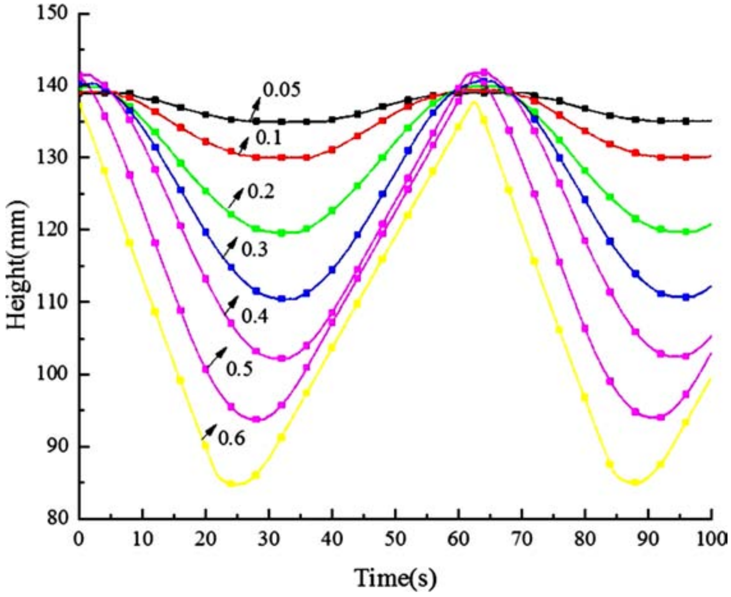


Figure 27. Changes in height at different volume ratios.

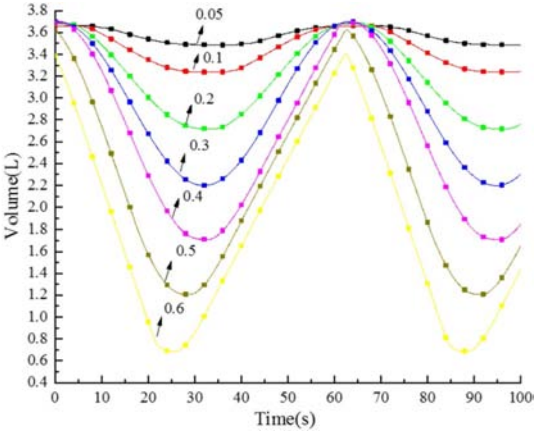


Figure 28. Changes in volume at different volume ratios.

Performance Analysis at Different Volume Ratios

To analyze the performance of the VVPR at different frequencies, sinusoidal signals with a working volume ratio of 0.1, and frequencies of 0.05 Hz and 0.067 Hz were selected for testing. It can be observed from Figures 29, 30, 31 and 32 that the height, volume, and pressure of the VVPR change with the displacement of the hydraulic cylinder. The working frequencies of the VVPR also changed with those of the hydraulic cylinder. At different working frequencies, the changes in the height, volume, and pressure of the VVPR were the same as those in the displacement of the cylinder. This is consistent with earlier simulation findings, suggesting that the performance of the VVPR is stable at different frequencies.

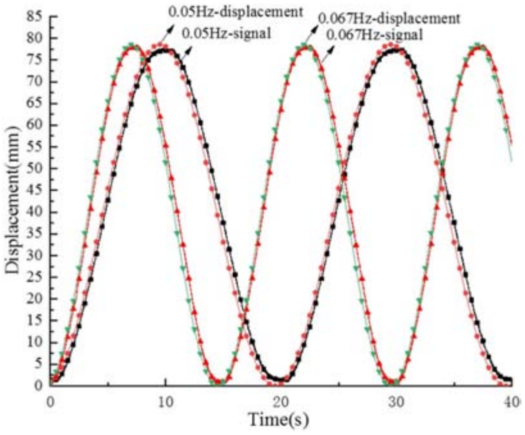


Figure 29. Displacements of the hydraulic cylinder.

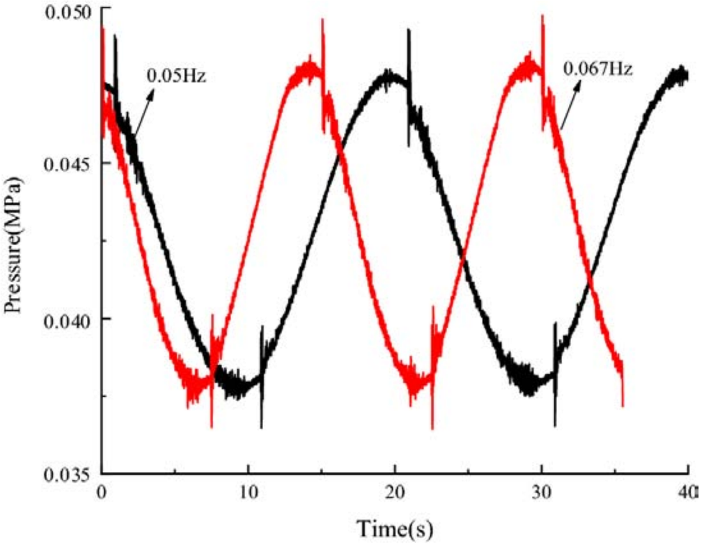


Figure 30. Changes in pressure at different frequencies.

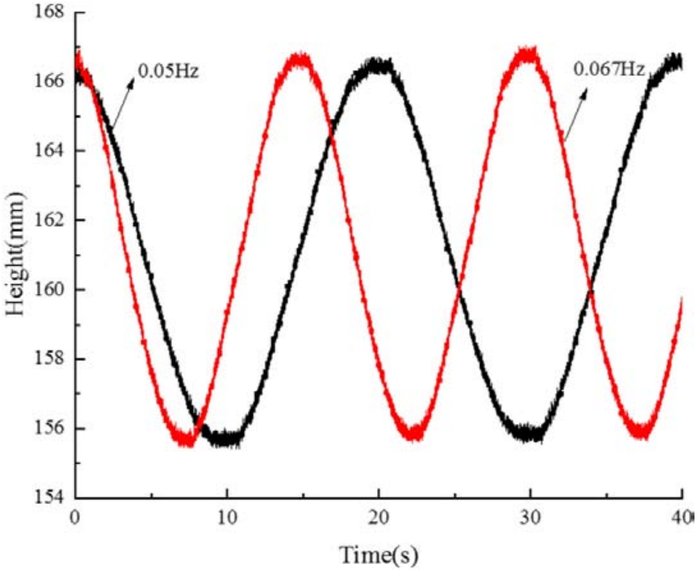


Figure 31. Changes in height at different frequencies.

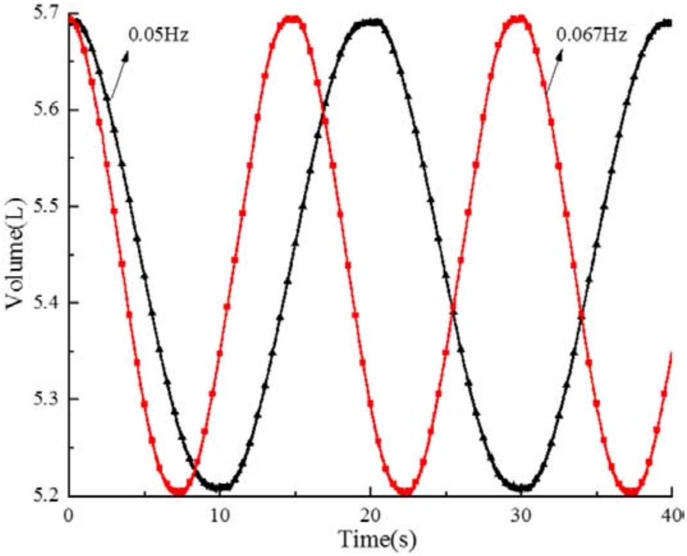


Figure 32. Changes in volume at different frequencies.

CONCLUSIONS

- A new type of non-metallic pressure hydraulic reservoir with variable volume was manufactured and analyzed in a closed-circuit asymmetric hydraulic system, and its main components and working principles were introduced.
- The rubber structural stiffness was obtained via FEA simulation, based on the Mooney-Rivlin model. Furthermore, mathematical models of the reservoir were established, including volume, flow, and force balance equations. Besides, MATLAB and AMESIM joint simulation models were built for the VVPR.
- Tests were conducted, and the results demonstrated that the pressure, height, and volume change with the displacement frequencies of the cylinder under different volume ratios and frequencies, which verified the accuracy of the mathematical models.
- The test volume-height curves were basically coincident, error of pressure repeatability was within 0.002 MPa, and its volume and pressure performance were stable by repeatability tests. The changes in pressure had no effect on the displacement output

characteristics and provided pressure for the pump inlet, which is beneficial for improving the service life of the pumps and the performance of the system.

- The designed VVPR would cut the volume and weight dramatically by approximately 98% and 94%, respectively, and could replace conventional open reservoirs in numerous applications. However, it also faced the challenge of heat generated by the closed hydraulic system.

Because this is the beginning of a research project on the VVPR, there is naturally much basic work that needs to be done. In addition, there are some limitations that need to be addressed to match the parameters of hydraulic systems, such as the natural frequency of the reservoir. Therefore, the improvement of this reservoir will continue on another matching closed-circuit hydraulic system to conduct dynamic tests in the time and frequency domains for further validation and intensive research.

ACKNOWLEDGEMENTS

The authors sincerely thank Professor Xiangdong Kong of Yanshan University for his critical discussions and reading during manuscript preparation.

AUTHORS' INFORMATION

Pei Wang, born in 1991, is currently a Ph.D. candidate at the *College of Mechanical Engineering, Yanshan University, China*. She received her M.S. degree in 2018 from *Yanshan University, China*. Her research interests include hydraulic system design, fluid transmission, and fluid control.

Jing Yao, born in 1978, is currently a professor at the *College of Mechanical Engineering, Yanshan University, China*. She received her Ph.D. degree in 2009 from *Yanshan University, China.* Her main research interests include hydraulic system control and lightweight hydraulic component design.

Baidong Feng, born in 1994, is currently a master degree candidate at the *College of Mechanical Engineering, Yanshan University, China*. He received his bachelor's degree in 2018 from *Binzhou University, China*. His research interests include hydraulic transmission and hydraulic control.

Mandi Li, born in 1993, is currently a PhD candidate at the *College of Mechanical Engineering, Yanshan University, China*. She received her M.S. degree in 2018 from the *Northeastern University*, China. Her research interests include hydraulic engineering, fluid transmission, fluid simulations, and experiments.

Dingyu Wang, born in 1998, is currently a PhD candidate at the *College of Mechanical Engineering, Yanshan University, China*. He received his bachelor's degree in 2020 from *Yanshan University, China*. His research interests include hydraulic system and machine design.

AUTHOR CONTRIBUTIONS

JY and PW were in charge of the whole trial; PW and ML wrote the manuscript; and BF and DW assisted with sampling and laboratory analyses. All authors have read and approved the fnal manuscript.

FUNDING

Supported by the National Key Research and Development Program of China (Grant No. 2018YFB2000700) and National Natural Science Foundation of China (Grant No. 51890811).

REFERENCES

1. Q J Gao. Adaptive control system of underground hydraulic supports. *Energy and Energy Conservation*, 2020(10): 135-136.
2. Y X Guo, L Zhang. Hybrid power drive system of large die forging hydraulic press. *Forging & Stamping Technology*, 2020, 45(10): 124-129.
3. Y E Wang, Y Liu, Y Ma, et al. Application of hydraulic steering system for heavy-duty mobile robot. *Hydraulics Pneumatics & Seals*, 2020, 40(9): 24-25.
4. Q F He, X H Chen, C J Yao, et al. Fault diagnosis expert system of construction machinery based on flowchart knowledge representation. *Machine Tool & Hydraulics*, 2019, 47(17): 216-219.
5. D J Yang, X B Jin, Z C Yang, et al. Optimization of sensor layout in aero-hydraulic system based on MINIP model. *Measurement & Control Technology*, 2020, 39(5): 49-53.
6. X D Kong, Q X Zhu, J Yao, et al. Reviews of lightweight development of hydraulic components and systems for high-level mobile equipment. *Journal of Yanshan University*, 2020, 44(3): 203–217.
7. L F Jiao, X H Lu. Improvement and reliability modeling analysis of aviation hydraulic oil tank sealing structure. *Lubrication Engineering*, 2015, 40(7): 115-120.
8. L Li, W S Sun, P Gao. Analysis and prevention of common faults about aircraft hydraulic oil Tank. *New Technology & New Process*, 2014(2): 122-124.
9. R J Liu, F Xu, X Cheng. Test selection and analysis of the hydraulic tank sealing ring of a Helicopter. *Helicopter Technique*, 2019(2): 33-36.
10. T Li, H L Yang, D B Han. Design and analysis of reservoir for civil aircraft hydraulic system. *Chinese Hydraulics & Pneumatics*, 2017(2): 101-106.
11. Y Zhang, Y M Wen, S. Wang. Research and experiment of multifunction aero hydraulic tank. *Chinese Hydraulics & Pneumatics*, 2018(11): 104-107.
12. X P OuYang, B Q Fan, H Y Yang, et al. A novel multi-objective optimization method for the pressurized reservoir in hydraulic

robotics. *Journal of Zhejiang University-Science A(Applied Physics & Engineering)*, 2016, 17(6): 454-467.

13. D X Zhao, T Jia, Y X Cui. Design and constant pressure characteristics of a ship-borne pressure tank. *Tsinghua Univ (Sci &. Technol)*, 2019, 59(4): 306-313.
14. G F Qi, J J Zhang, J G Sun. Miniaturization trend of the hydraulic fuel tank and the new trend of development. *Machine Tool & Hydraulics*, 2011, 39(24): 66–68, 104.
15. W J Xiao, H Hu, L J Wei. Hydraulic tank of a fighter: fault analysis and improvement. *Chinese Hydraulics & Pneumatics*, 2013(10): 58-61.(in Chinese)
16. T Zhang, X L Xiao. Troubleshooting of pipeline blockage for the aircraft tank booster system. *Hydraulics Pneumatics & Seals*, 2019, 39(6): 68-70.
17. D W Li, Z Z Zuo, Z Liu. Design improvement for hydraulic reservoir pressurization system of a certain type of aircraft. *Chinese Hydraulics & Pneumatics*, 2017(11): 105-108.
18. W B Qu, C W Han, B Z Feng, et al. Design and application of a closed tank in a pitch hydraulic system. *Chinese Hydraulics & Pneumatics*, 2016, (5): 74-77.
19. B Guo, G L Zhang, J G Zhang. The development of the underground scraper closed tank. *Chinese Hydraulics & Pneumatics*, 2014(1): 83-84.
20. W Y Kang. A kind of pressure hydraulic tank. CN201320558868.4 [P]. 2013-09-09.
21. Z J Feng, J Chen, S C Jiang, et al. Bladder hydraulic tank: CN1804408 [P]. 2006-07-19.
22. Q G Han, Y B Zhang, S Z Yang. Analysis of closed oil tank of interlocking control hydraulic system. *Metallurgical Equipment*, 2016(4): 78-80.
23. T J Li, L Y He, Z Wang, et al. Flexible variable volume vacuum tank: CN207225926U [P]. 2018-04-13.
24. Argo-Hytos. Tank Solutions [EB/OL]. https://www.argo-hytos.com/cn/products/tank-solutions.html, 2020.
25. Fuel Safe. Auxiliary fuel cell bladder tanks for use in airplanes, UAV, Helicopters [EB/OL]. https://fuelsafe.com/aircraft-fuel-bladders, 2020-10-01.

26. Tertle-Pac. Collapsible aircraft ferry tanks [EB/OL] https://www.turtlepac.com/products/collapsible-aircraft-ferry-tanks, 2020-10-01.

27. C Seguin. Variable volume reservoir: US6981523, 2006.

28. B Kim, S B Lee, J Lee. A comparison among Neo-Hookean model, Mooney-Rivlin model, and Ogden model for chloroprene rubber. *International Journal of Precision Engineering and Manufacturing*, 2012(13): 759-764.

29. C Bi, Z M Chen, L B Zhang, et al. Mathematical Model and Simulation Analysis of Hydraulic Bladder Accumulator. *Aerospace Manufaturing Technology*, 2017(2):11-15.

30. M Dong, X T Luan, J L Liang, et al. Dynamis Characteristics Analysis of Absorbing Pulsation for Bladder Accumulator. *Chinese Hydraulics & Pneumatics*, 2019(5):109-116.

Chapter

8

DYNAMIC MODELLING AND NATURAL CHARACTERISTIC ANALYSIS OF CYCLOID BALL TRANSMISSION USING LUMPED STIFFNESS METHOD

Peng Zhang, Bingbing Bao, and Meng Wang

Department of Mechanical Engineering, Anhui University of Technology, Maanshan 243000, China

ABSTRACT

The vibration of robot joint reducer is the main factor that causes vibration or motion error of robot system. To improve the dynamic precision of robot system, the cycloid ball transmission used in robot joint is selected as study object in this paper. An efficient dynamic modelling method is presented—lumped stiffness method. Based on lumped stiffness method, a translational–torsional coupling dynamics model of cycloid ball transmission system is established. Mesh stiffness variation excitation, damping of system are all intrinsically considered in the model. The dynamic equation of

Citation: (APA): Zhang, P., Bao, B., & Wang, M. (2017). Dynamic modelling and natural characteristic analysis of cycloid ball transmission using lumped stiffness method. Robotics and Biomimetics, 4(1), 1-10. (10 pages).

system is derived by means of relative displacement relationship among different components. Then, the natural frequencies and vibration modes of the derivative system are presented by solving the associated eigenvalue problem. Finally, the influence of the main structural parameters on the natural frequency of the system is analysed. The present research can provide a new idea for dynamic analysis of robot joint reducer and provide a more simplify dynamic modelling method for robot system with joint reducer.

Keywords: Lumped stifness method, Robot joint reducer, Natural frequencies, Vibration modes

INTRODUCTION

The main inducement of vibration of high-speed robot is robot joint reducer, and therefore, the dynamic research for robot joint reducer is necessary. At present, domestic and overseas scholars have made many deeply research on cycloid ball planetary transmission, including structure principle [1, 2], engagement principle [3], mechanical property [4, 5], and transmission accuracy [6, 7]. However, the dynamic analysis of it has rarely been reported. This paper effectively establishes a simple dynamic model of cycloid ball planetary transmission, which matches with engineering practice. After that, the characteristics of cycloid ball planetary transmission are analysed, and some improvement measures are presented with the purpose of reducing vibration and providing new ideas for robot dynamic analysis.

For the moment, the dynamic models of planetary gear mainly include purely rotational model [8, 9] and translational–torsional coupling model [10, 11]. In purely rotational model, the component's torsional degree of freedom is only considered. The model is simple because there are few factors are considered. Translational–torsional coupling model also includes the component's translational degrees of freedom. Compared with purely rotational model, translational–torsional coupling model is more complex, and solving is more difficult. Therefore, it is usually used in theoretical analysis. The result of Ref. [12] shows that when the ratio of support stiffness to mesh stiffness is greater than 10, the simplified purely rotational model and translational–torsional coupling model have some equivalence in the inherent characteristics. For cycloid ball planetary transmission, the translational–torsional coupling model is established, and the inherent characteristic is analysed in Refs. [13, 14]. But the modelling methods are too complex and difficult, especially for a large degree of freedom dynamic system.

In view of that, this paper uses the effectively and simple modelling method—lumped stiffness method to establish the translational–torsional coupling model of cycloid ball planetary transmission. Then, the natural frequencies and vibration modes are revealed by solving dynamic equations of system with the purpose of providing guidance for system design.

LUMPED STIFFNESS MODELLING

Structure

The structure of cycloid ball planetary transmission is shown in Fig. 1. Cycloid ball engagement pairs consist of hypocycloid groove in the left end face of planetary disc, epicycloid groove in the right end face of centre disc, and balls between two discs.

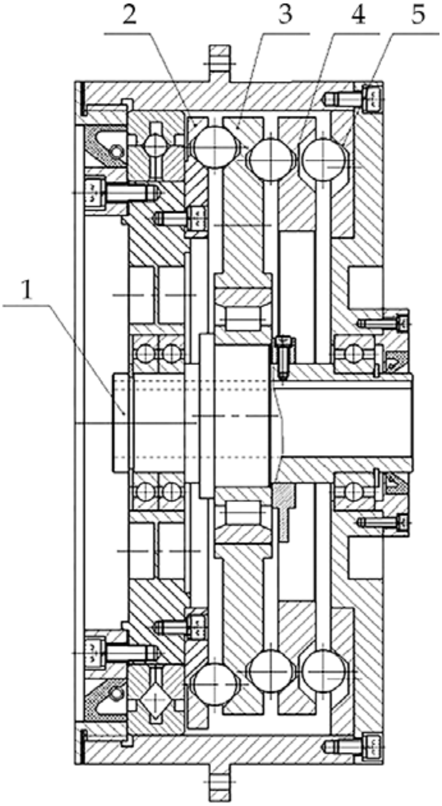


Figure 1. The structure of cycloid ball planetary transmission. 1. Input shaft 2. Centre disc 3. Planetary disc 4. Cross-disc 5. End cover disc.

This paper uses cross-ball equal-speed mechanism as output structure for the requirements of robot joint. Cross-ball equal-speed mechanism is made up with the horizontal taper grooves in the right end face of planetary disc, the horizontal taper grooves in the left end face of cross-disc, the taper grooves in the left end face of cross-disc, the taper grooves in the right end face of end cover disc, and balls among three discs. In this paper, cross-ball equal-speed mechanism is proposed, and centre disc is treated as output disc.

Lumped Stiffness Model

To simplify the dynamic model, an efficient dynamic modelling method—lumped stiffness method is proposed based on the lumped mass method. The basic thought of lumped stiffness method is as follows: first, the total meshing component force along axis direction will be obtained through mechanical analysis; second, the maximum deformation of meshing point is considered as global deformation, and the component of global deformation along axis direction can be presented; finally, the ratio of total meshing component force to global component deformation along axis direction will be obtained. Obviously, the ratio is lumped stiffness. Compared with the traditional modelling method, the advantages of lumped stiffness method are as follows: nonlinear stiffness, time-varying curvature, and time-varying load have been integrated into the lumped stiffness model and not directly reflected in the dynamic model; the dynamic model will be established and solved easily. The lumped stiffness model of cycloid ball meshing pair and cross-ball meshing pair is, respectively, solved using lumped stiffness method.

The mechanical model of cycloid ball engagement pairs is shown in Fig. 2. Reference [4] shows that the total meshing force of y axis is zero, but the total meshing force of x axis exists. Therefore, only the lumped stiffness model of x axis is needed.

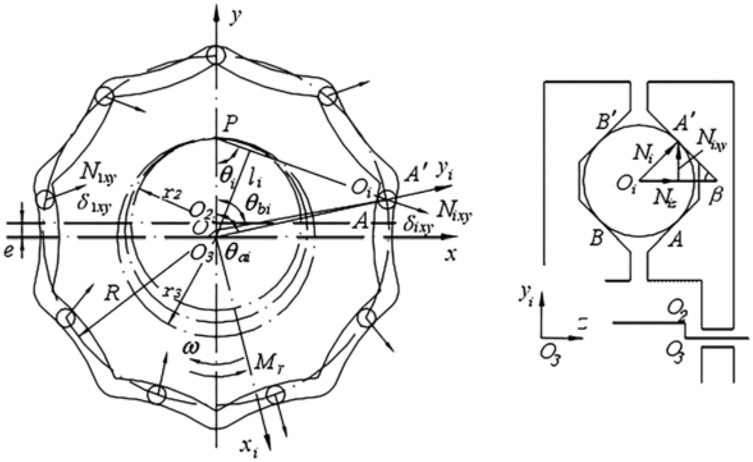


Figure 2. Mechanical model of cycloid ball meshing pairs.

According to the mechanical model, the stiffness model of cycloid ball meshing pairs can be established as shown in Fig. 3. Figure 3a shows the traditional stiffness model of cycloid ball meshing pairs, (b) shows the lumped stiffness model of cycloid ball meshing pairs. Obviously, the distribution of meshing force is complex. If the meshing forces are not effectively synthesized in modelling, the complexity of modelling will increase. Hence, lumped stiffness model is more convenient and simple compared to traditional stiffness model.

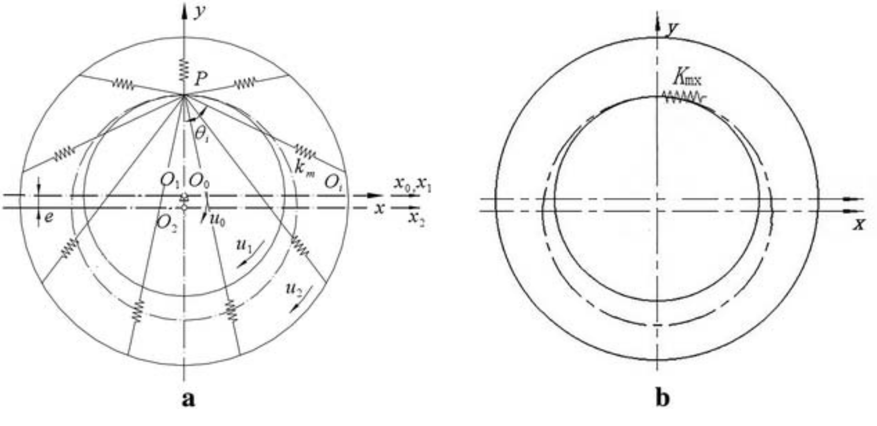


Figure 3. Stiffness model of cycloid ball meshing pairs. **a** Traditional stiffness model. **b** Lumped stiffness model.

According to the thought of lumped stiffness method, the lumped stiffness of *x* axis is

$$\begin{aligned} K_{mx} &= \frac{\left(\sum_{i=1}^{Z_m} N_i \cos\beta\right)|_x}{(\delta_1, \delta_2, \ldots, \delta_{Z_m})_{\max} \cos\beta|_x} \\ &= \frac{\left(\overrightarrow{k_m\delta_1} + \overrightarrow{k_m\delta_2} + \cdots + \overrightarrow{k_m\delta_{Z_m}}\right)|_x}{(\delta_1, \delta_2, \ldots, \delta_{Z_m})_{\max}|_x} \\ &= \frac{\sum_{i=1}^{Z_m} k_m\delta_m \sin^2\theta_i}{a\delta_{\mathrm{m}}} = \frac{1}{a}k_m \sum_{i=1}^{Z_m} \sin^2\theta_i \end{aligned} \tag{1}$$

where θ_i represents the angle between the normal line of the *i* meshing point and the *y* axis. δ_m is the maximum deformation in theory, which corresponding to the special location. δ_{imax} is the maximum deformation at any time during the operation. k_m is the meshing stiffness of single cycloid ball meshing pair. Z_m is the number of ball. β is the half angle of cycloid groove. a is deformation coefficient, $a = a' \cdot \overline{\sin\theta_i}$, $a' = \overline{\delta_{i\max}}/\delta_m$, where $\overline{\sin\theta_i}$ is the average value of the corresponding change interval.

In addition, the torsional angular displacement of discs in cycloid ball meshing pair is generated by meshing displacement. More importantly, the direction of meshing displacement and meshing force are identical. Hence, for the convenience of calculation, the torsional angular displacement is substituted by torsional linear displacement along the direction of meshing force. The lumped torsional stiffness is substituted by lumped stiffness of *x* axis.

Figure 4 shows the mechanical model of cross-ball meshing pair. The mechanical property of cross-ball meshing pair is analysed in Ref. [15], and its results proposed that the taper grooves along radial direction undertake most of the load, and the taper grooves perpendicular to radial direction hardly undertake the load. In this paper, three taper grooves along the radial direction are arranged on cross-ball equal-speed mechanism with the purpose of improving the bearing capacity. Meanwhile, the taper grooves of side-by-side arranged undertake equivalent load that is $Q_{1xy} = Q_{2xy} = Q_{3xy}$, $Q_{5xy} = Q_{6xy} = Q_{7xy}$

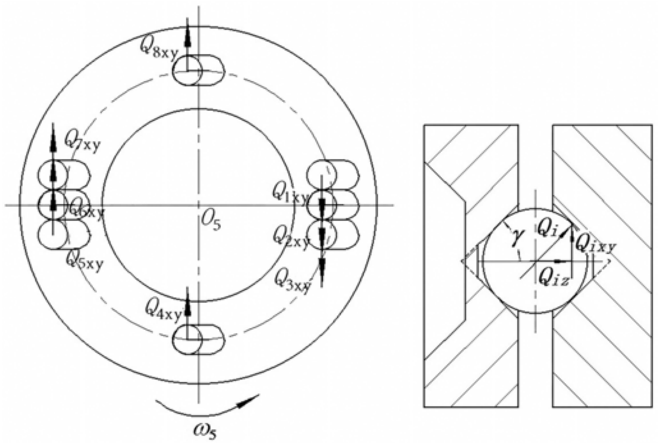


Figure 4. Mechanical model of cross-ball meshing pair.

The force analysis shows that the total meshing force along *x* axis and *y* axis of cross-ball meshing pair is zero. Hence, there is no need to obtain corresponding lumped stiffness except lumped torsional stiffness. The solution thought is shown as follows: the maximum torsional angular displacement is divided by resultant moment. The stiffness model of cross-ball meshing pair is shown in Fig. 5. (a) shows the traditional stiffness model, and (b) shows the lumped stiffness model. Similarly, the distribution of meshing force is complex. The meshing forces are effectively synthesized in the lumped stiffness model, which is beneficial for dynamic modelling.

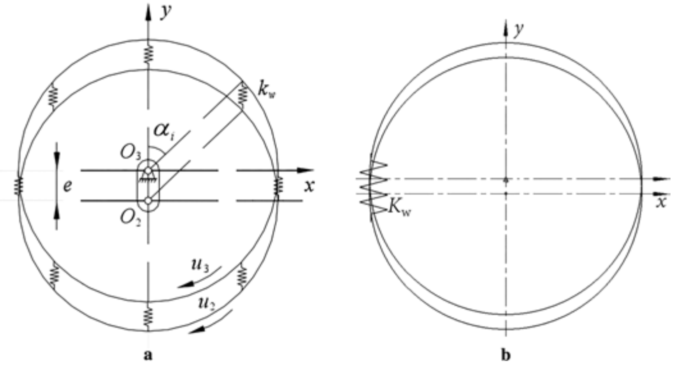


Figure 5. Stiffness model of cross-ball meshing pair. **a** Traditional stiffness model. **b** Lumped stiffness model.

Specifically, the lumped torsional stiffness of cross-ball equal-speed mechanism is

$$\begin{aligned} K_w &= \frac{\sum M}{\Delta\alpha_{i\max}} = \frac{R\sum\limits_{i=1}^{Z_w} Q_i\cos\beta}{R\delta_{i\max}\cos\beta} \\ &= \frac{3\left[Q_{2xy}\left(R-\frac{e}{2}\cos\phi\right)+Q_{5xy}\left(R+\frac{e}{2}\cos\phi\right)\right]}{\left(R+\frac{e}{2}\cos\phi\right)\delta_{6xy}} \\ &= 6k_w \end{aligned} \qquad (2)$$

where R is the distribution circle radius of taper grooves; ϕ is the angle between the straight lines formed by the components and the cross guide rod in the equivalent mechanism of cross-ball equal-speed mechanism; e is eccentric distance of input shaft.

TRANSLATIONAL–TORSIONAL COUPLING MODEL

Dynamic Model

To press close to the physical reality and avoid the complexity of mathematical treatment, the following simplifications and assumptions are made in the dynamic modelling:

- Balls are regarded as elastic element because of the small quality;
- Balls are pure rolling in the grooves, and the influence of friction force is ignored;
- The backlash can be eliminated by clearance screw mechanism, and the influence of backlash nonlinearity is ignored;
- The cross-disc is in a floating state, and the effects of cross-disc are not counted.

For the convenience of description of the relationship between the components of cycloid ball planetary transmission, this paper adopts a servo reference system of eccentric shaft (input shaft). Thus, the geometric centre of the input shaft is the coordinate origin. The coordinate system rotates at the speed of input shaft. According to the force analysis, a planar problem is considered where input shaft, centre disc, and planetary disc have two degrees of freedom: one translational around its own axis and one rotational along the *x* axis. End cover disc has one translational degree of freedom. In total, the model has seven degrees of freedom. Figure 6 shows the translational—torsional coupling model of cycloid ball planetary transmission. The sequence number of the components in Fig. 6 is consistent with the sequence number in Fig. 1.

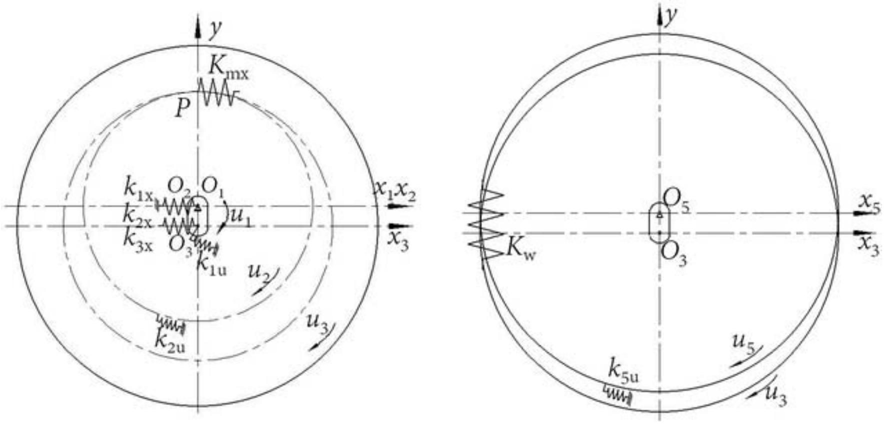


Figure 6. Translational–torsional coupling model of cycloid ball planetary transmission.

Relative Displacement between Components and Dynamic Equations

The relative displacement between components is clear because the system has fewer components. The specific contents are shown as follows:

- Relative displacement between centre disc and planetary disc

$$\delta_{23} = x_2 - x_3 + u_2 - u_3 \tag{3}$$

- Relative displacement between end cover disc and planetary disc

$$\delta_{53} = u_5 - u_3 \tag{4}$$

- Relative displacement between planetary disc and input shaft

$$\delta_{13} = x_1 - x_3 - u_1 \tag{5}$$

The differential equation of system can be obtained using Newton's second law:

$$\begin{cases} m_1(\ddot{x}_1 - \omega_1^2 x_1) + k_{3x}\delta_{13} + c_{3x}\dot{\delta}_{13} + k_{1x}x_1 + c_{1x}\dot{x}_1 = 0 \\ \frac{J_1}{e^2}\ddot{u}_1 - k_{3x}\delta_{13x} - c_{3x}\dot{\delta}_{13x} + k_{1u}u_1 + c_{1u}\dot{u}_1 = \frac{T_i}{e} \\ m_2(\ddot{x}_3 - \omega_1^2 x_2) + K_{mx}\delta_{23} + C_{mx}\dot{\delta}_{23} + k_{2x}x_2 + c_{2x}\dot{x}_2 = 0 \\ \frac{J_2}{r_2^2}\ddot{u}_2 + K_{mx}\delta_{23} + C_{mx}\dot{\delta}_{23} + k_{2u}u_2 + c_{2u}\dot{u}_2 = -\frac{T_o}{r_2} \\ m_3(\ddot{x}_3 - \omega_1^2 x_3) - K_{mx}\delta_{23} - C_{mx}\dot{\delta}_{23} - k_{3x}\delta_{13x} - c_{3x}\dot{\delta}_{13x} = 0 \\ \frac{J_3}{r_3^2}\ddot{u}_3 - K_{mx}\delta_{23} - C_{mx}\dot{\delta}_{23} - K_w\delta_{53} - C_w\dot{\delta}_{53} = 0 \\ \frac{J_5}{r_5^2}\ddot{u}_5 + K_w\delta_{53} + C_w\dot{\delta}_{53} + k_{5u}u_5 + c_{5u}\dot{u}_5 = 0 \end{cases} \tag{6}$$

where J_i is the moment of inertia of component i (i = 1, 2, 3, 5); m_i is the mass of component i (i = 1, 2, 3); r_i is the pitch radius of component i (i = 1, 2, 3), $r_3 = r_5$; c_{ix} is the lateral brace damping coefficient of component i (i = 1, 2, 3); c_{iu} is the torsion brace damping coefficient of component i(i = 1, 2, 5); C_{mx} is the meshing damping coefficient of cycloid ball meshing pair; C_w is the meshing damping coefficient of cross-ball meshing pair; T_i is the input torque of input shaft; T_o is the input torque of output disc(centre disc).

The formula (6) is arranged in matrix form:

$$\boldsymbol{M}\ddot{\boldsymbol{X}} + (\boldsymbol{C_b} + \boldsymbol{C_m})\dot{\boldsymbol{X}} + (\boldsymbol{K_b} + \boldsymbol{K_m} + \boldsymbol{K_\omega})\boldsymbol{X} = \boldsymbol{F} \tag{7}$$

$$\boldsymbol{X} = [x_1, u_1, x_2, u_2, x_3, u_3, u_5]^{\mathrm{T}}$$

$$\boldsymbol{M} = \mathrm{diag}\left(m_1, J_1/e^2, m_2, J_2/r_2^2, m_3, J_3/r_3^2, J_5/r_5^2\right)$$

$$\boldsymbol{F} = [0, T_{\mathrm{i}}/e, 0, -T_o/r_2, 0, 0, 0]^{\mathrm{T}}$$

$$\boldsymbol{K_b} = \mathrm{diag}(k_{1x}, k_{1u}, k_{2x}, k_{2u}, k_{3x}, 0, k_{5u})$$

$$\boldsymbol{C_b} = \mathrm{diag}(c_{1x}, c_{1u}, c_{2x}, c_{2u}, c_{3x}, 0, c_{5u})$$

$$\boldsymbol{K_\omega} = \mathrm{diag}\left(-m_1\omega_1^2, 0, -m_2\omega_1^2, 0, -m_3\omega_1^2, 0, 0\right)$$

$$\boldsymbol{K_m} = \left[\begin{array}{ccccccc} k_{2x} & -k_{2x} & 0 & 0 & -k_{2x} & 0 & 0 \\ & k_{2x} & 0 & 0 & k_{2x} & 0 & 0 \\ & & K_{mx} & K_{mx} & -K_{mx} & -K_{mx} & 0 \\ & & & K_{mx} & -K_{mx} & -K_{mx} & 0 \\ & & & & K_{mx} & K_{mx} & 0 \\ & & & & & K_{mx}+K_w & -K_w \\ \text{sym} & & & & & & K_w \end{array}\right]$$

$$\boldsymbol{C_m} = \left[\begin{array}{ccccccc} c_{2x} & -c_{2x} & 0 & 0 & -c_{2x} & 0 & 0 \\ & c_{2x} & 0 & 0 & c_{2x} & 0 & 0 \\ & & C_{mx} & C_{mx} & -C_{mx} & -C_{mx} & 0 \\ & & & C_{mx} & -C_{mx} & -C_{mx} & 0 \\ & & & & C_{mx} & C_{mx} & 0 \\ & & & & & C_{mx}+C_w & -C_w \\ \text{sym} & & & & & & C_w \end{array}\right]$$

where X is generalized coordinate array; M is generalized mass matrix; F is external excitation array; K_b, K_m, K_ω are support stiffness matrix, mesh stiffness matrix, and centripetal stiffness matrix; C_b, Cm are support damping matrix and meshing damping matrix. The elements C_{mx} and C_w in the matrix C_m have the following form:

$$C_{mx} = \frac{1}{a} c_m \sum_{i=1}^{Z_m} \sin^2 \theta_i \tag{8}$$

$$C_w = 6c_w \tag{9}$$

K_m is time-varying matrix because the lumped stiffness K_{mx} is a time-varying element with the parameter θ_i. To solve the problem conveniently, the θ_i is converted to input shaft angle and the higher-order term is omitted.

$$\begin{aligned} K_{mx} &= \frac{1}{a} k_m \sum_{i=1}^{Z_m} \sin^2 \theta_i \\ &= \frac{k_m Z_m}{2a} + \frac{k_m Z_m}{2a}\left(1 - K^{-2}\right) K^{Z_m} \cos \omega_1 t \end{aligned} \tag{10}$$

where K is short width coefficient of cycloid ball planetary transmission.

The time-varying element of formula (10) is omitted. After the time-invariant K_{mx} is substituted into the mesh stiffness matrix, the dynamic equation of derivative system can be obtained:

$$\boldsymbol{M\ddot{X}} + (\boldsymbol{C_b} + \boldsymbol{C'_m})\boldsymbol{\dot{X}} + (\boldsymbol{K_b} + \boldsymbol{K'_m} + \boldsymbol{K_\omega})\boldsymbol{X} = \boldsymbol{F} \tag{11}$$

In addition, the mechanical model and stiffness modelling method in this paper are different from Refs. [13, 14], but the mathematical model of cycloid ball planetary transmission is identical.

NATURAL CHARACTERISTIC ANALYSIS

Natural Frequency and Principal Mode

The natural characteristic of cycloid ball planetary transmission can be presented by solving the eigenvalue problem of derivative system. The eigenvalue problem of formula (11) is

$$(\boldsymbol{K}_b + \boldsymbol{K}'_m + \boldsymbol{K}_\omega)\varphi_i - \omega_i^2 \boldsymbol{M}\varphi_i = 0 \tag{12}$$

where ω_i is the *i* order natural circular frequency of system; ϕ_i is the *i* order principal mode of system, $\varphi_i = \left[\varphi_{1x}^{(i)}, \varphi_{1u}^{(i)}, \varphi_{2x}^{(i)}, \varphi_{2u}^{(i)}, \varphi_{3x}^{(i)}, \varphi_{3u}^{(i)}, \varphi_{5u}^{(i)}\right]$

Without loss of generality, take the cycloid ball planetary transmission used in robot joint as an example, the dynamic characteristics are simulated and analysed. The cross ball equal-speed mechanism is arranged in front of the cycloid ball meshing pair in the prototype. In other words, end the cover disc is fixed and the central disc is used as output component. The speed of input shaft is 1000 *r*/min; the meshing stiffness of single cycloid ball meshing pair is 2.87×10^7 N/m; the meshing stiffness of single cross-ball meshing pair is 4.44×10^7 N/m, and the deformation coefficient *a* is 0.9978. Other basic parameters are shown in Table 1.

Table 1. Essential parameters of cycloid ball planetary transmission

Essential parameter	**Input shaft**	**Centre disc**	**Planetary disc**	**End cover disc**
Number of teeth *Z*		38	40	
Mass/kg	0.854	0.924	2.185	1.447
Moment of inertia J_i/(kg m^2)	2.69×10^{-4}	6.69×10^{-3}	1.44×10^{-2}	1.19×10^{-2}
Pitch radius r_i/m	2.5×10^{-3}	4.75×10^{-2}	5×10^{-2}	5×10^{-2}
Radial stiffness k_{ix}/(N m^{-1})	5.85×10^8	5.85×10^8	5.85×10^8	
Torsional stiffness k_{iu}/(N m^{-1})	0	0		1×10^9

By solving the formula (12), the natural frequencies and the principal modes of the system are obtained as shown in Table 2. All natural frequencies are single. The first-order natural frequency is 0, which represents the rigid motion of system. The vibration modes corresponding to the other six-order natural frequencies are both translational vibration and torsional vibration. Furthermore, the approximate results of natural frequencies and principal modes of cycloid ball planetary transmission can be obtained when prototype data in this paper are plugged into the dynamic model of literature [13].

Table 2. Natural frequencies and principal modes of cycloid ball planetary transmission

Natural frequency f_i/(Hz)		**0**	**708.6**	**1309.8**	**2552.8**	**2705.3**	**5927.8**	**6614.9**
Principal mode ϕ_i	$\phi_{1x}^{(i)}$	0	− 0.0479	0.1962	− 0.2455	− 0.1651	− 0.8094	− 0.6225
	$\phi_{1u}^{(i)}$	− 1	− 0.2214	0.2464	− 0.0562	− 0.0312	0.0562	− 0.052
	$\phi_{2x}^{(i)}$	0	0.0432	− 0.1024	− 0.3741	− 0.3328	0.6434	0.6365
	$\phi_{2u}^{(i)}$	1	− 2.8938	1.85	1.1826	0.8576	0.7569	0.8729
	$\phi_{3x}^{(i)}$	1	− 0.4340	2.3444	− 2.7125	− 1.7785	0.0847	− 2.2035
	$\phi_{3u}^{(i)}$	0	− 0.2674	− 0.2689	− 0.0445	− 0.1416	− 0.0581	− 0.0665
	$\phi_{5u}^{(i)}$	0	− 0.0608	− 0.0759	− 0.2846	0.3459	0.0029	0.0025

Parametric Influence of Natural Frequency

It is necessary to analyse the change regulation of natural frequency relative to parameters of system with the purpose of avoiding vibration. In this paper, based on translational–torsional coupling model, the natural frequency curves of each order are obtained by calculating eigenvalue problem with consideration of main parameters, as shown in Fig. 7, 8, 9 and 10.

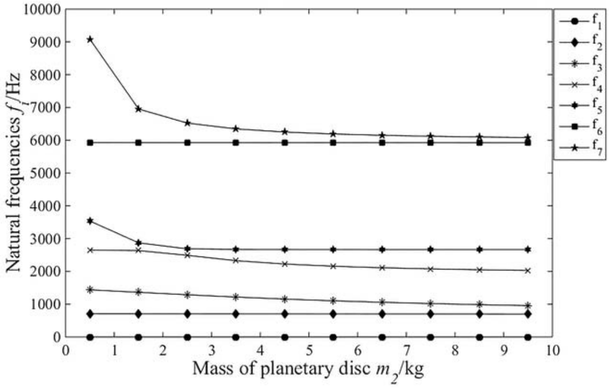


Figure 7. The influence of planetary disc mass on the natural frequencies.

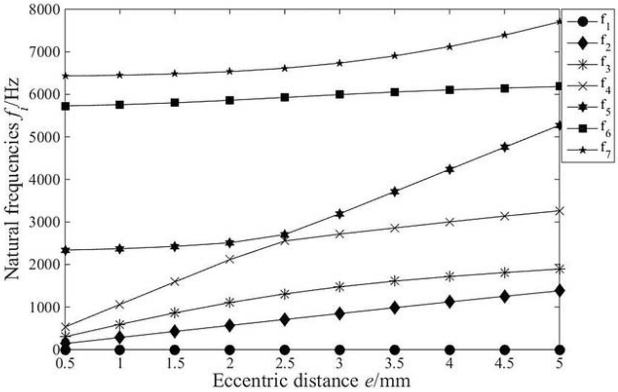


Figure 8. The influence of eccentric distance on the natural frequencies.

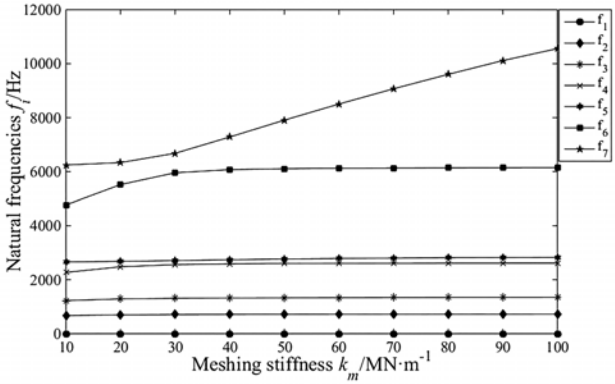


Figure 9. The influence of meshing stiffness on the natural frequencies.

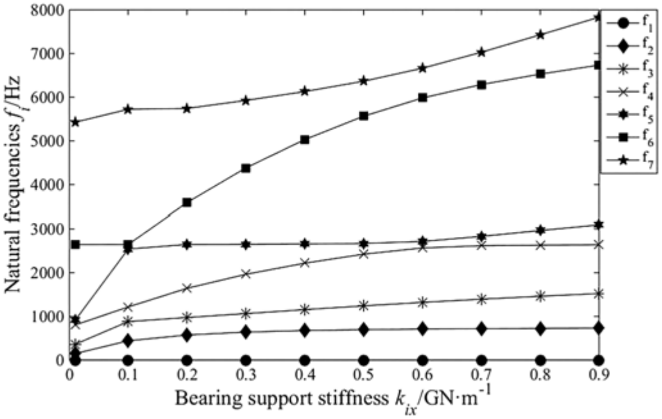


Figure 10. The influence of bearing support stiffness on the natural frequencies.

As shown in Fig. 7, when the mass of the planetary disc is less than 2.5 kg, the fifth- and seventh-order natural frequencies decrease sharply, and other orders are weakly affected. When the mass of the planetary disc is bigger than 2.5 kg, all the natural frequencies have barely budged.

As shown in Fig. 8, all the natural frequencies increase gradually with the increase in the eccentric, except for the first order. When the eccentricity is 2.5 mm, mode transition appears between the fourth- and fifth-order natural frequencies. At the point of mode transition, the subtle change in parameters will lead to drastic change in natural frequencies. Hence, the sensitive points of parameters should be avoided in the design to avoid drastic change in transmission characteristics.

As shown in Fig. 9, the mesh stiffness has little effect on the first 5 orders natural frequencies of system. The sixth-order and seventh-order natural frequencies increase with the increase in meshing stiffness. When meshing stiffness increases to 3×10^7 N/m, the sixth order natural frequency remains constant, but the seventh order natural frequency rises dramatically.

As shown in Fig. 10, the bearing support stiffness has certain influence on the natural frequencies, except for the first order. When the bearing support stiffness is less than 1×10^8 N/m, the natural frequencies increase obviously with the increase in bearing support stiffness, especially the fifth order; when the bearing support stiffness is greater than 1×10^8 N/m, the sixth- and seventh-order natural frequencies increase significantly. Modal transition phenomenon occurs in the fifth- and sixth-order natural frequencies when bearing support stiffness is 1×10^8 N/m, which should be avoided in the optimization design of system.

CONCLUSION

- To improve the motion accuracy of robot system, the cycloid steel ball planetary transmission used in robot joint is selected as research object. An efficient dynamic modelling method is presented—lumped stiffness method. A translational–torsional coupling model is modelling, and the natural characteristics of system are revealed.
- All natural frequencies of system are single. The first-order natural frequency is 0, which represents the rigid motion of system. The vibration modes corresponding to the other six-order natural frequencies are both translational vibration and torsional vibration.
- The number of eccentricity distance and bearing support stiffness may lead to the modal transition phenomenon. The sensitive points of parameters should be avoided as far as possible in the optimization design of system.

AUTHORS' CONTRIBUTIONS

PZ created lumped stiffness method. BB established the dynamic model, made the natural characteristics analysis, and wrote the manuscript. PZ supervised the research. Both authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

This research study was supported financially by National Natural Science Foundation of China (Grant No. 51405003). We acknowledge and thank their support.

REFERENCES

1. Terada H, Makino H, Imase K. Fundamental analysis of cycloid ball reducer (3rd report). JSPE. 1995;61(12):1075–9.
2. Terada H, Makino H, Imase K. Fundamental analysis of cycloid ball reducer (4th report). JSPE. 1997;63(6):834–8.
3. An ZiJun Q, Zhigang ZR. Research on tooth shape synthesis of cycloid ball transmission. J Mech Eng. 1996;32(5):41–6.
4. Peng Z, Zijun A, Zuomei Y. Research on nonlinear mechanical properties for engagement pair of cycloid ball planetary transmission. Eng Mech. 2010;27(3):186–92.
5. ZiJun A. Force and strength analysis on end face engagement cycloid steel ball planetary transmission. J Mech Transm. 2003;27(4):29–31.
6. Zijun A, Ruixue H. Parameter analysis of tooth profile error of cycloid steel ball planetary transmission. J Mech Transm. 2007;31(2):63–5.
7. Qingkun X, Junping Z. Error analysis cycloid ball planetary reducer based on the Monte Carlo. J Northwest A&F Univ. 2008;36(7):224–8.
8. Guo Y, Parker RG. Purely rotational model and vibration modes of compound planetary gears. Mech Mach Theory. 2010;45:365–77.
9. Shiyu W, Ce Z. Natural mode analysis of planetary gear trains. Chin Mech Eng. 2005;16(16):1461–5.
10. Yimin S, Jun Z. Inherent characteristics of 3K-II spur planetary gear trains. J Mech Eng. 2009;45(7):23–8.
11. Sun T, HaiYan H. Nonlinear dynamics of a planetary gear system with multiple clearances. Mech Mach Theory. 2003;38:1371–90.
12. Kahraman A. Free torsional vibration characteristics of compound planetary gear sets. Mech Mach Theory. 2001;36:953–71.
13. Peng Z, Zijun A. Dynamics model and natural characteristics of cycloid ball planetary transmission. Chin Mech Eng. 2014;25(2):157–62.
14. Ronggang Y, Zijun A. Analysis of free vibration of cycloid ball planetary transmission. Chin Mech Eng. 2016;27(14):1883–91.
15. Zhang P, Bao B. Mechanical property analysis and finite element simulation of cross steel equal-speed mechanism. Chongqing Proc Int Conf Power Transm. 2016;2016:87–92.

Chapter

9

MODELLING OF FLOWSLIDES AND DEBRIS AVALANCHES IN NATURAL AND ENGINEERED SLOPES: A REVIEW

Sabatino Cuomo

Geotechnical Engineering Group (GEG), University of Salerno, Via Giovanni Paolo II, 132 84084 Fisciano, Italy

BACKGROUND

The landslides of the flow-type are dangerous and also challenging to study. A wide literature has been investigating the principal mechanisms governing each stage in which these phenomena can be ideally subdivided: failure, post-failure and propagation. However, holistic contributions and general overviews are very rare. In addition, a number of numerical methods have been issued and validated so that new chances exist to efficiently model those threats. The paper focuses on two classes of rainfall-induced landslides of the flow-type, namely debris flows and debris avalanches. The principal

Citation: (APA): Cuomo, S. (2020). Modelling of flowslides and debris avalanches in natural and engineered slopes: a review. Geoenvironmental Disasters, 7(1), 1-25. (25 pages).

numerical methods are reviewed for modelling the landslide initiation and propagation and are later used for analyzing a series of benchmark slopes and real case histories which are successfully simulated.

Results

The rainfall from ground surface and water spring from the bedrock are key factors for slope instability. Pore water pressure plays a relevant role also during the propagation stage. The entrainment of further material makes the propagation patterns complex due to lateral spreading and slow-down of the front of flows. It is shown that the used models are capable to provide useful indications even for combined channelized and unchannelized flows.

Conclusions

Notwithstanding the complexity of flow-like landslides and the related challenges in modelling, the understanding and forecasting of such natural hazards is achievable with a satisfactory confidence. Among the key factors, rainfall, pore water pressure and bed entrainment deserves a special attention. Further improvements are expectable as the numerical models are becoming more efficient. Thus, more accurate descriptions of local effects will be possible and also additional mechanisms will be eventually analyzed.

Keywords: Rainfall, Landslide, Flow, Modelling, Countermeasure

INTRODUCTION

The geomechanical modelling of hillslope instability phenomena has been posing challenges to scientists for many decades. Indeed, most of the difficulties arise from the significant kinematic differences between the different stages of a landslide namely failure, post-failure and propagation.

Hillslopes generally undergo small deformations in the so-called pre-failure stage. The failure stage (Cuomo, **2006**), in turn, may consist in the formation of a continuous shear surface through the entire soil mass (Leroueil, **2001**) where large soil deformations mainly concentrate and it is usually referred as “localized” failure. In some cases, plastic strains can affect large amount of soil originating a so-called “diffuse” failure (Darve and Laoufa, **2000**; Pastor et al., **2004**). In both cases, the failure stage leads to large displacements.

The post-failure stage is characterized by the rapid generation of large plastic strains and the consequent sudden acceleration of the failed soil mass (Hungr, **2004**) and it discriminates among different types of phenomena (Cascini et al., **2010**), i.e. slide, slides turning into flows and flowslides. A slide occurs when limit equilibrium condition is gradually reached along a shear zone so that unbalance between driving and resisting forces is moderate and the unstable mass does not accelerate abruptly. The transition from a slide to a flow is typically caused by cascading effects of local failure and variation in slope geometry. The initial stress state is changed abruptly and no chance exists for the slope to be stable anymore. While the previous two categories are independent on the soil constitutive behaviour, flowslides are related to static liquefaction (Sladen et al., 1985; Chu et al., 2003) or soil mechanical instability phenomena (Darve and Laouafa, 2000), which are both even challenging to be modelled. It is important noting that large acceleration of the failed mass are typical of "flows" and "flowslides", as labelled later on.

The propagation stage includes the movement of the failed mass from the source to the deposition area, where a new equilibrium configuration is possible and depends on both the amount of moving material and slope geometry. In the case of slides, the failed mass experiences displacements of one or two orders of magnitude lower than the landslide source dimensions. Conversely, for flows and flowslides the run-out distances are up to two orders of magnitude higher than the length of the landslide source (Cascini et al., 2011a, 2011b, 2016, 2019).

These mentioned differences are even more exacerbated in the case of the so-called flow-like landslides, which in most of the cases originate from shallow landslides. Two categories deserve special attention: debris flows and debris avalanches. Debris Flows (DF) propagate in V-shaped channels, where large amount of water is available during heavy rainstorms so that the propagating mass may fluidize before stopping (Cascini et al., 2014). Relevant examples of DFs are available from British Columbia (Canada), Cina, France, Hong Kong, Japan, Oregon (USA) and Switzerland (Braun et al., 2017, 2018;Iverson, 1997; Pastor et al., 2007a, b; Crosta et al., 2009; Hungr and McDougall, 2009; Quan Luna et al., 2012). It is worth noting that channelised landslides can be classified as 'flowslides' (Hungr et al., 2001) when liquefaction occurs in the source areas; otherwise, they can be simply referred to as 'debris flows' (Hungr et al., **2001**). Debris Avalanche (DA) is defined as "very rapid to extremely rapid shallow flow of partially or fully saturated debris on a steep slope, without confinement in an established channel" (Hungr et al., **2001**). Avalanche formation is mostly related to bed entrainment (Cascini et al., **2013a**, **b**; Cuomo et al., **2014**). As an example of DA, the 1999 Nomash River debris avalanche (Vancouver Island, British Columbia, Canada) mobilized a volume of 3×10^5 m^3 at the source, whereas the erosion processes yielded nearly the same volume, with an average erosion depth of 8 m measured along 25° to 35° steep slopes (Hungr and Evans, **2004**; Hungr et al., **2005**).

To overcome the numerical difficulties of modelling the soil displacements across different orders of magnitude, the failure analysis of hillslope is generally treated separately from the propagation stage with different numerical methods. While this twofold approach allows the solution of relevant technical problems, it can avoid the full understanding of the instability mechanism as a whole and, some time, can produce inaccurate results. So, a number of emerging methods have been proposed.

This paper focuses on shallow soil deposits along steep slopes, rainfall-induced instability and unsaturated coarse-grained soils. In doing that, ideal slopes, laboratory experiments (such as centrifuge tests) and real case histories will be examined. This is because evidences from laboratory and field are both fundamental. Firstly, the mechanisms of triggering, slide -flow transition and propagation will be reviewed. Then, the numerical models will be discussed. The aim of the paper is to provide a general overview of the current potentialities for modelling such challenging phenomena. Related to that, it is chosen to subdivide the numerical results in two categories: natural and artificial slopes. While such distinction is not relevant from a mechanical viewpoint, it makes sense if one thinks that mitigation structures and design procedures need often if not always the support of numerical analyses. Most of the conceptual concepts of this paper have been previously reviewed. The same applies also to the numerical models. What is lacking in the literature is an overall discussion of both the issues in a single paper. Of course, only some types of landslides have been taken into consideration. Also a limited set of models have been used. However, the choice of the challenging category of flow-like landslides allows exploring at once the failure, post-failure and propagation stages. Correspondingly, very different mechanisms are analyzed. The novelty of this work is to combine different concepts and rielaborate previous numerical results in a more general framework. For the sake of generality, both reduced-sclae experimental tests, real case histories and idealized slope schemes have been considered in the paper.

The work is organized as follows: a background about the fundamental mechanisms is firstly provided; the numerical methods are presented; then, applications are presented for both natural slopes and engineered slopes.

BACKGROUND

Triggering Mechanisms

The failure onset induced by rainfall is strictly related to the increase of pore water pressures and the consequent reduction of mean effective stresses (Anderson and Sitar, 1995; Alonso et al., 1996; Iverson et al., 1997). Within shallow soil deposits (Fig. 1), the increase of pore water pressures can be generated by rainfall that directly infiltrates the slope surface (Tsaparas et al., 2002; Futai et al., 2004) and propagate in depth through groundwater flow patterns related to the stratigraphical setting of the slope (Ng and Shi, 1998). Sometimes pore water pressure regime is also affected by the hydrogeological features of the underlying bedrock (Johnson and Sitar, 1990; Montgomery et al., 1997; Matsushi et al., 2006; Cascini et al., 2008) which can impose severe hydraulic boundary conditions such as water springs at the bottom of the soil deposits.

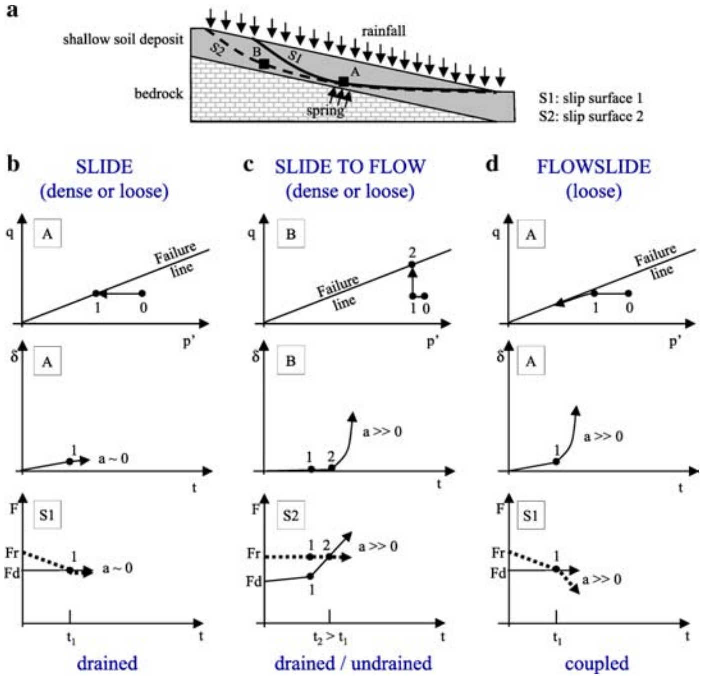


Figure 1. Slope scheme (**a**) and stress paths, displacements and F (Forces as F_r: resisting forces, and F_d: driving forces) typical of rainfall-induced (**b**) slide, (**c**) slide to flow, and (**d**) flowslide triggered in shallow landslides inside colluvial hollows (Cascini et al., **2010**).

For the so-called “slides”, the soil mechanical behaviour is controlled, in drained conditions, by the hydrologic response up to the failure onset. Resisting force (F_r) gradually decreases down to the value of the driving force (F_d) along a slip surface where the spring from the bedrock is located (later named spring zone, point A in Fig. **1**a) (Fig. **1**b, t = t_1). At this time lapse, the soil strength is fully mobilized. This means that the *p*’-*q* point reaches the failure line. Drained conditions are kept, both for loose and dense soils, during the post-failure stage and both displacement (δ) and acceleration (a) are low (Fig. **1**b, t > t_1). In loose saturated soils, the above process is associated with a volume reduction. A so-called “flowslide” occurs (Fig. **1**d, t > t_1), if the pore-water pressure cannot freely dissipate, so that partially or totally undrained conditions develop during the post-failure stage. Particularly, in the spring zone (point A in Fig. **1**), pore water pressures build up and the soil cannot sustain the imposed deviatoric stress *q* and failed soil mass accelerates (Fig. **1**d, t > t_1), leading to a catastrophic failure. Slides can turn into “flows” (Fig. **1**c) as a consequence of complex mechanisms. First, a decrease in shear strength can occur due to local hydraulic boundary conditions such as in the spring zone (t=t_1). Upslope (point B), the mobilised shear stresses increases, both in loose and dense soils (t>t_1) and a further slide can occur (t=t_2), which is characterised by a high initial acceleration. Consequently a flow is generated (t>t_2).

Major differences between a flowslide (Fig. **1**d) and a slide turning into flow (Fig. **1**c) can be also outlined focusing on pore water pressures at failure. For the analysed schemes, pore water pressures reach the highest values in the spring zone (point A) due to both rainfall and local hydraulic boundary conditions (i.e. spring from bedrock) and the lowest values upslope the spring zone (point B), where can be still negative at failure (Fig. **1**c) and slides turning into flow can also occur in portions of the slope characterised by unsaturated conditions.

Another important aspect to be taken into account is the type of failure. In the cases sketched in Figs. **1**b-c, drained failure takes place at the critical state line, and failure can be “localized” (Pastor et al., **2002**, **2004**). It means that soil deviatoric strains mostly concentrate in a thin shear zone. On the contrary, fully or partially undrained post-failure stage of very loose materials (Fig. **2**d) is diffuse (Darve and Lauoafa, **2000**; Merodo et al., **2004**). It entails that a soil volume is yielded. The difference relates mostly to pore water pressure generation, which has to be carefully considered by using suitable constitutive and mathematical models. Notwithstanding the previous differences, it can be stated that for all the landslide typologies of Fig. 1, the eventual sudden acceleration of the failed mass (post-failures stage) is a consequence rather than a cause of the slope instability process, as experimentally demonstrated by Eckersley (**1990**) and Chu et al. (**2003**). This means that the failure and post-failure stages can be separately analysed.

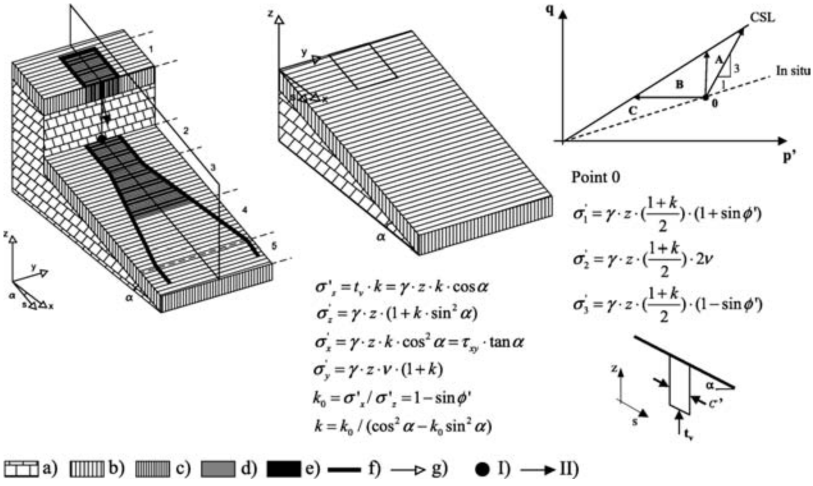

1
2
3
4
5
z
y
x
s
α
q
CSL
A
3
1
In situ
B
0
C
p'
Point 0
$\sigma_1' = \gamma \cdot z \cdot (\frac{1+k}{2}) \cdot (1+\sin\phi')$
$\sigma_2' = \gamma \cdot z \cdot (\frac{1+k}{2}) \cdot 2\nu$
$\sigma_3' = \gamma \cdot z \cdot (\frac{1+k}{2}) \cdot (1-\sin\phi')$
$\sigma'_s = t_v \cdot k = \gamma \cdot z \cdot k \cdot \cos\alpha$
$\sigma_z' = \gamma \cdot z \cdot (1 + k \cdot \sin^2\alpha)$
$\sigma_x' = \gamma \cdot z \cdot k \cdot \cos^2\alpha = \tau_{xy} \cdot \tan\alpha$
$\sigma_y' = \gamma \cdot z \cdot \nu \cdot (1+k)$
$k_0 = \sigma'_x / \sigma'_z = 1 - \sin\phi'$
$k = k_0 / (\cos^2\alpha - k_0 \sin^2\alpha)$
c'
t_v
a)
b)
c)
d)
e)
f)
g)
I)
II)

Figure 2. Schematic of an open slope prone to debris avalanche and stress paths relative to the triggering stage. General features: **a**) bedrock, **b**) stable soil deposit, **c**) failed soil, **d**) propagating failed mass, **e**) entrained material, **f**) boundary of debris avalanche, **g**) propagation pattern. Triggering factors: **I**) spring from bedrock, **II**) impact loading. Zone 1–2: triggering. Zone 3: thrust of failed material and/or soil entrainment. Zone 4: soil entrainment. Zone 5: propagation. Stress paths for: drained impact (zone A), undrained impact (zone B), spring from bedrock (zone C), liquefaction (zone B and/or C), and thrust of failed mass on stable soils (zone B or to be determined). (Cascini et al., **2013a**).

In the scientific literature, distinct triggering mechanisms are indicated for the inception of debris avalanches: i) the impact of failed soil masses on stable deposits, ii) direct rainfall infiltration from the ground surface, locally facilitated by anthropogenic factors such as mountain roads and tracks, iii) karst spring from bedrock as observed for pyroclastic soils in southern Italy, iv) runoff from bedrock outcrops as evidenced for shallow landslides in cohesionless soils of the Eastern Italian Alps and v) multiple failures in the landslides source areas. The scientific literature also indicate that: i) all these triggering mechanisms originate small translational slides; ii) the failed mass increases its volume inside triangular-shaped areas during the so-called “avalanche formation” which is mostly explained referring to soil liquefaction induced by impact loading; iii) soil erosion along the landslide propagation path may also play a paramount role.

Two different stages can be individuated for debris avalanches, i.e. the failure stage and the avalanche formation stage: the former includes all the triggering mechanisms which cause the soil to fail; the latter is associated to the increase of the unstable volume. Referring to these stages, four different zones can be distinguished (Fig. **2**). Zone 1 corresponds to small failures which occur at natural or anthropogenic discontinuities of soil deposits, bedrock outcrops and cut slopes, respectively. Zone 2 is the impact zone of the previously mentioned failed masses that usually corresponds to water supplies from bedrock (either karst spring or water runoff at bedrock outcrops); if the Zone 1 is absent, Zone 2 is the source area of small landslides triggered by water supplies from bedrock. Zone 3 corresponds to distinct mechanisms: thrust of the failed mass upon the downslope stable material and/or soil entrainment due to the propagating mass. Zone 4 exclusively corresponds to soil entrainment. It is worth noting that while zone 1 and 2 are few tens of metres large, the width of zone 3 and 4 is not known a priori and its forecasting is a challenging task.

With reference to the stages and zones in Fig. **1**, the mechanics of debris avalanches can be well analysed referring to the scheme of infinite slope (Fig. **2**) and to the stress invariants q and p’. Particularly, in-situ initial conditions (before the debris avalanche has been triggered) at the zones 2 and 3 of Fig. **1** depend on soil saturation degree (S_r) and are represented by the stress point 0 of Fig. **2**. In dry condition ($S_r=0$) the principal stress directions ($\sigma'_{i=1,2,3}$) are known (Lambe and Whitman, **1979**; Iverson et al., **1997**) and the normal stress values σ'_z, σ'_y and σ'_s can be easily obtained if the lateral earth pressure coefficient k_0 refers to stress conditions at rest (Jaky, **1944**). Particularly, σ'_z increases with soil depth while both σ'_y and σ'_s increase with slope angle. In the case of steep slopes, equilibrium conditions require high soil friction angles which correspond to low values of k_0 and σ'_y; consequently, the associated (p’, q) points have a high stress ratio $\eta=q/p'$ and they lie very close to the failure criterion. For saturated soil condition ($S_r=1$), the soil unit weight (γ_{sat}) and the deviatoric stress (q) are higher than in the previous case while the mean effective stress (p’) can be either higher or lower, depending on soil unit weight (γ_{sat}) and pore water pressure (p_w). Therefore, for saturated soil condition ($S_r=1$), the (p’, q) stress points can be even closer to the failure line than for dry condition ($S_r=0$). For unsaturated soil condition ($S_r<1$), the suction (s) determines higher mean effective stresses (p’) than in saturated condition and a shear strength envelope with a positive apparent cohesion intercept (Fredlund et al., **1978**); thus, the stress points (*p*’, *q*) are more distant from the failure criterion than in saturated soil conditions.

When an impact loading occurs (see zone 2 of Fig. **2**), it mainly corresponds to an increase of deviatoric stresses; the stress paths are inside the zone A of Fig. **2** (for drained conditions) or in the zone B of Fig. **2** (for undrained conditions). In the latter case, the stress path may rapidly approach the failure criterion. However, the assumption of drained or undrained conditions can be more or less acceptable depending on loading velocity and soil conductivity and the hydro-mechanical coupling between the solid skeleton and pore fluid may play a crucial role, as discussed later. Other triggering factors such as direct rainfall infiltrating the slope ground surface, karst springs from bedrock or runoff from upslope bedrock outcrops induce stress paths in the zone C of the q-p’ plot of Fig. **2**; in these cases, fully drained conditions can be reasonably assumed (Cascini et al., **2010**).

For the avalanche formation, remarks can be also outlined referring to the zone 3 of Fig. **1**. Particularly, the occurrence of soil liquefaction is strongly related to the initial stress state in the q-p’ plane (Fig. **2**) and mechanical features of soils, thus corresponding to stress paths moving in the zone B and/or C of the q-p’ plot of Fig. **2**. Analogously, the thrust of an unstable mass upon downslope stable soils cause an increase of deviatoric stresses and a stress path moving in the zone B of q-p’ plot of Fig. **2**. On the other hand, soil entrainment phenomena depends on the kinematic features of the propagating mass which are, in turn, related to: i) initial volume, ii) rheological behaviour and iii) hillslope topography.

Mechanisms for the Transformation of a Slide into a Flow

Post-failure stage is a fundamental topic since it discriminates different types of phenomena. In fact, it is quite evident that the chance for a landslide to achieve high velocities depends on: i) the initial acceleration of the failed mass and ii) subsequent transformation in to a landslide of the flow type.

Anyway, the acceleration of the failed mass during the post-failure stage is associated to different mechanisms. Many Authors outline that the development of total or partial undrained conditions as the main cause of high pore-water pressures upon shearing. In particular, for loose unsaturated soils, volumetric collapse is discussed by Olivares & Damiano (**2007**), Yasufuku et al. (**2005**), Bilotta et al. (**2006**) and it is observed in constant-shear-drained triaxial tests upon wetting (Anderson and Riemer, **1995**; Dai et al. **1999**; Chu et al. **2003**; Olivares & Damiano, **2007**). For loose saturated soils, static liquefaction is introduced by Wang et al. (**2002**), Olivares & Damiano (**2007**), Van Asch et al. (**2006**) and observed in undrained triaxial tests (Lade **1992**; Yamamuro and Lade **1998**; Chu et al. **2003**) as well as in undrained ring shear tests under controlled strain rates (Wang et al. **2002**). Particularly, the build-up of pore pressures is shown to be relevant for soils having low relative density index (Eckersley **1990**; Iverson **2000**; Wang and Sassa **2001**), fine content (Wang and Sassa **2003**), low hydraulic conductivity (Iverson et al. **1997**; Lourenco et al. **2006**) and subjected to high deformation rate (Iverson et al. **1997**).

The most of the above findings are obtained through laboratory tests such as isotropically consolidated undrained triaxial tests (ICU) (Chu et al., **2003**), anisotropically consolidated undrained triaxial tests (ACU) (Eckersley, **1990**), constant shear-drained triaxial tests (CSD) (Chu et al., **2003**) even though strain localisation is more important under plane-strain or the 3D conditions compared to triaxial conditions, as recently discussed by Wanatowski and Chu (**2007**, **2012**). It is worth noting that all laboratory tests refer to idealized drainage conditions.

On the other hand, a direct measurement of pressures and displacements in real slopes is easy only for: i) sites monitored during the occurrence of landslides, ii) artificially induced failure in real slopes. In both cases, once the failure has occurred, the measurements cannot be repeated anymore at the same conditions.

Further insights derive from direct observation of pore water pressures and stresses in landslides artificially induced in slope models at a reduced scale (also called flume tests). Through this approach, information can be obtained on failure and post-failure (Eckersley, **1990**); however, these experiments are expensive and since they reproduce the real processes at a greatly reduced scale they may be irrespective of the full-scale slope behaviour. For instance, a large difference in stress levels may exist between model and prototype; in particular, the eventual capillary suction is out of proportion with its self-weight stress, allowing the model slope to remain steeper than would be possible at higher effective stress levels. Nevertheless, complex groundwater conditions, such as downward rainfall infiltration from ground surface and/or a downwards/upwards water spring from the bedrock to the tested soil layer, can be analysed through these tests (Lourenco et al., **2006**).

A more recent approach is based on centrifuge tests which reproduce stress levels similar to those experienced by a real slope. Centrifuge tests - except for some drawbacks such as the high costs and the availability of sophisticated equipments - combine the advantages of highly instrumented slopes (such as full/reduced scale models) with the potential of geometrical configurations realistically reproducing the in-situ conditions. Particularly, Take et al. (**2004**) point out that the transition from slide to flow is caused by local failures producing a variation in the slope geometry. This mechanism is related to transient localized pore-water pressures that are not associated to the development of undrained conditions, but originated by the combination of particular hydraulic boundary conditions and stratigraphical settings. Experimental evidences show that the transition from slide to flow can occur both in loose and dense soils and that it can also correspond to decreasing pore-water pressures during the post-failure stage. These results have been later confirmed also by other researchers through small-scale flume tests (Lourenco et al. **2006**) or centrifuge tests (Lee et al., 2008, Ng, 2009; among others).

Based on previous considerations, mathematical modelling may be outlined as a powerful tool because, in principle, it can be used to investigate a wide variety of different scenarios even though the modelling of the post-failure stage is still poorly addressed in the literature such as in the case of earthquake (Pastor et al. **2004**) or static perturbations (Laouafa and Darve **2002**).

Mechanisms of the Propagation Stage

Velocities, heights and percentages of water and debris are 3D spatially distributed quantities in a landslide of the flow type (Hungr et al., **2001**); they may be distributed either along the path or in a vertical direction. The propagation stage is difficult to analyse as relevant parameters such as viscosity, soil friction angle or other rheological parameters and pore water pressures cannot be easily measured in full-scale examples and direct measurements are rarely available for real cases.

These analytical difficulties exist even for channelised landslides (Hungr et al., **2001**) that, independent of the triggering mechanisms occurring in the source areas (Fig. **3**), propagate in ‘V’ shaped channels with steep flanks. For instance, Fig. **3**a shows a landslide source area located at the upper limit of the channel, as in the case of zero order basins (Cascini et al., **2008**); alternatively, the landslide source area may be lateral to the upper limit of the channel (Fig. **3**b). In either case, the propagation stage can be schematised as follows: i) at the entry of the channel, the height and velocity of the propagating mass increase and this effect is worse if two or more propagating masses join together; ii) along the channel, a great amount of material is available for bed entrainment during heavy rainstorms, and the channel may also provide water to the propagating masses, which will then fluidise, even without static liquefaction in the landslide source areas; iii) at the exit of the channel, the mass may stop or propagate further; in the latter case, the propagation direction is not known a priori and mass bifurcation may occur along secondary branches with different run-out distances travelled along each path; and iv) at the piedmont, deposition takes place where the channel terminates, i.e. where the longitudinal slope angle sharply decreases and cross sections are progressively wider and less deep (Fig. **3**).

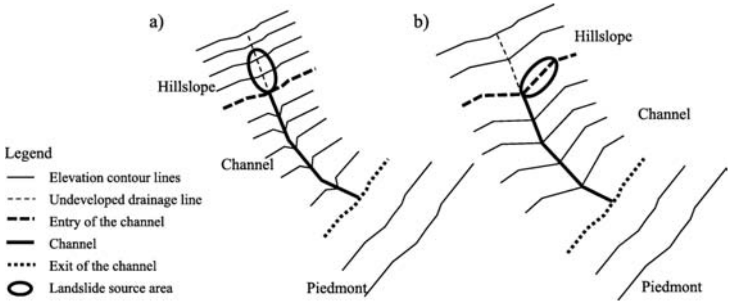


Figure 3. Schemes of propagation patterns for channelised flows, with the source area (**a**) along the channel or (**b**) located aside.

The prediction of propagation pattern(s), run-out distances, velocities and heights of propagating mass can reduce losses as it provides a means for i) defining the hazardous areas and estimating the intensity of the hazard and ii) working out the information for the identification and design of appropriate mitigation strategies (Fell et al., **2008**).

Bed entrainment - also called erosion or basal erosion - is the process that causes an increase in the volume of flow-like landslides (Savage and Hutter, **1991**; Pastor et al., **2009**) owing to the inclusion of soil, debris and trees uprooted from the ground surface. In principle, the entrainment process can be simply analysed by referring to the entrainment rate (e_r), defined as the time derivative of the ground surface elevation (z), over which the landslide propagates. It is generally agreed that the entrainment is positive if z diminishes, i.e. $e_r=-\delta z/\delta t$. However, the entrainment rate (e_r) depends on several variables: the flow structure (i.e., percentage of solid and fluid in the mixture), the density and size of the solid particles, the saturation degree of the base soil along the landslide path, the slope angle, and how close to failure the effective stresses are at the bed of the propagating mass. Bed entrainment is a crucial process increasing the landslide volume and modifying the mass velocities along the whole landslide path(s), as shown in Fig. **4**. It is worth noting that landslide volume promotes the travel distance (Rickenmann, **2009**), whereas bed entrainment absorbs momentum from the sliding/propagating mass and should reduce the run-out distance. However, this interplay also depends on other factors. In fact, mass velocities (and heights) determine the capability of a landslide to entrain further material and, in turn, the total entrained volume. Consequently, the percentages of water and debris change over time and so the mass rheology does. Bed entrainment also affects pore water pressures in different ways depending on slope morphology, e.g. confined/not confined flow. Therefore, bed entrainment and propagation are coupled processes that should be analysed within a unified mathematical framework. However, this interplay is not clearly addressed and modelled in the current literature.

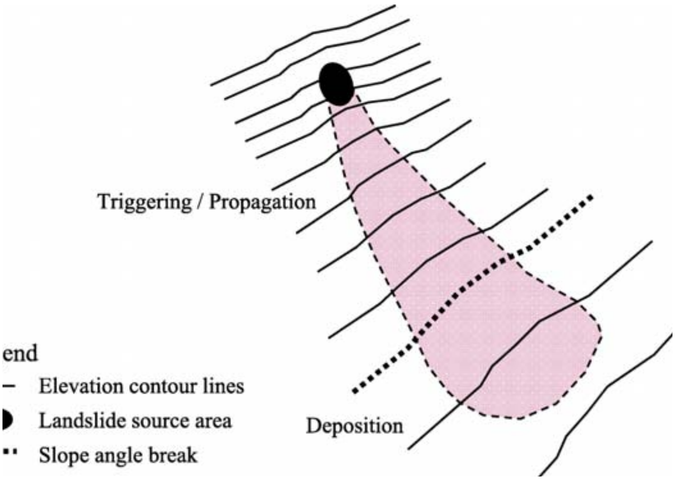


Figure 4. Scheme for a Debris Avalanche (DA) developing as an unchannelised flow along an open slope.

Based on these key factors, many formulations for the entrainment rate have been proposed in the literature, and a comprehensive review of the entrainment models has been provided by Pirulli and Pastor (**2012**) and Cascini et al. (**2014**). Here, it is worth noting that most of the formulations indicate a direct proportionality between the entrainment rate (e_r) and the flow velocity (v) and/or the flow depth (h). Moreover, it is recognised that the occurrence of bed entrainment implies that: i) velocity and height of the flowing mass are modified, ii) pore water pressure at the base of the flow is altered, and iii) the rheology (i.e., the features and mechanical behaviour) of the flow could be modified as well if the flowing mass and the entrained materials are very different. Indeed, the entrainment process is very complex, and former contributions have been proposed based on tests (flume, centrifuge, or full scale) of differently sized, generally smaller than 10 m^3, propagating volumes (Iverson et al., **2011**) or numerical modelling of real debris flows (Cascini et al., **2014**) or historical debris avalanches (Cuomo et al., **2014**).

To provide further insight on the topic, this paper will focus also on real-scale landslides, particularly on debris avalanches and debris flows, and related cascading effects (Chen et al., **2006**; Crosta et al., **2009**; Pirulli and Pastor, **2012**). Figure **5** provides a sketch for a DA evolving into a single DF (Fig. **5**a); a single DA generating multiple DFs (Fig. **5**b); and several DAs and DFs evolving in a single huge DF or in multiple surges delayed in time (Fig. **5**c). An example of scenario "a)", the 1995 Izoard pass debris flow (southern French Alps) is characterized by an erosion thickness up to 5 m at the top of a 25° to 30° steep channel (Lake et al., 1998). An interesting debris avalanche, which bifurcated into two debris flows (scenario "b)"), occurred in Tsing Shan (Hong Kong) in 1990. Bed entrainment greatly increased the landslide volume, from 150 to 1600 m^3, because of the very steep slope (approximately 40°) and the abundance of colluvial material along the slope (King **2001a**, **b**). The scenario "c)" is typical of high mountain ridges of China and Canada (Hungr and Evans, **2004**).

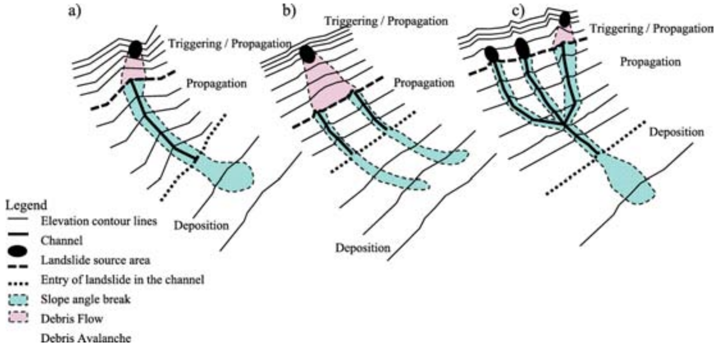


Figure 5. Schemes of combined flows in different slope configurations: **a**) *DA* turning into a *DF*, **b**) *DA* turning into two *DFs*, **c**) multiple *DAs* and *DFs* joining into a big *DF*.

METHODS

Alternatives for Landslide Modelling

Generally speaking, two main different approaches can be referred: the Lagrangian description and the Eulerian one. In the Lagrangian description, the computation points are linked to the material which is deforming and this category includes the well-known Finite Element Method (FEM) in the small displacement Lagrangian description, the Discrete Element Method (DEM) and Smoothed Particles Hydrodynamics (SPH). Among these methods, Lagrangian FEM analyses have been extensively used in solid mechanics to simulate small strains accumulated prior of failure (pre-failure stage) and at failure onset (during the failure stage) based on solid-like constitutive laws, as reviewed by Duncan (**1996**). Nevertheless, the FEM with Lagrangian description does not allow the description of the flow of the soil until deposition because of the tendency of the mesh to become more and more distorted. Concerning the DEM, this method is proper for modelling the behaviour of granular materials in small and large deformations: for example it is successfully applied to model granular flows with comparison against experimental laboratory evidences (Favier et al., **2009**, Faug et al., **2011**). However, with this method it is hard to handle large domains of space or time because of the high numerical cost necessary to compute the particle connectivity. In particular, in the field of landslides the continuity of media and kinematic fields can be often assumed and thus the benefit of DEM method is drastically reduced. Lastly, the SPH is developed in a continuum mechanics framework and it does not show important limitations apart from some drawbacks with boundary conditions. Up to now, this method has been mostly applied to landslide propagation problems (Pastor et al., **2009**) and only recently for analyzing static equilibrium problems (Fukagawa et al., **2011**).

In the Eulerian description that is common in fluid mechanics, the nodes are fixed and that is why this solution is usually the best one to model a fluid-like material with large deformations in the propagation stage: there exists, for instance, the FEM with an Eulerian formulation. However, the material properties are advected across the fixed computational grid. Such a procedure causes a spurious (numerical) diffusion of history variables (e.g. plastic strains) and interfaces of heterogeneous material setting are smoothed in space through time.

Finally, some mixed methods are available which try to combine the advantages of the two main descriptions. For example, the Arbitrary Lagrangian Eulerian method (ALE) avoids the mesh tangling by allowing computation points to move but additional advection terms are required to handle transport of quantities related to the mesh and thus drawbacks of the pure Eulerian approach appear again. Still for large and complex deformation processes, such as those involved in slope stability problems, ALE cannot avoid mesh distortion and hence computation is stopped.

In fully Lagrangian FEM, SPH and DEM all the computational points coincide with material points (Fig. **6**); the latter ones are not tracked in ALE and in Eulerian FEM. In order to get over the difficulties of the classical numerical methods in this framework, as an alternative the Finite Element Method with Lagrangian Integration Point (FEMLIP) (Moresi et al., **2002**, **2003**), is proposed as derived from the Particle In Cell method (Sulsky et al., **1995**). Similar concept is that behind the Material Point Method (MPM), which has been recently applied to a number of different slope stability and landslide cases (Wang et al., **2016**, **2018**; Ghasemi et al., **2018**, **2019**; Cuomo et al., **2019a**, **b**). Both methods and others similar available in the literature are based on a kinematic dissociation between the material points and the computational nodes of the finite element Eulerian mesh. For a given material configuration, the material points are used as integration points on one element. The resolution of the equilibrium equation at the nodes gives a velocity field. At the end of each step, the velocity is interpolated from the nodes to the material points which are moved accordingly throughout the fixed mesh up to a new configuration. Since all material properties including internal variables are stored at material points, they are accurately tracked during the advection process. Actually, thanks to this dissociation between mesh nodes and material points, such approach benefits both from the ability of an Eulerian FEM (the mesh is kept fixed) to support large transformations, and from the possibility of a Lagrangian FEM to track internal variables during the material movement. This method is – in principle – suitable to deal with: i) static equilibrium of elasto-plastic materials in the pre-failure stage, ii) large deformations upon failure, iii) large displacements during the propagation stage while still tracking the history of material properties.

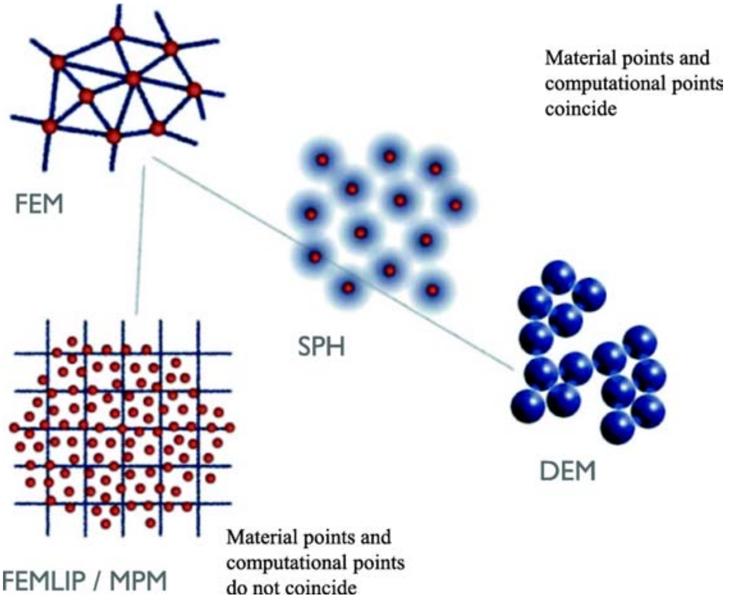

FEM
SPH
DEM
FEMLIP / MPM
Material points and computational points coincide
Material points and computational points do not coincide

Figure 6. Scheme of different methods for the analysis of soil deformation and slope failure.

Models for Landslide Triggering Simulation

Several approaches are currently available for the slope analysis and they allow separately modelling the failure, post-failure and propagation stages of hillslope instability phenomena. They are divided in three broad classes depending on the amount of deformation taken into account.

Particularly, the failure stage can be analysed using many standard engineering methods that, anyway, disregard the deformations prior to failure and during the failure stage; these methods are usually called Limit Equilibrium Methods (LEM, later on) which include the Infinite Slope Method (ISM, hereafter) and the so-called slice methods proposed under distinct hypotheses by many Authors (Morgenstern & Price, **1965**; Janbu, **1954**; among others). In these methods the constitutive law of the material is rigid-perfectly plastic and thus the displacements along the slip surface cannot be assessed.

The result of LEM analysis is the so-called factors of safety (FS), which has been extensively used to satisfactorily tackle a number or real cases in the last decades. Particularly, both failures and stable conditions computed via LEM have been fully confirmed by in-situ evidences of natural or man-made slopes (Leroueil, **2004**, among others).

More sophisticated approaches are available which allow computing soil deformations and displacements in boundary value problems. In order to properly reproduce the previously described typologies of shallow landslides, it is necessary to use a (i) mathematical model describing the coupling between pore fluids and soil skeleton, (ii) a suitable constitutive relationship able to describe the unsaturated soil behaviour, and (iii) a numerical model where (i) and (ii) are implemented. To the authors' knowledge, these have not been done yet in a full satisfactory manner, and until such tools are available, simplified models have to be carefully used.

This paper uses the mathematical framework derived from the fundamental contributions of Zienkiewicz et al. (**1980**, **1999**). This framework can be profitably used to simulate the landslide failure and post-failure stages. It is assumed that the soil consists of a solid skeleton and two fluid phases, water and air, which fills the voids. The movement of the fluid is considered as composed of two parts, the movement of soil skeleton and motion of the pore water relative to it. The total stress tensor acting on the mixture is decomposed into a hydrostatic pore pressure term and an effective stress tensor acting on soil skeleton, which can be also extended to the case of unsaturated soils. In the balance of momentum equation for the mixture the acceleration of water relative to soil grains is neglected. Whereas, the deformation of soil skeleton, the deformation of soil grains caused by pore pressure, the deformation of pore water caused by pore pressure, and the increase of water storage are considered. The Darcy law is used to describe water flow through the soil skeleton, although other alternatives can be chosen. In above, the acceleration terms of the pore water relative to soil skeleton are neglected, and the space derivatives of accelerations are assumed to be small. Finally, the model is completed with kinematic relations linking velocities to rate of deformation tensor and a suitable constitutive equation. More details are provided by Pastor et al. (**2004**). In the next sections the "GeHoMadrid" code will be used combined to either standard constitutive models like Drucker Prager (DP) or advanced constitutive models like that proposed by Pastor and Zienkiewicz (**1990**), (PZ).

Models for the Simulation of Landslide Propagation

Several methods have been developed to analyse the landslide propagation.

Empirical methods are based on field observations and identify relationships between landslide volume, local morphology, presence of obstructions and landslide run-out distance. The availability of landslide datasets has encouraged statistical (bivariate and multivariate) analyses that point out indexes directly (or indirectly) related to landslide mobility. To date, these empirical models have provided an estimation of run-out distance that can be correlated to i) the amount of the unstable volume (Corominas, **1996**) and ii) the features of landslide source areas (e.g. width/length ratio, depth of slip surface) or slope morphology (Cascini et al., **2011b**). These approaches are commonly used for the back-analysis of case histories; they capture the global observed behaviour (high mobility) of these landslides, but disregard crucial local effects (e.g. diversions and/or bifurcations).

Analytical methods simulate the landslide propagation using physical-based equations derived from solid and fluid dynamics (Pastor et al., **2009**; Pirulli and Sorbino, **2008**; Hungr and Mc Dougall, **2009**). Thus, velocity and height are provided alternatively at (i) each point of a given domain in Eulerian-formulated models or (ii) at each point of the propagating mass for Lagrangian approaches. The three main categories of the Lagrangian approaches are i) block (‘lumped mass’) models, ii) two-dimensional models that look at a typical section profile of the slope, neglecting the width dimension, and iii) three-dimensional models treating the flow of a landslide over an irregular 3D terrain. Most of the models belonging to the latter two categories are simplified by integrating the internal stresses in either the vertical or bed-normal direction to obtain a form of St. Venant equation. Then, the governing equations are solved using numerical methods such as finite difference (O’Brien et al., **1993**), finite element (Pastor et al., **2002**), finite volume (Pirulli and Sorbino, **2009**; Pirulli and Pastor, **2012**) or smooth particle hydrodynamic (SPH) (Pastor et al., **2009**). Among the governing equations, the rheological model poses important scientific and practical difficulties; rheological parameters can often only be obtained from the back-analysis of case histories, and thus simple models are preferred because a limited number of parameters can be constrained more easily. It is worth noting that only a few models schematise the propagating mass as a mixture of solid grains and pores, thus providing information on pore water pressures in space and time (Pastor et al., **2011**).

Different hypotheses have been formulated for the entrainment onset: i) velocity threshold, ii) dependency on slope angle (Brufau et al., **2000**, Egashira et al., **2000**, **2001**; Papa et al., **2004**); and iii) correlation with landslide volume (Chen et al., **2006**). Alternatively, the so-called ‘erosion rate' (e_r) may be invoked, which is defined as the time derivative of the ground surface elevation and is equal to the time derivative of the soil depth of the propagating mass when other causes are not in play. The erosion rate can be modelled as proportional to the product of velocity (v) and propagation height (h). In this case, it is convenient to refer to the 'landslide grow rate' (E_r), which is independent of the flow velocity. Once assigned an E_r, the amount of bed entrainment depends on both the height and velocity of the propagating mass at each point of the landslide path. The terms e_r and E_r are related by the equation $e_r = E_r \cdot h \cdot v$. Takahashi et al. (**1991**) relate E_r to two factors: the solid concentration of the propagating mass and the availability of solid particles along the landslide path. However, Hungr (**1995**) relates E_r to the initial and final landslide volume and to the travelled distance (L) in the following way: $E_r \cdot = \ln(V_{\text{final}}/V_{\text{initial}})/L$, where V_{initial} is the volume entering an erodible zone of the slope, and V_{final} is the sum of the initial volume and the entrained material. In more complex formulations the growth rate depends on the solid concentration, slope angle and shear strength of the eroded material (Ghilardi et al., **2001**).

Analytical approaches to bed entrainment analysis require a proper rheological (or constitutive) model for the behaviour of the interface between the propagating landslide and the ground surface. Bed entrainment is also related to flow structure, density, size of particles and how close to failure the effective stresses at the ground surface are. To the authors' knowledge, there are very few analytical models for bed entrainment in the current literature. Medina et al. (**2008**) relate e_r to factors such as i) landslide velocity, ii) shear stress mobilised at the base of the propagating mass, iii) slope angle, and iv) unit weight of the propagating material. Quan Luna et al. (**2012**) proposed a 1D analytical model for erosion assessment based on limit equilibrium considerations and the generation of excess pore water pressure through undrained loading of the in-situ bed material; similar approaches could provide fully realistic results if extended to 3D conditions. Analytical approaches have rarely been implemented in numerical codes and thus their application to real case histories is still limited.

Finally, mixed methods combine analytical methods for propagation and empirical methods for bed entrainment. Mixed methods have been recently applied by Hungr and McDougall (**2009**) and Pastor et al. (**2009**) to landslides of the flow type. All of these contributions refer to the empirical erosion law of Hungr (**1995**) and it is worth comparing their back-analysed values of E_r, which span a wide range of values due to differences in site conditions and soil properties. However, the estimated entrainment coefficients in these analyses also depend on both the chosen rheological model and the calibration procedure for the rheological parameters. Therefore, further applications of numerical approaches to real case histories are necessary to better assess the potential bed entrainment during the landslide propagation stage. Therefore, in this study a relevant case history from Southern Italy is analysed.

The 'GeoFlow_SPH' model proposed by Pastor et al. (**2009**) is applied here below. The model is based on the theoretical framework of Hutchinson (**1986**) and Pastor et al. (**2002**) and schematises the propagating mass as a mixture of a solid skeleton saturated by water; the unknowns are the velocity of the solid skeleton (v) and the pore water pressure (p_w).

The governing equations are i) the balance of the mass of the mixture combined with the balance of the linear momentum of the pore fluid, ii) the balance of the linear momentum of the mixture, iii) the rheological equation relating the soil stress tensor to the deformation rate tensor, and iv) the kinematical relations between the deformation rate tensor and the velocity field. From this, we derive a propagation–consolidation model by assuming that pore water pressure dissipation takes place along the normal to ground surface, and the velocity of the solid skeleton and pressure fields can be split into the sum of two components related to two processes: propagation and consolidation (for further details see Pastor et al., **2009**). The initial pore water pressure is taken into account through the relative height of the water, h_w^{rel}, which is the ratio of the height of the water table to the soil thickness, and the relative pressure of the water p_w^{rel}, that is to say the ratio of pore-water pressure to liquefaction pressure. Estimates of both parameters can be obtained from the analysis of the triggering stage, and they play an important role in the propagation stage of a flow-like landslide (Cuomo et al., **2014b**). In the model here used, the vertical distribution of pore water pressure is approximated using a quarter cosinus shape function, with a zero value at the surface and zero gradient at the basal surface (Pastor et al., **2009**), and the time-evolution of the basal pore water pressure (p^b_w) relates to the consolidation factor (c_v).

As many flow-like landslides have small average depths in comparison to their lengths or widths, the above equations can be integrated along the vertical axis and the resulting 2D depth-integrated model presents an excellent balance of accuracy and simplicity. The GeoFlow_SPH model also accounts for bed entrainment along the landslide path, and the elevation of the ground surface consistently decreases over time. In addition, different empirical erosion laws can be implemented in the GeoFlow_SPH model (e.g. Hungr, **1995**; Blanc, **2011**; Egashira et al. **2000**, **2001**; Blanc et al., **2011**; Blanc and Pastor, **2011**, **2012a** and **2012b**). The simple yet effective law proposed by Hungr (**1995**) is used mainly to achieve results comparable to those available in the literature. Hungr (**1995**) relates E_r to the initial and final landslide volume and to L; the entrained material is assumed to have nil velocity and nil pore water pressure when entrained by the propagating mass.

In the GeoFlow_SPH model, the Smoothed Particle Hydrodynamics (SPH) method is used; this method discretises the propagating mass through a set of moving 'particles' or 'nodes'. Information, i.e. unknowns and their derivatives, is linked to the particles, and the SPH discretisation consists of a set of ordinary differential equations whose details are provided by Pastor et al. (**2009**). The accuracy of the numerical solution and the level of approximation for engineering purposes depend on how the nodes are spaced and how the digital terrain model (DTM) is detailed, as recently reviewed by Pastor and Crosta (**2012**) and Cuomo et al. (**2013**).

RESULTS AND DISCUSSION FOR NATURAL SLOPES

Failure of Shallow Soil Covers

A first example of landslide triggering simulation is provided for three different combinations of shallow covers potentially unstable due to rainfall from ground surface combined to water spring from the bedrock. This is a recurrent site condition in several geoenvironmental contexts.

Three infinite slope schemes are referred and parametric analyses are performed with typical slope angles (35 degrees), depths (4.5 m) and stratigraphical settings (Fig. **7**) provided by the in-situ evidences (Cascini, **2004**). Particularly, the three schemes well averages the stratigraphy of pyroclastic covers located around the Vesuvius volcano (Naples, Italy), like the Pizzo d'Alvano massif, where impressive flow-like landslides occurred in 1998. The geomechanical modelling of the pore water pressure is performed for the time period from January 1, 1998 to May 5, 1998, by using the commercial finite element code SEEP/W (Geoslope, **2005**). The main soil parameters are here summarised: i) Ashy soil A (porosity, n: 0.66; saturated unit weight, γ_{sat}: 15.7 kN/m^3; saturated conductivity, k_{sat}: 10^{-6} m/s; friction angle, φ': 35°; effective cohesion, c': 10 kPa); ii) Ashy soil B (n: 0.58; γ_{sat}: 13.1 kN/m^3; k_{sat}: 10^{-5} m/s; φ': 37°; c': 0 kPa); iii) Pumice soils (n:

0.69; γ_{sat}: 13.1 kN/m^3; k_{sat}: 10^{-4} m/s; φ’: 37°; c’: 0 kPa). More details on the characterization of pyroclastic soils can be found in Cascini et al. (**2010**).

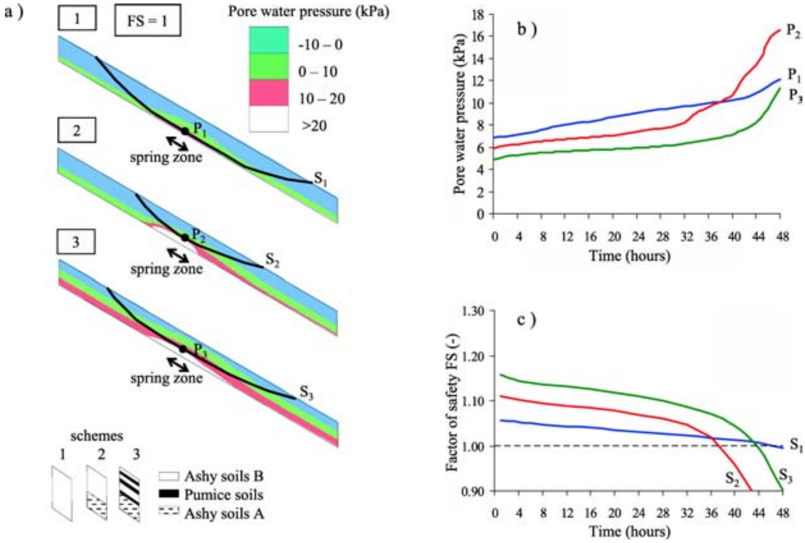


Figure 7. Results of numerical modelling for seepage and slope stability analysis. **a**) Pore water pressure computed at failure (FS: Factor of Safety equal to 1) for different slope schemes (1–3), **b**) pore water pressure versus time at a representative point along the slip surface for each slope, **c**) Factor of Safety versus time for the slip surfaces (S1-S3) (Cascini et al., **2010**).

The adopted FEM mesh consists in 3755 quadrilateral elements with lengths and heights respectively smaller than 1.0 m and 0.5 m. As initial conditions, suction values are assumed respectively equal to 5 kPa, all over the slope section. Daily rainfall intensity is applied as flux boundary condition at the ground surface for the period January 1, 1998 - May 3, 1998; hourly rainfall intensities are assigned for the last 2 days (May 4–5). At the contact between the pyroclastic deposit and the limestone bedrock, an impervious condition is assumed except for the zone where the spring from the bedrock is located (Fig. **7**). Here, a flux condition is considered with a flux value of 1.67 × 10^{-5} m^3/s, starting from 2nd or 3th May 1998. Using the computed pore pressures values, slope stability conditions are evaluated. To this aim, the limit equilibrium methods proposed by Janbu (**1954**) and Morgenstern and Price (**1965**) are adopted and the corresponding factor of safety values are computed by using the commercial SLOPE/W code (Geoslope, **2005**). For all the involved soils, a rigid-perfectly plastic constitutive model is referred considering, in both saturated and unsaturated conditions, the extended Mohr-Coulomb failure criterion proposed by Fredlund et al. (**1978**) with geotechnical properties listed in Table 1. The numerical results of the parametric analysis (Fig. **7**) indicate that rainfall infiltration from ground surface and spring from the bedrock increase the pore water pressures up to the slide occurrence (Fig. **2**b), independently from the assumed stratigraphical setting and for any shear strength value listed in Tab. 1. Different stratigraphical settings and mechanical properties of pyroclastic deposits anyhow determine different depths of the slip surfaces from the ground surface (Cascini et al., **2005**).

The successful application of the uncoupled approach based on the use of unsaturated transient seepage analysis and limit equilibrium slope stability analysis is worth of twofold comments. On one hand, this approach is relative simple to apply and based on the use of codes easily available for researchers and practitioners. On the other hand, the main physical processes and the key factors are properly taken into account so that a satisfactory interpretation of complex slope stability problems is obtained.

The main limitation of such type of application is that no information can be derived about the post-failure events. Will the soil liquefy or not? There will be any transformation into a flow? No answer will be obtained to these relevant questions, unless other approaches are used.

Transformation of a Slide into a Flow

The observation of soil liquefaction, slope fluidization, and similar phenomena is seldom observed or quantitatively measured in the field. For this reason, it is very useful to refer to laboratory slope experiments. Until few years ago, the option of small-scaled slopes is the only chance to consider. More recently, centrifuge tests allowed having almost a 1:1 correspondence between the prototype and real boundary value problems. The centrifuge tests of Take et al. (**2004**) are here analysed using the GeHoMadrid code. In the numerical analyses an unstructured mesh is used with triangular elements on average not larger than 0.4 m. A null pore water pressure values is assumed at point E - corresponding to the water table level observed at failure during the tests - to reproduce the raising of the water table in the upper soil layer. In the FEM analysis, pore water pressure is allowed to change in space and time, starting from an initial value of -5 kPa throughout the slope model. This is adequately taken into account referring to Bishop's stresses (for details see Pastor et al., **2002**). However, for sake of simplicity, numerical analyses are performed in the hypothesis of fully saturated conditions and the used version of the PZ constitutive model fits this hypothesis (Pastor et al., **1990**; Merodo et al., **2004**). Of course, the analyses could be extended to the case of unsaturated conditions but this is beyond the scope of the present paper. The soil mechanical properties are either taken from GEO (1999), Ng et al. (**2004**) and Take et al. (**2004**), e.g. $\gamma_{sat} = 14$ kN/m^3, e = 0.32 (Dense soil) or 0.62 (Loose soil), M_g and M_f, or indirectly estimated/calibrated, e.g. $k_{sat} = 10^{-4}$ m/s, E, η, H_0, comparing the experimental evidences and the numerical results. It is worth noting that two different values of M_f = 0.825 (Dense soil) or 0.550 (Loose soil) are assumed which derive from different values of relative soil density while the same critical friction angle (M_g = 1.375) and bulk modulus (K_{ev0} = 11.5 e^3 kPa) are considered for both loose and dense soils. This strong assumption is aimed at emphasizing in a limit case the role played by soil porosity as a fundamental factor for slope behaviour upon failure and beyond. The details of such soil characterization are given in Cascini et al. (**2013b**), and also more insights about the calibration of the constitutive model parameters are given in Cuomo et al. (**2018**).

Hydro-mechanical coupled quasi-static analyses are performed to take into account the coupling between the solid skeleton and pore fluid. The simulated plastic strains significantly differ in the case of loose and dense soil (Fig. **8**) for both the value (larger for loose soil) and extent of the affected zone. In the case of loose soil, "diffuse" plastic strains are simulated, firstly at the toe of the slope, and then they involve a larger amount of the slope as time elapses. For dense soil, plastic strains appear firstly at the toe of the slope and then they are "localized" along a slip surface where plastic strains accumulate as the process evolves. The above mentioned differences depend only on relative density being the other mechanical properties equal in the two cases. However, apart from the different type of failure, i.e. diffuse or localized, a different time evolution is also outlined. For loose soil, the failure stage is shorter because higher excess pore water pressures rapidly accumulate in the slope until it fails. Conversely, in the case of dense soil, both the pre-failure stage (mainly corresponding to elastic strains) and the failure stage are longer in time.

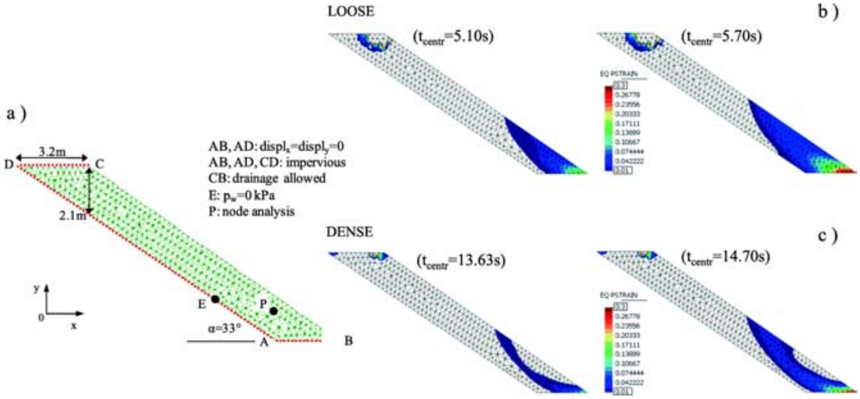


Figure 8. Modelling of centrifuge tests. FEM mesh used for computation (**a**), equivalent plastic strains computed at different time lapses for loose (**b**) and dense (**c**) soil (Cascini et al., **2013b**).

In this case, the use of a sophisticated hydro-mechanical approach combined to an advanced constitutive model is mandatory to reproduce the transformation of a slide into a flow. The use of such approach can highlight how a contractive loose soil slope undergoes a significant build up of pore water pressure due to a rainfall-induced soil volume change (Cascini et al., **2013b**). Based on limit equilibrium analysis, the slope would be stable while using a more adequate approach, such that used here, the slope will fail. More details on such scenarios are given in Cascini et al. (**2013b**).

The main limitation of the approach showed here is that the equations are all written in the framework of "small deformations", which means that once the deformed slope configuration becomes too much distorted compared to the original slope, the simulation stops or the numerical results are unreliable.

Modelling the Propagation Stage of Debris Flows (DFs)

The SPH model is here applied to a real case of two debris flows converging inside the same valley channel. A 3 × 3 m DTM is used as input for the GeoFlow_SPH model, as it accurately reproduces the topographical/morphological conditions of the sites before the event and the anthropogenic streets/channels (5–10 m large). The extent of the landslide source areas and the initial depths of the propagating masses are obtained from detailed landslide inventory maps and soil thickness maps at the 1:5000 scale (Cascini et al., **2005**; Cuomo, **2006**). Specifically, the landslide source areas have lengths of 250–400 m and widths of 50–200 m, and initial soil thicknesses in the range of 3–4.5 m. At point '1' of Fig. **10**, the eroded depths are 1–2 m and the piedmont areas, shown in red, indicate the piedmont areas hit by the flowslides. To set up the numerical simulations, 2936 and 4598 points are considered within the two landslide source areas of Sarno. In each zone, the points are spaced at 3 m at the beginning of the computation. Furthermore, the two propagating masses are released at once from the source areas, thus disregarding the possibility of multiple/delayed failures. The frictional rheological law is used, with the rheological properties ($tan\phi_b = 0.4$) selected first by referring to Pastor et al. (**2009**) and then with different hypotheses considered for the erodible areas (A_{er}), and E_r. In particular, the propagation path is divided into three zones: hillslope, channel and piedmont. The numerical simulations consider erosion in channel and piedmont zones, or only in the channel zone. Moreover, different E_r values, ranging from 9×10^{-4} to $1.3 \times 10^{-3}\,m^{-1}$, are used to back-analyse the case studies. The initial pore water pressure normalised to soil liquefaction pressure ($p_w^{rel} = p_w / \gamma_{sat} \cdot h$), where p_w^{rel} is the so-called 'normalised pore water pressure' and γ_{sat} is the soil unit weight) is assumed equal to 1.0 inside the landslide source area. An automatic adaptive time stepping is used for time discretisation (Pastor et al., **2002**) with time steps shorter than 0.8 s. The Runge-Kutta algorithm is used for numerical time integration, as suggested by Pastor et al. (**2009**).

The results show that the bed entrainment greatly modifies the landslide propagation pattern (Fig. **9**). In case 1, only the channel is erodible and the simulated landslide travels mainly at the right-hand side. In case 2, the propagating mass entrains material along the whole propagation path and the simulated bed entrainment causes the material deposition and reduces the landslide run-out distance. In both cases, SPH modelling provides distinct propagation areas, similar run-out distances and run-outs shorter than the observed one of about 400 m. Assuming the highest E_r value (case 3), the field evidence is poorly reproduced, as in the model the landslide stops at the exit of the channel where a thick deposit is simulated and the propagation path observed at the piedmont in the case study is not captured. However, the results show that bed entrainment slightly modifies the duration of the whole propagation/deposition stage (45 to 50 s for cases 1–3). Moreover, the comparison with the case 4 highlights the important role played by the initial height of the water table ($h_w^{rel} = 0.4$ instead of 0.25).

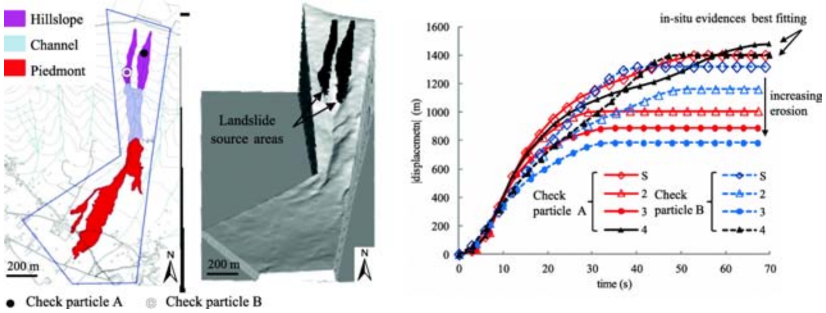


Figure 9. The case of two debris flows converging in the same valley (Cascini et al., **2014**).

The numerical results satisfactorily reproduce the in-situ evidence for both the run-out distance and the extent of the propagation-deposition zones. The simulated phenomenon lasted about 60 s, which is in agreement with Pastor et al. (**2009**) and eyewitness accounts of inhabitants (Cascini et al., **2005**).

Modelling the Propagation Stage of Debris Avalanches (DAs)

Lateral spreading combined to the bed entrainment is another fundamental mechanism governing the propagation stage. To assess the roles of entrainment, frictional basal resistance and pore water pressure in the lateral spreading of the propagating mass, an ideal slope is parametrically analysed. The slope consists of two planes dipping at i_1 and i_2 (Fig. **10**). The failed volume is located at the uppermost edge of the upper slope, inside the source area. The propagation area of a debris avalanche is analysed with reference to the semi-apical angle (β) computed from the lateral boundary of the debris avalanche to its axis at the source area. Other important features, such as the angle of reach (α) formed by the line connecting the uppermost point of the landslide crown scarp to the tip of the mass deposit in a longitudinal section, are not investigated here, as they also depend on piedmont characteristics (Cascini et al., **2011b**).

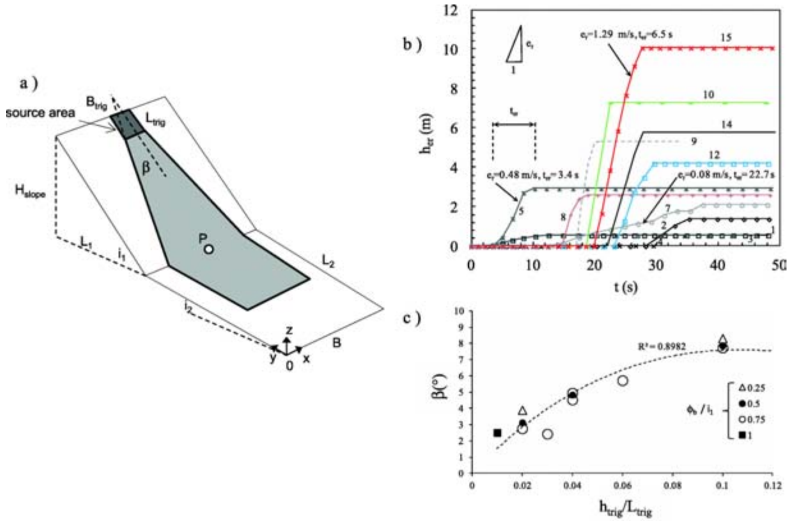


Figure 10. Schematic (**a**) of an open slope affected by a debris avalanche: modeled eroded thickness (**b**), and lateral spreading (**c**) (Cuomo et al., **2014**).

Several analyses of frictional-like materials are performed by varying the morphometric features of the hillslope (i_1, i_2, H_{slope}), the geometrical aspect ratio of the source area (B_{trig}, L_{trig}, h_{trig}) and the main rheological parameter (the friction angle of the propagating mass, $\square_b$). A fixed value for landslide growth rate (E_r) is used to account for the entrainment phenomena. The results indicate that the greater the ratio of the triggering soil height to the length of the source area (h_{trig} / L_{trig}), the greater the lateral spreading (β), with a maximum of 8.3°. The 8.3° maximum corresponds to a triggering soil height of 5 m. Such a high h_{trig} value is likely to occur in Zone 2 of the slope shown in Fig. **2** due to the impact of material falling from a bedrock scarp.

Figure **10** provides an example of these results with L_{trig}, B_{trig}, $\square_b$, c_v and E_r fixed at 50 m, 40 m, 10.2°, 0.01 m^2/s and 8.2×10^{-3} m^{-1}, respectively. The semi-apical angle (β) increases from 1.3 to 5.2° until the ratio B_{trig}/L_{trig} reaches 0.5, and then β reduces to a minimum value of 3.2°, independent of relative pore water pressure (p_w^{rel}).

It is also worth showing the time trend in simulated eroded depths at point ‘P’, at the boundary between slope and piedmont. The final eroded depths (h_{er}) range from 1 m to 10 m, with an erosion rate (e_r) ranging from 0.08 to 1.29 m/s, and an erosion time (t_{er}, defined as the time in which bed entrainment occurs at a given point of the slope) ranging from 3.4 to 22.7 s.

The eroded depths simulated at the boundary between slope and piedmont show two key characteristics: (i) they are the product of a combination of slope morphology, features of the triggering area, rheology and bed entrainment; and (ii) they range between 0.03 and 10.07 m for a wide array of debris avalanches in coarse-grained soils. Therefore, the results of the benchmark cases facilitate assessing the roles and interplay of entrainment, rheology and pore water pressure, and provide theoretical values for apical angle (β), erosion rate (e_r), eroded depth (h_{er}) and erosion time (t_{er}) in highly idealised cases. Using these results, the analysis of relevant case histories in the following sections can be approached with confidence.

A debris avalanche triggered at the uppermost part of a hillslope may propagate into a well established channel or even spread into two or more valleys. The latter case is recorded on 5 May 1998 at the Pizzo d'Alvano massif (about 1000 m high), in the Cortadonica basin. A debris avalanche is triggered at 745 m a.s.l., enlarged along the hillslope at a semi-apical angle (β) of about 7°, travelled for 510 m, then divided in two wide valleys. It propagated over a total run-out distance of 1.95 km up to the piedmont area at 65 m a.s.l.. The numerical analysis of this case is performed using a 3 × 3 m Digital Elevation Model. The topography is reproduced by means of a mesh of 35,520 squares. The initial mass is schematised into a set of 639 SPH points, 1 m spaced, with a uniform soil height of 1–2 m over the impact zone (data from Cascini et al., 2008). A frictional model is used to analyse the rheological behaviour of the unstable mass, based on the rheological parameters used by Pastor et al. (2009) to back-analyse an important channelised landslide that occurred during the May 1998 event in a neighbouring mountain basin. The landslide growth rate is assumed to be in the range $1.3 \times 10^{-4} \div 8.2 \times 10^{-2}$, which is similar to the rate of the Nocera Inferiore landslide, due to important similarities between either morphometric hillslope features or soil mechanical parameters in the two areas under study (Cascini et al., 2013a).

The results shown in Fig. 11 provide a satisfactory simulation of the observed behaviour of the landslide, especially in terms of the lateral boundary of the debris avalanche and the splitting of its initial mass into two channels. The estimated landslide growth rate is 4.0×10^{-3}.. The simulated erosion rate (e_r) is 0.57 m/s and the simulated erosion time (t_{er}) is 2.5 s. All of the results achieved for the Cortadonica debris avalanche show that the greater the friction angle or erosion growth rate, the higher the simulated eroded heights (h_{er}); similarly, if the consolidation coefficient (c_v) increases, the depth of erosion increases. Moreover, it is shown that bed entrainment decreases if the water-table height increases (h_w^{rel}). These results are consistent with those obtained for the previous benchmark cases and other case histories.

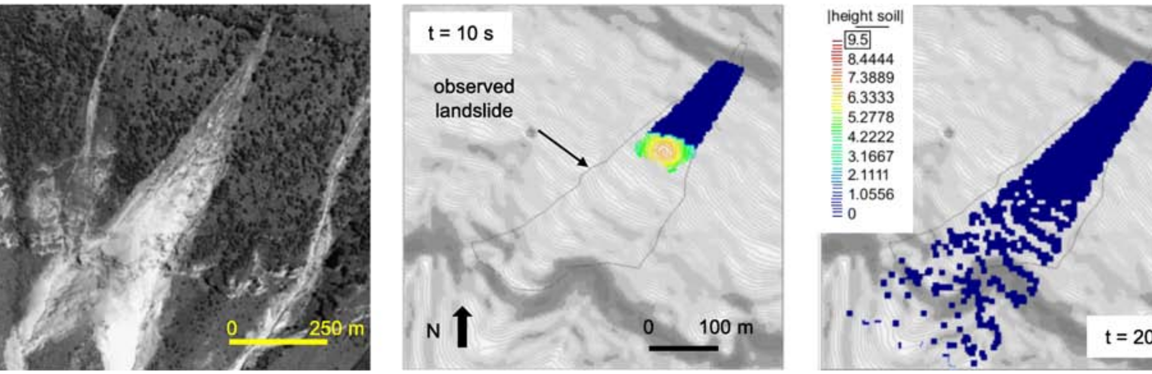


Figure 11. Modelling of the propagation height of a debris avalanche at different time lapses for the case history of Cortadonica catchment (Italy) (Cascini et al., 2013a).

Modelling the Propagation Stage of Combined Flows

Different types of flows can occur in nearby locations and nearly at the same time, so that multiple soil volumes can join and propagate together. The numerical modelling is here conducted for a series of very small (1088 m^3) to medium-sized (11,630 m^3) landslides. They were recorded at Bracigliano site, approximately at 2 p.m. on May 5, 1998, along the hills to the northwest of town (Monte Faitaldo and Monte Foresta), where the largest landslide occurred (950 m a.s.l.). Different triggering mechanisms and types of source areas are identified by Cascini et al. (2008), including the following: *M1*, colluvial hollows with convergent sub-superficial groundwater circulation and temporary springs from bedrock; *M2*, triangular areas at open slopes associated with outcropping or buried bedrock scarps; *M6*, areas shaped like short and thick spoons situated at either the base of the convex–concave hillslopes or along the flanks of the inner gorges. Among these different source areas, Cascini et al. (2013a) identified two debris avalanches. The source areas are generally at elevations between 800 and 900 m, and slope failures involved, in some cases, the entire thickness of the pyroclastic cover. The numerous detachments induced debris flows that converged in one main gully, exiting in urban roads and causing loss of life and widespread damage to buildings. The flows reached high water content owing to the runoff along the channels and urban roads: this explains the unusually large shape of the deposition zone and the long run-out distance of the flow, which reached the near city of Siano. The rheological properties and bed entrainment rate (*K*) are calibrated to best fit the extent of the propagation area, from the uppermost slopes to the urbanized area located at the piedmont. Here, a dense network of paved roads and narrow streets is present, and an adequate modelling of landslide propagation would require a finer DTM and, for instance, very accurate specific information about the hydraulic works; this is certainly beyond the scope of this paper. Particularly, the attention is focused on the upper part of the right-hand side of Monte Foresta, where 2 debris avalanches (M2) and 9 debris flows (M1) were triggered between 800 and 900 m a.s.l.. In fact, the eyewitnesses and in situ evidence shows that from the left-hand side of the catchment, a flood arrived, which caused the enlargement of the landslide body within the urban zone.

The numerical analyses are based on a DTM of 939,330 squares, each 3 × 3 m in size. The 11 unstable masses are schematized into 11 sets of SPH computational points for a total of 2905 points. The initial soil height in source areas is 2.5 m or 1.5 m, depending on the triggering mechanisms M1 and M2, respectively. The rheological parameters are the same as those chosen for the numerical simulations of the Cortadonica catchment, owing to the proximity of the two sites and to compare the results. Two opposite results are simulated (Fig. **12**): a) the landslides do not reach the piedmont, if the water table height in the source areas (h_w^{rel}) is assumed to be lower than 0.4 or if the bed entrainment rate (*K*) is higher than 0.007; b) the landslide overcomes the left-hand side boundary of the propagation path if h_w^{rel} is higher than 0.4 or *K* is lower than 0.007. This entails that $h_w^{rel} = 0.4$ and $K = 0.007$ are found to be the best-fitting values for the propagation back-analysis of the event. In addition, $tan(\phi_b)$, p_w^{rel} and c_v are the same as the nearby Cortadonica catchment. This is also a consistency check of the parameters used for rheology and bed entrainment at the three sites.

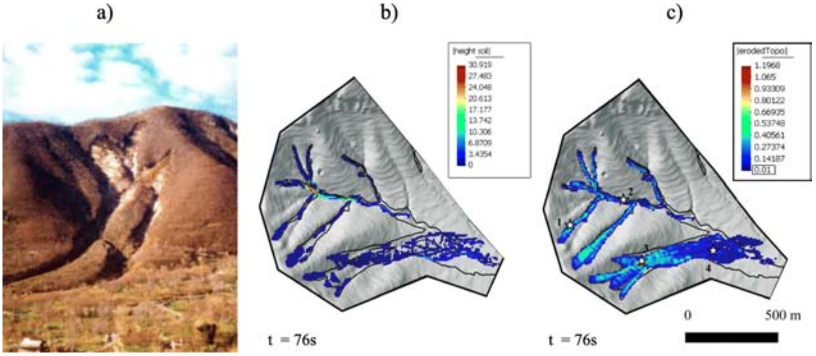


Figure 12. a Case history of combined flows at Bracigliano site, (**b**) deposition depths and (**c**) eroded depths (Cuomo et al., **2016**).

Particularly, the simulated propagation heights shown in Fig. **13** reproduce quite well the landslide lateral boundaries and the path observed in the field, with all masses propagating in distinct channels and stopping at the uppermost boundary of the urban area. It is also important to note that the eroded thicknesses have high spatial variation (Fig. **12**), with the maximum erosion depth simulated at the right-hand side of the boundary of the catchment, as observed in situ. As in the previous cases, the simulated entrainment increases when moving from the top to the toe of the massif, but then entrainment drastically decreases where steepness diminishes. In fact, a lower slope angle has a direct effect on reducing the bed entrainment and also an indirect effect because a lower slope angle favours the lateral spreading of flow - as shown in Cuomo et al., **2014**) - and causes a general reduction in the propagation heights, which once again diminishes the bed entrainment. Finally, it is worth noting that the plots of h_{er} versus t–t_{flow} (Fig. **12**) are inconsistent with the values ($0.08\,\text{m/s} < e_r < 1.29\,\text{m/s}$) indicated by Cuomo et al. (**2014**) for DAs; this is clearly because the simulated events are a combination of several DFs and two DAs.

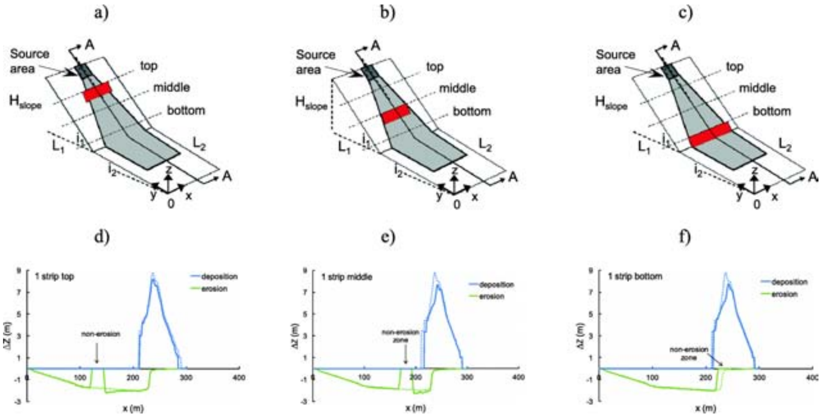


Figure 13. Schemes of installation of erosion control zones (**a-c**) and respective results (**d-f**) in terms of elevation change (**d-f**) along the slope.

RESULTS AND DISCUSSION FOR ENGINEERED SLOPES

Analysis of the Effect of Erosion Control

One potential solution to reduce the volume of debris avalanches is represented by the construction of erosion control installation. The desired effect is to eliminate the bed entrainment at designed locations. Several analyses are carried on a schematic open slope, consisting of two planes with inclines to the horizon i_1 and i_2, respectively (Fig. **13**). Different arrangements of non-erodible zone are considered. Three relevant combinations are proposed for the numerical simulations, as those depicted in Fig. **13**. In all the cases, inside a non-erodible zone, large as the distance from the uppermost baffle to the lowermost one, the prevention of entrainment is guaranteed. On the other hand, the overall benefit and the eventual side effects are here evaluated through the numerical modeling.

The source area is located at the uppermost edge of the upper slope. The numerical analysis are performed using a 1 m × 1 m Digital Terrain Model (DTM). The initial mass is schematized into a set of 544 SPH points, 1 m spaced, with a uniform soil thickness of 1 m over the failure zone. A frictional rheological model is used with parameters taken from literature ($tan\phi_b=0.5$, $h_w^{rel}=0.4$, $p_w^{rel}=0.5$, $c_v=10^{-2}\,m^2\,s^{-1}$, $K=0.03$). A sensitivity analysis is also conducted changing both the slope inclination (30–40°) and the initial volume (500, 5000, 10,000 or 15,000 m^3).

In order to quantify the reduction in the eroded soil thickness, a longitudinal section is represented (Fig. **13**), where we can see the longitudinal profile of the slope before and after the flow propagated, with also the erosion heights represented. The entrainment rate (Δz/x) is almost the same upslope and downslope the erosion control areas. It means that the landslide dynamic is poorly modified. This is observed in all the combinations. The erosion control has major local effect while smaller general consequence on the landslide.

Based on that, one can say that the higher the extent of the non-erodible areas, the higher the benefit of this countermeasure, as the volume reduction relates exclusively to the extent of the treated area. Of course, the closer the intervention is to the toe of the slope, the less is the erosion, as the faster is the landslide when it reaches the control work. This last observation is valid for the examined cases of relatively "short" slopes, some 300–400 m long. The benefit in terms of volume reduction is negligible for small-medium sized landslides, and does not exceed the 18% for the biggest ones here considered.

Modelling the Benefits of Artificial Baffles

Other types of control works are more focused to change the dynamics of flow propagation. The destructiveness of a debris avalanche can be mitigated, for instance, by obstacles along the flow path as they can slow or even stop the flow. This kind of obstacles can be natural, for example big trees or boulders, or artificial such as rigid or flexible barriers, or concrete columns known as baffles. Along the flow path two rows of rectangular obstacles have been positioned. Different combinations of these obstacles changing both disposition and position are considered. The presence of the obstacles is taken into account in the simulation by considering nil normal velocity along the obstacle boundaries. The same rheological properties of the previous section are here considered.

The first analysis is carried out to understand how these obstacles and their position can change the dynamics of the debris avalanche. Referring to the schematic slopes of Fig. **14**, three different cases are analysed different for the distance of the obstacles from the source area (*L*). We will have, therefore, the first case with the obstacles in the upper zone of the slope (Fig. **14**a), the second one with the obstacles in the middle of the slope (Fig. **14**c) and in the final case they are positioned near the break of the slope (Fig. **14**d). Moreover for the first case, a reverse position of the obstacles is also analysed, with two obstacles in the first row impacted and three in the second one (Fig. **14**b).

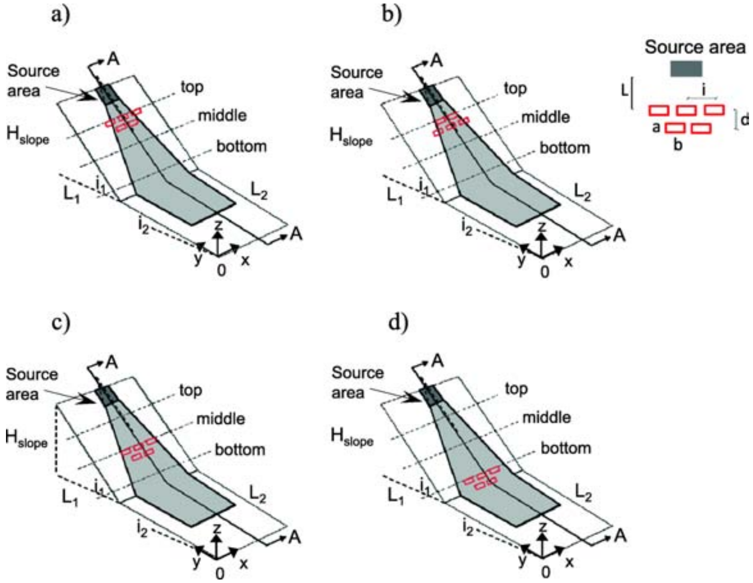


Figure 14. Different configurations (**a-d**) of multiple artificial baffles along an open slope.

The installation of the baffles highly changes i) the eroded depths along the slope, ii) the runout and, iii) the final deposition thicknesses. Regarding the former issue, it is worth noting that the debris avalanche entrains material at a nearly constant rate (Δz/x) in the upper part of the slope. Then, a drastic reduction of entrainment occurs at the baffle location, as expected. More interestingly, the debris avalanche starts to entrain material again downslope the baffles. The material is entrained at a lower rate downslope (i.e. after the interaction with) the baffles. It means that the baffles completely modify the dynamics of landslide as desired. Of course, this drastic change would be positive in case the runout and deposition at the toe of the slope are both reduced. Such expectations are confirmed in Fig. **15**. The runout is decreased and some of the soil volume is trapped behind the barrier for any baffle combination.

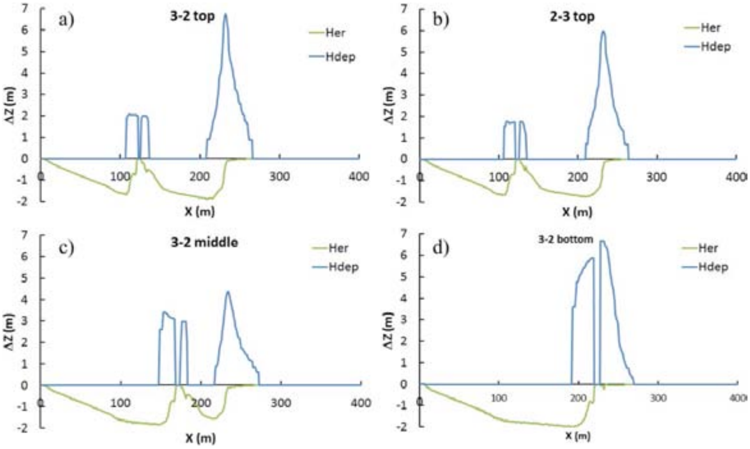


Figure 15. Effects (**a-d**) of the artificial baffles on the eroded and deposition depths along an open slope.

The decrease of velocity due to the impact of the flow on the obstacles, the anti-erosion effect and the capacity of the obstacles to contain part of the flow, permit a reduction in the final mobilised volume. To evaluate this issue, the amplification factor A_f is introduced as the rate of the final mobilised volume (V_f) to the initial volume (V_i).

The landslide cumulated volume is plotted versus and also the final amplification factor is reported (Fig. **16**), so that it is possible evaluating how much the landslide increases in different cases, and compared with natural slope (without obstacles).

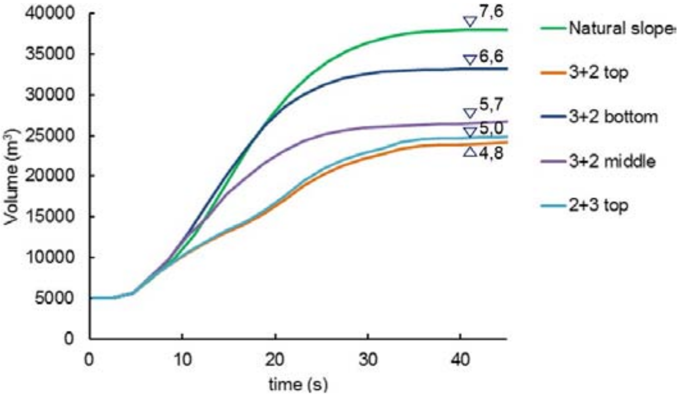


Figure 16. Landslide volume amplification for a natural slope and another equipped with baffles.

Analysis of Artificial Barriers Installed at the Piedmont Areas

In combination or as an alternative to the previous mitigation works, a barrier installed at the toe of the slope can be considered. A schematic open slope is firstly analysed, which is composed of two differently inclined planes and a debris avalanche triggered at the uppermost portion of the slope. The computational scheme and the soil properties are taken from Cuomo et al. (**2014**), who extensively investigated the role of the several factors involved in the propagation stage of a debris avalanche.

The slopes are inclined at 30° or 40° with different lengths (horizontal projection) L_1 (Fig. **17**). The piedmont zone is flat or gently inclined (10° steep) with length L_2. The length and the width of the source area are L_{trig} and B_{trig}, respectively, and H_{trig} is the initial height of soil inside the source area. A selection of the several numerical simulations are with $L_1 = 230$ m, $L_2 = 500$ m, the width of the slope (B) equal to 800 m, and the slope height (H_{slope}) equal to 222 m or 130 m for $\alpha_p = 10°$ or $\alpha_p = 0°$, respectively. The DTM cell size is equal to 1.1 m for both slopes, inclined with 40° and 30°. One or more barriers are added in the piedmont zone. Each barrier is 5 m high (H), with top width (b) equal to 3 m, the upslope raceway (a) 3 m wide, and both lateral scarps inclined at 60°. The Type I barrier has a trapezoidal shape; the Type II barrier is similar but with an additional step (H/2 high, and large as b) located upslope. In the simulations, the first barrier is in the piedmont zone, specifically 10 m (x = 240 m) or 25 m (x = 255 m) or 50 m (x = 280 m) downslope the divide between the slope and the piedmont.

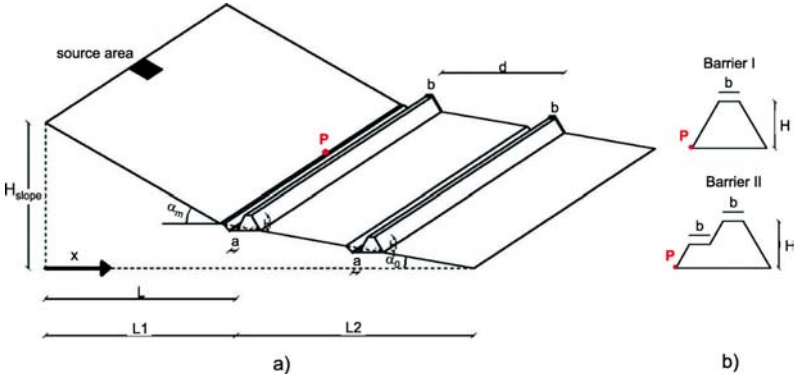


Figure 17. Scheme of the artificial barriers considered at the toe of the slope: **a**) overview, **b**) cross sections (Cuomo et al., **2019c**).

Different sets of soil properties, such as the soil unit weight (γ), the friction angle ($tan\phi_b$=0.30 or 0.52), the initial height of water table divided by the soil thickness (h_w^{rel}=0.40 or 0.75), the initial value of relative pore water pressure (p_w^{rel}=0.5 or 1.0), the dimensions of the source area (L_{trig}=25 m or 100 m; B_{trig}=10 m or 50 m), and the initial height of the flow (h_{trig}=1.0 or 4.0 m) are taken from Cuomo et al. (**2014**), resembling the features of catastrophic events that occurred in Southern Italy, such as those of Cervinara in 1999 (Cascini et al., **2011a**), and Nocera Inferiore in 2005 (Cuomo et al., **2014**).

The computational points are initially spaced 1.1 m and the time step is 0.5 s. Two parameters are referred for the flow propagation analysis, namely the Index of Piedmont Runout Reduction ($\mathrm{I_{PRR}}$) and the Index of Lateral Spreading ($\mathrm{I_{LS}}$), which read as: $I_{PRR}=PR_{eng}/PR_{nat}, I_{LS}=W_{eng}/W_{nat}$, where PR_{eng} is the Piedmont Runout distance travelled by the flow inside the piedmont zone engineered with barriers, PR_{nat} is the runout inside the piedmont zone for the natural slope, W_{eng} is the maximum lateral width of the flow behind the barrier for the engineered slope, and W_{nat} is the analogous feature of the flow computed at the same point for the natural slope.

A value of $\mathrm{I_{PRR}}<1.0$ is desirable, and the lower $\mathrm{I_{PRR}}$, the better the efficiency of the barrier. $\mathrm{I_{PRR}}$ also depends on where the barriers are located. A barrier favours the flow material to spread laterally and it is expected that $\mathrm{I_{LS}}>1.0$. For multiple barriers, $\mathrm{I_{LS}}$ is computed with the highest $\mathrm{W_{eng}}$ obtained for each barrier. The computed values of I_{PRR} and I_{LS} are reported in Fig. **18** for all cases. Four zones can be individuated in the plots: 1) $I_{PRR}<1.0$ and $I_{LS}<1.0$, i.e. both the runout and the width decrease, meaning that the barrier is effective. This is an unlikely condition; 2) $\mathrm{I_{PRR}}<1.0$ and $\mathrm{I_{LS}}>1.0$, i.e. the runout diminishes while the width increases, meaning the barrier is still effective. This is a very likely condition; 3) $\mathrm{I_{PRR}}>1.0$ and $\mathrm{I_{LS}}>1.0$, i.e. both runout and the width increase and thus the barrier is ineffective in terms of reduction of runout; 4) $\mathrm{I_{PRR}}>1.0$ and $\mathrm{I_{LS}}<1.0$, i.e. there is a reduction of width and an increase of runout, so that the barrier is ineffective. However, this condition is unrealistic. For two barriers, we considered the maximum width of flow in the plane-view. The computed runout is always reduced with one or two barriers, irrespective of overtopping. In general, runout can be reduced to 70% (Case S3) with a maximum increase of lateral spreading of 5% compared to the natural slope. Furthermore, the barrier type differently influences the area affected by the flow. In particular, $\mathrm{I_{PRR}}$ decreases, passing from Type I to Type II for the same position of the barriers (Case S4 and Case S5, or Case R17 and Case R18). The barrier type does not influence

I_{PRR} for barriers located very far from the landslide source area (Cases S6 and S7, R24 and R25).

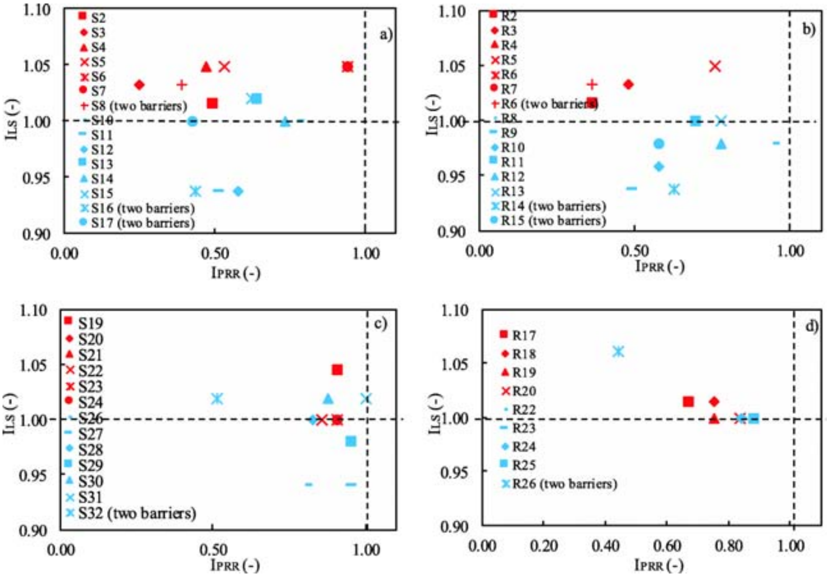


Figure 18. Effects of artificial barriers on Piedmont Runout Reduction (I_{PRR}) and Lateral Spreading (I_{LS}) (Cuomo et al., **2019c**).

CONCLUSIONS

Numerical modelling is a powerful tool to understand and forecast heights and velocities, given that all these variables change very rapidly and are spatially distributed. This is even truer considering that unrevealed propagation patterns have been observed in real case histories and small-scaled laboratory experiments. Notwithstanding the complexity of flow-like landslides and the related challenges for modelling, the understanding and forecasting of such natural hazards is achievable with a satisfactory confidence. Among the key factors, rainfall, pore water pressure and bed entrainment deserves a special attention. Thus, the paper provided a number of examples regarding that. Further improvements are expectable as the numerical models are becoming more efficient. Thus, more accurate descriptions of local effects will be possible and also additional mechanisms will be possibly analysed.

On the other hand, control works and engineering countermeasures represent one option for risk mitigation and disasters reduction. In this sense, intervention along the slopes or at the piedmont areas may be conceived depending on many other aspects such as, for instance, the feasibility of concrete structures, the costs, and the acceptance of resident populations. The paper compares three mitigation options in a relatively small set of simplified cases. More investigation could be useful to generalize the range of mitigation opportunities also for real case histories.

More in general, the recent increased understanding of those tremendous hazards, and the availability of accurate simulation instruments should also increase the awareness of specialists and populations about the fact that the mitigation of geoenvironmental disasters is not an optional topic to be considered, but a fundamental issue to be mandatorily tacked by the new generations.

Abbreviations

ALE:	Arbitrary Lagrangian Eulerian method
c':	effective cohesion
c_v**:**	consolidation coefficient
DA :	Debris Avalanche
DEM:	Discrete Element Method
DF :	Debris Flow
E_r**:**	erosion coefficient
e_r**:**	entrainment rate
FEM:	Finite Element Method
FEMLIP:	Finite Element Method with Lagrangian Integration Point
h :	flow depth
h_w^{rel}**:**	height of water table normalized to soil thicknees
K :	bed entrainment parameter
k_{sat}**:**	saturated conductivity
LEM:	Limit Equilibrium Methods
MPM:	Material Point Method
n :	porosity
p':	effective isotropic stress
p^b_w**:**	basal pore water pressure

p_w^{rel} :	ratio of pore water pressure to liquefaction pressure
q :	deviatoric stress
s :	suction
SPH:	Smoothed Particles Hydrodynamics
S_r :	saturation degree
v :	flow velocity
γ_{sat} :	saturated unit weight
φ':	friction angle
ϕ_b :	basal friction angle

ACKNOWLEDGEMENTS

Prof. Manuel Pastor (Universidad Politecnica de Madrid, Spain) and co-workers are much acknowledged for precious suggestions and fruitful cooperation and also for having provided the “GeHoMadrid” code and the “GeoFlow_SPH” code used for the numerical simulations. The results shown in the paper comes from a long-lasting research developed by the Geotechnical Engineering Group (GEG) of University of Salerno. Fundamental contributions were given over the years by Prof. Leonardo Cascini, Prof. Giuseppe Sorbino, Dr. Luca Piciullo, Dr. Vittoria Capobianco, Dr. Ilaria Rendina, Eng. Giuseppe Claudio Castorino and Eng. Stefano Petrosino.

Authors’ Contributions

Sabatino Cuomo is responsible for the data collection, numerical modelling and the corresponding passages in the manuscript. The author read and approved the final manuscript.

Funding

The research was supported by a number of Italian Research Projects funded by the Italian Education and Research Ministry such as: Project FARB 2017 “Numerical modelling and inverse analysis for flow-like landslides”; Project FARB 2014 “Large area analysis of triggering and propagation landslide susceptibility for flow-like landslides”; Project FARB 2012 “New Frontiers of advanced numerical simulation of destructive landslides”.

REFERENCES

1. Alonso, E., A. Gens, A. Lloret, and C. Delahaye. 1996. *Effect of rain infiltration on the stability of slopes*, 241–249. Paris: Alonso & Delage eds.
2. Anderson, A., and N. Sitar. 1995. Analysis of rainfall-induced debris flow. *Journal of Geotechnical Engineering* 121 (7): 544–552.
3. Anderson, S.A., and M.F. Riemer. 1995. Collapse of saturated soil due to reduction in confinement. *Journal of Geotechincal Engineering ASCE* 121 (2): 216–220.
4. Bilotta, E., V. Foresta, and G. Migliaro. 2006. Suction controlled laboratory tests on undisturbed pyroclastic soil: Stiffnesses and volumetric deformations. In *Proc. international conference on unsaturated soils, 2–6 April, carefree, Arizona USA, 1*, 849–860.
5. Blanc, T. 2011. *A SPH depth integrated model with pore pressure coupling for fast landslides and related phenomena*. PhD Thesis (Madrid), 292.
6. Blanc, T., and M. Pastor. 2011. A stabilized Smoothed Particle Hydrodynamics, Taylor-Galerkin algorithm for soil dynamics problems. *International Journal for Numerical and Analytical Methods in Geomechanics* Published online in Wiley Online Library (**wileyonlinelibrary.com**). **https://doi.org/10.1002/nag.1082**.
7. Blanc, T., and M. Pastor. 2012a. A stabilized Runge Kutta, Taylor smoothed particle hydrodynamics algorithm for large deformation problems in dynamics. *International Journal for Numerical Methods in Engineering* 91 (issue 13): 1427–1458.
8. Blanc, T., and M. Pastor. 2012b. A stabilized fractional step, Runge Kutta Taylor SPH algorithm for coupled problems in Geomechanics. *Computer Methods in Applied Mechanics and Engineering* 221 (222): 41–53.
9. Blanc, T., M. Pastor, V. Drempetic, and B. Haddad. 2011. Depth integrated modelling of fast landslides propagation. *European Journal of Environmental and Civil Engineering* 15: 51–72.
10. Braun, A., S. Cuomo, S. Petrosino, X. Wang, and L. Zhang. 2018. Numerical SPH analysis of debris flow run-out and related river damming scenarios for a local case study in SW China. *Landslides* 15 (3): 535–550.

11. Braun, A., X. Wang, S. Petrosino, and S. Cuomo. 2017. SPH propagation back-analysis of Baishuihe landslide in south-western China. *Geoenvironmental Disasters* 4 (1): 2.
12. Brufau, P., P. Garcìa-Navarro, P. Ghilardi, L. Natale, and F. Savi. 2000. 1D mathematical modelling of debris flow. *Journal Rech Hydraul* 38: 435–446.
13. Cascini, L. 2004. The flowslides of may 1998 in the Campania region, Italy: The scientific emergency management. *Italian Geotechnical Journal* 2: 11–44.
14. Cascini, L., S. Cuomo, and A. De Santis. 2011a. Numerical modelling of the December 1999 Cervinara flow-like mass movements (Southern Italy). *Italian Journal of Engineering Geology and Environment*: 635644.
15. Cascini, L., S. Cuomo, and M. Della Sala. 2011b. Spatial and temporal occurrence of rainfall-induced shallow landslides of flow type: A case of Sarno-Quindici, Italy. *Geomorphology* 126: 148–158.
16. Cascini, L., S. Cuomo, A. Di Mauro, M. Di Natale, S. Di Nocera, and F. Matano. 2019. *Multidisciplinary analysis of combined flow-like mass movements in a catchment of Southern Italy*, 1–18. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards.
17. Cascini, L., S. Cuomo, and D. Guida. 2008. Typical source areas of May 1998 flow-like mass movements in the Campania region, Southern Italy. *Engineering Geology* 96 (3-4): 107–125.
18. Cascini, L., S. Cuomo, and M. Pastor. 2013a. Geomechanical modelling of debris avalanches inception. *Landslides* 10 (6): 701–711.
19. Cascini, L., S. Cuomo, M. Pastor, and I. Rendina. 2016. SPH-FDM propagation and pore water pressure modelling for debris flows in flume tests. *Engineering Geology* 213: 74–83.
20. Cascini, L., S. Cuomo, M. Pastor, and C. Sacco. 2013b. Modelling the post-failure stage of rainfall-induced landslides of the flow type. *Canadian Geotechnical Journal* 50 (9): 924–934.
21. Cascini, L., S. Cuomo, M. Pastor, and G. Sorbino. 2010. Modeling of rainfall-induced shallow landslides of the flow-type. *Journal of Geotechnical and Geoenvironmental Engineering* 136 (1): 85–98.

22. Cascini, L., S. Cuomo, M. Pastor, G. Sorbino, and L. Piciullo. 2014. SPH run-out modelling of channelized landslides of the flow type. *Geomorphology* 214: 502–513.
23. Cascini, L., S. Cuomo, and G. Sorbino. 2005. Flow-like mass movements in pyroclastic soils: Remarks on the modelling of triggering mechanisms. *Italian Geotechnical Journal* 4: 11–31.
24. Chen, H., G.B. Crosta, and C.F. Lee. 2006. Erosional effects on runout of fast landslides, debris flows and avalanches: A numerical investigation. *Geotechnique* 56: 305–322.
25. Chu, J., S. Leroueil, and W.K. Leong. 2003. Unstable behaviour of sand and its implications for slope instability. *Canadian Geotechnical Journal* 40: 873–885.
26. Corominas, J. 1996. The angle of reach as a mobility index for small and large landslides. *Canadian Geotechnical Journal* 33: 260–271.
27. Crosta, G.B., S. Imposimato, and D.G. Roddeman. 2009. Numerical modelling of entrainment/deposition in rock and debris-avalanches. *Engineering Geology* 109: 135–145.
28. Cuomo, S. 2006. *Geomechanical modelling of triggering mechanisms for flow-like mass movements in pyroclastic soils*, 274. Italy: PhD dissertation at the University of Salerno.
29. Cuomo, S., A. Di Perna, P. Ghasemi, M. Martinelli, and M. Calvello. 2019a. Combined LEM and MPM analyses for the simulation of a fast moving landslide in Hong Kong. In *Proc. of II International Conference on the Material Point Method for modelling soil–water–structure interaction. 8–10 January 2019, University of Cambridge, UK*.
30. Cuomo, S., P. Ghasemi, M. Martinelli, and M. Calvello. 2019b. Simulation of Liquefaction and Retrogressive Slope Failure in Loose Coarse-Grained Material. *International Journal of Geomechanics* 19 (10): 04019116.
31. Cuomo, S., S. Moretti, and S. Aversa. 2019c. Effects of artificial barriers on the propagation of debris avalanches. *Landslides* 16 (6): 1077–1087.
32. Cuomo, S., M. Moscariello, D. Manzanal, M. Pastor, and V. Foresta. 2018. Modelling the mechanical behaviour of a natural unsaturated pyroclastic soil within generalized plasticity framework. *Computers and Geotechnics* 99: 191–202.

33. Cuomo, S., M. Pastor, V. Capobianco, and L. Cascini. 2016. Modelling the space–time evolution of bed entrainment for flow-like landslides. *Engineering Geology* 212: 10–20.

34. Cuomo, S., M. Pastor, L. Cascini, and G.C. Castorino. 2014. Interplay of rheology and entrainment in debris avalanches: A numerical study. *Canadian Geotechnical Journal* 51 (11): 1318–1330.

35. Cuomo, S., M. Pastor, S. Vitale, and L. Cascini. 2013. Improvement of irregular DTM for SPH modelling of flow-like landslides. Proc. of XII International Conference on Computational Plasticity. Fundamentals and Applications (COMPLAS XII), E Oñate*,* DRJ Owen, D Peric and B Suárez*.* 3–5 September 2013, Barcelona, Spain*.* ISBN: 978–84–941531-5-0, 512–521.

36. Dai, F., C.F. Lee, S. Wang, and Y. Feng. 1999. Stress-strain behaviour of a loosely compacted volcanic-derived soil and its significance to rainfall-induced fill slope failures. *Engineering Geology* 53: 359–370.

37. Darve, F., and F. Laouafa. 2000. Instabilities in granular materials and application to landslides. *Mechanics of Cohesive frictional Materials* 5 (8): 627–652.

38. Duncan, J.M. 1996. State of the art: Limit equilibrium and finite element analysis of slopes. *Journal of Geotechnical Engineering, ASCE* 122 (7): 557–596.

39. Eckersley, D. 1990. Instrumented laboratory flowslides. *Géotechnique* 40: 489–502.

40. Egashira, S., N. Hondab, and T. Itohc. 2001. Experimental study on the entrainment of bed material into debris flow. *Physics and Chemistry of the Earth, Part C* 26: 645–650.

41. Egashira, S., T. Itoh, and H. Takeuchi. 2000. Transition mechanism of debris flows over rigid bed to over erodible bed. *Physics and Chemistry of the Earth, Part B* 26: 169–174.

42. Faug, T., P. Caccamo, and B. Chanut. 2011. Equation for the force experienced by a wall overflowed by a granular avalanche: Experimental verification. *Physical Review* 84 (051301): 1–18.

43. Favier, L., D. Daudon, F.V. Donzé, and J. Mazars. 2009. Predicting the drag coefficient of a granular flow using the discrete element method. *Journal of Statistical Mechanics: Theory and Experiment* **https://doi.org/10.1088/1742-5468/2009/06/P06012**.

44. Fell, R., J. Corominas, Ch. Bonnard, L. Cascini, E. Leroi, and W.Z. Savage. 2008. On behalf of the JTC-1 joint technical committee on landslides and engineered slopes. Guidelines for landslide susceptibility, hazard and risk zoning for land use planning. *Engineering Geology* 102: 85–98.
45. Fredlund, D.G., N.R. Morgenstern, and R. A Widger. 1978. The shear strength of unsaturated soils. Canadian Geotechnical Journal 15: 313–321.
46. Fukagawa, R., K. Sako, H.H. Bui, and J.C. Wells. 2011. Slope stability analysis and discontinuous slope failure simulation by elasto-plastic smoothed particle hydrodynamics (SPH). *Geotechnique* 61 (7): 565–574.
47. Futai, M.M., W.A. Lacerda, and M.S.S. Almeida. 2004. Evolution of gully processes in unsaturated soils. In *Landslides: Evaluation and Stabilization, Lacerda, Ehrlich, Fontoura &and Sayao (eds)*, vol. 2, 1019–1025.
48. Geoslope. 2005. *User's guide. GeoStudio 2004, Version 6.13*. Calgary: Geo-Slope Int. Ltd.
49. Ghasemi, P., S. Cuomo, A. Di Perna, M. Martinelli, and M. Calvello. 2019. MPM-analysis of landslide propagation observed in flume test. In *Proc. of II International Conference on the Material Point Method for modelling soil–water–structure interaction. 8–10 January 2019, University of Cambridge, UK*.
50. Ghasemi, P., M. Martinelli, S. Cuomo, and M. Calvello. 2018. MPM modelling of static liquefaction in reduced-scale slope. *Numerical Methods in Geotechnical Engineering IX* 2: 1041–1046.
51. Ghilardi, P., L. Natale, and F. Savi. 2001. Modeling debris flow propagation and deposition. *Physics and Chemistry of the Earth, Part C: Solar, Terrestrial & Planetary Science* 26 (9): 651–656.
52. Hungr, O. 1995. A model for the runout analysis of rapid flow slides, debris flows, and avalanches. *Canadian Geotechnical Journal* 32: 610–623.
53. Hungr, O. 2004. Flow slides and flows in granular soils. In *Proc. of the Int. Workshop Flows 2003 - Occurrence and Mechanisms of Flows in Natural Slopes and Earthfill, Sorrento, patron Ed*.

54. Hungr, O., J. Corominas, and E. Eberhardt. 2005. Estimating landslide motion mechanism, travel distance and velocity. In *Landslide Risk Management*, 99–128.

55. Hungr, O., and S.G. Evans. 2004. Entrainment of debris in rock avalanches: An analysis of a long run-out mechanism. *Geological Society of America Bulletin* 116 (9–10): 1240–1252.

56. Hungr, O., S.G. Evans, M.J. Bovis, and J.N. Hutchinson. 2001. A review of the classification of landslides of the flow type. *Environmental and Engineering Geoscience VII* 3: 221–238.

57. Hungr, O., and S. McDougall. 2009. Two numerical models for landslide dynamic analysis. *Computers & Geosciences* 35: 978–992.

58. Hutchinson, J.N. 1986. A sliding-consolidation model for flow slides. *Canadian Geotechnical Journal* 23: 115–126.

59. Iverson, R.M. 2000. Landslide triggering by rain infiltration. *Water Resources Research* 367: 1897–1910.

60. Iverson, R.M., M.E. Reid, and R.G. LaHusen. 1997. Debris flow mobilization from landslides. *Annual Review of Earth and Planetary Sciences* 25: 85–138.

61. Iverson, R.M., M.E. Reid, M. Logan, R.G. LaHusen, J.W. Godt, and J.P. Griswold. 2011. Positive feedback and momentum growth during debris-flow entrainment of wet bed sediment. *Nature Geoscience* 4 (2): 116.

62. Jaky, J. 1944. The coefficient of earth pressure at rest. *Journal of the Society of Hungarian Architects and Engineers*: 355–358.

63. Janbu, N. 1954. Application of composite slip surface for stability analysis. In *European Conference on Stability Analysis, Stockholm, Sweden*.

64. Johnson, K.A., and N. Sitar. 1990. Hydrologic conditions leading to debris-flow initiation. *Canadian Geotechnical Journal* 27 (6): 789–801.

65. King, J.P. 2001a. The Tsing Shan debris flow and debris flood. Landslide study report LSR 2/2001. In *Geotechnical Engineering Office, Civil Engineering and Development Department, The Government of the Hong Kong Special Administrative Region, Hong Kong, People's Republic of China*.

66. King, J.P. 2001b. The 2000 Tsing Shan debris flow and debris flood. Landslide study report no. LSR 3/2001. In *Geotechnical Engineering*

Office, Civil Engineering and Development Department, The Government of the Hong Kong Special Administrative Region, Hong Kong, People's Republic of China.

67. Lade, P.V. 1992. Static instability and liquefaction of loose fine sandy slopes. *Journal of Geotechnical Engineering, ASCE* 118 (1): 51–71.
68. Lambe, T.W., and R.V. Whitman. 1979. *Soil mechanics*, 553. York: Wiley.
69. Laouafa, F., and F. Darve. 2002. Modelling of slope failure by a material instability mechanism. *Computers and Geotechnics* 29: 301–325.
70. Leroueil, S. 2001. Natural slopes and cuts: Movement and failure mechanisms. *Geotechnique* 51 (3): 197–243.
71. Leroueil, S. 2004. Geotechnics of slopes before failure. *Landslides: Evaluation and Stabilization, Lacerda, Ehrlich* 1: 863–884 Fontoura and Sayao (eds).
72. Lourenco, S., K. Sassa, and H. Fukuoka. 2006. Failure process and hydrologic response of a two layer physical model: Implications for rainfall-induced landslides. *Geomorphology* 731-2: 115–130.
73. Matsushi, Y., T. Hattanji, and Y. Matsukura. 2006. Mechanisms of shallow landslides on soil-mantled hillslopes with permeable and impermeable bedrocks in the Boso peninsula, Japan. *Geomorphology* 76: 92–108.
74. Medina, V., M. Hurlimann, and A. Bateman. 2008. Application of FLATModel, a 2D finite volume code, to debris flows in the northeastern part of the Iberian Peninsula. *Landslides* 5: 127–142.
75. Merodo, J.F., M. Pastor, P. Mira, L. Tonni, M.I. Herreros, E. Gonzalez, and R. Tamagnini. 2004. Modelling of diffuse failure mechanisms of catastrophic landslides. *Computer Methods in Applied Mechanics and Engineering* 193 (27–29): 2911–2939.
76. Montgomery, D.R., W.E. Dietrich, R. Torres, S. P Anderson, J.T. Heffner, and K. Loague. 1997. Piezometric response of a steep unchanneled valley to natural and applied rainfall. Water Resources Research, 33, 91–109.
77. Moresi, L.N., F. Dufour, and H.-B. Muhlhaus. 2002. Mantle convection modeling with viscoelastic/brittle lithosphere: Numerical methodology and plate tectonic modeling. *Pure and Applied Geophysics* 159 (10): 2335–2356.

78. Moresi, L.N., F. Dufour, and H.-B. Muhlhaus. 2003. A Lagrangian integration point finite element method for large deformation modelling of viscoelastic geomaterials. *Journal of Computational Physics* 184: 476–497.

79. Morgenstern, N.R., and V.E. Price. 1965. The analysis of the stability of general slip surfaces. *Geotechnique* 15 (1): 79–93.

80. Ng, C.W.W., W.T. Fung, C.Y. Cheuk, and L. Zhang. 2004. Influence of stress ratio and stress path on behaviour of loose decomposed granite. *ASCE Journal of Geotechnical and Geoenvironmental Engineering* 130 (1): 36–44.

81. Ng, C.W.W., and Q. Shi. 1998. A numerical investigation of the stability of unsaturated soil slopes subjected to transient seepage. *Computers and Geotechnics* 22 (1): 1–28.

82. O'Brien, J.S., P.Y. Julien, and W.T. Fullerton. 1993. Two-dimensional water flood and mudflow simulation. *Journal of Hydraulic Engineering* 119 (2): 244–261.

83. Olivares, L., and E. Damiano. 2007. Postfailure mechanics of landslides: Laboratory investigation of Flowslides in pyroclastic soils. *Journal of Geotechnical and Geoenvironmental Engineering, ASCE* 1331: 51–62.

84. Papa, M., S. Egashira, and T. Itoh. 2004. Critical conditions of bed sediment entrainment due to debris flow. *Natural Hazards and Earth System Sciences* 4: 469–474.

85. Pastor, M., T. Blanc, M.J. Pastor, M. Sanchez, B. Haddad, P. Mira, J.A. Fernandez Merodo, M.I. Herreros, and V. Drempetic. 2007a. A SPH depth integrated model with pore pressure coupling for fast landslides and related phenomena. In *International Forum on Landslides Disaster Management*, ed. K. Ho and L. Li, 987–1014.

86. Pastor, M., A.H.C. Chan, P. Mira, D. Manzanal, J.A. Fernández Merodo, and T. Blanc. 2011. Computational geomechanics: The heritage of Olek Zienkiewicz. *International Journal for Numerical Methods in Engineering* 87: 457–489.

87. Pastor, M., and G.B. Crosta. 2012. Landslide runout: Review of analytical/empirical models for subaerial slides, submarine slides and snow avalanche. In *Numerical modelling. Software tools, material models, validation and benchmarking for selected case studies* Deliverable D1.7 for SafeLand project **http://www.safelandfp7.eu/results/Documents/D1.7_revised.pdf**.

88. Pastor, M., J.A. Fernández Merodo, M.I. Herreros, P. Mira, E. González, B. Haddad, M. Quecedo, L. Tonni, and V. Drempetic. 2007b. Mathematical, constitutive and numerical Modelling of catastrophic landslides and related phenomena. *Rock Mechanics and Rock Engineering* 411: 85–132.
89. Pastor, M., J.A. Fernandez-Merodo, E. Gonzalez, P. Mira, T. Li, and X. Liu. 2004. Modelling of landslides: (I). Failure mechanisms. In *Degradations and Instabilities in Geomaterials, CISM Course and Lectures No. 461*, ed. F. Darve and I. Vardoulakis, 287–317. Springer-Verlag.
90. Pastor, M., B. Haddad, G. Sorbino, S. Cuomo, and V. Drempetic. 2009. A depth integrated coupled SPH model for flow-like landslides and related phenomena. *International Journal for Numerical and Analytical Methods in Geomechanics* 33 (2): 143–172.
91. Pastor, M., M. Quecedo, J.A. Fernández Merodo, M.I. Herreros, E. González, and P. Mira. 2002. Modelling tailing dams and mine waste dumps failures. *Geotechnique LII* 8: 579–592.
92. Pastor, M., O.C. Zienkiewicz, and A.H.C. Chan. 1990. Generalized plasticity and the modelling of soil behaviour. Int. J. Numer. And anal. *Methods in Geomechanics* 14: 151–190.
93. Pirulli, M., and M. Pastor. 2012. Numerical study on the entrainment of bed material into rapid landslides. *Geotechnique* 62 (11): 959–972.
94. Pirulli, M., and G. Sorbino. 2008. Assessing potential debris flow runout: A comparison of two simulation models. *National Hazards Earth System Sciences* 8: 961–971.
95. Quan Luna, B., A. Remaître, T.W.J. van Asch, J.P. Malet, and C.J. van Westen. 2012. Analysis of debris flow behavior with a one dimensional run-out model incorporating entrainment. *Engineering Geology* 128: 63–75.
96. Rickenmann. 2009. Empirical relationships for debris flows. *National Hazards* 19: 47–77.
97. Savage, S.B., and K. Hutter. 1991. The dynamics of avalanches of granular materials from initiation to runout. Part I: Analysis. *Acta Mechanica* 86 (1–4): 201–223.
98. Sladen, J.A., R.D. D'Hollander, and J. Krahn. 1985. The liquefaction of sands, a collapse surface approach. *Canadian Geotechnical Journal* 22: 564–578.

99. Sulsky, D., S.-J. Zhou, and H.L. Schreyer. 1995. Application of a particle-in-cell method to solid mechanics. *Computer Physics Communications* 87: 236–252.

100. Takahashi, T. 1991. Debris flows, IAHR Monograph. A.A. Balkema, Rotterdam, pp. 165.

101. Take, W.A., M.D. Bolton, P.C.P. Wong, and F.J. Yeung. 2004. Evaluation of landslide triggering mechanisms in model fill slopes. *Landslides* 1: 173–184.

102. Tsaparas, I., H. Rahardjo, D.G. Toll, and E.C. Leong. 2002. Controlling parameters for rainfall-induced landslides. *Computers and Geotechnics* 1: 1–27.

103. Van Asch, Th.W.J., J.P. Malet, and L.P.H. van Beek. 2006. Influence of landslide geometry and kinematic deformation to describe the liquefaction of landslides: Some theoretical considerations. *Engineering Geology* 88: 59–69.

104. Wanatowski, D., and J. Chu. 2007. Static liquefaction of sand in plane-strain. *Canadian Geotechnical Journal* 44 (3): 299–313.

105. Wanatowski, D., and J. Chu. 2012. Factors affecting pre-failure instability of sand under plane-strain conditions. *Geotechnique* 62 (2): 121–135.

106. Wang, B., P.J. Vardon, and M.A. Hicks. 2016. Investigation of retrogressive and progressive slope failure mechanisms using the material point method. *Computers and Geotechnics* 78: 88–98.

107. Wang, B., P.J. Vardon, and M.A. Hicks. 2018. Rainfall-induced slope collapse with coupled material point method. *Engineering Geology* 239: 1–12.

108. Wang, F.W., K. Sassa, and G. Wang. 2002. Mechanism of a long-runout landslide triggered by the august 1998 heavy rainfall in Fukushima prefecture, Japan. *Engineering Geology* 63: 169–185.

109. Wang, G., and K. Sassa. 2001. Factors affecting rainfall induced landslides in laboratory flume tests. *Géotechnique* 51: 587–600.

110. Wang, G., and K. Sassa. 2003. Pore-pressure generation and movement of rainfall-induced landslides: Effects of grain size and fine-particle content. *Engineering Geology* 69: 109–125.

111. Yamamuro, J.A., and P.J. Lade. 1998. Steady-state concepts and static liquefaction of silty sands. *ASCE J Geotech Geoenviron Eng* 1249: 868–878.

112. Yasufuku, N., H. Ochiai, and D. Hormdee. 2005. An empirical relationship for evaluating collapsible settlements of volcanic ash sandy soil. In *Advanced experimental unsaturated soil mechanics*, ed. Tarantino, Romero, and Cui, 265–272.

113. Zienkiewicz, O.C., A.H.C. Chan, M. Pastor, B.A. Shrefler, and T. Shiomi. 1999. *Computational Geomechanics*. Wiley.

114. Zienkiewicz, O.C., C.T. Chang, and P. Bettess. 1980. Drained, undrained, consolidating dynamic behaviour assumptions in soils. *Geotechnique* 30: 385–395.

Chapter

10

ON SOME WAVELET SOLUTIONS OF SINGULAR DIFFERENTIAL EQUATIONS ARISING IN THE MODELING OF CHEMICAL AND BIOCHEMICAL PHENOMENA

Mo Faheem[1], Arshad Khan[1], and E.R. El-Zahar[2,3]

[1]Department of Mathematics, Jamia Millia Islamia, New Delhi, 110025, India.
[2]Department of Mathematics, College of Sciences and Humanities in Al-Kharj, Prince Sattam bin Abdulaziz University, Alkharj 11942, Saudi Arabia.
[3]Department of Basic Engineering Science, Faculty of Engineering, Menoufia University, Shebin El-Kom 32511, Egypt.

ABSTRACT

This paper is concerned with the Lane–Emden boundary value problems arising in many real-life problems. Here, we discuss two numerical schemes based on Jacobi and Bernoulli wavelets for the solution of the governing equation of electrohydrodynamic flow in a circular cylindrical conduit, nonlinear heat conduction model in the human head, and non-isothermal

Citation: (APA): Faheem, M., Khan, A., & El-Zahar, E. R. (2020). On some wavelet solutions of singular differential equations arising in the modeling of chemical and biochemical phenomena. Advances in Difference Equations, 2020(1), 1-23. (23 pages).

reaction–diffusion model equations in a spherical catalyst and a spherical biocatalyst. These methods convert each problem into a system of nonlinear algebraic equations, and on solving them by Newton's method, we get the approximate analytical solution. We also provide the error bounds of our schemes. Furthermore, we also compare our results with the results in the literature. Numerical experiments show the accuracy and reliability of the proposed methods.

MSC: 65L05; 65T60

Keywords: Jacobi wavelet; Bernoulli wavelet; Collocation grids

INTRODUCTION

The solution of Emden–Fowler type equation is vital because of its numerous applications in engineering and technical problems. There are several phenomena like astrophysics, aerodynamics, stellar structure, chemistry, biochemistry, and many others (see [19, 38, 40, 41]) which can be modeled by the Lane–Emden equation of shape operator w given by [18]

$$u''(z) + \frac{w}{z}u'(z) + f(u,z) = 0, \quad w > 0. \tag{1.1}$$

A number of research papers are inclined toward the numerical solution of such type of differential equations. The numerical methods for the solution of Lane–Emden equation based on B-spline have been studied in [24, 30–32]. Homotopy analysis methods and iterative schemes for fast convergence and accuracy of solutions of singular and doubly singular BVPs have been developed in [21, 22, 26, 33]. Roul et al. have dealt with the solution of a class of two-point nonlinear singular boundary value problems with Neumann and Robin boundary conditions by deploying a high order compact finite difference method [25]. A least square recursive approach together with convergence analysis for solving Lane–Emden type initial value problems has been developed in [27], in which they simply reduce the solution of the original initial value problem to the solution of an integral equation. The B-spline method fails to provide a satisfactory approximation in the presence of singularity; on the other hand, the Adomian decomposition methods (ADM) fail to establish a convergent series solution to strongly nonlinear BVPs. To overcome these shortcomings, Roul came up with the combination of ADM and B-spline collocation methods for accurate solution, see [23]. Madduri and Roul developed a fast converging iterative scheme for the solution of a system of Lane–Emden equations converting them into equivalent Fredholm integral equations and treating them with homotopy analysis method [14]. In this paper, we discuss and solve some mathematical models of the chemical and biochemical phenomena using wavelet methods.

Model of Electrohydrodynamic (EHD) Flow in a Circular Cylindrical Conduit

The effect of the electric and magnetic field on fluid has been studied by many researchers. Phenomena involving the conversion of electrical and magnetic energy into kinetic energy are known as electrohydrodynamics (EHD) and magnetohydrodynamics (MHD). The effect of the electric field on fluids gives extra means of controlling flow conditions and has various technical applications such as EHD thruster, EHD flow, heat transfer enhancement, EHD drying and evaporation, and functional electrostatic bowler (EHD pump). EHD pump has been designed for semiconductor cooling [5], electrospray mass spectrometry, and electrospray nanotechnology [45]. The MHD flow has a wide range of applications in the fields of chemistry and biology, for instance, the fabrication in cancer tumor therapy resulting hypothermia, decreasing bleeding in the state of acute injuries, magnetic resonance visualizing, and various other diagnostic experiments [3]. Magneto-hybrid nanofluids flow via mixed convection past a radiative circular cylinder was studied in [4]. The EHD flow of a fluid is modeled by a set of partial differential equations, which can be reduced to an ordinary differential equation as in [16], and results in the following Emden–Fowler type of equation:

$$u''(z) + \frac{1}{z}u'(z) + H^2\left(1 - \frac{u}{1-\alpha u}\right) = 0, \tag{1.2}$$

subject to the boundary conditions

$$u'(0) = 0, \qquad u(1) = 0, \tag{1.3}$$

where $u = -\frac{\overline{u}}{KE_0\alpha}, \alpha = \frac{K}{j_0}\frac{\partial p}{\partial z} - 1$.

Here, the pressure gradient $\frac{\partial p}{\partial z}$ is a constant that measures the nonlinearity and $H = \sqrt{\frac{j_0 a^2}{\mu K^2 E_0}}$ is the Hartmann number [16]. A schematic diagram of EHD flow is given in Fig. 1.

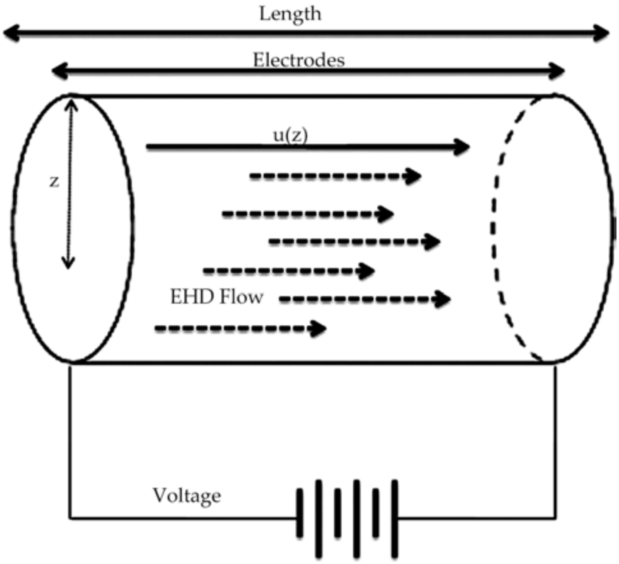


Figure 1. Schematic diagram of EHD flow in a circular cylindrical conduit.

Equation (1.2) is a strong nonlinear differential equation having a singularity at $z = 0$. Finding the exact solution to this problem is quite complicated, and therefore the development and use of numerical techniques for the solution of this problem play an important role. Only few numerical methods are available for the solution of (1.2). For instance, Mastroberardino developed homotopy analysis method [15], Ghasemi et al. used least square method [6], Mosayebidorcheh applied Taylor series [17], and Roul et al. gave a new iterative algorithm [28] for the solution of strongly nonlinear singular boundary value problems.

Nonlinear Heat Conduction Model in Human Head

Biomechanics is the area of science in which mechanics laws and formulae are used to study the behavior of the human body. The heat flow in the human body is quivering and vital field that helps to analyze the human heat stress at various temperatures. The human head is the only organ in the human body that controls different parts and functions in the body. The authors in [37] and [13] studied the effect of digital mobile phone emission on the human brain and concluded that the cellular phone waves can cause several brain problems, like exciting the brain cell, weakening the neural behavior, and possible disruption in the functionality of the nervous system. Ketley [11] points out the neuropsychological squeal of digital mobile phone exposure in humans. Similarly, the thermal effect of wave and radiation from digital phones on the human nervous system and brain is studied in [7, 12, 39, 44].

The following Emden-type equation is used to model the distribution of heat source in the human head [2]:

$$u''(z) + \frac{2}{z}u'(z) + \frac{p(u)}{\gamma} = 0, \quad 0 < z < 1, \tag{1.4}$$

subject to the boundary conditions

$$u'(0) = 0, \qquad -\nu u'(1) = \mu(u - u_k), \tag{1.5}$$

where $p(u)$ is the heat production rate per unit volume, u is the absolute temperature, z is the radial distance from the center. Figure 2 shows the schematic diagram of human heat conduction model.

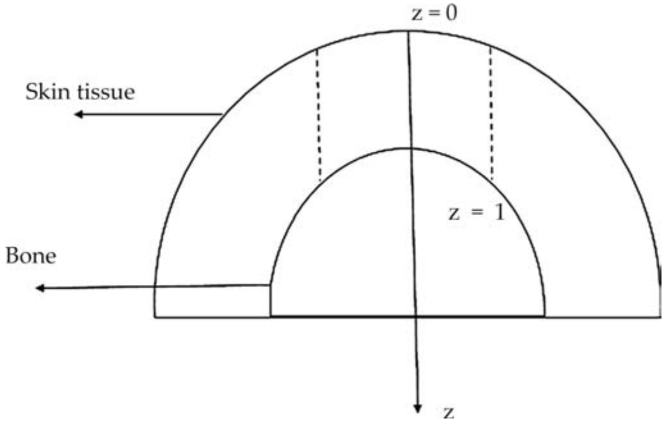


Figure 2. Schematic diagram of human heat conduction model.

Many researchers have shown their interest in solving this model numerically. For example, Wessapan et al. [43] derived a numerical algorithm of specific absorption rate and heat transfer in the human body to leakage electromagnetic field. Keangin et al. [9] gave an analysis of heat transfer in liver tissue during microwave ablation using single and two double slot antennae. Wessapan and Rattanadecho [42] used a three-dimensional human head model for simulating the heat distribution by applying 3-D finite element mesh (see Fig. 3).

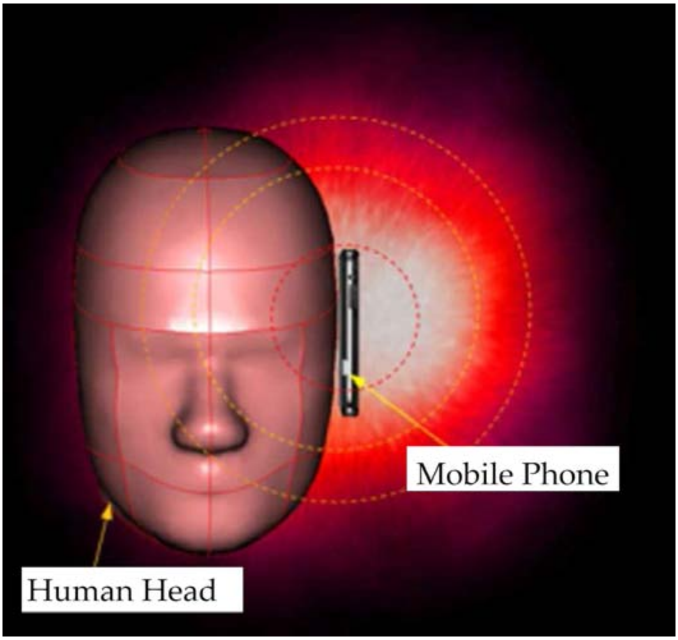


Figure 3. Human head exposed to mobile phone radiation [42].

Mathematical Model of Spherical Catalyst Equation

The following Lane–Emden equation is used to model the dimensionless concentration of chemical species which occur in a spherical catalyst [19]:

$$u''(z) + \frac{2}{z}u'(z) - \rho^2 u(z)e^{\left(\frac{\sigma\beta(1-u(z))}{1+\beta(1-u(z))}\right)} = 0, \tag{1.6}$$

subject to the boundary conditions

$$u'(0) = 0, \quad u(1) = 1, \tag{1.7}$$

where ρ^2, σ, and β denote the Thiele modulus, dimensionless activation energy, and dimensionless heat of reaction, respectively, and are given by

$$\rho^2 = \frac{\kappa_{ref}\mathrm{R}^2}{\mathrm{D}}, \qquad \sigma = \frac{\mathrm{E}}{\mathrm{R_g T_s}}, \qquad \beta = \frac{(-\Delta\mathrm{H})\mathrm{DC_{As}}}{\mathrm{KT_s}}, \qquad z = \frac{\mathrm{R}}{\mathrm{r}}, \quad \text{and} \quad u = \frac{\mathrm{C_A}}{\mathrm{C_{As}}} \quad [19].$$

The effectiveness factor of spherical pellet is defined as [34]

$$\eta = \frac{3}{\rho^2}u'(z) \quad \text{at } z = 1.$$

Mathematical Model of Spherical Biocatalyst Equation

The following Lane–Emden equation is used for modeling the spherical biocatalyst equation [34]:

$$u''(z) + \frac{2}{z}u'(z) - \rho^2 \frac{(1+\beta)u(z)}{1+\beta u(z)} = 0, \tag{1.8}$$

subject to the boundary conditions

$$u'(0) = 0, \qquad u(1) = 1, \tag{1.9}$$

where ρ^2, σ, and β denote the Thiele modulus, dimensionless activation energy, and dimensionless heat of reaction, respectively, and are given by

$$\rho^2 = \frac{-r_{As}R^2}{DD_{As}}, \qquad \beta = \frac{C_{As}}{K_m}, \qquad z = \frac{R}{r}, \quad \text{and} \quad u = \frac{C_A}{C_{As}} \quad [34].$$

The effectiveness factor of spherical pellet is defined as [34]

$$\eta = \frac{3}{\rho^2}u'(z) \quad \text{at } z = 1.$$

The schematic diagram of spherical biocatalyst is shown in Fig. 4.

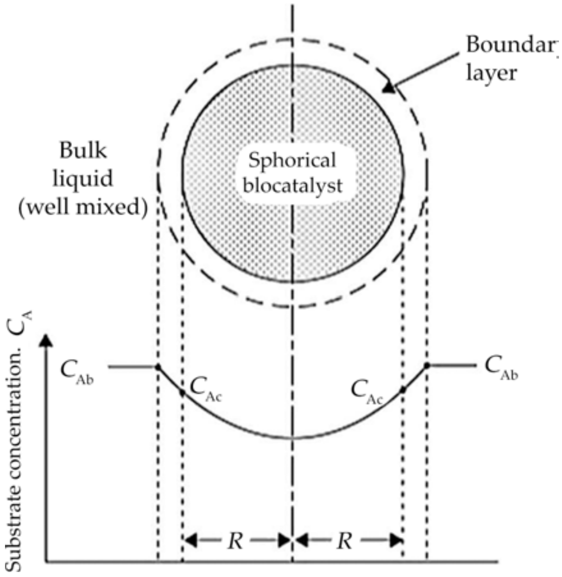


Figure 4. Schematic diagram of spherical biocatalyst.

Several numerical techniques have been adopted for solving non-isothermal reaction–diffusion model equations. For instance, Singh [34] applied optimal homotopy analysis method, and Jamal and Khuri [8] used Green's function and fixed point iteration approach for solving such type of equations. Rach et al. [19] reduced this model equation into an equivalent Volterra integral equation and then solved it by coupling the modified Adomian decomposition method and the Volterra integral technique.

Jacobi wavelet is the family of wavelets reduced into Legendre wavelet, Chebyshev wavelet, and Gegenbauer wavelet for the specific value of κ and ω. There are a lot of research papers available for the solution of ordinary and partial differential equations using Jacobi and Bernoulli wavelets, for instance, see [1, 10, 20, 46]. In this study, we introduce two methods based on Jacobi and Bernoulli wavelets for solving models of electrohydrodynamic flow in a circular cylindrical conduit, nonlinear heat conduction model in the human head, spherical catalyst equation, and spherical biocatalyst equation. These wavelets transform these model equations into a system of nonlinear algebraic equations, and on solving them, we get the unknown wavelet coefficients. With the help of these coefficients, we get the approximate analytical solution that is valid over all the problem domain, not only at grid points. The outline of this paper is as follows: The second section describes the Jacobi wavelet, function approximation by Jacobi wavelet, and integration of Jacobi wavelet. Similarly, the third section describes the Bernoulli wavelet, function approximation by Bernoulli wavelet, and integration of Bernoulli wavelet. In the fourth section, the wavelet approximation method for all the above models is given. In the fifth section, we state some theoretical proof for error bounds of our methods. In the sixth section, the numerical experiments confirm that our methods converge fast.

JACOBI WAVELET

Jacobi Polynomials

Jacobi polynomials, which are often called hypergeometric polynomials, are denoted by $\mathrm{J}_m^{\kappa,\omega}(z)$ and can be defined by the following explicit formula:

$$\mathrm{J}_m^{\kappa,\omega}(z) = \frac{\Gamma(\kappa+m+1)}{m!\Gamma(\kappa+\omega+m+1)}\sum_{i=0}^{m}\binom{m}{i}\frac{\Gamma(\kappa+\omega+i+m+1)}{\Gamma(\kappa+i+1)}\left(\frac{z-1}{2}\right)^i.$$

Some first few Jacobi polynomials are given by

$$\mathrm{J}_0^{\kappa,\omega}(z) = 1,$$

$$\mathrm{J}_1^{\kappa,\omega}(z) = \kappa + 1 + (\kappa + \omega + 2)\frac{z-1}{2},$$

$$\mathrm{J}_2^{\kappa,\omega}(z) = \frac{(\kappa+1)(\kappa+2)}{2} + (\kappa+2)(\kappa+\omega+3)\frac{z-1}{2} + \frac{(\kappa+\omega+3)(\kappa+\omega+4)}{2}\left(\frac{z-1}{2}\right)^2, \quad \ldots.$$

These polynomials are orthogonal on [−1, 1] with respect to the weight $(1 - z)^{\kappa}(1 - z)^{\omega}$ and satisfy the following properties:

$$\mathrm{J}_m^{\kappa,\omega}(-1) = (-1)^m \binom{m+\omega}{m},$$

$$\mathrm{J}_m^{\kappa,\omega}(-z) = (-1)^m \mathrm{J}_m^{\omega,\kappa}(z),$$

$$\mathrm{J}_m^{\kappa,\kappa}(z) = \begin{cases} \frac{\Gamma(m+\kappa+1)\Gamma(\frac{m}{2}+1)}{\Gamma(\frac{m}{2}+\kappa+1)\Gamma(m+1)} \mathrm{J}_{\frac{m}{2}}^{\kappa,\frac{-1}{2}}(2z^2-1), & \text{if } m \text{ is even,} \\ \frac{\Gamma(m+\kappa+2)\Gamma(\frac{m}{2}+1)}{\Gamma(\frac{m}{2}+\kappa+1)\Gamma(m+2)} z\mathrm{J}_{\frac{m}{2}}^{\kappa,\frac{1}{2}}(2z^2-1), & \text{if } m \text{ is odd.} \end{cases}$$

$$\int_{-1}^{1} (1-z)^{\kappa}(1+z)^{\omega} \mathrm{J}_m^{\kappa,\omega}(z)\mathrm{J}_n^{\kappa,\omega}(z)\,dz = \frac{2^{\kappa+\omega+1}}{2m+\kappa+\omega+1}\frac{\Gamma(m+\kappa+1)\Gamma(m+\omega+1)}{\Gamma(m+\kappa+\omega+1)m!}\delta_{nm},$$

where δ_{nm} is Kronecker delta.

$$\begin{aligned} &2m(m+\kappa+\omega)(2m+\kappa+\omega-2)J_m^{\kappa,\omega}(z) \\ &\quad = (2m+\kappa+\omega-1)\big((2m+\kappa+\omega)(2m+\kappa+\omega-2)z \\ &\qquad + \kappa^2 - \omega^2\big)\mathrm{J}_{m-1}^{\kappa,\omega}(z) - 2(m+\alpha-1)(m+\omega-1)(2m+\kappa+\omega)\mathrm{J}_{m-2}^{\kappa,\omega}(z). \end{aligned}$$

Jacobi Wavelet of Shifted Jacobi Polynomial

Jacobi wavelet of the shifted Jacobi polynomial defined on six arguments *k*, *n*, *κ*, *ω*, *m*, *z* is denoted by $\mathcal{J}(k,n,\kappa,\omega,m,z) = \mathcal{J}_{n,m}^{\kappa,\omega}(z),$ and can be defined on [0, 1) as follows [1]:

$$\mathcal{J}^{\kappa,\omega}_{n,m}(z) = \begin{cases} 2^{\frac{k}{2}} \mu^{\kappa,\omega}_m \mathrm{J}^{\kappa,\omega}_m(2^k z - 2n + 1), & \text{if } z \in [\xi_1, \xi_2), \\ 0, & \text{otherwise,} \end{cases} \tag{2.1}$$

where $\xi_1 = \frac{n-1}{2^{k-1}}, \xi_2 = \frac{n}{2^{k-1}},$ and $\mu^{\kappa,\omega}_m = \sqrt{\frac{(2m+\kappa+\omega+1)\Gamma(2m+\kappa+\omega+1)m!}{2^{\kappa+\omega+1}\Gamma(m+\kappa+1)\Gamma(m+\omega+1)}}.$

Equivalently, for any positive integer *k*, Jacobi wavelet can also be defined as follows:

$$\mathcal{J}^{\kappa,\omega}_{i}(z) = \begin{cases} 2^{\frac{k}{2}} \mu^{\kappa,\omega}_m \mathrm{J}^{\kappa,\omega}_m(2^k z - 2n + 1), & \text{if } z \in [\xi_1, \xi_2), \\ 0, & \text{otherwise,} \end{cases} \tag{2.2}$$

where *i* is wavelet number determined by $i = n + 2^{k-1}m$, where $n = 0, 1, 2,$ … and $m = 0, 1, 2, \ldots, M - 1$, where *m* is degree of polynomial. *M* can be determined by $M = \frac{N}{2^{k-1}},$ where $k = 1, 2, \ldots$.

Function Approximation by Jacobi Wavelet

Let $\{\mathcal{J}^{\kappa,\omega}_{1,0}, \ldots, \mathcal{J}^{\kappa,\omega}_{1,M-1}, \mathcal{J}^{\kappa,\omega}_{2,0}, \ldots, \mathcal{J}^{\kappa,\omega}_{2,M-1}, \mathcal{J}^{\kappa,\omega}_{2^{k-1},0}, \ldots, \mathcal{J}^{\kappa,\omega}_{2^{k-1},M-1}\}$ be a set of Jacobi wavelets.

Any function $f(z) \in L^2\,[0, 1)$ can be expressed in terms of Jacobi wavelet as follows [1]:

$$f(z) = \sum_{n=1}^{\infty}\sum_{m=0}^{\infty} a_{n,m}\mathcal{J}^{\kappa,\omega}_{n,m}(z) = \sum_{i=1}^{\infty} a_i \mathcal{J}^{\kappa,\omega}_i(z).$$

For approximation, we truncate this series for a natural number *N*, and we get

$$f(z) \approx \sum_{n=1}^{2^{k-1}}\sum_{m=0}^{M-1} a_{n,m}\mathcal{J}^{\kappa,\omega}_{n,m}(z) = \sum_{i=1}^{N} a_i \mathcal{J}^{\kappa,\omega}_i(z) \tag{2.3}$$

$$= a^T \mathcal{J}(z), \tag{2.4}$$

where *a* and $\mathcal{J}(z)$ are matrices of order *N* × 1 given by

$$a = [a_{1,0}, a_{1,1}, \ldots, a_{1,M-1}, a_{2,0}, a_{2,1}, \ldots, a_{2,M-1}, \ldots, a_{2^{k-1},0}, \ldots, a_{2^{k-1},M-1}]^T$$

$$= [a_1, a_2, \ldots, a_N]^T, \tag{2.5}$$

$$\mathcal{J}(z) = \left[\mathcal{J}_{1,0}^{\kappa,\omega}(z), \ldots, \mathcal{J}_{1,M-1}^{\kappa,\omega}(z), \mathcal{J}_{2,0}^{\kappa,\omega}(z), \ldots, \mathcal{J}_{2,M-1}^{\kappa,\omega}(z), \mathcal{J}_{2^{k-1},0}^{\kappa,\omega}(z), \ldots, \mathcal{J}_{2^{k-1},M-1}^{\kappa,\omega}(z)\right]$$

$$= \left[\mathcal{J}_1^{\kappa,\omega}(z), \ldots, \mathcal{J}_N^{\kappa,\omega}(z)\right], \tag{2.6}$$

where the coefficient a_i can be determined by $a_i = \langle f(z), \mathcal{J}_i^{\kappa,\omega}(z)\rangle = \int_0^1 f(z)\overline{\mathcal{J}_i^{\kappa,\omega}(z)}\,dz.$

Integration of Jacobi Wavelet

Let $\mathcal{J}_i^1(z), \mathcal{J}_i^2(z),$ and $\mathcal{J}_i^3(z)$ be the first, second, and third integration of Jacobi wavelet from 0 to z respectively. These integrations can be determined as follows:

$$\mathcal{J}_{1,i}^{\kappa,\omega}(z) = \begin{cases} 2^{\frac{-k}{2}}\mu_m^{\kappa,\omega}(\frac{1}{(m+\kappa+\omega)})\{\mathrm{J}_{m+1}^{\kappa-1,\omega-1}(\hat{z}) - \mathrm{J}_{m+1}^{\kappa-1,\omega-1}(-1)\}, & \xi_1 \le z < \xi_2, \\ 2^{\frac{-k}{2}}\mu_m^{\kappa,\omega}(\frac{1}{(m+\kappa+\omega)})\{\mathrm{J}_{m+1}^{\kappa-1,\omega-1}(1) - \mathrm{J}_{m+1}^{\kappa-1,\omega-1}(-1)\}, & \xi_2 \le z \le 1, \end{cases} \tag{2.7}$$

$$\mathcal{J}_{2,i}^{\kappa,\omega}(z) = \begin{cases} 2^{\frac{-3k}{2}}\mu_m^{\kappa,\omega}(\frac{1}{(m+\kappa+\omega)})\{(\frac{1}{(m-2+\kappa+\omega)})\{\mathrm{J}_{m+2}^{\kappa-2,\omega-2}(\hat{z}) & \\ \quad - \mathrm{J}_{m+2}^{\kappa-2,\omega-2}(-1)\} - (1+\hat{z})\mathrm{J}_{m+1}^{\kappa-1,\omega-1}(-1)\}, & \xi_1 \le z < \xi_2, \\ 2^{\frac{-3k}{2}}\mu_m^{\kappa,\omega}(\frac{1}{(m+\kappa+\omega)})\{(\frac{1}{(m-2+\kappa+\omega)})\{\mathrm{J}_{m+2}^{\kappa-2,\omega-2}(1) & \\ \quad - \mathrm{J}_{m+2}^{\kappa-2,\omega-2}(-1)\} - 2\mathrm{J}_{m+1}^{\kappa-1,\omega-1}(-1) & \\ \quad + (\hat{z}-1)\{\mathrm{J}_{m+1}^{\kappa-1,\omega-1}(1) - \mathrm{J}_{m+1}^{\kappa-1,\omega-1}(-1)\}\}, & \xi_2 \le z \le 1, \end{cases} \tag{2.8}$$

where $\hat{z} = 2^k z - 2n + 1$.

BERNOULLI WAVELET

Bernoulli Polynomials

Bernoulli polynomials are denoted by $\text{ß}_m(z)$, where m is the degree of polynomials and can be defined by the following explicit formula:

$$\text{ß}_m(z) = \sum_{i=0}^{m} \binom{m}{i} \text{ß}_{m-i} z^i,$$

where $ß_k, k = 0, 1, 2, \ldots m,$ are the Bernoulli numbers. Another explicit formula for these polynomials is given by

$$ß_m(z) = \sum_{i=0}^{m} \frac{1}{i+1} \sum_{j=0}^{i} (-1)^j \binom{i}{j} (z+j)^m.$$

The first few Bernoulli polynomials are given by

$$ß_0(z) = 1, \qquad ß_1(z) = z - \frac{1}{2}, \qquad ß_2(z) = z^2 - z + \frac{1}{6}, \qquad ß_3(z) = z^3 - \frac{3}{2}z^2 + \frac{1}{2}z.$$

Bernoulli polynomial satisfies the following properties:

$$ß_m(1) = (-1)^m ß_m(0),$$

$$ß_{2m+1}(0) = 0, \qquad ß_{2m-1}\left(\frac{1}{2}\right) = 0,$$

$$ß_m(1-z) = (-1)^m ß_m(z),$$

$$ß_m(z+1) - ß_m(z) = mz^{m-1},$$

$$\int_0^1 ß_n(z) ß_m(z)\, dz = (-1)^{n-1} \frac{n!m!}{(m+n)!} ß_{m+n}.$$

Bernoulli polynomials can be calculated by the following recursive formula: $ß'_m(z) = m ß_{m-1}(z)$.

Bernoulli Wavelet

Bernoulli wavelet defined on four arguments *k*, *n*, *m*, *z* is denoted by $\mathcal{B}(k,n,m,z) = \mathcal{B}_{n,m}(z)$ and can be defined on [0,1) as follows [10]:

$$\mathcal{B}_{n,m}(z) = \begin{cases} 2^{\frac{k-1}{2}} \overline{ß}_m(2^{k-1}z - n + 1), & \xi_1 \le z \le \xi_2, \\ 0, & \text{elsewhere,} \end{cases} \tag{3.1}$$

where $\xi_1 = \frac{n-1}{2^{k-1}}$ and $\xi_2 = \frac{n}{2^{k-1}}$ and

$$\overline{\text{ß}_m}(z) = \begin{cases} 1, & m = 0, \\ \frac{1}{\sqrt{\frac{(-1)^{m-1}(m!)^2}{(2m)!}\text{ß}_{2m}}}\text{ß}_m(z), & m > 0, \end{cases} \tag{3.2}$$

where ß_{2m} is the Bernoulli number.

On the interval [0,1), for any positive integer k, Bernoulli wavelet can also be defined as follows:

$$\mathcal{B}_i(z) = \begin{cases} 2^{\frac{k-1}{2}}\overline{\text{ß}}_m(2^{k-1}z - n + 1), & \xi_1 \le z \le \xi_2, \\ 0, & \text{elsewhere.} \end{cases} \tag{3.3}$$

Here, i is wavelet number and can calculated by the relation $i = n + 2^{k-1}m$, where $n = 0, 1, 2, \ldots$ and $m = 0, 1, 2, \ldots, M - 1$, m is degree of polynomials. For $k = 1, 2, \ldots, M$ can be found by $N = 2^{k-1}M$.

Function Approximation by Bernoulli Wavelet

Let $\{\mathcal{B}_{1,0}, \ldots, \mathcal{B}_{1,M-1}, \mathcal{B}_{2,0}, \ldots, \mathcal{B}_{2,M-1}, \mathcal{B}_{2^{k-1},0}, \ldots, \mathcal{B}_{2^{k-1},M-1}\}$ be a set of Bernoulli wavelets.

Any function $f(z) \in L^2$ [0,1) can be expressed in terms of Bernoulli wavelet as follows [46]:

$$f(z) = \sum_{n=1}^{\infty}\sum_{m=0}^{\infty} b_{n,m}\mathcal{B}_{n,m}(z) = \sum_{i=1}^{\infty} b_i\mathcal{B}_i(z).$$

For approximation, we truncate this series for a natural number N, and we get

$$f(z) \approx \sum_{n=1}^{2^{k-1}}\sum_{m=0}^{M-1} b_{n,m}\mathcal{B}_{n,m}(z) = \sum_{i=1}^{N} b_i\mathcal{B}_i(z) \tag{3.4}$$

$$= b^T\mathcal{B}(z), \tag{3.5}$$

where b and $\mathcal{B}(z)$ are matrices of order $N \times 1$ given by

$$b = [b_{1,0}, b_{1,1}, \ldots, b_{1,M-1}, b_{2,0}, b_{2,1}, \ldots, b_{2,M-1}, \ldots, b_{2^{k-1},0}, \ldots, b_{2^{k-1},M-1}]^T$$

$$= [b_1, b_2, \ldots, b_N]^T, \tag{3.6}$$

$$\begin{aligned}\mathcal{B}(z) &= \left[\mathcal{B}_{1,0}(z),\ldots,\mathcal{B}_{1,M-1}(z),\mathcal{B}_{2,0}(z),\ldots,\mathcal{B}_{2,M-1}(z),\mathcal{B}_{2^{k-1},0}(z),\ldots,\mathcal{B}_{2^{k-1},M-1}(z)\right]\\ &= \left[\mathcal{B}_1(z),\ldots,\mathcal{B}_N(z)\right].\end{aligned} \tag{3.7}$$

The coefficient b_i is calculated by $b_i = \langle f(z), \mathcal{B}_i(z)\rangle = \int_0^1 f(z)\overline{\mathcal{B}_i(z)}\,dz.$

Integration of Bernoulli Wavelet

Let $\mathcal{B}_{1,i}(z)$ and $\mathcal{B}_{2,i}(z)$ be the first and second integration of Bernoulli wavelet from 0 to z, respectively. These integration can be determined as follows:

$$\mathcal{B}_{1,i}(z) = \begin{cases} 2^{\frac{-k+1}{2}}\zeta(\frac{1}{m+1})\{\text{ß}_{m+1}(\hat{z}) - \text{ß}_{m+1}(0)\}, & \xi_1 \le z < \xi_2,\\ 2^{\frac{-k+1}{2}}\zeta(\frac{1}{m+1})\{\text{ß}_{m+1}(1) - \text{ß}_{m+1}(0)\}, & \xi_2 \le z \le 1, \end{cases} \tag{3.8}$$

$$\mathcal{B}_{2,i}(z) = \begin{cases} 2^{\frac{-3k+3}{2}}\zeta(\frac{1}{m+1})\{(\frac{1}{m+2})\{\text{ß}_{m+2}(\hat{z}) - \text{ß}_{m+2}(0)\} - (\hat{z})\text{ß}_{m+1}(0)\}, & \xi_1 \le z < \xi_2,\\ 2^{\frac{-3k+3}{2}}\zeta(\frac{1}{m+1})\{(\frac{1}{m+2})\{\text{ß}_{m+2}(1) - \text{ß}_{m+2}(0)\} - 2\text{ß}_{m+1}(0) & \\ \quad + (\hat{z}-1)\{\text{ß}_{m+1}(1) - \text{ß}_{m+1}(0)\}\}, & \xi_2 \le z \le 1, \end{cases} \tag{3.9}$$

where $\hat{z} = 2^{k-1}z - n + 1$ and $\zeta = \dfrac{1}{\sqrt{\frac{(-1)^{m-1}(m!)^2}{(2m)!}\text{ß}_{2m}}}$.

METHODS FOR SOLUTION

In this section, we discuss the methods for the solution of the models described above. The following notations have been introduced:

$$\phi_{1,i}(z) = \int_0^z \phi_i(z)\,dz, \tag{4.1}$$

$$\phi_{2,i}(z) = \int_0^z \phi_{1,i}(z)\,dz, \tag{4.2}$$

$$\Phi_{1,i} = \int_0^1 \phi_{1,i}(z)\,dz, \tag{4.3}$$

$$\Phi_{2,i} = \int_0^1 \phi_{2,i}(z)\,dz. \tag{4.4}$$

Method for Solution of Model of Electrohydrodynamic Flow in a Circular Cylindrical Conduit

We can express the second derivative of (1.2) in terms of wavelet series as follows:

$$u''(z) = \sum_{i=1}^{N} c_i \phi_i(z). \tag{4.5}$$

Integrating (4.5) twice from 0 to *z*, we get

$$u'(z) = \sum_{i=1}^{N} c_i \phi_{1,i}(z) + u'(0), \tag{4.6}$$

$$u(z) = \sum_{i=1}^{N} c_i \phi_{2,i}(z) + zu'(0) + u(0). \tag{4.7}$$

Using boundary conditions (1.3) in (4.6)–(4.7), we get

$$u'(z) = \sum_{i=1}^{N} c_i \phi_{1,i}(z), \tag{4.8}$$

$$u(z) = \sum_{i=1}^{N} c_i \phi_{2,i}(z) + u(0). \tag{4.9}$$

Putting $z = 1$ in (4.9) and after simplifying, we get

$$u(0) = -\sum_{i=1}^{N} c_i \Phi_{2,i}. \tag{4.10}$$

Therefore equation (4.9) becomes

$$u(z) = \sum_{i=1}^{N} c_i \left(\phi_{2,i}(z) - \Phi_{2,i} \right). \tag{4.11}$$

Putting the values of $u(z)$, $u'(z)$, and $u''(z)$ from equations (4.5), (4.8), (4.11) in equation (1.2) and collocating at $z = z_l = \frac{l-0.5}{N}$, where $l = 1, 2, \ldots, N$, yields the following system of nonlinear equations:

$$\sum_{i=1}^{N} c_i\phi_i(z_l) + \frac{1}{z_l}\sum_{i=1}^{N} c_i\phi_{1,i}(z_l) + \mathrm{H}^2\left(1 - \frac{\sum_{i=1}^{N} c_i(\phi_{2,i}(z_l) - \Phi_{2,i})}{1 - \alpha\sum_{i=1}^{N} c_i(\phi_{2,i}(z_l) - \Phi_{2,i})}\right) = 0. \tag{4.12}$$

On solving this system of nonlinear equations by Newton's method, we get the unknown wavelet coefficients c_i's. After putting these c_i's in equation (4.11), we get the approximate solution.

Method for Solution of Nonlinear Heat Conduction Model in the Human Head

We can approximate the second derivative of equation (1.4) in terms of wavelet series as follows:

$$u''(z) = \sum_{i=1}^{N} c_i\phi_i(z). \tag{4.13}$$

Integrating (4.13) twice from 0 to z, we get

$$u'(z) = \sum_{i=1}^{N} c_i\phi_{1,i}(z) + u'(0), \tag{4.14}$$

$$u(z) = \sum_{i=1}^{N} c_i\phi_{2,i}(z) + zu'(0) + u(0). \tag{4.15}$$

Using boundary conditions (1.5) in (4.14)–(4.15), we get

$$u'(z) = \sum_{i=1}^{N} c_i\phi_{1,i}(z), \tag{4.16}$$

$$u(z) = \sum_{i=1}^{N} c_i\phi_{2,i}(z) + u(0). \tag{4.17}$$

Putting $z = 1$ in (4.16)–(4.17) and multiplying (4.16) by v and (4.17) by μ and after solving these equations for $u(0)$, we get

$$u(0) = u_\kappa - \frac{1}{\mu}\sum_{i=1}^{N} c_i(\nu\Phi_{1,i} + \mu\Phi_{2,i}). \tag{4.18}$$

Therefore equation (4.17) becomes

$$u(z) = \sum_{i=1}^{N} c_i\left(\phi_{2,i}(z) - \frac{1}{\mu}(\nu\Phi_{1,i} + \mu\Phi_{2,i})\right) + u_\kappa. \tag{4.19}$$

Putting the values of $u''(z)$, $u'(z)$, and $u(z)$ from equations (4.13), (4.16), (4.19) in equation (1.4) and collocating at $z = z_l = \frac{l-0.5}{N}$, where $l = 1, 2, \ldots, N$, yields the following system of nonlinear equations:

$$\sum_{i=1}^{N} c_i\phi_i(z_l) + \frac{2}{z_l}\sum_{i=1}^{N} c_i\phi_{1,i}(z_l) + \frac{p(\sum_{i=1}^{N} c_i(\phi_{2,i}(z_l) - \frac{1}{\mu}(\nu\Phi_{1,i} + \mu\Phi_{2,i})) + u_\kappa)}{\gamma} = 0. \tag{4.20}$$

After solving this system of nonlinear equations by Newton's method, we get the unknown wavelet coefficients. On putting these coefficients in equation (4.19), we get the approximate wavelet solutions of nonlinear heat conduction model in the human head.

Method for Solution of Spherical Catalyst Equation

We can approximate the second derivative of equation (1.6) in terms of wavelet series as follows:

$$u''(z) = \sum_{i=1}^{N} c_i\phi_i(z). \tag{4.21}$$

Integrating (4.21) twice from 0 to z, we get

$$u'(z) = \sum_{i=1}^{N} c_i\phi_{1,i}(z) + u'(0), \tag{4.22}$$

$$u(z) = \sum_{i=1}^{N} c_i\phi_{2,i}(z) + zu'(0) + u(0). \tag{4.23}$$

Using boundary conditions (1.7) in (4.22)–(4.23), we get

$$u'(z) = \sum_{i=1}^{N} c_i\phi_{1,i}(z), \tag{4.24}$$

$$u(z) = \sum_{i=1}^{N} c_i\phi_{2,i}(z) + u(0). \tag{4.25}$$

Putting $z = 1$ in (4.25) and after simplification, we get

$$u(0) = 1 - \sum_{i=1}^{N} c_i\Phi_{2,i}. \tag{4.26}$$

Therefore equation (4.25) becomes

$$u(z) = \sum_{i=1}^{N} c_i\left(\phi_{2,i}(z) - \Phi_{2,i}\right) + 1. \tag{4.27}$$

Putting the values of $u''(z)$, $u'(z)$, and $u(z)$ from equations (4.21,4.24,4.27) in equation (1.6) and collocating at $z = z_l = \frac{l-0.5}{N}$, where $l = 1, 2, \ldots, N$, yields the following system of nonlinear equations:

$$\sum_{i=1}^{N} c_i\phi_i(z_l) + \frac{2}{z_l}\sum_{i=1}^{N} c_i\phi_{1,i}(z_l) - \rho^2\left(\sum_{i=1}^{N} c_i\left(\phi_{2,i}(z_l) - \Phi_{2,i}\right) + 1\right)e^{\{\frac{\gamma\beta(1-(\sum_{i=1}^{N} c_i(\phi_{2,i}(z_l)-\Phi_{2,i})+1))}{1+\beta(1-(\sum_{i=1}^{N} c_i(\phi_{2,i}(z_l)-\Phi_{2,i})+1))}\}} = 0. \tag{4.28}$$

After solving this system of nonlinear equations by Newton's method, we get the unknown wavelet coefficients c_i's. On putting these c_i's in equation (4.27), we get the approximate wavelet solutions of spherical catalyst equation.

Method for Solution of Spherical Biocatalyst Equation

The same procedure has been implemented as in case of spherical catalyst equation. Substituting the values of $u''(z)$, $u'(z)$, and $u(z)$ from equations (4.21), (4.24), (4.27) in equation (1.8) and collocating at $z = z_l = \frac{l-0.5}{N}$, where $l = 1, 2, \ldots, N$, yields the following system of nonlinear equations:

$$\sum_{i=1}^{N} c_i\phi_i(z_l) + \frac{2}{z_l}\sum_{i=1}^{N} c_i\phi_{1,i}(z_l) - \rho^2\frac{(1+\beta)(\sum_{i=1}^{N} c_i(\phi_{2,i}(z_l)-\Phi_{2,i})+1)}{1+\beta(\sum_{i=1}^{N} c_i(\phi_{2,i}(z_l)-\Phi_{2,i})+1)} = 0. \tag{4.29}$$

Solving this system of nonlinear equations, we get the unknown wavelet coefficients c_i's. After putting the values of c_i's in equation (4.27), we get the approximate wavelet solutions of spherical biocatalyst equation.

ERROR BOUNDS

Lemma 5.1

Let $u(z) \in C^M\,[0, 1]$ *with* $|u^M(z)| \leq \lambda$, $\forall_z \in (0, 1)$; $\alpha > 0$ *and* $u(z) \simeq \sum_{n=1}^{2^{k-1}}\sum_{m=0}^{M-1} a_{n,m}\phi_{n,m}(z) = \sum_{i=0}^{N} a_i\phi_i(z),$ *where* $\phi_{n,m}(z)$ *is Jacobi or Bernoulli wavelet. Then* $|a_i| \leq \alpha_M 2^{-m(M+\frac{1}{2})}\lambda.$

Lemma 5.2

Let$u(z)\in C^M[0,1]$*and*$u(z) \simeq \sum_{n=1}^{2^{k-1}}\sum_{m=0}^{M-1} a_{n,m}\phi_{n,m}(z) = \sum_{i=0}^{N} a_i\phi_i(z)$*.Let* $\varepsilon_m(z)$ *be the error of approximation. Then* $|\varepsilon_m(z)| \leq \alpha_M\frac{2^{-mM}}{1-2^{-m}}(\frac{M}{2}-1)\lambda\alpha_\Phi.$

Proof

$a_{n,m}\phi_{n,m}(z), \; z \in R$

$\varepsilon_m(z) = u(z) - P_{V_m}u(z) = \sum_{m=M}^{\infty}\sum_{n=2^{k-1}+1}^{\infty}$ and

$|\varepsilon_m(z)| \leq \sum_{m=M}^{\infty}\sum_{n=2^{k-1}+1}^{\infty} |a_{n,m}\phi_{n,m}(z)|, z \in R$

Using Lemma 5.1, we get

$$\left|\varepsilon_m(z)\right| \leq \sum_{m=M}^{\infty}\sum_{n=2^{k-1}+1}^{\infty} \alpha_M 2^{-m(M+\frac{1}{2})}\max_{z\in I_n^m}\left|u^M(z)\right|2^{\frac{m}{2}}\alpha_\phi,$$

where $\alpha_\phi = \max_{z\in I_n^m}|\phi(2^m z - n)|$.

$$\left|\varepsilon_m(z)\right| \leq \sum_{m=M}^{\infty}\alpha_M 2^{-mM}(2M-1)\max_{z\in I_n^m}\left|g^M(z)\right|\alpha_{\phi G}.$$

For very large $m, |\varepsilon_m(z)| \leq \alpha_M \frac{2^{-mM}}{1-2^{-m}}(2M-1)\lambda\alpha_\phi,$ where $\lambda = \max_{z \in I_n^m} |g^M(z)|$.

Theorem 5.3

Let u(z)u(z)*be the exact solution of* (1.2), (1.4), (1.6), *and* (1.8) *and* $u_N(z)$ *be the approximate solution*, *and let* $\varepsilon_m(z)$ *be the error of approximation*. *Then* $|\varepsilon_m(z)| = O(2^{-mM})|$.

Proof

Here, we calculate the error bounds for solution of (1.2). The same procedure can be applied for equations (1.4), (1.6), and (1.8).

The error is given by

$$\left|\varepsilon_m(z)\right| = \left|u(z) - u_N(z)\right| = \left|\sum_{m=M}^{\infty} \sum_{n=2^{k-1}+1}^{\infty} a_{n,m}\left(\phi_{n,m}^2(z) - \Phi_{n,m}^1\right)\right|, \tag{5.1}$$

where $\phi_{n,m}^2(z)$ is the second integration of $\phi_{n,m}(z)$ from 0 to z and $\Phi_{n,m}^1$ denotes the second integration of $\phi_{n,m}(z)$ from 0 to 1. Therefore,

$$\left|\varepsilon_m(z)\right| \leq \sum_{m=M}^{\infty} \sum_{n=2^{k-1}+1}^{\infty} \left|a_{n,m}\left(\phi_{n,m}^2(z) - \Phi_{n,m}^1\right)\right|. \tag{5.2}$$

Using Lemma 5.1, we get

$$\left|\varepsilon_m(z)\right| \leq \sum_{m=M}^{\infty} \sum_{n=2^{k-1}+1}^{\infty} \alpha_M 2^{-m(M+\frac{1}{2})} \left|\left(\phi_{n,m}^2(z) - \Phi_{n,m}^1\right)\right| \tag{5.3}$$

$$= \sum_{m=M}^{\infty} \sum_{n=2^{k-1}+1}^{\infty} \alpha_M 2^{-m(M+\frac{1}{2})} 2^{\frac{m}{2}} \left|\left(\phi^2(2^m z - n) - \Phi_{n,m}^1\right)\right| \tag{5.4}$$

$$\leq \sum_{m=M}^{\infty} \alpha_M 2^{-mM}(2M-1)\lambda\gamma_\phi, \tag{5.5}$$

where $\gamma_{\phi} = \max_{z \in I_n^m} |(\phi^2(2^m z - n) - \Phi^1_{n,m})|$. Hence $|\varepsilon_m(z)| = O(2^{-mM})$
.

It is clear that each of the Jacobi and Bernoulli wavelet methods has an exponential rate of convergence/spectral accuracy.

NUMERICAL SIMULATION

In this section, we solve the examples of electrohydrodynamic model flow in a circular cylindrical conduit, nonlinear heat conduction model in the human head, spherical catalyst equation, and spherical biocatalyst equation. For the sake of comparison, the resultant approximate analytical solution has been used to find the solution at any point in the interval [0, 1]. We have chosen the initial guess as a zero vector of length *N*. We have used optimality tolerance $=10^{-06}$ and function tolerance $=10^{-03}$ in stopping criteria for Newton's method.

Numerical Treatment of EHM Equation

We applied Bernoulli wavelet series method (BWSM) and Jacobi wavelet series method (JWSM) $\left(\kappa = -\frac{3}{7}, \omega = -\frac{1}{8}\right)$ for the solution of (1.2). First we study the effect of nonlinearity (α) on the velocity profile at small value of the Hartmann number (*H*) and observe that as we increase *H*, the velocity profile becomes flatter near to the center, see Fig. 5. For small value of *H*, the velocity profile almost remains parabolic with change in α, see Fig. 6. We also study the influence of different *H* with fixed αand see that a strong boundary layer is build up in velocity for a large value of *H*, see Figs. 7 and 8. We see that BWSM result for fixed values $H^2 = 2, 100$ with different value of $\alpha = 0.1, 0.5, 1$ and for fixed values of $\alpha = 0.1, 1$ with different value of $H^2 = 0.5, 2, 16, 36, 49, 64$ agrees with the result of SSNM (sixth-order spline numerical method), see Figs. 10, 12, 3, and 4 of [29]. The numerical solution by BWSM and JWSM for different values of H^2 is given in Tables 1 and 2, respectively. The absolute residual errors and CPU time for different values of *J* are shown in Table 3.

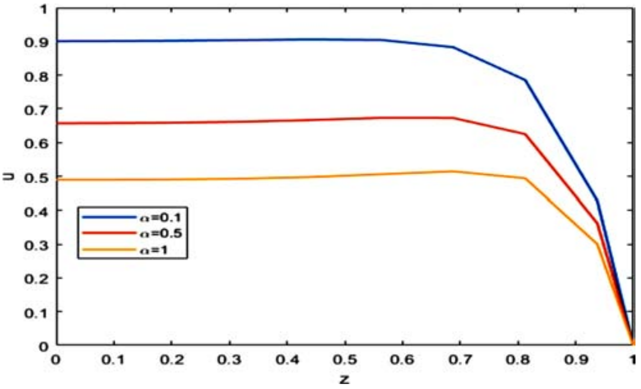


Figure 5. Graph of BWSM solution of EHD equation for *J* = 3, *M* = 8 and *k* = 1, H^2 =100.

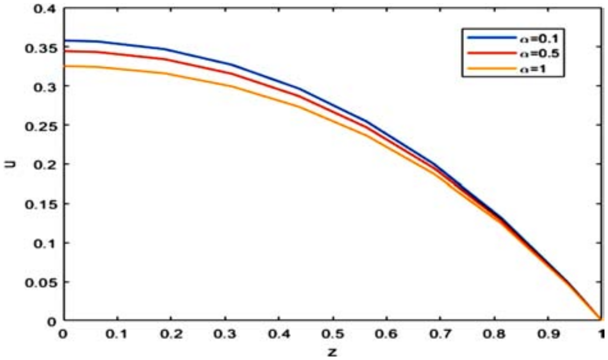


Figure 6. Graph of JWSM solution of EHD equation for *J* = 3, *M* = 8 and *k* = 1, H^2 = 2.

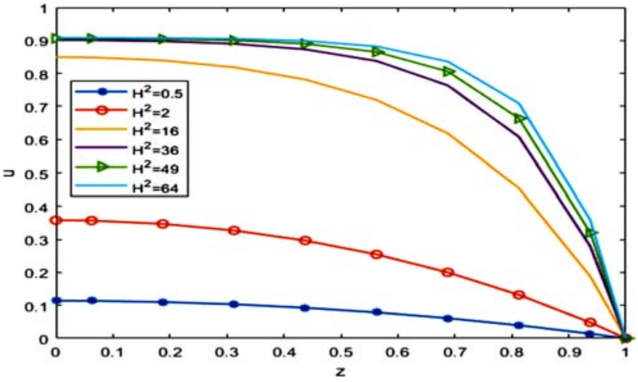


Figure 7. Graph of BWSM solution of EHD equation for *J* = 3, *M* = 8 and *k* = 1, α = 0.1.

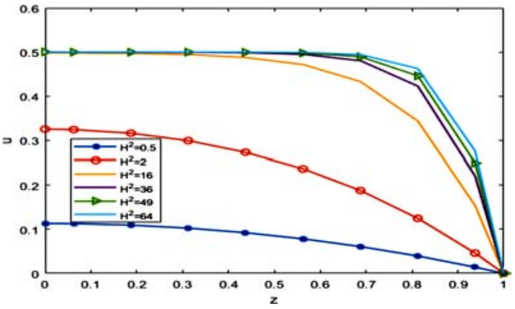


Figure 8. Graph of JWSM solution of EHD equation for $J = 3$, $M = 8$ and $k = 1$, $\alpha = 1$.

Table 1. BWSM solution of EHM equation for $\alpha = 0.1$ with $J = 3$, $M = 8$, $k = 1$

z	$\boldsymbol{H^2 = 0.5}$	$\boldsymbol{H^2 = 2}$	$\boldsymbol{H^2 = 16}$	$\boldsymbol{H^2 = 36}$	$\boldsymbol{H^2 = 49}$
0	0.1141	0.3583	0.8498	0.9010	0.9062
0.0625	0.1137	0.3571	0.8487	0.9006	0.9060
0.1875	0.1102	0.3472	0.8394	0.8976	0.9044
0.3125	0.1033	0.3272	0.8186	0.8898	0.9001
0.4375	0.0928	0.2966	0.7816	0.8730	0.8898
0.5625	0.0788	0.2546	0.7196	0.8377	0.8652
0.6875	0.0611	0.2002	0.6182	0.7636	0.8060
0.8125	0.0396	0.1320	0.4537	0.6075	0.6635
0.9375	0.0142	0.0483	0.1881	0.2795	0.3207
1	0	0	0	0	0

Table 2. JWSM solution of EHD equation for $\alpha = 1$ with $J = 3$, $M = 8$, $k = 1$

z	$\boldsymbol{H^2 = 0.5}$	$\boldsymbol{H^2 = 2}$	$\boldsymbol{H^2 = 16}$	$\boldsymbol{H^2 = 36}$	$\boldsymbol{H^2 = 49}$
0	0.1132	0.3255	0.4984	0.5003	0.5004
0.0625	0.1128	0.3244	0.4982	0.5000	0.5000
0.1875	0.1094	0.3163	0.4973	0.4999	0.4999
0.3125	0.1025	0.2995	0.4946	0.4997	0.4999
0.4375	0.0922	0.2733	0.4879	0.4988	0.4996
0.5625	0.0783	0.2366	0.4718	0.4951	0.4980
0.6875	0.0607	0.1877	0.4331	0.4805	0.4896
0.8125	0.0394	0.1248	0.3440	0.4230	0.4463
0.9375	0.0141	0.0460	0.1540	0.2204	0.2495
1	0	0	0	0	0

Table 3. Maximum absolute residual errors of EHD equation for α = 1 and H^2 = 0.5

J	JWSM	CPU time	BWSM	CPU time
3	5.5511e−15	0.38 seconds	4.2466e−14	0.9 seconds
4	5.1070e−15	0.49 seconds	9.9920e−16	2 seconds

Numerical Treatment of Nonlinear Heat Conduction Model in the Human Head

Consider equation (1.4) along with boundary conditions (1.5) and $p(u) = e^{-u}$, γ = 1, ν = 1, μ = 2, and $u_k = 0$ and get the following Emden–Fowler type equation:

$$u''(z) + \frac{2}{z}u'(z) + e^{-u} = 0, \qquad (6.1)$$

and the boundary conditions become

$$u'(0) = 0; \qquad u'(1) + 2u(1) = 0. \qquad (6.2)$$

We used Bernoulli and Jacobi wavelets for solving this problem. The calculation has been done by taking $\kappa = -\frac{1}{4}$ and $\omega = -\frac{1}{3}$ in Jacobi wavelet. A comparison of our results with the results of Haar solution [35] and ADM [36] is given in Table 4. We show the absolute residual errors and CPU time for different J in Table 5. Figures 9 and 10 show the BWSM and JWSM solution at J = 3 for different values of γ, respectively.

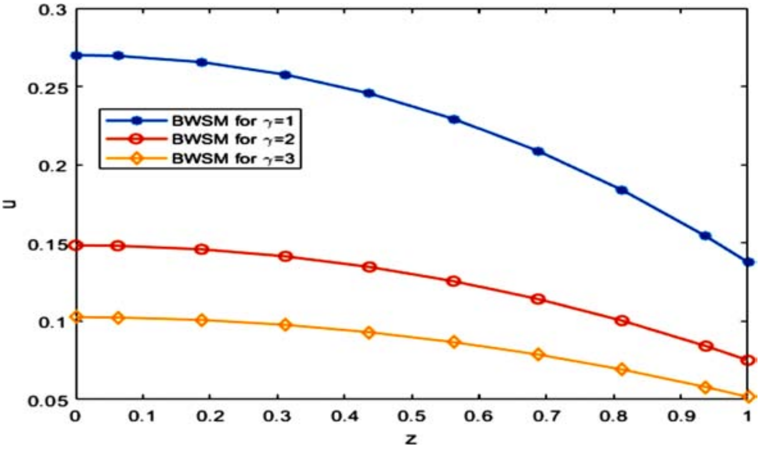


Figure 9. Graph of BWSM solution of nonlinear heat conduction model in human head equation for different γ with J = 3, M = 8, and k = 1.

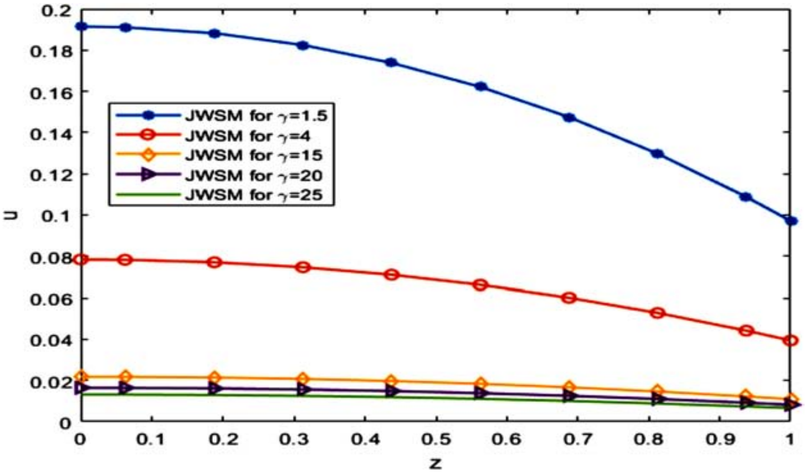


Figure 10. Graph of JWSM solution of nonlinear heat conduction model in human head equation for different γ with $J = 3$, $M = 8$, and $k = 1$.

Table 4. Numerical solution of nonlinear heat conduction model in the human head for γ = 1 with $J = 3$, $M = 8$, $k = 1$

z	BWSM	JWSM	Haar [35]	ADM [36]
0	0.2700	0.2700	–	–
0.1	0.2688	0.2688	0.26866	0.26862
0.2	0.2649	0.2649	0.26484	0.26480
0.3	0.2585	0.2585	0.25845	0.25841
0.4	0.2495	0.2495	0.24945	0.24943
0.5	0.2379	0.2379	0.23782	0.23781
0.6	0.2236	0.2236	0.22349	0.22349
0.7	0.2065	0.2065	0.20640	0.20641
0.8	0.1866	0.1866	0.18646	0.18648
0.9	0.1637	0.1637	0.16356	0.16359

Table 5. Maximum absolute residual errors of nonlinear heat conduction model for γ = 1

J	JWSM	CPU time	BWSM	CPU time
3	1.6263e−12	0.4 seconds	2.1283e−13	1.4 seconds
4	8.7896e−13	0.5 seconds	8.6597e−15	2 seconds

Numerical Treatment of Spherical Catalyst Equation

Consider equation (1.6) with (1.7) by taking β = 1, ρ = 1. We have performed BWSM and JWSM $\left(\kappa = \frac{1}{5}, \omega = -\frac{1}{6}\right)$ for the solution of (1.6). The influence of different values of activation energy is shown in Figs. 11 and 12. The numerical solution of (1.6) for σ = 0.5, 1, 1.5 is given in Table 6. We compare our results with the results of OHAM [29] in Table 7.

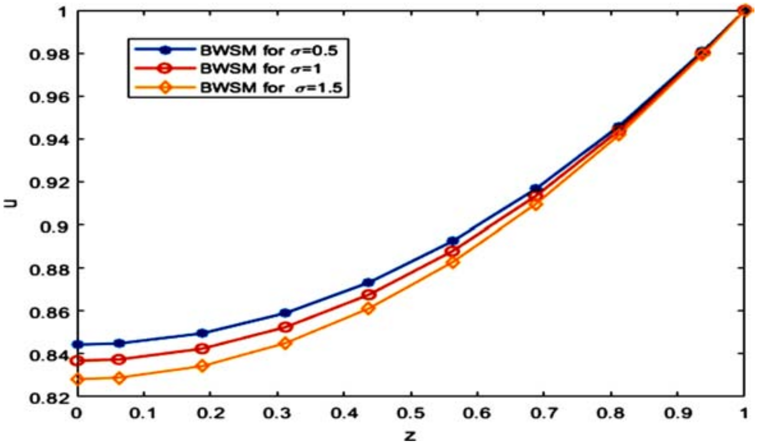


Figure 11. Graph of BWSM solution of spherical catalyst equation for $J = 3$, $M = 8$ and $k = 1$, σ = 0.5, 1, 1.5.

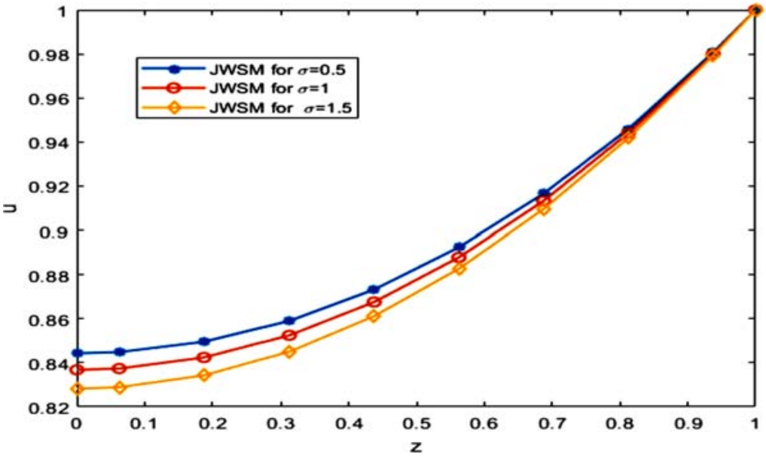


Figure 12. Graph of JWSM solution of spherical catalyst equation for $J = 3$, $M = 8$ and $k = 1$, σ = 0.5, 1, 1.5.

Table 6. Numerical solution of spherical catalyst equation for β = 1, ρ = 1 with $J = 3$, $M = 8$, $k = 1$

z	**BWSM (σ=0.5)**	**BWSM (σ=1)**	**JWSM (σ=1.5)**	**OHAM [34] (σ=0.5)**	**OHAM [34] (σ=1.5)**
0	0.8443	0.8368	0.8282	–	–
0.1	0.8458	0.8384	0.8299	0.8457	0.8299
0.2	0.8503	0.8432	0.8351	0.8502	0.8351
0.3	0.8579	0.8512	0.8437	0.8578	0.8437
0.4	0.8685	0.8625	0.8558	0.8684	0.8557
0.5	0.8822	0.8771	0.8713	0.8822	0.8712
0.6	0.8991	0.8950	0.8902	0.8991	0.8902
0.7	0.9193	0.9162	0.9126	0.9193	0.9126
0.8	0.9428	0.9407	0.9384	0.9427	0.9383
0.9	0.9696	0.9687	0.9675	0.9696	0.9675

Table 7. Maximum absolute residual errors of spherical catalyst equation for β = ρ = 1 with $J = 2$

σ	**OHAM [34]**	**JWSM**	**CPU time**	**BWSM**	**CPU time**
0.5	5.66e−05	1.1768e−14	0.4 seconds	2.5158e−13	0.5 seconds
1	4.75e−05	1.0628e−12	0.5 seconds	2.9277e−12	0.5 seconds
1.5	4.28e−04	1.6083e−11	0.4 seconds	7.4860e−12	0.5 seconds

Numerical Treatment of Spherical Biocatalyst Equation

Consider equation (1.8) with (1.9) by fixing β = 2. We have performed BWSM and JWSM $\left(\kappa = \frac{-3}{5}, \omega = -\frac{1}{8}\right)$ for the solution of (1.8). The influence of different values of Thiele modulus is shown in Figs. 13 and 14. The numerical solution of (1.6) for ρ = 1, 1.5, 2 is given in Table 8. We compare our results with the results of OHAM [34] in Table 9.

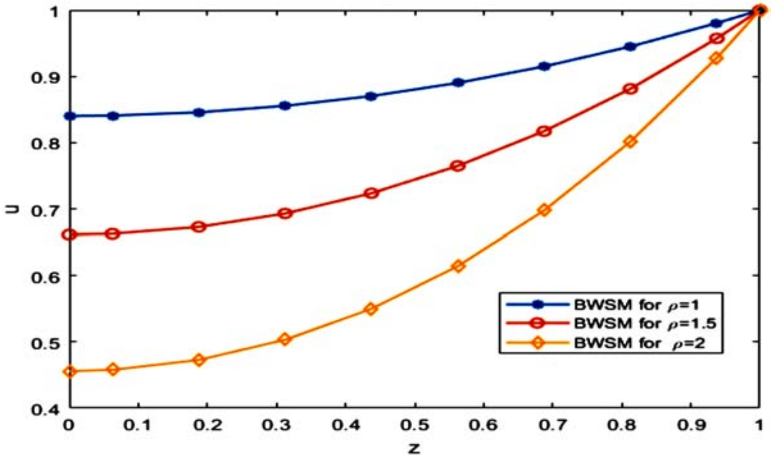


Figure 13. Graph of BWSM solution of spherical biocatalyst equation for $J = 3$, $M = 8$ and $k = 1$, $\beta = 2$.

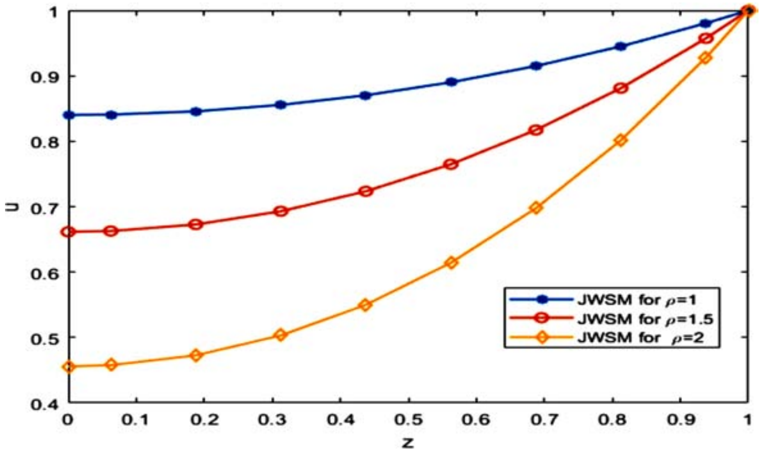


Figure 14. Graph of JWSM solution of spherical biocatalyst equation for $J = 3$, $M = 8$ and $k = 1$, $\beta = 2$.

Table 8. Numerical solution of spherical biocatalyst equation for $\beta = 2$ with $J = 3$, $M = 8$, $k = 1$

z	**BWSM ($\rho = 1$)**	**BWSM ($\rho = 1.5$)**	**JWSM ($\rho = 2$)**	**OHAM [34] ($\rho = 1$)**	**OHAM [34] ($\rho = 1.5$)**
0	0.8401	0.6615	0.4559	–	–
0.1	0.8417	0.6647	0.4607	0.8417	0.6646

0.2	0.8464	0.6743	0.4751	0.8464	0.6743
0.3	0.8543	0.6905	0.4994	0.8542	0.6904
0.4	0.8653	0.7132	0.5342	0.8653	0.7132
0.5	0.8795	0.7428	0.5798	0.8795	0.7427
0.6	0.8970	0.7793	0.6372	0.8969	0.7793
0.7	0.9177	0.8230	0.7070	0.9177	0.8230
0.8	0.9418	0.8742	0.7902	0.9417	0.8741
0.9	0.9692	0.9331	0.8875	0.9691	0.9330

Table 9. Maximum absolute residual errors of spherical biocatalyst equation for β = 2 with *J* = 2.

ρ	**OHAM [34]**	**JWSM**	**CPU time**	**BWSM**	**CPU time**
1	1.21e−06	8.8425e−08	0.4 seconds	5.6621e−15	0.5 seconds
2	1.98e−06	7.8504e−08	0.6 seconds	6.2004e−10	0.6 seconds
3	6.02e−04	2.2560e−13	0.5 seconds	1.9376e−12	0.5 seconds

CONCLUSION

In this paper, we have studied EHD flow in a charged circular cylinder conduit, nonlinear heat conduction model in the human head, non-isothermal reaction–diffusion model equations in a spherical catalyst, and non-isothermal reaction–diffusion model equations in a spherical biocatalyst which are modeled by Lane–Emden type equations having strong nonlinearity. We have solved these models by two numerical methods based on Jacobi and Bernoulli wavelets. These wavelet methods solved Lane–Emden type equations by converting them into a system of nonlinear equations. In the study of EHD flow, we observed that the effects of Hartmann number and nonlinearity have an important impact. Further we also compare our results with the results of SSNM [29], Haar [35], ADM [36], and OHAM [34]. The graphs show the efficiency of our methods. Moreover, the present semi-analytical numerical methods have lower computational cost than ADM, Haar, and OHAM, since in our methods there is no need for symbolic successive integration which is computationally higher than numerical methods.

ACKNOWLEDGEMENTS

The authors are indebted to the anonymous reviewer for his helpful, valuable comments and suggestions in the improvement of the manuscript. First author is thankful to Council of Scientific and Industrial Research (CSIR), Govt. of India, for providing Junior Research Fellowship.

REFERENCES

1. Azodi, H.D.: Numerical solution of fractional-order sir epidemic model via Jacobi wavelets. J. Int. Math. Virtual Inst. 10(1), 183–197 (2020)
2. Duggan, R.C., Goodman, A.M.: Pointwise bounds for a nonlinear heat conduction model of the human head. Bull. Math. Biol. 48(2), 229–236 (1986)
3. El-Kabeir, S.M.M., El-Zahar, E.R., Modather, M., Gorla, R.S.R., Rashad, A.M.: Unsteady MHD slip flow of a ferrofluid over an impulsively stretched vertical surface. AIP Adv. 9(4), 045112 (2019)
4. El-Zahar, E.R., Rashad, A.M., Saad, W., Seddek, L.F.: Magneto-hybrid nanofluids flow via mixed convection past a radiative circular cylinder. Sci. Rep. 10(1), 1–13 (2020)
5. Fylladitakis, E.D., Theodoridis, M.P., Moronis, A.X.: Review on the history, research and application of electrohydrodynamics. IEEE Trans. Plasma Sci. 42(2), 358–375 (2014)
6. Ghasemi, S.E., Hatami, M., Ahangar, G.R.M., Ganji, D.D.: Electrohydrodynamic flow analysis in a circular cylindrical conduit using least square method. J. Electrost. 72, 47–52 (2014)
7. Hossmann, K.-A., Hermann, D.M.: Effects of electromagnetic radiation of mobile phones on the central nervous system. Bioelectromagnetics 24(1), 49–62 (2003)
8. Jamal, B., Khuri, S.A.: Non-isothermal reaction–diffusion model equations in a spherical biocatalyst:Green's function and fixed point iteration approach. Int. J. Appl. Comput. Math. 5(4), 120 (2019). https://doi.org/10.1007/s40819-019-0704-1
9. Keangin, P., Rattanadecho, P., Wessapan, T.: An analysis of heat transfer in liver tissue during microwave ablation using single and 2 double slot antenna. Int. Commun. Heat Mass Transf. 38, 757–766 (2011)
10. Keshavarz, E., Ordokhani, Y., Razzaghi, M.: Numerical solution of nonlinear mixed Fredholm–Volterra integro-differential equations of fractional order by Bernoulli wavelets. Comput. Methods Differ. Equ. 7(2), 163–176 (2019)
11. Ketley, V., Wood, A.W., Spoung, J., Stough, C.: Neuropsychological sequelae of digital mobile phone exposure in humans. Neuropsychologia 44(10), 1843–1848 (2006)
12. Lin, L.: Cataracts and personal communication radiation. IEEE Microw. Mag. 4, 26–32 (2003)

13. Lindholm, H., Alanko, T., Rintamäki, H.: Thermal effects of mobile phone RF fields on children: a provacation study. Prog. Biophys. Mol. Biol. 107(3), 399–403 (2011)
14. Madduri, H., Roul, P.: A fast-converging iterative scheme for solving a system of Lane–Emden equations arising in catalytic diffusion reactions. J. Math. Chem. 57(2), 570–582 (2019)
15. Mastroberardino, A.: Homotopy analysis method applied to electrohydrodynamic flow. Commun. Nonlinear Sci. Numer. Simul. 16, 2730–2736 (2011)
16. Mckee, S., Watson, R., Cuminato, J.A., Caldwell, J., Chen, M.S.: Calculation of electro-hydrodynamic flow in a circular cylindrical conduit. Z. Angew. Math. Mech. 77, 457–465 (1997)
17. Mosayebidorcheh, S.: Taylor series solution of the electrohydrodynamic flow equation. J. Mech. Eng. Technol. 1(2), 40–45 (2013)
18. Rach, R., Duan, J.S., Wazwaz, A.M.: Solving coupled Lane–Emden boundary value problems in catalytic diffusion reactions by the Adomian decomposition method. J. Math. Chem. 52(1), 255–267 (2014)
19. Rach, R., Duan, J.S., Wazwaz, A.M.: On the solution of non-isothermal reaction–diffusion model equations in a spherical catalyst by the modified Adomian method. Chem. Eng. Commun. 202(8), 1081–1088 (2015)
20. Rong, L.J., Phang, C.: Jacobi wavelet operational matrix of fractional integration for solving fractional integro-differential equation. J. Phys. Conf. Ser. 693, 012002 (2016)
21. Roul, P.: An improved iterative technique for solving nonlinear doubly singular two-point boundary value problems. Eur. Phys. J. Plus 131(6), 1–15 (2016)
22. Roul, P.: Doubly singular boundary value problems with derivative dependent source function: a fast-converging iterative approach. Math. Methods Appl. Sci. 42(1), 354–374 (2019)
23. Roul, P.: A new mixed MADM-collocation approach for solving a class of Lane–Emden singular boundary value problems. J. Math. Chem. 57(3), 945–969 (2019)
24. Roul, P.: A fourth-order non-uniform mesh optimal B-spline collocation method for solving a strongly nonlinear singular boundary value problem describing electrohydrodynamic flow of a fluid. Appl. Numer.

Math. (2020). https://doi.org/10.1016/j.apnum.2020.03.018

25. Roul, P., Goura, V.P., Agarwal, R.: A compact finite difference method for a general class of nonlinear singular boundary value problems with Neumann and Robin boundary conditions. Appl. Math. Comput. 350, 283–304 (2019)
26. Roul, P., Madduri, H.: A new highly accurate domain decomposition optimal homotopy analysis method and its convergence for singular boundary value problems. Math. Methods Appl. Sci. 41(16), 6625–6644 (2018)
27. Roul, P., Madduri, H., Agarwal, R.: A fast-converging recursive approach for Lane–Emden type initial value problems arising in astrophysics. J. Comput. Appl. Math. 359, 182–195 (2019)
28. Roul, P., Madduri, H., Kassner, K.: A new iterative algorithm for a strongly nonlinear singular boundary value problem. J. Comput. Appl. Math. 351, 167–178 (2019)
29. Roul, P., Madduri, H., Kassner, K.: A sixth-order numerical method for a strongly nonlinear singular boundary value problem governing electrohydrodynamic flow in a circular cylindrical conduit. Appl. Math. Comput. 350, 416–433 (2019)
30. Roul, P., Thula, K.: A fourth-order B-spline collocation method and its error analysis for Bratu-type and Lane–Emden problems. Int. J. Comput. Math. 96(1), 85–104 (2019)
31. Roul, P., Thula, K., Agarwal, R.: Non-optimal fourth-order and optimal sixth-order B-spline collocation methods for Lane–Emden boundary value problems. Appl. Numer. Math. 145, 342–360 (2019)
32. Roul, P., Thula, K., Goura, V.P.: An optimal sixth-order quartic B-spline collocation method for solving Bratu-type and Lane–Emden-type problems. Math. Methods Appl. Sci. 42(8), 2613–2630 (2019)
33. Roul, P., Warbhe, U.: A new homotopy perturbation scheme for solving singular boundary value problems arising in various physical models. Z. Naturforsch. A 72(8), 733–743 (2017)
34. Singh, R.: Optimal homotopy analysis method for the non-isothermal reaction–diffusion model equations in a spherical catalyst. J. Math. Chem. 56(9), 2579–2590 (2018)
35. Singh, R., Guleria, V., Singh, M.: Haar wavelet quasilinearization method for numerical solution of Emden–Fowler type equations. Math. Comput. Simul. 174, 123–133 (2020)

36. Singh, R., Kumar, J.: An efficient numerical technique for the solution of nonlinear singular boundary value problems. Comput. Phys. Commun. 185(4), 1282–1289 (2014)

37. Tullis, T.K., Bayazitoglu, Y.: Analysis of relaxation times on the human head using the thermal wave model. Int. J. Heat Mass Transf. 67, 1007–1013 (2013)

38. Van Gorder, R.A.: Exact first integrals for a Lane–Emden equation of the second kind modeling a thermal explosion in a rectangular slab. New Astron. 16(8), 492–497 (2011)

39. Wainwright, P.: Thermal effects of radiation from cellular telephones. Phys. Med. Biol. 45(8), 2363–2372 (2000)

40. Wazwaz, A.M.: The variational iteration method for solving new fourth-order Emden–Fowler type equations. Chem. Eng. Commun. 202(11), 1425–1437 (2015)

41. Wazwaz, A.M.: Solving the non-isothermal reaction–diffusion model equations in a spherical catalyst by the variational iteration method. Chem. Phys. Lett. 679, 132–136 (2017)

42. Wessapan, T., Rattanadecho, P.: Numerical analysis of specific absorption rate and heat transfer in human head subjected to mobile phone radiation: effects of user age and radiated power. J. Heat Transf. 134, 121101 (2012)

43. Wessapan, T., Srisawatdhisukul, S., Rattanadecho, P.: Numerical analysis of specific absorption rate and heat transfer in the human body exposed to leakage electromagnetic field at 915 MHz and 2450 MHz. ASME J. Heat Transfer. 133, 051101 (2011)

44. Wessapan, T., Srisawatdhisukul, S., Rattanadecho, P.: Specific absorption rate and temperature distributions in human head subjected to mobile phone radiation at different frequencies. Int. J. Heat Mass Transf. 55(1–3), 347–359 (2012)

45. Xiaopeng, C., Jiusheng, C., Xiezhen, Y.: Advances and applications of electrohydrodynamics. Chin. Sci. Bull. 48, 1055–1063 (2003)

46. Zogheib, B., Tohidi, E., Shateyi, S.: Bernoulli collocation method for solving linear multidimensional diffusion and wave equations with Dirichlet boundary conditions. Adv. Math. Phys. 2017, Article ID 5691452 (2017). https://doi.org/10.1155/2017/5691452

Chapter

11

A MATHEMATICAL ANALYSIS OF HOPF-BIFURCATION IN A PREY-PREDATOR MODEL WITH NONLINEAR FUNCTIONAL RESPONSE

Assane Savadogo[1], Boureima Sangaré[1], and Hamidou Ouedraogo[2]

[1]Department of Mathematics and Informatics, UFR/ST, UNB, 01 BP 1091 Bobo Dsso 01, Bobo Dioulasso, Burkina Faso.
[2]Department of Mathematics and Informatics, UNB, Bobo Dioulasso, Burkina Faso.

ABSTRACT

In this paper, our aim is mathematical analysis and numerical simulation of a prey-predator model to describe the effect of predation between prey and predator with nonlinear functional response. First, we develop results concerning the boundedness, the existence and uniqueness of the solution. Furthermore, the Lyapunov principle and the Routh–Hurwitz criterion are applied to study respectively the local and global stability results. We also establish the Hopf-bifurcation to show the existence of a branch of nontrivial

Citation: (APA): Savadogo, A., Sangaré, B., & Ouedraogo, H. (2021). A mathematical analysis of Hopf-bifurcation in a prey-predator model with nonlinear functional response. Advances in Difference Equations, 2021(1), 1-23. (23 pages).

periodic solutions. Finally, numerical simulations have been accomplished to validate our analytical findings.

MSC: 65L12; 65M20; 65N40

Keywords: Prey-predator system; Hopf-bifurcation; Global stability; Numerical simulations

INTRODUCTION

The study of the dynamics relationship of the prey-predator system has long been and will continue to be one of the dominant subjects in both ecology and mathematical ecology due to its universal existence and importance. In recent decades, mathematics has had a huge impact as a tool for modeling and understanding biological phenomena. Mathematical modeling of the population dynamics of a prey-predator system is an important objective of mathematical models in biology, which has attracted the attention of many researchers [1–4]. Many authors, such as Holling 1959 [5], Getz 1984, and Arditi and Ginzburg 1989 [6, 7], studied the prey-predator system with various functional responses. These different types of functional responses present a key element for understanding the dynamics of these populations. The main questions concerning population dynamics concern the effects of interaction in the regulation of natural populations, the reduction of their size, the reduction of their natural fluctuations, or the destabilization of the equilibria in oscillations of the states of the population [8–13]. The predator-prey relationship is important to maintain the balance between different animal species. Without predators, some prey species would force other species to disappear due to competition. Without prey, there would be no predators. The main feature of predation is therefore a direct impact of the predator on the prey population.

It is in this line of thought that we are interested here in the study of the dynamics of prey-predator populations with an alternative food resource for predators, meaning that the predator population can survive if there is no prey. Our objective is to understand what is the impact of predation on the dynamics of prey and predator species, in order to avoid any extinction of the two species.

Several authors have studied the prey-predator model with logistics growth in both species. Haque in [14] proposed a prey-predator model with logistic growth in both species and a linear functional response. The author assumed that the predator has logistic growth rate since it has sufficient resources for alternative foods; and it is argued that alternative food sources may have an important role in promoting the persistence of predator-prey systems. Guin in [15] studied a prey-predator model with logistic growth in both species and using ratio-dependent functional for predators.

Motivated by the above works, we consider the following predator-prey model [14]:

$$\begin{cases} \frac{dN}{dt} = r_1(1 - \frac{N}{K_1})N - bNP, & N(0) = N_0 > 0, \\ \frac{dP}{dt} = r_2(1 - \frac{P}{K_2})P + ebNP, & P(0) = P_0 > 0, \end{cases} \tag{1.1}$$

where

- $N(t)$ and $P(t)$ stand for the prey and predator density, respectively, at time t.
- r_1, K_1, b, e are positive constants that stand for prey intrinsic growth rate, the prey carrying capacity of the environment, predation rate per unit of time, and conversion rate, respectively.
- The term bNP models prey consumption due to predation.
- r_2, K_2 represent respectively the predator intrinsic growth rate and the predator carrying capacity of the environment.

Some similar models have appeared in the recent literature [14, 15]. We remark that the main new distinctive feature is the inclusion of Holling function response of type II. Thus, by incorporating Holling function response of type II, we describe the predation strategy. Indeed, many researchers suggested that Holling type II response is the characteristic of predators. It determines the stability and bifurcation dynamics of the model. Usually, the feeding rate of predator is saturated, so it is more realistic to consider prey dependence functional response. Our model differs from the one of [14], since in the latter the term of predation is linear. In fact, we consider Holling function response of type II defined by

$$\phi(x) = \frac{B\alpha_0 x}{1 + B\alpha_1 x},$$

where

- α_0 and α_1 represent respectively the search and capture time of the prey,
- B is the predation rate per unit of time.

Indeed, physiological absorption capabilities of prey are limited, and even if a large number of prey is available, a predator will not be able to absorb prey numbers beyond this limit, it is more realistic to design a response function with a saturation effect with the density of prey. Thus, Holling function response of type II is more appropriate. In order to sustain the coexistence of ecosystem species, it is very important to control some key demographic parameters.

The paper is organized as follows. In Sect. 2, we present the general mathematical model of the prey-predator system. Section 3 provides the mathematical analysis of the model established in Section 2. We perform some numerical simulations to support our main results in Section 4. A final discussion concludes the paper.

MATHEMATICAL MODEL FORMULATION

In this section, we proceed to the construction of the prey-predator model by taking into account the fact that the predator has an alternative source of food. Our main goal is to modify system (1.1) in order to describe the effect of predation on the prey. Our task here is to analyze the impact of predation on a predator-prey community [16–19].

The following hypotheses hold for our models:

H1H1**::** prey populations follow a logistic growth in the absence of the predator;

H2H2**::** functional response of the predator is Holling type II;

H3H3**::** the predator has an alternative source of food.

The system modeling the evolution over time of prey and predators is given by

$$\begin{cases} \frac{dN}{dt} = \psi(N) - \phi(N)P, \\ \frac{dP}{dt} = g_0(P) + g_1(N,P), \end{cases} \tag{2.1}$$

where

- ψ, ϕ, g_0, and g_1 are positive functions and C^∞;
- $\psi(N)$, $g_0(P)$ is a growth function of prey and predator population, respectively;
- $\phi(N)$ is the amount of prey consumed by a predator per time unit;

- $g_1(N, P)$ represents the rate of conversion of the prey into the predator.

The model presented here is general, and it is necessary to make choices, particularly for the functions $g_0(P)$, $g_1(N, P)$, $\phi(N)$, and $\psi(N)$. Then we make the following choices:

- $\psi(N) = r_1(1 - \frac{N}{K_1})N$ represents the dynamics of the prey population governed by the logistic equation when there is no predator;
- $g_0(P) = r_2(1 - \frac{P}{K_2})P$ represents the logistic growth of the predator population when there is no prey;
- $\phi(N) = \frac{\delta_1 N}{1+\delta_2 N}$ represents the functional response of the predator which is Holling type II;
- $g_1(N,P) = \frac{\omega NP}{1+\delta_2 N}$ represents the quantity of prey consumed by predators.

Consequently, we obtain the following nonlinear differential system defined by

$$\begin{cases} \frac{dN}{dt} = r_1(1 - \frac{N}{K_1})N - \frac{\delta_1 NP}{1+\delta_2 N}, & N(0) = N_0 > 0, \\ \frac{dP}{dt} = r_2(1 - \frac{P}{K_2})P + \frac{e\delta_1 NP}{1+\delta_2 N}, & P(0) = P_0 > 0, \end{cases} \tag{2.2}$$

where

- $r_1, r_2 > 0$ are respectively the prey and predator growth rates;
- $K_1, K_2 > 0$ represent respectively the carrying capacity of the prey and the predator;
- δ_1 and δ_2 represent respectively predator search and satiety rates;
- $e = \frac{\omega}{\delta_1}$ represents the conversion rate of prey biomass into predatory biomass, with $0 < e < 1$;
- $\frac{\delta_1 NP}{1+\delta_2 N}$ represents the quantity of prey taken by predators per unit of time;
- $\frac{\omega NP}{1+\delta_2 N}$ represents the amount of prey consumed by predators per unit of time;

- $\frac{(\delta_1-\omega)NP}{1+\delta_2 N}$ is a residual term and represents the quantity of prey taken by predators and which did not contribute to the growth of predators.

Thus, we obtain the following interaction diagram: Fig. 1.

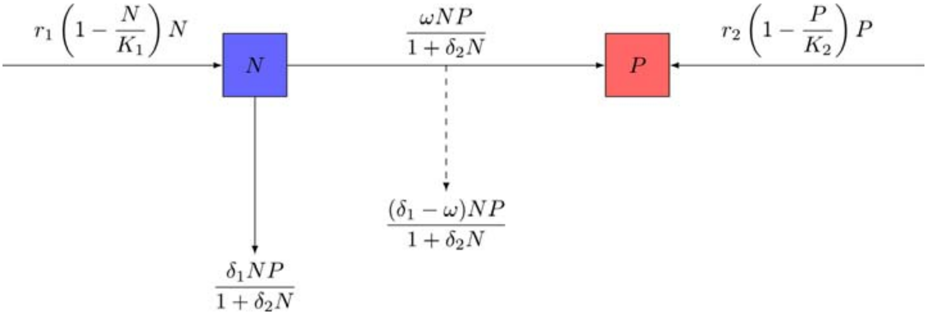

$r_1\left(1-\frac{N}{K_1}\right)N$
N
$\frac{\omega NP}{1+\delta_2 N}$
P
$r_2\left(1-\frac{P}{K_2}\right)P$
$\frac{(\delta_1-\omega)NP}{1+\delta_2 N}$
$\frac{\delta_1 NP}{1+\delta_2 N}$

Figure 1. Interaction diagram for the prey-predator model.

Using the above assumptions and according to Figure 1, at any time t > 0, the dynamics of the system can be represented by the following set of differential equations:

$$\begin{cases} G_1(N,P) = \frac{dN}{dt} = r_1(1-\frac{N}{K_1})N - \frac{\delta_1 NP}{1+\delta_2 N}, & N(0) = N_0 > 0, \\ G_2(N,P) = \frac{dP}{dt} = r_2(1-\frac{P}{K_2})P + \frac{\omega NP}{1+\delta_2 N}, & P(0) = P_0 > 0. \end{cases} \tag{2.3}$$

MATHEMATICAL ANALYSIS

This section deals with mathematical analysis including the stability and the bifurcation analysis of system (2.3) [2, 8, 15, 20–22].

Then we rewrite model (2.3) in the following form:

$$\dot{X}(t) = G\big(X(t)\big),$$

where $X(t) = (N(t), P(t))^T$ and G is defined on R^2 by

$$G(X) = \begin{pmatrix} G_1(N,P) \\ G_2(N,P) \end{pmatrix} = \begin{pmatrix} r_1(1-\frac{N}{K_1})N - \frac{\delta_1 NP}{1+\delta_2 N} \\ r_2(1-\frac{P}{K_2})P + \frac{\omega NP}{1+\delta_2 N} \end{pmatrix}.$$

The preliminary results concern the existence, positiveness, and boundedness of solutions of system (2.3).

Existence, Positiveness, and Boundedness of Solutions

From the biological point of view, it is important to show the existence, positivity, and boundedness of the solution of system (2.3) [9, 19, 23, 24].

Proposition 1

System (2.3) *admits a unique global solution* (N(t), P(t)) *defined on the interval* [0, $T_{\max}$[. *Moreover*, *the set*

$$A := \left\{(N,P) \in \mathbb{R}_+^2 / 0 \leq N \leq K_1, 0 \leq P \leq K_p\right\} \quad \textit{with}\ K_p = K_2\left(1 + \frac{\omega}{r_2\delta_2}\right)$$

is positively invariant and absorbing for system (2.3).

Proof

Indeed,

- the theorem of Cauchy–Lipschitz [11] assures the existence and uniqueness of local solution of system (2.3) on [0, $T_{\max}$[given the regularity of the functions involved in the model.
- Now, let us show that the set $A = \{(N,P) \in \mathbb{R}_+^2 / 0 \leq N \leq K_1, 0 \leq P \leq K_p\}$ is positively invariant and absorbing for system (2.3).

Let us show that

$$A_0 = \left\{(N,P) \in \mathbb{R}_+^2 / 0 \leq N \leq K_1\right\} \quad \text{and} \quad B = \left\{(N,P) \in \mathbb{R}_+^2 / 0 \leq P \leq K_p\right\}$$

are positively invariant and absorbing for system (2.3).

Let us prove that

$$A_1 = \left\{(N,P) \in \mathbb{R}^2 / N \geq 0\right\} \tag{3.1}$$

is positively invariant. Indeed, let f_1 be the function defined on R^2 by f_1(N, P) = −N. We have

$$\nabla f_1(N,P) = \begin{pmatrix} -1 \\ 0 \end{pmatrix} \neq 0_{\mathbb{R}^2} \quad \text{and} \quad \langle \nabla f_1 | G \rangle = 0 \times r_2\left(1 - \frac{P}{K_2}\right)P = 0.$$

Thus, $\langle \nabla f1|G \rangle \leq 0$ on $\{(N, P) \in R^2/N = 0\}$, where $\langle | \rangle$ is the usual scalar product.

Therefore, A_1 is positively invariant.

Proceeding in the same way, with $f_2(N, P) = -N$, we show that

$$A_2 = \left\{(N,P) \in \mathbb{R}^2/P \geq 0\right\} \tag{3.2}$$

is positively invariant.

Let us show that

$$A_3 = \left\{(N,P) \in \mathbb{R}^2/N \leq K_1\right\}. \tag{3.3}$$

Indeed, let f_3 be defined on R^2 by $f_3(N, P) = N - K_1$. We have

$$\nabla f_3(N,P) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \neq 0_{\mathbb{R}^2} \quad \text{and} \quad \langle \nabla f_1 | G \rangle = -\frac{\delta_1 K_1 P}{1+\delta_2 K_1}.$$

Thus, $\langle \nabla f_1 | G \rangle \leq 0$ on $\{(N, P) \in R^2/N = K_1\}$. Therefore, A_3 is positively invariant.

According to (3.1) and (3.3), A_0 is positively invariant.

Now, we aim to show that the set A_0 is absorbing. The variable N satisfies the inequality

$$\frac{dN}{dt} \leq r_1\left(1 - \frac{N}{K_1}\right)N,$$

and by the principle of comparison, we deduce that $\lim_{t\to+\infty}\sup N(t) \leq K_1$. Hence, for $\epsilon > 0$, there exists $T > 0$ such that $N(t) \leq \sup_{t\geq T} N(t) \leq K_1 + \epsilon$; as ϵ is arbitrary, we deduce that A_0 is absorbing.

Now, we aim to show that $P \leq K_p$. Indeed,

$$\frac{dP}{dt} = r_2\left(1 - \frac{P}{K_2}\right)P + \frac{\omega NP}{1+\delta_2 N} \quad \text{and} \quad \frac{e\delta_1 N}{1+\delta_2 N} < \frac{e\delta_1}{\delta_2},$$

thus we have the following differential inequality:

$$\frac{dP}{dt} \leq \left(r_2\left(1 - \frac{P}{K_2}\right) + \frac{e\delta_1}{\delta_2}\right)P, \quad \forall t \geq t_0. \tag{3.4}$$

According to the comparison principle, we deduce that

$$\lim_{t\to+\infty} \sup(P(t)) \leq K_2\left(1 + \frac{\omega}{r_2\delta_2}\right) = K_p. \tag{3.5}$$

According to (3.2) and (3.5), *B* is positively invariant and absorbing.

From the above result, the set defined by $A = \{(N, P) \in \mathbb{R}_+^2 / 0 \leq N \leq K_1, 0 \leq P \leq K_p\}$ is positively invariant and absorbing.

To show the global existence of solutions, we must show that the solutions of the system are bounded. In the previous demonstration we have established that *N* and *P* are bounded. Thus, we can conclude that the solutions of system (2.3) exist globally.

For the study of system (2.3), we restrain a set defined by

$$A = \left\{(N, P) \in \mathbb{R}_+^2 / 0 \leq N \leq K_1, 0 \leq P \leq K_p\right\}.$$

Stability Analysis of the Equilibria

In this section, we analyze the local and global stability of different equilibrium.

Trivial Equilibrium Points of the Model

The trivial stationary states of system (2.3) are given in the following proposition [8, 11–13].

Proposition 2

The equilibrium states are as follows:

- $E_0 = (0, 0)$, *the predators and prey are extinct*. *This equilibrium is always admissible*.
- $E_1 = (0, K_2)$, *the prey is extinct*. *This equilibrium is always admissible*.
- $E_2 = (K_1, 0)$, *the predator is extinct*. *This equilibrium is always admissible*.

Proof

Indeed, to get the equilibrium points, we solve the following system:

$$\begin{cases} r_1(1 - \frac{N}{K_1})N - \frac{\delta_1 NP}{1+\delta_2 N} = G_1(N, P) = 0, \\ r_2(1 - \frac{P}{K_2})P + \frac{\omega NP}{1+\delta_2 N} = G_2(N, P) = 0. \end{cases} \tag{3.6}$$

- We have $G_1(0, 0) = G_2(0, 0) = 0$. Thus, $E_0 = (0, 0)$ is the trivial equilibrium point.
- In the same way, $G_1(0, K_2) = G_2(0, K_2) = 0$. Then $E_1 = (0, K_2)$ is an equilibrium point of system (2.3).
- We also have $G_1(K_1, 0) = G_2(K_1, 0) = 0$. Then $E_2 = (K_1, 0)$ is an equilibrium point of system (2.3).

The local stability analysis of a trivial equilibrium point is given by the following proposition.

Proposition 3

- *E_0 and E_2 are always unstable.*
- *E_1 is locally asymptotically stable if $\delta_1 > \frac{r_1}{K_2}$, with extinction for the prey population and stability for the predator population. If $\delta_1 < \frac{r_1}{K_2}$, E_1 is unstable with stability for the predator. In addition, if $\delta_1 = \frac{r_1}{K_2}$, E_1 is a stable non-hyperbolic point.*

Proof

Indeed, let us determine the eigenvalues of the Jacobian matrix associated with each equilibrium point $E_i = 0, 1, 2$. The Jacobian matrix of system (2.3) is

$$\mathrm{DG}(X) = \begin{pmatrix} r_1\left(1 - \frac{2N}{K_1}\right) - \frac{\delta_1 P}{(1+\delta_2 N)^2} & -\frac{\delta_1 N}{1+\delta_2 N} \\ \frac{\omega P}{(1+\delta_2 N)^2} & r_2\left(1 - \frac{2P}{K_2}\right) + \frac{\omega N}{1+\delta_2 N} \end{pmatrix},$$

where $X(t) = (N(t), P(t))^T$.

- For $E_0 = (0, 0)$, the associated Jacobian matrix is $\mathrm{DG}(E_0) = \begin{pmatrix} r_1 & 0 \\ 0 & r_2 \end{pmatrix}$. The eigenvalues are $r_1 > 0$ and $r_2 > 0$. Then E_0 is always unstable. In this case, we have instability of the prey and the predator.
- For $E_2 = (K_1, 0)$, the Jacobian matrix of system (2.3) evaluated at E_2 is

$$\mathrm{DG}(E_2) = \begin{pmatrix} -r_1 & -\frac{\delta_1 K_1}{1+\delta_2 K_2} \\ 0 & r_2 + \frac{\omega K_1}{1+\delta_2 K_1} \end{pmatrix}.$$

The eigenvalues are $\lambda_1 = -r_1 < 0$ and $\lambda_2 = r_2 + \frac{\omega K_1}{1+\delta_2 K_1} > 0$. Then E_2 is unstable with stability for the prey population and instability for the predator population.

- For $E_1 = (0, K_2)$, the associated Jacobian matrix is

$$\mathrm{DG}(E_1) = \begin{pmatrix} r_1(1 - \frac{\delta_1 K_2}{r_1}) & 0 \\ \omega K_2 & -r_2 \end{pmatrix}.$$

The associated characteristic polynomial is given by

$$P_{\mathrm{DG}(E_1)}(\lambda) = \left(r_1\left(1 - \frac{\delta_1 K_2}{r_1}\right) - \lambda\right)(-r_2 - \lambda).$$

Then the eigenvalues of $\mathrm{DG}(E_1)$ are $\lambda_1 = r_1(1 - \frac{\delta_1 K_2}{r_1})$ and $\lambda_2 = -r_2 < 0$. If $\delta_1 > \frac{r_1}{K_2}$, then $\lambda_1 < 0$, therefore the system is locally asymptotically stable with extinction for the prey population and stability for the predator population. If $\delta_1 < \frac{r_1}{K_2}$, then $\lambda_1 > 0$, therefore E_1 is unstable, so we have stability for the predator population.

If $\frac{\delta_1 K_2}{r_1} = 1$, then the equilibrium E_1 is a stable non-hyperbolic point. Indeed, to study the stability of E_1, we will use the center manifold theorem [25].

The eigenvalues of $\mathrm{DG}(E_1)$ are $\lambda_1 = 0$ and $\lambda_2 = -r_2$, and the eigenspace associated with those eigenvalues is

$$W^0 = \left\langle\left(\frac{r_2}{\omega K_2}; 1\right)\right\rangle,$$

$$W^{-r_2} = \langle(0; 1)\rangle.$$

According to the center manifold theorem [25], there exists a center manifold which is tangent to W^0 at the point E_1.

Denote $x = N$ and $y = P$. The center manifold in this case is given by

$$W^c = \left\{y = h(x)/h(0) = K_2, h'(0) = \frac{\omega K_2}{r_2}\right\},$$

where $h(x) = h(0) + h'(0)x + h''(0)x^2 + \mathcal{O}(x^3)$ is analytical at the neighborhood of the origin. Denoting $a = \frac{1}{K_2}$ and $b = \frac{1}{K_1}$, we have

$$\begin{aligned}
&y = h(x)\\
&\Leftrightarrow \quad \dot{y} = \dot{x}h'(x)\\
&\Leftrightarrow \quad r_2 h(x)\big(1 - ah(x)\big) + \frac{\omega x h(x)}{1 + x\delta_2} - \left(r_1 x(1 - bx) - \frac{\delta_1 x h(x)}{1 + x\delta_2}\right)h'(x) = 0.
\end{aligned}$$

By plugging the function *h* by its expression in the previous equation and by grouping, we get:

$$\begin{aligned}
&r_2 h(0)\big(1 - ah(0)\big) + \big(r_2\delta_2 h(0)\big(1 - ah(0)\big) + r_2 h'(0)\big(1 - 2ah(0)\big) + \omega h(0)\\
&\quad + h'(0)\big(\delta_1 h(0) - r_1\big)\big)x + \big(r_1 h'(0)(b - \delta_2) + \big(h'\big)^2(\delta_1 - ar_2)\\
&\quad - 2h''(0)\big(r_1 - \delta_1 h(0)\big) + r_2\delta_2 h'(0)\big(1 - 2ah(0)\big)\\
&\quad + r_2 h''(0)(1 - 2a)\big)x^2 + \mathcal{O}\big(x^3\big) = 0.
\end{aligned} \tag{$*$}$$

Since $r_1 = \delta_1 K_2$, then $r_1 - \delta_1 K_2 = 0$, by using $(*)$, we deduce that

$$h'(0) = \frac{\omega h(0)}{r_2} = \frac{\omega K_2}{r_2},$$

$$h''(0) = h'(0)\frac{r_1(b_1 - \delta_2) + h'(0)(\delta_1 - ar_2) - r_2\delta_2}{r_2}.$$

The equation reduced to the center manifold is

$$\dot{x} = r_1 x(1 - ax) - \frac{\delta_1 x(h(0) + h'(0)x + h''(0)x^2)}{1 + \delta_2 x} = f(x).$$

A study of the sign of *f* in the neighborhood of 0 gives the following result:

- If $x < 0$, then $f(x) < 0$;
- If $x > 0$, then $f(x) < 0$. So the equilibrium point $E_1 = (K, 0)$ is stable in *A*.

Coexistence Equilibria Point of the Model

To determine the coexistence equilibrium $E_3 = (N^*, P^*)$ of system (2.3), we solve the following system:

$$\begin{cases} r_1(1-\frac{N}{K_1})N - \frac{\delta_1 NP}{1+\delta_2 N} = G_1(N,P) = 0, \\ r_2(1-\frac{P}{K_2})P + \frac{\omega NP}{1+\delta_2 N} = G_2(N,P) = 0. \end{cases} \tag{3.7}$$

Assume that $N^*, P^* > 0$. Dividing $G_1(N^*, P^*)$ by N^*, we obtain

$$P^* = \frac{r_1}{\delta_1}\left(1+\delta_2 N^*\right)\left(1-\frac{N^*}{K_1}\right) \quad \text{if } N^* < K_1.$$

By plugging P^* in $G_1(N^*, P^*)$, we get the cubic equation in N^* as follows:

$$\left(N^*\right)^3 + \theta_2\left(N^*\right)^2 + \theta_1 N^* + \theta_0 = 0, \tag{3.8}$$

where

$$\theta_0 = \frac{K_1(\delta_1 K_2 - r_1)}{r_1\delta_2^2},$$

$$\theta_1 = \frac{r_1 r_2 + r_2\delta_1\delta_2 K_1 K_2 + \omega\delta_1 K_1 K_2 - 2r_1 r_2\delta_2 K_1}{r_1 r_2 \delta_2^2} = \frac{1}{\delta_2}\left(\frac{1}{\delta_2} - 2K_1\right) + \frac{\delta_1 K_1}{r_1\delta_2}K_p,$$

$$\theta_2 = \frac{2r_1 r_2\delta_2 - r_1 r_2\delta_2^2 K_1}{r_1 r_2\delta_2^2} = \frac{2}{\delta_2} - K_1.$$

The following result gives the existence of a coexistence equilibrium point [8, 19, 22, 26, 27].

Theorem 1

Set $\Delta' = (K_1 + \frac{1}{\delta_2})^2 - \frac{3\delta_1 K_1}{r_1\delta_2}K_p$, $\kappa_0 = \frac{r_2\Delta_1}{\delta_1\delta_2\Delta_2}$, *and* $\kappa_1 = \frac{\delta_2\Delta_1'}{\Delta_2'}$.

1. *System* (2.3) *has no feasible coexistence equilibria if either*
   (i) $r_1 < \delta_1 K_2$ *and* $\Delta' \le 0$;
   (ii) $r_1 < \delta_1 K_2$, $\frac{r_1}{\delta_1 K_1}(2K_1 - \frac{1}{\delta_2}) > K_p$, $\delta_2 K_1 < 2$, $\Delta' > 0$, *and* $\kappa_1 > 1$;
   (iii) $r_1 < \delta_1 K_2$, $\frac{r_1}{\delta_1 K_1}(2K_1 - \frac{1}{\delta_2}) > K_p$, $\delta_2 K_1 < 2$, $\Delta' > 0$, *and* $\kappa_0 > 1$.
2. *System* (2.3) *has a unique feasible coexistence equilibrium* E3E3 *if either*
   (i) $r_1 > \delta_1 K_2$ *and* $\Delta' \le 0$;
   (ii) $r_1 > \delta_1 K_2$, $\frac{r_1}{\delta_1 K_1}(2K_1 - \frac{1}{\delta_2}) > K_p$, $\delta_2 K_1 < 2$, $\Delta' > 0$, *and* $\kappa_1 < 1$;
   (iii) $r_1 > \delta_1 K_2$, $\frac{r_1}{\delta_1 K_1}(2K_1 - \frac{1}{\delta_2}) > K_p$, $\delta_2 K_1 > 2$, $\Delta' > 0$, $\kappa_0 < 1$, *and* $\kappa_1 < 1$;

(iv) $r_1 > \delta_1 K_2, \frac{r_1}{\delta_1 K_1}(2K_1 - \frac{1}{\delta_2}) > K_p, \delta_2 K_1 > 2, \Delta' > 0, \kappa_0 > 1,$ *and* $\kappa_1 > 1$.

3. *System* (2.3) *has two distinct feasible coexistence equilibria if either*
   (i) $r_1 < \delta_1 K_2, \frac{r_1}{\delta_1 K_1}(2K_1 - \frac{1}{\delta_2}) > K_p, \delta_2 K_1 < 2, \Delta' > 0,$ *and* $\kappa_1 < 1$, *with* $E_3^- \ll E_3^+$;
   (ii) $r_1 < \delta_1 K_2, \frac{r_1}{\delta_1 K_1}(2K_1 - \frac{1}{\delta_2}) > K_p, \delta_2 K_1 > 2, \Delta' > 0, \kappa_0 > 1,$ *and* $\kappa_1 < 1$, *with* $E_4^- \ll E_4^+$.
4. *System* (2.3) *has three distinct feasible coexistence equilibria if*

$$r_1 > \delta_1 K_2, \qquad \frac{r_1}{\delta_1 K_1}\left(2K_1 - \frac{1}{\delta_2}\right) > K_p, \qquad \delta_2 K_1 > 2, \qquad \Delta' > 0,$$

$$\kappa_0 > 1, \quad and \quad \kappa_1 < 1.$$

Proof

Indeed, consider the following cubic equation:

$$L(X) = X^3 + \theta_2 X^2 + \theta_1 X + \theta_0 = 0. \qquad (3.9)$$

We have

$$L(0) = \theta_0$$

and

$$L(K_1) = \frac{\delta_1 K_1 K_2(\omega K_1 + r_2(1 + K_1))}{r_1 r_2 \delta_2^2} > 0,$$

$$L(X_1) = \frac{r_2 \Delta_1 - \delta_1 \delta_2 \Delta_2}{r_1 r_2 \delta_1 \delta_2^3}$$

and

$$L(X_2) = \frac{\delta_2 \Delta_1' - \Delta_2'}{r_1 r_2 \delta_1 \delta_2^3},$$

where

$$\begin{aligned}\Delta_1 &= 22r_1\delta_1\delta_2 + 27r_1K_1\delta_1\delta_2K_1K_2 + 27\delta_1^2\delta_2K_1K_2\\ &\quad + 9\delta_1\delta_2^3K_1^2K_2 + 2r_1\delta_1\delta_2(1+\delta_2K_1)^2\sqrt{\Delta'},\\ \Delta_2 &= 5r_1r_2\delta_2^3K_1^3 + 87r_1r_2K_1 + 18r_1r_2\delta_2K_1 + 6K_1K_2(r_2\delta_2+\omega)\big(3+\delta_1\sqrt{\Delta'}\big),\\ \Delta_1' &= 2r_1r_2\delta_1K_1 + 9\delta_1r_2\delta_2^2K_1^2K_2 + 12r_1r_2\delta_1\delta_2K_1 + 6\delta_1K_1K_2(\omega+r_2\delta_2)\sqrt{\Delta'},\\ \Delta_2' &= 2r_1r_2\delta_1\delta_2(1+\delta_2K_1)\sqrt{\Delta'} + 2r_1r_2\delta_1\delta_2^3K_1 + 18r_1r_2\delta_1\delta_2^2K_1^2 + 18\delta_1\delta_2K_1K_2(\omega+r_2\delta_2)\end{aligned}$$

with Δ' defined from

$$L'(X) = 3X^2 + 2\theta_2X + \theta_1 = 0 \tag{3.10}$$

as

$$\begin{aligned}\Delta' &= \theta_2^2 - 3\theta_1\\ &= \left(K_1 + \frac{1}{\delta_2}\right)^2 - \frac{3\delta_1K_1K_2}{r_1\delta_2}\left(1 + \frac{\omega}{r_2\delta_2}\right)\\ &= \left(K_1 + \frac{1}{\delta_2}\right)^2 - \frac{3\delta_1K_1}{r_1\delta_2}K_p.\end{aligned}$$

There are two cases:

- If $\frac{r_1\delta_2}{3\delta_1K_1}(K_1 + \frac{1}{\delta_2})^2 \leq K_p,$ then $\Delta' \leq 0$. Thus, L is increasing on $]0, K_1[$.
  - According to $r_1 < \delta_1K_2$, we have $\theta_0 > 0$ with $L(0) \times L(K_1) > 0$. Hence, $L(X) > 0$ $\forall N \in]0, K_1[$. Thus, equation (3.9) has no real roots on $]0, K_1$ [and there are no feasible coexistence equilibria for system (2.3).
  - If $r_1 > \delta_1K_2$, we have $\theta_0 < 0$ with $L(0) \times L(K_1) < 0$. Thus, equation (3.9) has a unique positive root. Then system (2.3) has a unique feasible coexistence equilibrium.
- If $\frac{r_1\delta_2}{3\delta_1K_1}(K_1 + \frac{1}{\delta_2})^2 > K_p,$ then $\Delta' > 0$. Therefore, equation (3.10) has two roots

$$X_1 = \frac{-\theta_2 - \sqrt{\theta_2^2 - 3\theta_1}}{3} \quad \text{and} \quad X_2 = \frac{-\theta_2 + \sqrt{\theta_2^2 - 3\theta_1}}{3},$$

where

$$\begin{cases} X_1 + X_2 = -\frac{2\theta_2}{3}, \\ X_1 X_2 = \frac{\theta_1}{3}. \end{cases}$$

- If $\delta_2 K_1 < 2$ and $\frac{r_1}{\delta_1 K_1}(2K_1 - \frac{1}{\delta_2}) > K_p,$ then we have respectively $\theta_2 > 0$ and $\theta_1 < 0$. Thus, we obtain also $X_1 < 0 < X_2$.

Using the fact that $\Delta' < (K_1 + \frac{1}{\delta_2})^2,$ we have

$$\begin{aligned} X_2 - K_1 &= \frac{-\theta_2 + \sqrt{\Delta'}}{3} - K_1 \\ &\leq -\frac{1}{3}\left(K_1 + \frac{1}{\delta_2}\right). \end{aligned} \tag{3.11}$$

So $X_2 < K_1$. Hence L is decreasing on $]0, X_2[$ and increasing on $[X_2, K_1[$.

- If $r_1 < \delta_1 K_2$ and $\kappa_1 > 1$, then we have respectively $\theta_0 > 0$ and $L(X_2) > 0$, then $L(X) = 0$ has no root. Therefore, system (2.3) has no coexistence equilibria.
- If $r_1 > \delta_1 K_2$ and $\kappa_1 < 1$, then we get respectively $\theta_0 < 0$ and $L(X_2) \leq< 0$. Thus, equation (3.9) has one positive root with $\beta_2 = L(X_2)$ is a minimum. Therefore, system (2.3) has a unique coexistence equilibrium.
- If $r_1 < \delta_1 K_2$ and $\kappa_1 < 1$, then we have respectively $\theta_0 > 0$ and $L(X_2) < 0$ with $L(K_1) > 0 > L(X_2)$. Thus, equation $L(X) = 0$ has two distinct real positive roots, one is N_1^- in $]0, X_2[$, and the other N_1^+ in $]X_2, K_1[$. Each root corresponds to a distinct feasible coexistence equilibrium. Therefore, system (2.3) has two coexistence equilibria $E_3^-(N_1^-, P^*)$ and $E_3^+(N_1^+, P^*)$ with $N_1^- < N_1^+$.

(a) If $\delta_2 K_1 > 2$ and $\frac{r_1}{\delta_1 K_1}(2K_1 - \frac{1}{\delta_2}) > K_p,$ then we have respectively $\theta_2 < 0$ and $\theta_1 < 0$. Thus, we obtain $X_1 < 0 < X_2$ with $X_2 < K_1$. By a similar argument as previously, we obtain the same result.

(b) If $\delta_2 K_1 > 2$ and $\frac{r_1}{\delta_1 K_1}(2K_1 - \frac{1}{\delta_2}) < K_p,$ then we get respectively $\theta_2 < 0$ and $\theta_1 > 0$. Thus, we obtain $0 < X_1 < X_2 < K_1$. Consequently, L is increasing on $]0, X_1]$ and $[X_2, K_1[$ and decreasing on $]X_1, X_2[$.

- According to $r_1 < \delta_1 K_2$, we have $\theta_0 > 0$. If $\kappa_0 > 1$ and $\kappa_1 > 1$, then we have respectively $L(X_1) > 0$ and $L(X_2) > 0$. So equation (3.9) has no real positive roots, and there are no feasible coexistence equilibria.
- According to $r_1 < \delta_1 K_2$, and if $\kappa_0 > 1$ and $\kappa_1 < 1$, then we have respectively $L(X_1) > 0$ and $L(X_2) < 0$. Thus, $L(X) = 0$ has two distinct real positive roots, one is N_2^- in $]X_1, X_2[$, and the other N_2^+ in $]X_2, K_1[$. Therefore, system (2.3) has two coexistence equilibria $E_4^-(N_2^-, P^*)$ and $E_4^+(N_2^+, P^*)$ with $N_2^- < N_2^+$.
- According to $r_1 > \delta_1 K_2$, and if $\kappa_0 < 1$ and $\kappa_1 < 1$, we have respectively $L(X_1) < 0$ and $L(X_2) < 0$. Consequently, $L(X) = 0$ has one root.
- According to $r_1 > \delta_1 K_2$, and if $\kappa_0 > 1$ and $\kappa_1 < 1$, we have respectively $L(X_1) > 0$ and $L(X_2) < 0$. Thus, $L(X) = 0$ has three roots. Therefore, system (2.3) has three coexistence equilibria.
- According to $r_1 > \delta_1 K_2$, and if $\kappa_0 < 1$ and $\kappa_1 > 1$, we have respectively $L(X_1) < 0$ and $L(X_2) > 0$. Thus, $L(X) = 0$ has one root.

The local stability analysis of coexistence equilibrium is given by the following theorem [8, 11, 12].

Theorem 2

If condition (2) *of Theorem* 1*is satisfied, and moreover the following condition holds*:

$$\frac{2r_1(1+\delta_2 N^*)(K_1 - N^*)}{\delta_1 K_1 K_2} > 1, \tag{3.12}$$

then the coexistence equilibrium $E_3 = (N^*, P^*)$ *is locally asymptotically stable.*

Proof

Indeed, the Jacobian matrix of system (2.3) evaluated at the point E_3 is given by

$$\mathrm{DG}(E_3) = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

where

$$A_{11} = \frac{r_1(K_1 - 2N^*)(1 + \delta_2 N^*) - r_1(K_1 - N^*)}{K_1(1 + \delta_2 N^*)},$$

$$A_{12} = \frac{-\delta_1 N^*}{(1 + \delta_2 N^*)},$$

$$A_{21} = \frac{\omega r_1(K_1 - N^*)}{\delta_1 K_1(1 + \delta_2 N^*)},$$

$$A_{22} = \frac{r_2(1 + \delta_2 N^*)(\delta_1 K_1 K_2 - 2r_1(1 + \delta_2 N^*)(K_1 - N^*)) + \omega N^* \delta_1 K_1 K_2}{\delta_1 K_1 K_2(1 + \delta_2 N^*)}.$$

The characteristic polynomial is therefore $P(\lambda) = \lambda^2 - B_1\lambda + B_2 = 0,$ with

$$B_1 = \mathrm{tr}\big(\mathrm{DG}(E_3)\big) = \frac{\delta_1[M_3 K_2(r_1(M_2 - N^*) + r_2 K_1) - r_1 r_2 K_2] - 2r_1 r_2 M_2 M_4}{AM_3},$$

$$B_2 = \det\big(\mathrm{DG}(E_3)\big)$$
$$= \frac{r_1 r_2 \delta_1 K_2(\delta_1 K_1 K_2 M_3 - 2r_1 M_4) \times (\delta_2 N^*(N^* - K_1) - N^*)}{\mathcal{A} K_1 M_3^2} + \frac{\omega r_1 N^* M_2}{K_1 M_3^2},$$

where

$$M_1 = K_1 + (\delta_2 K_1 - 2)N^*, \qquad M_2 = K_1 - N^*, \qquad M_3 = 1 + \delta_2 N^*,$$
$$M_4 = -\delta_2^2 (N^*)^3 + \delta_2(\delta_2 K_1 - 2)(N^*)^2 + (2\delta_2 K_1 - 1)N^* + K_1 = (1 + \delta_2 N^*)^2 (K_1 - N^*),$$

and $\mathcal{A} = \delta_1 K_1 K_2$.

By a simple calculation, we get $B_1 = A_{11} + A_{22}$ and $B_2 = A_{11}A_{22} - A_{12}A_{21}$. According to (3.12), we get $B_1 < 0$ and $B_2 > 0$. By applying the Routh–Hurwitz criterion, E_3 is locally asymptotically stable.

The following theorem gives the global stability [8, 9, 13, 19, 22, 28, 29].

Theorem 3

If condition (2) *of Theorem* 1*is satisfied*, *then the coexistence equilibrium* $E_3 = (N^*, P^*)$ *is globally asymptotically stable in the following subset of* $\mathbb{R}^2_+$:

$$B = \left\{(N,P) \in \mathbb{R}_+^2 / N \geq N^*, P \geq P^*, N \geq \frac{1}{\delta_2}\left(\frac{K_1}{r_1(1+\delta_2 N^*)} - 1\right)\right\}.$$

Proof

Indeed, we construct a Lyapunov candidate function defined by

$$V(N,P) = h_1(N) + h_2(P),$$

with $h_1(N) = a_1 \int_{N^*}^{N} \frac{\zeta - N^*}{\zeta} d\zeta, h_2(P) = a_2 \int_{P^*}^{P} \frac{\zeta - P^*}{\varepsilon} d\varepsilon,$ and $(a_1, a_2) \in \mathbb{R}_+^{*2}$ to be determined. It is easy to see that $V(N^*, P^*) = 0,$ and for all $(N,P) \neq (N^*,P^*),\ V(N,P) > 0$. So V is well defined.

The time derivative of $V(N, P)$ along the solutions of system (2.3) is

$$\begin{aligned}\dot{V}(N,P) &= a_1(N-N^*)\left[r_1\left(1-\frac{N}{K_1}\right) - \frac{\delta_1 P}{1+\delta_2 N}\right] \\ &\quad + a_2(P-P^*)\left[r_2\left(1-\frac{P}{K_2}\right) + \frac{\omega N}{1+\delta_2 N}\right].\end{aligned}$$

After simplification, we can write

$$\begin{aligned}\dot{V}(N,P) &= \frac{-a_2 r_2}{K_2}(P-P^*)^2 + (a_2\omega - \delta_1 a_1 - a_1\delta_1\delta_2 N_*)\frac{(N-N^*)(P-P^*)}{(1+\delta_2 N)(1+\delta_2 N^*)} \\ &\quad + (N-N^*)^2\left[-\frac{a_1 r_1}{K_1} + \frac{a_1\delta_1\delta_2 P^*}{(1+\delta_2 N^*)(1+\delta_2 N)}\right].\end{aligned} \tag{3.13}$$

Taking $a_1 = 1$ and $a_2 = \frac{\delta_1 M_3}{\omega}$, finally we obtain

$$\dot{V}(N,P) = (N-N^*)^2\left[-\frac{r_1}{K_1} + \frac{\delta_1\delta_2 P^*}{(1+\delta_2 N^*)(1+\delta_2 N)}\right] - \frac{r_2\delta_1(1+\delta_2 N^*)}{K_2\omega}(P-P^*)^2$$

for all $(N,P) \neq (N^*,P^*)$. The coefficients $(N - N^*)^2$ and $(P - P^*)^2$ are positive. By using the fact that $N \geq \frac{1}{\delta_2}(\frac{K_1}{r_1(1+\delta_2 N^*)} - 1),$ we have $\dot{V}(N,P) < 0$.

In addition $\dot{V}(N,P) = 0$ if and only if $(N, P) = (N^*, P^*)$. By using LaSalle's invariance principle, $E_3 = (N^*, P^*)$ is globally asymptotically stable on B.

The following proposition gives the necessary and sufficient conditions of stability in case there is more than one equilibrium point [13, 19, 22, 27, 30]. Let us define the quadratic function

$$\pi(N) = -2r_1\delta_2 N^2 + 2r_1(\delta_2 K_1 - 1)N + K_1(2r_1 - \delta_1 K_2). \tag{3.14}$$

For $N^* \in]0, K_1[$,

$$\pi\left(N^*\right) > 0 \quad \Leftrightarrow \quad \frac{2r_1(1+\delta_2 N^*)(K_1 - N^*)}{\delta_1 K_1 K_2} > 1. \tag{3.15}$$

Theorem 4

- *According to condition* 3, (i) *of Theorem* 1, *and if* $\delta_2 K_1 < 1$ *and* $r_1 > \frac{\delta_1 K_2}{2}$, *then the coexistence equilibrium* $E_3^-(N_1^-, P^*)$ *is locally asymptotically stable and* $E_3^+(N_1^+, P^*)$ *is unstable.*
- *According to condition* 3, (ii) *of Theorem* 1, *if* $\delta_1 < \frac{5r_1}{\delta_2}(\delta_2 K_1 - 1)$ *and* $r_1 > \frac{\delta_1 K_2}{2}$, *then the coexistence equilibrium* $E_4^-(N_2^-, P^*)$ *is locally asymptotically stable and* $E_4^+(N_2^+, P^*)$ *is unstable.*

Proof

Indeed,

(i) Consider the function π(*N*) defined by (3.14), we have

$$\pi'(N) = -4r_1\delta_2 N + 2r_1(\delta_2 K_1 - 1), \quad \forall N \in]0, K_1[. \tag{3.16}$$

Also, we have $\pi(0) = K_1(2r_1 - \delta_1 K_2) > 0$ if $r_1 > \frac{\delta_1 K_2}{2}$, $\pi(K_1) = -\delta_1 K_1 K_2$. According to $\delta_2 K_1 < 1$, we obtain $\pi'(N) < 0 \ \forall N \in]0, K_1[$. Consequently, there exists δ such that π(δ) = 0 with $0 < N_1^- < \delta < N_1^+ < K_1$. Thus, $\pi(N_1^-) > 0$ and $\pi(N_1^+) < 0$.

By using $\pi(N_1^-) > 0,$ we obtain

$$\frac{2r_1(1+\delta_2 N_1^-)(K_1 - N_1^-)}{\delta_1 K_1 K_2} > 1. \tag{3.17}$$

Thus, inequality (3.17) verifies the condition of stability given by (3.15). As result, $\mathrm{tr}(E_3^-) < 0$ and $\det(E_3^-) > 0$. Consequently, $E_3^-(N_1^-, P^*)$ is locally asymptotically stable.

By using $\pi(N_1^+) < 0,$ we obtain

$$\frac{2r_1(1+\delta_2 N_1^+)(K_1 - N_1^+)}{\delta_1 K_1 K_2} < 1.$$

(3.18)

Thus, inequality (3.18) does not check the condition of stability given by (3.15). Consequently, $E_3^+(N_1^+, P^*)$ is unstable.

(ii) By using equation (3.16), there exists $N_0 \in]0, K_1[$ such that $\pi'(N_0) = 0$ with $N_0 = \frac{\delta_2 K_1 - 1}{2\delta_2}$. Thus, π is increasing on $]0, N_0[$ and decreasing on $]N_0, K_1[$. By simple computation, we get $\pi(N_0) = 5r_1K_1 - (\frac{r_1\delta_2K_1}{2} + \frac{5r_1}{2\delta_2} + \delta_1K_1K_2)$. By using $\delta_2 K_1 > 2$, we get $\delta_1 < \frac{5r_1}{\delta_2}(\delta_2K_1 - 1)$. According to $r_1 > \frac{\delta_1K_2}{2}$ and $\delta_1 < \frac{5r_1}{\delta_2}(\delta_2K_1 - 1),$ we get respectively $\pi(0) > 0$ and $\pi(N_0) > 0$. Consequently, there exists δ_0 such that $\pi(\delta_0) = 0$ with $0 < N_2^- < \delta_0 < N_2^+ < K_1$. Thus, $\pi(N_2^-) > 0$ and $\pi(N_2^+) < 0$. Consequently, $E_4^-(N_2^-, P^*)$ is locally asymptotically stable and $E_4^+(N_2^+, P^*)$ is unstable.

Bifurcation analysis

In this subsection, we define the conditions of Hopf-bifurcations and the critical values of Hopf bifurcations. Here, δ_1 is taken as a bifurcation parameter [10, 15, 31].

Theorem 5

If condition (2), (ii) *of Theorem* 1*is satisfied and if the following conditions are satisfied*:

$$\delta_1 < \frac{2r_1r_2M_4}{M_3K_2(r_1(M_2 - N^*) + r_2K_1) - r_1r_2K_2},$$

(3.19)

$$B_1^2(\delta_1) \geq 4B_2(\delta_1),$$

(3.20)

then a Hopf-bifurcation occurs at the value $\delta_1 = \delta_{1c}$, *where*

$$\delta_{1c} = \frac{2r_1r_2M_4}{M_3K_2(r_1(M_2 - N^*) + r_2K_1) - r_1r_2K_2}.$$

(3.21)

Proof

Indeed, assuming that $N-N^{*}\simeq e^{xt}, P-P^{*}\simeq e^{xt}$, we get the following characteristic equation corresponding to the Jacobian matrix DG(E_3) evaluated at $E_3=(N^{*},P^{*})$:

$$x^2-B_1(\delta_1)x+B_2(\delta_1)=0, \tag{3.22}$$

where

$$B_1(\delta_1)=\mathrm{tr}\big(\mathrm{DG}(E_3)\big)=\frac{\delta_1[M_3K_2(r_1(M_2-N^{*})+r_2K_1)-r_1r_2K_2]-2r_1r_2M_2M_4}{AM_3},$$

$$B_2(\delta_1)=\det\big(\mathrm{DG}(E_3)\big)$$
$$=\frac{r_1r_2\delta_1K_2(\delta_1K_1K_2M_3-2r_1M_4)\times(\delta_2N^{*}(K_1-N^{*})-N^{*})}{\mathcal{A}K_1M_3^2}+\frac{\omega r_1N^{*}M_2}{K_1M_3^2}.$$

If conditions (3.20) and (3.21) are respectively satisfied, we have respectively $B_1(\delta_1)=0$ and $B_2(\delta_1)>0$, then the eigenvalues will be purely complex at $\delta_1=\delta_{1c}$ with

$$x=\frac{B_1(\delta_1)+\sqrt{B_1^2(\delta_1)-4B_2(\delta_1)}}{2}\quad\text{or}\quad x=\frac{B_1(\delta_1)-\sqrt{B_1^2(\delta_1)-4B_2(\delta_1)}}{2}.$$

Replacing $x=x_1+ix_2$ into (3.22), we have $(x_1^2-x_2^2)-B_1(\delta_1)x_1+B_2(\delta_1)+i(2x_1x_2-B_1(\delta_1)x_2)=0$, and separating real and complex parts, we obtain

$$\begin{cases} x_1^2-x_2^2-B_1(\delta_1)x_1+B_2(\delta_1)=0, & \text{(4.a)}\\ 2x_1x_2-B_1(\delta_1)x_2=0. & \text{(4.b)}\end{cases}$$

Now, we verify the transversality condition.

Considering Re(x) = 0 and differentiating (4.b) with respect δ_1, we get

$$\left(\frac{dx_1}{d\delta_1}\right)_{\delta_1=\delta_{1c}}=\frac{r_1r_2K_1K_2M_3(K_2(1-\delta_{1c})+2M_2M_4)}{2(\delta_{1c}K_1K_2M_3)^2}\neq 0.$$

As result, system (2.3) admits a Hopf-bifurcation at $\delta_1=\delta_{1c}$ corresponding to E_3.

Remark 1

Since $B_1(\delta_{1c}) = 0$ and $(\frac{dB_1}{d\delta_1})_{\delta_1=\delta_{1c}} = 2(\frac{dx_1}{d\delta_1})_{\delta_1=\delta_{1c}} \neq 0, B_1(\delta_1) > 0,$ if condition (3.12) is satisfied and according to the Routh–Hurwitz criterion, E_3 is locally asymptotically stable. In addition, for $\delta_1 = \delta_{1c}$, a Hopf-bifurcation occurs. For $\delta_1 > \delta_{1c}$, E_3 approaches a periodic solution.

NUMERICAL EXPERIMENTS AND BIOLOGICAL EXPLANATIONS

In this section, we present a sequence of numerical simulations in order to support our mathematical results and to analyze the effect of predation on the dynamics of the two species. We use MATLAB technical computer software [8, 12, 32]. The values of the parameters are given in Tables 1 and 2.

Table 1. Parameter values used for the numerical simulation

Parameters	Values	References
r_1	0.1	estimated
r_2	1.8	estimated
K_1	10	estimated
K_2	1800	estimated
δ_1	0.01	estimated
δ_2	0.01	[33]
ω	0.015	[33]

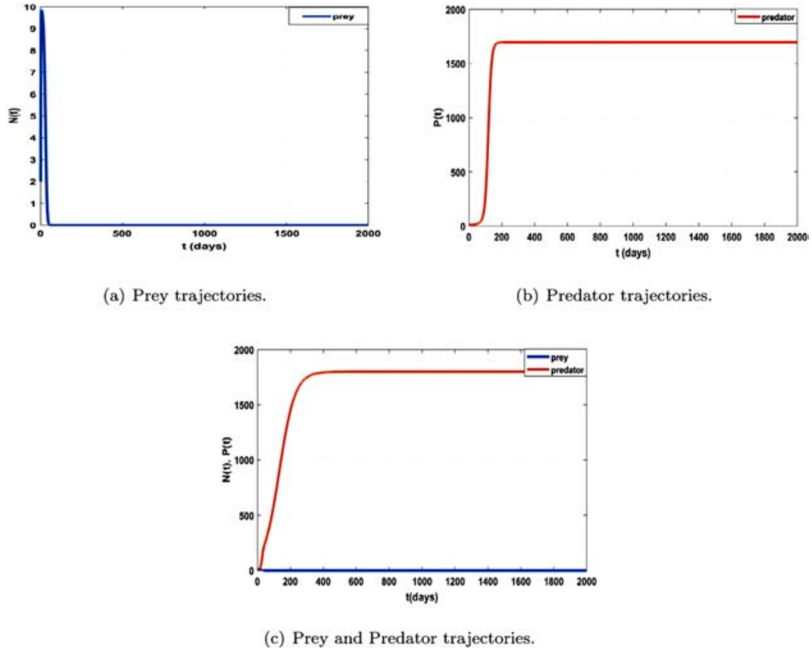
Table 2. Parameter values used for the numerical simulation

Parameters	Values	References
r_1	1.8	estimated
r_2	0.01	estimated
K_1	200	estimated

K_2	30	estimated
δ_1	0.01	[33]
δ_2	0.029	estimated
α	0.0015	estimated

Global Behavior of System (2.3)

Here, we are interested in the predation effect on the dynamics of the two species in order to follow its impact over time. Figure 2 shows the behavior of system (2.3) around E_1 and the parameter values used are given in Table 1. We observe the stability of the predator population and the extinction of the prey population for the predation parameter $\delta_1 = 0.01 > 5.510^{-5}$. This result supports (ii) of Proposition 3. This result confirms that the predator population can survive even if the prey dies out.



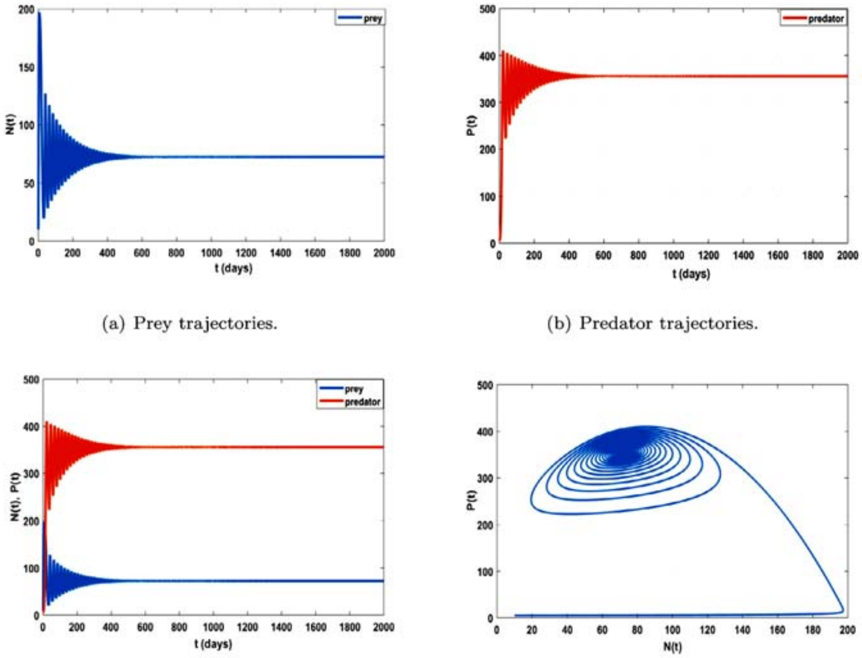
(a) Prey trajectories.

(b) Predator trajectories.

(c) Prey and Predator trajectories.

Figure 2. Evolution of system (2.3) around $E_1 = (0, 1800)$.

Now, we examine the behavior of system (2.3) around the coexistence equilibrium. We take $N > 96$, and the parameter values used are given in Table 2. We observe that system (2.3) converges globally towards the coexisting equilibrium E_3 = (70, 354.1) (see Figure 3(a)–(b)–(c)). The existence of center (Figure 3(d)) confirms the existence and the global asymptotic stability of the coexisting equilibrium. It means that the prey population exists despite the predation. Thus, we talk about the phenomenon of subsistence. That illustrates the result of our Theorem 3.



(a) Prey trajectories.

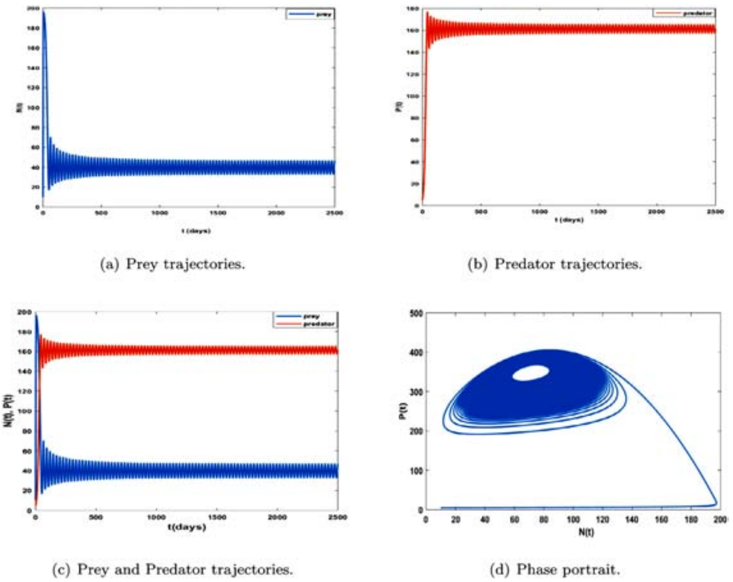
(b) Predator trajectories.

(c) Prey and Predator trajectories.

(d) phase portrait.

Figure 3. Global asymptotic stability of the coexisting equilibrium of system (2.3) around E_3 = (70, 354.51).

If we increase the value of δ_1 = 0.033 and keep the other parameters fixed in Table 2, from Figure 4, we observe that the equilibrium E_3 loses its stability. This result confirms Theorem 2. In next subsection Figure 5 shows the Hopf-bifurcation of system (2.3) around E_3 at $\delta_1 = \delta_{1c}$.



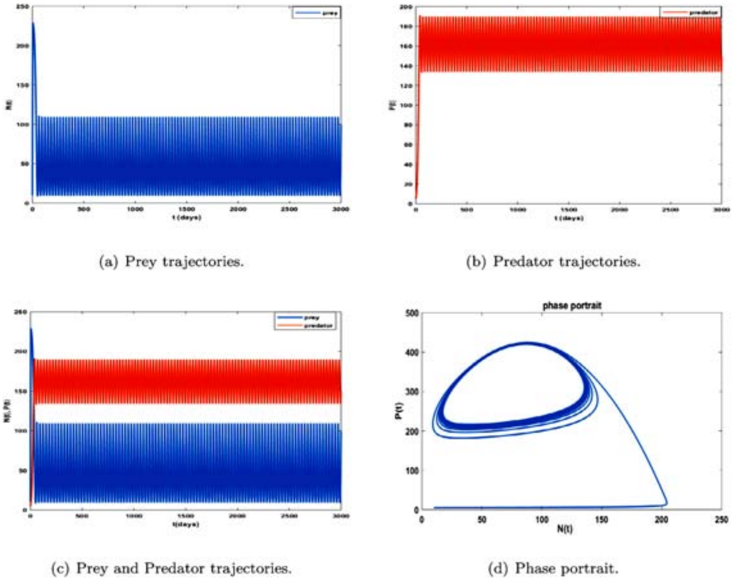
(a) Prey trajectories.

(b) Predator trajectories.

(c) Prey and Predator trajectories.

(d) Phase portrait.

Figure 4. Local asymptotic stability of the coexisting equilibrium of system (2.3) corresponding to $\delta_1 = 0.033$.



(a) Prey trajectories.

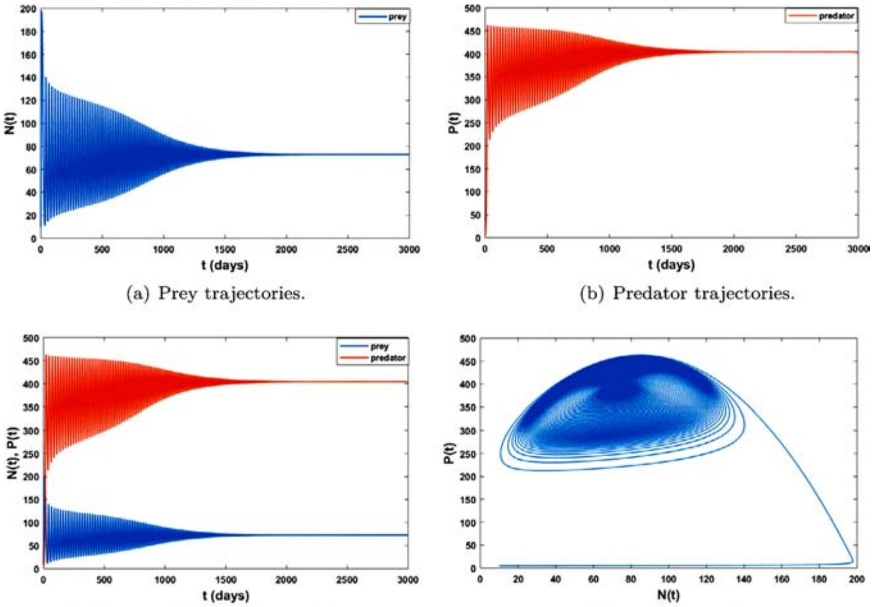
(b) Predator trajectories.

(c) Prey and Predator trajectories.

(d) Phase portrait.

Figure 5. Dynamics of the trajectories showing the existence of limit cycle arising from the Hopf-bifurcation of system (2.3) around $E_3 = (N^*, P^*)$ with $\delta_1 = \delta_{1c} = 0.0636$.

We continue our numerical simulations when the system admits two coexistence equilibria in order to look the behavior of the system around E_3^- and E_+^3. By increasing the value δ_1 to $\delta_1 = 0.023$, we observe that system (2.3) converges globally towards the coexisting equilibrium $E_3^- = (75, 400)$ (see Figure 6(a)–(b)–(c)). By increasing the parameter of predation δ_1 to $\delta_1 = 0.044$, we observe the loss of stability of the coexistence equilibrium E_3^- (see Figure 7). This is in accordance with the mathematical results established in Theorem 4.



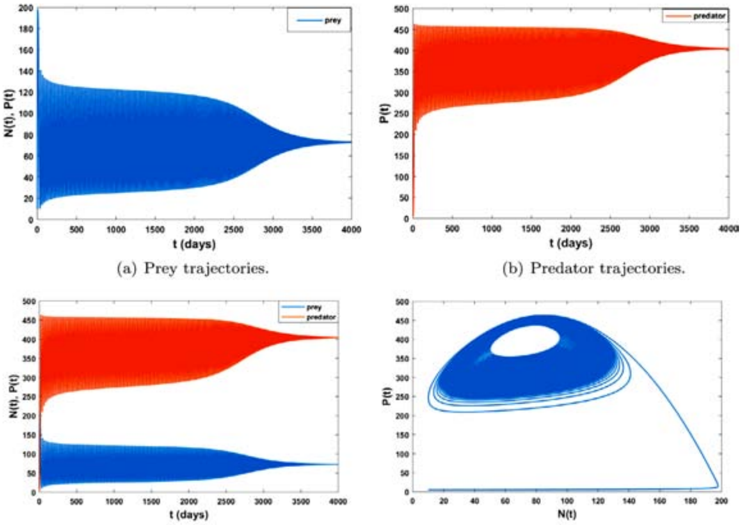
(a) Prey trajectories.

(b) Predator trajectories.

(c) Prey and Predator trajectories.

(d) phase portrait.

Figure 6. Global asymptotic stability of the coexisting equilibrium of system (2.3) around $E_3^- = (75, 400)$.

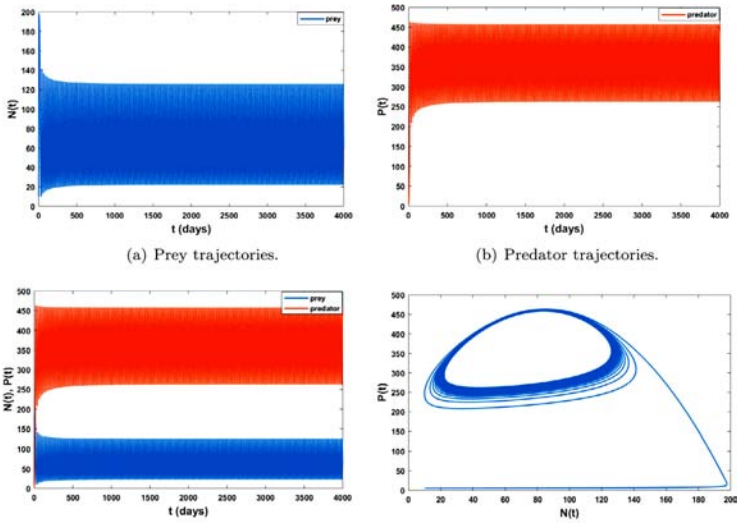


(a) Prey trajectories. (b) Predator trajectories.

(c) Prey and Predator trajectories. (d) phase portrait.

Figure 7. Local asymptotic stability of the coexisting equilibrium E_3^- of system (2.3) corresponding to $\delta_1 = 0.044$.

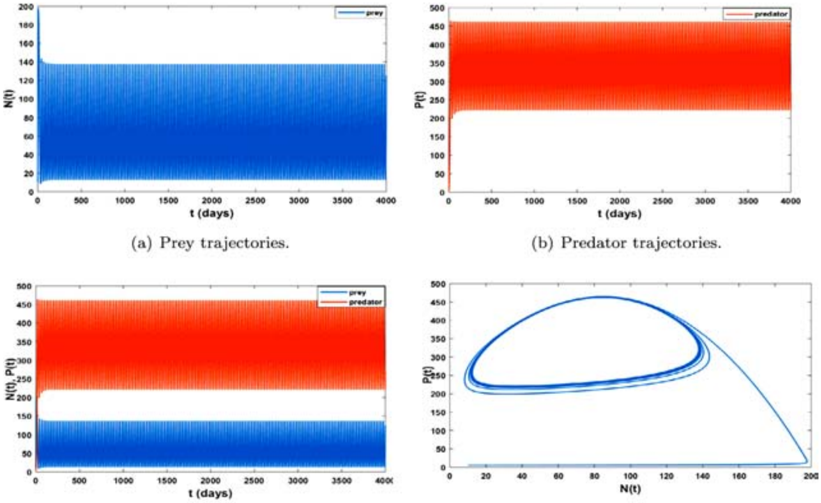
At the same time, we observe the instability of the coexistence equilibrium E_3^+ showing the existence of a limit cycle illustrated by Figures 8 and 9.



(a) Prey trajectories. (b) Predator trajectories.

(c) Prey and Predator trajectories. (d) phase portrait.

Figure 8. Limit cycle behavior of the solution of system (2.3) at the coexisting equilibrium E_3^+ corresponding to $\delta_1 = 0.065$.



(a) Prey trajectories.

(b) Predator trajectories.

(c) Prey and Predator trajectories.

(d) phase portrait.

Figure 9. Limit cycle behavior of the solution of system (2.3) at the coexisting equilibrium E_3^+ corresponding to $\delta_1 = 0.067$.

Analysis of Hopf-Bifurcation Diagram

We continue our numerical analysis in this subsection to observe the dynamics behavior of the system by considering the predation parameter. Now, if we consider the critical value $\delta_{1c} = 0.0636$, Figure 5((c)–(d)) shows that the coexisting equilibrium $E_3 = (N^*, P^*)$ is unstable, and we have a limit cycle arising from the Hopf-bifurcation. Theorem 5 then holds.

Remark 2

The biological interpretation of the Hopf-bifurcation is that the prey coexists with the predator, exhibiting oscillatory equilibrium behavior [10, 11]. Indeed, we observe that if the predation threshold $\delta_1 > \delta_{1c}$, we have periodic fluctuation of the prey and predator species: Figures 5(c) and 5(d) show the existence of a limit cycle resulting from the Hopf-bifurcation. This highlights an extinction of the population of prey (at risk) if predation exceeds a certain threshold.

CONCLUSION

The effect of predation in the dynamics of the prey-predator model plays an essential role in the equilibrium of the ecosystem, because it allows natural mechanisms of regulation of species. It is for this reason that in this paper we proposed and analyzed a nonlinear mathematical model to describe the dynamics of the populations of prey and predators, taking into account the effect of predation. The formulation of the model derived from an ODE system by considering Holling function response of type II to represent the interaction between the prey and the predator. The mathematical results allowed us first to establish the positivity of the solutions indicating the existence of the population, as well as the bornitude to explain the natural control of the growth due to the restriction of the resources. In addition, we established the conditions of existence of the coexistence equilibria. Under certain conditions of the predation rate, we were able to establish the local stability of the coexistence equilibrium. In order to show the long-term coexistence of prey and predator species, we established the global stability of the coexistence equilibrium via an appropriate Lyapunov function under certain conditions of the model parameters. Moreover, we have described the conditions of existence of the Hopf-bifurcation in order to analyze to what extent the trajectories will be influenced by changes in the predation rate.

Our numerical results gave interesting findings on the effect of predation on the dynamics of the prey-predator model and also allowed to validate our results established in the mathematical study. We have shown the dynamic behavior of our model under different values of the predation rate. Indeed, considering Fig. 2, under certain values of the predation rate, we note an extinction of the prey species and persistence of predators towards the carrying capacity. Staying in this same logic of variation of the predation rate and by considering the parameters fixed in Table 2, we obtain the global stability of coexistence equilibrium indicated by Figure 2(d); this also attests the results of Theorem 3. By increasing the value of δ_1, we lost the stability indicated in Figure 4(d); this phenomenon confirms our mathematical results established in Theorem 2. If we exceed the critical threshold of predation δ_{1c} found in Theorem 5, then we observe a periodic variation in the numbers of prey and predators indicated by Figures 5(a), (b), (c) and the existence of a limit cycle arising from the Hopf-bifurcation. In the light of these observations, we are led to conclude that the predation rate is a key parameter which governs our model and is useful for understanding the dynamics of species of prey and predators in the natural environment, and plays a regulator role of species.

Despite the important findings on this dynamic, in order to deepen our study, we plan to extend this work, taking into account the presence of infectious diseases in both species in order to look at the impact of this disease on the dynamics of the two species.

ACKNOWLEDGEMENTS

The authors would like to express their sincere thanks to the associate editor and the anonymous referees for their constructive comments and suggestions given in their different reports, which have helped us to significantly improve the quality of the paper.

FUNDING

This work has been prepared with the support of the Department of Mathematics, Nazi BONI University.

AUTHORS' CONTRIBUTIONS

All authors worked together to produce the results and read and approved the final manuscript.

REFERENCES

1. Akçakaya, H.R.: Population cycles of mammals: evidence for a ratio-dependent predation hypothesis. Ecol. Monogr. **62**(1), 119–142 (1989)
2. Bairagi, N., Roy, P., Chattopadhyay, J.: Role of infection on the stability of a predator-prey system with several response functions – a comparative study. J. Theor. Biol. **248**(1), 10–25 (2007)
3. Gakkhar, S., Singh, B., Naji, R.K.: Dynamical behavior of two predators competing over a single prey. Biosystems **90**(3), 808–817 (2007)
4. Samanta, G.P.: Analysis of a delay nonautonomous predator prey system with disease in the prey. Nonlinear Anal., Model. Control **15**(8), 97–108 (2010)
5. Holling, C.: The components of predation as revealed by a study of small-mammal predation of the European pine sawfly. Can. Entomol. **91**(5), 293–320 (1959)
6. Arditi, R., Ginzburg, L.R., Akçakaya, H.R.: Variation in plankton densities among lakes: a case for ratio-dependent predation models. Am. Nat. **138**(5), 1287–1289 (1991)
7. Arditi, R., Ginzburg, L.R.: Coupling in predator-prey dynamics: ratio-dependence. J. Theor. Biol. **139**, 311–326 (1989)
8. Koutou, O., Traoré, B., Sangaré, B.: Mathematical model of malaria transmission dynamics with distributed delay and a wide class of nonlinear incidence rates. Cogent Math. Stat. **5**(1), 1–25 (2018)
9. Koutou, O., Traoré, B., Sangaré, B.: Mathematical modeling of malaria transmission global dynamics: taking into account the immature stages of the vectors. Adv. Differ. Equ. **220**, 1–34 (2018)
10. Ouedraogo, H., Ouedraogo, W., Sangaré, B.: Mathematical analysis of toxin-phytoplankton-fish model with self-diffusion and cross-diffusion. Biomathematics **8**, 1911237 (2019)
11. Ouedraogo, H., Ouedraogo, W., Sangaré, B.: A mathematical model with a trophic chain predation based on the ODEs to describe fish and plankton dynamics. An. Univ. Craiova, Math. Comput. Sci. **46**(1), 164–177 (2019)
12. Traoré, B., Koutou, O., Sangaré, B.: A global mathematical model of malaria transmission dynamics with structured mosquito population and temperature variations. Nonlinear Anal., Real World Appl. **53**, 1–32 (2020)

13. Traoré, B., Koutou, O., Sangaré, B.: Global dynamics of a seasonal mathematical model of schistosomiasis transmission with general incidence function. J. Biol. Syst. **27**(1), 1–31 (2019)
14. Haque, M.: A predator-prey model with disease in the predator species only. Nonlinear Anal., Real World Appl. **11**(4), 2224–2236 (2010)
15. Guin, L.N.: Existence of spatial patterns in a predator-prey model with self- and cross-diffusion. J. Comput. Appl. Math. **226**, 320–335 (2014)
16. Das, K.P., Kundu, K., Chattopadhyay, J.: A predator-prey mathematical model with both populations affected by diseases. Ecol. Complex. **8**(1), 68–80 (2011)
17. Xiao, S.D.: Global analysis in predator-prey system with non monotonic functional response. SIAM J. Appl. Math. **61**(4), 1445–1472 (2001)
18. Haque, M.: A detailed study of the Beddington–DeAngelis predator-prey model. Math. Biosci. **234**(1), 1–16 (2011)
19. Savadogo, A., Ouedraogo, H., Sangaré, B., Ouedraogo, W.: Mathematical analysis of a fish-plankton eco-epidemiological system. Nonlinear Stud. **27**(1), 1–22 (2020)
20. Ajraldi, V., Pittavino, M., Venturino, E.: Modeling herd behavior in population systems. Nonlinear Anal., Real World Appl. **12**(4), 2319–2338 (2011)
21. Haque, M., Greenhalgh, D.: When predator avoids infected prey: a model based theoretical studies. Math. Med. Biol. **27**(1), 75–94 (2009)
22. Traoré, B., Sangaré, B., Traoré, S.: A mathematical model of malaria transmission in a periodic environment. J. Biol. Dyn. **12**(1), 400–432 (2018)
23. Ruan, S., Xiao, D.: Global analysis in a predator–prey system with nonmonotonic functional response. SIAM J. Appl. Math. **61**(4), 1445–1472 (2001)
24. Tewa, J.J., Djeumen, V.Y., Bowong, S.: Predator-prey model with Holling response function of type II and SIS infectious disease. Appl. Math. Model. **37**(7), 4825–4841 (2013)
25. Guckenheimer, J., Homes, P.: Nonlinear Oscillations, Dynamical System and Bifurcations of Vector Fields, pp. 140–141. Springer, Berlin (1983)
26. David, G., Ahmed, K.Q.J., Ahmed, A.F.: Eco-epidemiological model with fatal disease in the prey. Nonlinear Anal., Real World Appl. **53**, 103072 (2020)

27. Harrison, G.W.: Multiple stable equilibria in a predator-prey system. Bull. Math. Biol. **48**(2), 137–148 (1986)

28. Chiu, C.H.: Lyapunov functions for the global stability of competing predators. J. Math. Anal. Appl. **230**(1), 232–241 (1999)

29. Korobeinikov, A.: A Lyapunov function for Leslie-Gower predator-prey models. Appl. Math. Lett. **14**, 697–699 (2001)

30. Martin, S., Hervé, B., Yves, D.: On the use of the sterile insect release technique to reduce or eliminate mosquito populations. Appl. Math. Model. **68**(1), 443–470 (2019)

31. Ouedraogo, H., Ouedraogo, W., Sangaré, B.: Bifurcation and stability analysis in complex cross-diffusion mathematical model of phytoplankton-fish dynamics. J. Partial Differ. Equ. **32**(3), 1–22 (2019)

32. Anguelov, R., Dumont, Y., Lubuma, L.J.M., Shillor, M.: Comparison of some standard and nonstandard numerical methods for the MSEIR epidemiological model. AIP Conf. Proc. **1168**(2), 1209–1212 (2009)

33. Hsieh, Y.-H., Hsiao, C.-K.: Predator–prey model with disease infection in both populations. Math. Med. Biol. **25**, 247–266 (2008)

Chapter

12

MULTISCALE MODELLING TOOL: MATHEMATICAL MODELLING OF COLLECTIVE BEHAVIOUR WITHOUT THE MATHS

James A. R. Marshall, Andreagiovanni Reina, and Thomas Bose

Department of Computer Science, University of Sheffield, Sheffield, United Kingdom

ABSTRACT

Collective behaviour is of fundamental importance in the life sciences, where it appears at levels of biological complexity from single cells to superorganisms, in demography and the social sciences, where it describes the behaviour of populations, and in the physical and engineering sciences, where it describes physical phenomena and can be used to design distributed systems. Reasoning about collective behaviour is inherently difficult, as the non-linear interactions between individuals give rise to complex emergent dynamics. Mathematical techniques have been developed to analyse systematically collective behaviour in such systems, yet these

Citation: (APA): Marshall, J. A., Reina, A., & Bose, T. (2019). Multiscale modelling tool: Mathematical modelling of collective behaviour without the maths. PloS one, 14(9), e0222906. (16 pages).

frequently require extensive formal training and technical ability to apply. Even for those with the requisite training and ability, analysis using these techniques can be laborious, time-consuming and error-prone. Together these difficulties raise a barrier-to-entry for practitioners wishing to analyse models of collective behaviour. However, rigorous modelling of collective behaviour is required to make progress in understanding and applying it. Here we present an accessible tool which aims to automate the process of modelling and analysing collective behaviour, as far as possible. We focus our attention on the general class of systems described by reaction kinetics, involving interactions between components that change state as a result, as these are easily understood and extracted from data by natural, physical and social scientists, and correspond to algorithms for component-level controllers in engineering applications. By providing simple automated access to advanced mathematical techniques from statistical physics, nonlinear dynamical systems analysis, and computational simulation, we hope to advance standards in modelling collective behaviour. At the same time, by providing expert users with access to the results of automated analyses, sophisticated investigations that could take significant effort are substantially facilitated. Our tool can be accessed online without installing software, uses a simple programmatic interface, and provides interactive graphical plots for users to develop understanding of their models.

INTRODUCTION

Collective behaviour models, in which individuals interact and in doing so change state, describe a large variety of physical, biological, and social phenomena. One particularly general formulation is that of *reaction kinetics*, developed to describe the time evolution of chemical reactions, but also able to describe networks in molecular biology (*e.g.* [1]), collective behavioural phenomena such as decision-making in animal groups (*e.g.* [2]), demographic and ecological models such as predator-prey dynamics (*e.g.* [3]), epidemiological models (*e.g.* [3]), and social behaviour in human groups, such as opinion dynamics and economics (*e.g.* [4]). The generality of the reaction kinetics formalism is demonstrated by the fact that many of the aforementioned processes, although apparently quite different, are in fact described by the same dynamical equations; for example, the famous Lotka-Volterra equations were simultaneously developed in the description of a chemical reaction, and predator-prey dynamics [5, 6].

Modelling collective behaviour is essential to develop understanding, yet mathematical and computational modelling are skills than can be found in some disciplines much more than others. To understand commonalities and analogies across disciplines it would be beneficial to ensure a consistent standard of modelling is reached across all. However, it is unreasonable to expect all disciplines to ensure the same standard of mathematical training in their practitioners. Reaction kinetics have the advantage that they describe observations of a system in a very natural way, indeed the very way that experimental scientists tend to record those interactions. Reaction kinetics can also be transformed into mathematical equations according to a variety of procedures. The level of description attainable may vary, however. In their simplest form, mathematical models as Ordinary Differential Equations will assume infinitely large, well-mixed populations; this *mean-field* approach ignores fluctuations in subpopulation sizes due to the stochastic effects that small populations entail, and also ignores spatial heterogeneity and attendant sources of noise. Yet ODEs are analytically most tractable, and so enable general insights to be developed into the behaviour of an idealised version of the system of interest. By introducing finite population effects, noisy fluctuations around the mean-field solution can be studied; these can be approximated analytically, through the application of techniques from statistical mechanics, or numerically through efficient and probabilistically correct simulation of the Master Equation, which gives the continuous-time change in the probability density over the possible states of the system. These approaches are still idealisations, in that they ignore noise due to spatial effects, but they retain some tractability. Finally, one may analyse spatial sources of noise, by embedding a finite population in a spatial environment, such as a network, or a 2-dimensional plane or 3-dimensional volume. While in some cases analytic results may be possible, particularly in the case of networks, in general numerical simulation is required, sometimes referred to as Individual-Based Simulation or Agent-Based Simulation. This approach is therefore the most realistic, while also the least analytically tractable. In understanding the collective behaviour of some real-world system, therefore, the approach is generally to understand the simplest model of the system, then progressively introduce more realistic sources of noise in order to see if that behaviour is changed in important ways.

Taking all of the above points into consideration, we here present a Multiscale Modelling Tool, intended to simplify as much as possible the application of analytic and numerical techniques to descriptions of simple collective behaviour systems. The tool has the following objectives, and in the remainder of the paper we describe how these are achieved:

1. enable non-modellers to describe collective behaviour systems intuitively
2. enable a variety of analyses to be applied easily to such systems, accounting for increasingly realistic sources of noise
    a. infinite-population non-spatial noise-free dynamics
    b. non-spatial finite-population noisy dynamics
    c. spatial finite-population noisy dynamics
3. enable interactive exploration of analysis results
4. enable expert-level access to analysis results
5. minimise overheads for installation and use of the software

DESIGN AND IMPLEMENTATION

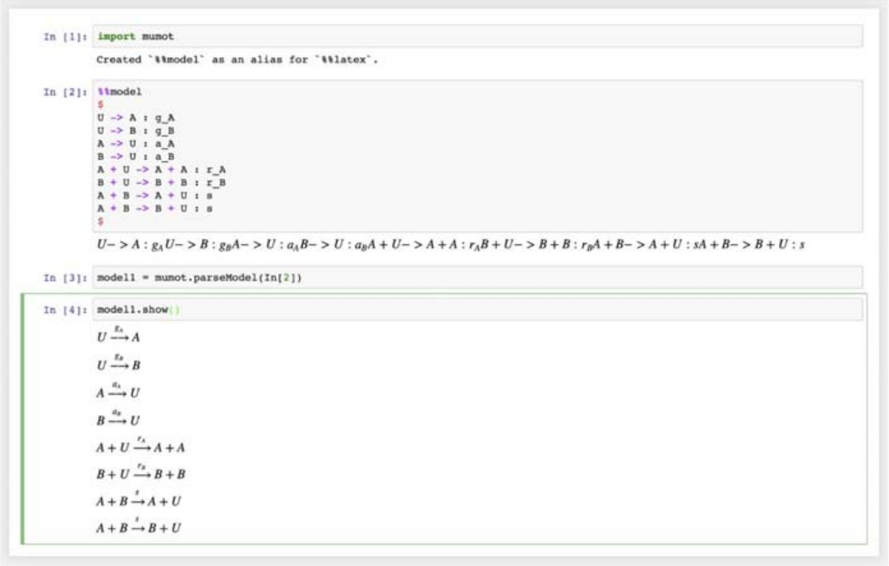
MuMoT (Multiscale Modelling Tool) is written in Python 3 [7] and designed to be run within Jupyter notebooks [8]. This enables MuMoT to be used in interactive notebook sessions using widgets, with explanations written in Markdown and $\LaTeX$ to develop interactive computational documents, particularly suited to communication of results and concepts in research or teaching environments. A Jupyter notebook server can be deployed with a MuMoT installation to allow users to work through a standard web browser, without the need to install client-side software, facilitating access and uptake; at the time of writing, the interactive MuMoT user manual can be executed in this mode via Binder [9] (see [10]). Despite being primarily designed for interactive use, MuMoT uses a variant of the Model, View, Controller design pattern [11] enabling a separation between model descriptions, analytic tools applied to models, and interactive widgets for manipulation of analyses; this enables MuMoT to be used non-interactively, for example with routines called directly from user code.

As MuMoT runs in Jupyter notebooks the user enters simple commands in notebook cells. Models are generated from intuitive textual descriptions, or from mathematical manipulation of previously-defined models, and most commands applicable to models result in interactive graphical output. To enable users to concentrate on presenting the key relevant concepts, users can partially or totally fix parameters in the resulting controllers, and have single controllers connected to multiple model views, with nesting of views if desired [10].

MuMoT's implementation, testing, and documentation seeks to adhere to the best standards for scientific software deployment [12, 13].

Specifying Collective Behaviour Models

Users describe models as simple textual rules, standard in the description of reaction kinetics. We refer to individuals as *reactants* which can be, for example, different classes of individuals as in the case of chemical molecules or members of different biological species, or individuals having different changeable states as in the case of voter models, or robot swarms. Rules describe which reactants interact with each other, the resulting reactants, and the rate at which such reactions occur. For example, Fig 1 shows the description of a model of collective decision-making in honeybee swarms [2, 14] within MuMoT, and how this is parsed into a mathematical object.



```
In [1]: import mumot
        Created `%%model` as an alias for `%%latex`.

In [2]: %%model
        $
        U -> A : g_A
        U -> B : g_B
        A -> U : a_A
        B -> U : a_B
        A + U -> A + A : r_A
        B + U -> B + B : r_B
        A + B -> A + U : s
        A + B -> B + U : s
        $
```

$U->A:g_AU->B:g_BA->U:a_AB->U:a_BA+U->A+A:r_AB+U->B+B:r_BA+B->A+U:sA+B->B+U:s$

```
In [3]: model1 = mumot.parseModel(In[2])

In [4]: model1.show()
```

$U \xrightarrow{g_A} A$

$U \xrightarrow{g_B} B$

$A \xrightarrow{a_A} U$

$B \xrightarrow{a_B} U$

$A + U \xrightarrow{r_A} A + A$

$B + U \xrightarrow{r_B} B + B$

$A + B \xrightarrow{s} A + U$

$A + B \xrightarrow{s} B + U$

Figure 1. Specification of a collective behaviour model. A model is described using simple textual rules denoting interactions between reactants in the system, and rates and which they are transformed into new combinations. Parsing this model description automatically results in a mathematical model object ready for analysis.

Models can also be created from the mathematical manipulation of other models; for example, it can be convenient to note that the frequency of one of the reactants can be determined from the frequencies of the remaining reactants, and the total system size, in any closed system where no reactant can be created or destroyed:

model2 = model1.substitute('U = N - A - B')

and to redefine rates in terms of other quantities, such as the qualities of potential nest sites in this example:

model3 = model2.substitute('a_A = 1/v_A, a_B = 1/v_B, g_A = v_A, g_B = v_B, r_A = v_A, r_B = v_B')

or even in terms of the mean and difference between those qualities [2, 14]:

model4 = model3.substitute('v_A = \mu + \Delta/2, v_B = \mu - \Delta/2')

Once parsed, a model exists as a mathematical object ready for analysis, as can be seen by asking to see the Ordinary Differential Equations (ODEs) that describe its time evolution:

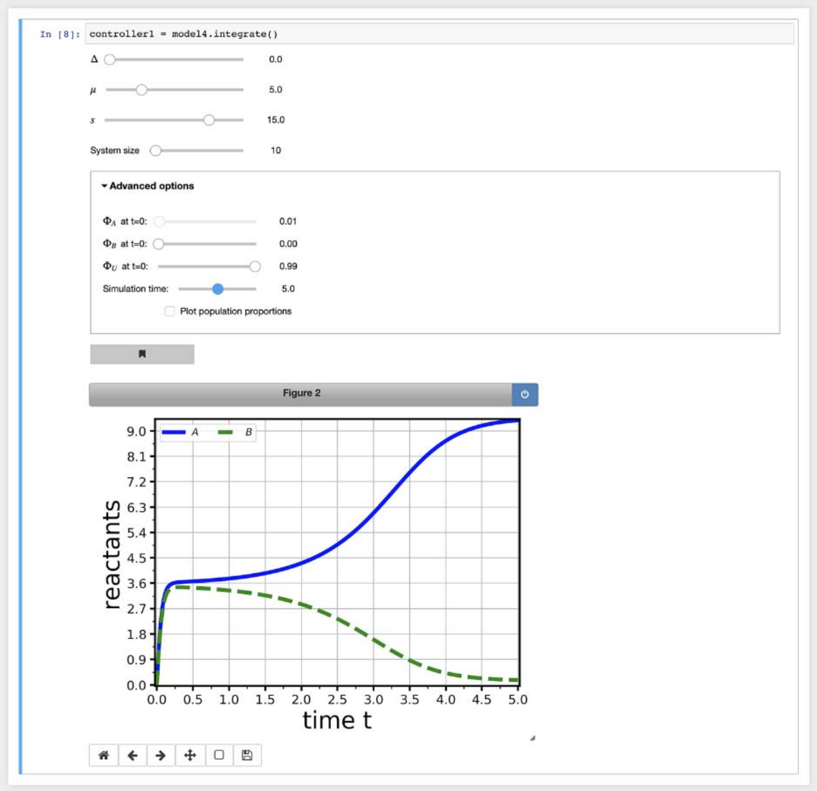
model4.showODEs()

which results in the following system of equations:

$$\begin{aligned}\frac{\mathrm{d}A}{\mathrm{d}t} &:= -ABs + A\left(\frac{\Delta}{2}+\mu\right)(-A-B+N) - \frac{A}{\frac{\Delta}{2}+\mu} + \left(\frac{\Delta}{2}+\mu\right)(-A-B+N),\\ \frac{\mathrm{d}B}{\mathrm{d}t} &:= -ABs + B\left(-\frac{\Delta}{2}+\mu\right)(-A-B+N) - \frac{B}{-\frac{\Delta}{2}+\mu} + \left(-\frac{\Delta}{2}+\mu\right)(-A-B+N).\end{aligned} \tag{1}$$

Eq 1 have been automatically derived from the rule-based description of the model we provided. Two techniques can be used to derive these ODEs, either a *mass action* heuristic similar to the one a mathematician would use to derive the ODEs, or a more involved statistical physics approach described in section 'Analysing noisy behaviour' (*e.g.* [15]). Both, however, have the same result.

Once a model has been parsed, a variety of analytic and numerical techniques can be applied to it. Many of these result in interactive graphical displays of the analysis, which users can manipulate to explore their model. For example, Fig 2 shows the result of performing a numerical integration on the model of Eq 1 within the notebook environment, using the integrate() command. Although not described in this paper, parameters can be fixed as desired to focus on a particular set of free parameters (*partial controllers*), and multiple views on the same model can be manipulated via a single controller (*multicontroller*). Users can also *bookmark* interesting parameter combinations to reproduce subsequently, and save the results from some views for analysis in external software packages. Such devices allow researchers and teachers to focus exploration and exposition of important concepts. Full details are given in the online user manual [10].



```
In [8]: controller1 = model4.integrate()
```

Δ 0.0

μ 5.0

s 15.0

System size 10

▾ Advanced options

Φ_A at t=0: 0.01

Φ_B at t=0: 0.00

Φ_U at t=0: 0.99

Simulation time: 5.0

□ Plot population proportions

Figure 2. Interactive manipulation of a model view via a controller. Most model analysis commands result in an interactive graphical display of that analysis on the model. Users can explore and visualise the effects of changing free model parameters, and other analysis-specific parameters, through manipulating interactive controls.

Analysing Noise-Free Infinite Population Behaviour

The most analytically tractable means of analysing collective behaviour are typically those that assume infinite populations; in this mean-field approach sources of intrinsic noise due to finite population effects are neglected, and space is also ignored. Thus understanding the noise-free dynamics of a collective behaviour system is normally the most fruitful starting point in dealing with any new system.

Numerical Integration of ODEs and Phase Portraits

The simplest way to approach the noise-free dynamics of a system is often to integrate the ODEs that describe it. To achieve this MuMoT provides the integrate() method, which makes use of the odeint interface to numerical integrators implemented in Python's SciPy package scipy.integrate [16]. Solutions are displayed as interactive graphical output (see for example Fig 2). Plots can be presented either in terms of absolute numbers, or of population proportions (i.e. the number of 'particles' for each *reactant* divided by the system size at $t = 0$).

The dynamics of a MuMoT model can also be studied by means of a *phase plane* analysis. To visualise the model's trajectories in a *phase portrait* the methods stream() and vector() can be applied. Both methods depict phase planes representing the time evolution of the system as a function of its state; in a vector plot arrows give the direction in which the system will move, and their lengths show how fast, whereas in a stream plot lines show the average change of the system over time in finer resolution, and their shading represents the speed of change. It is also possible to calculate and display fixed points and noise around these; the corresponding theory and computations are introduced below. Stream plot examples are shown in Fig 4. More detailed explanations can be found in the online user manual [10].

Bifurcations

Nonlinear dynamical systems may change behaviour qualitatively if model parameters are varied. To detect such transitions between different dynamic regimes MuMoT implements basic *bifurcation analysis* functionality by integrating with PyDSTool [17]. MuMoT's method enabling bifurcation analysis is called bifurcation(). Currently available is the detection of *branch points* (BPs) and *limit points* (LPs) of one-dimensional and two-dimensional systems; remember that a three-dimensional system may be reduced to a two-dimensional one using MuMoT's substitute() method. Detectable bifurcation points in MuMoT belong to the class of local codimension-one bifurcations. For example, BPs are observed for *pitchfork* bifurcations such as the one shown in Fig 5 (left panel). *Saddle-node* bifurcations are typical LPs (Fig 5 (middle and right panels)). For two-dimensional systems it may be desirable to directly compare the behaviour of both dynamical variables (or *state variables* as we call them within MuMoT) depending on a critical parameter in the same two-dimensional plot, where the bifurcation parameter is plotted on the horizontal axis. MuMoT allows users to plot single reactants as response variables, but also sums or differences of reactants, as illustrated in Fig 5 (left panel). For more information on the usage of bifurcation() we refer the reader to the online user manual [10].

When executing the bifurcation() method the following computations run behind the scenes. For a given parameter configuration, which includes the choice of the initial value of the bifurcation parameter, MuMoT attempts to determine all stationary states. If this is successful, MuMoT then starts the numerical continuation of each branch on which it found a stable fixed point. In case no stable fixed point could be detected, MuMoT numerically integrates the system using the initial conditions, and uses the final state at the end of the numerical integration as the starting point for the bifurcation analysis. If LPs or BPs were found those will be displayed and labelled in the bifurcation diagram. When MuMoT finds a BP it then tries to automatically start another continuation calculation along the other branch that meets the current branch at the BP. All curves that could be detected are displayed together at the end of the automated bifurcation analysis, colour-coded and shown with different line-styles to reflect the underlying stability properties of the corresponding stationary states. Fig 5 shows examples of different types of bifurcations that can be studied with MuMoT's bifurcation() method.

Analysing Noisy Behaviour

Any real-world system is subject to noise, hence the next step in analysing a collective behaviour system is to examine deviations from the mean-field solutions of the model under such noise. There are two primary sources of noise, that due to finite population size, and that due to spatial distribution of the population; MuMoT enables analysis of both.

Finite-Population Noise

We start with intrinsic noise, due to finite population size. In any finite system the number of interactions fluctuates around an average value and hence so do the numbers of agents in the states available. The following derivation is based on the classical textbook by van Kampen [18]. In analogy to a typical chemical reaction let us consider a system of interacting agents X_k with $k =$ 1, 2…, K being the different states agents might be in. Here X denotes the type of agent and the state represented by index k may be the commitment state. For example this could be a honeybee advertising a potential new nest site. The number of agents in state k is denoted n_k; when agents interact the numbers in any state k may change. Using integer stoichiometric coefficients denoted α_k and β_k the change of the system's state following interactions may be described by

$$\alpha_1 X_1 + \alpha_2 X_2 + \cdots \rightarrow \beta_1 X_1 + \beta_2 X_2 + \cdots \,, \tag{2}$$

where the left-hand side characterises the state before the interaction (reaction) and the right-hand side the state after the interaction (reaction). All interaction processes are affected by the total number of agents. To account for this, we introduce the system size $\overline{V}$ as a formal (auxiliary) parameter that is necessary for the following derivation.

The Master Equation

In order to sufficiently describe our system of interest, we need to compute the averaged macroscopic numbers and we also need to quantify the fluctuations around these averaged quantities. This may be achieved by means of the chemical Master equation, which can be written as follows [18]:

$$\begin{aligned}\frac{\partial P(\{n_k\};\,t)}{\partial t} &= \sum_i \left(r_+^{(i)} \overline{V} \left(\prod_k \mathbb{E}_k^{\alpha_k^{(i)} - \beta_k^{(i)}} - 1 \right) \prod_j \left(\frac{((n_j))^{\alpha_j^{(i)}}}{\overline{V}^{\alpha_j^{(i)}}} \right) P(\{n_k\};\,t) \right.\\ &\qquad \left. + r_-^{(i)} \overline{V} \left(\prod_k \mathbb{E}_k^{\beta_k^{(i)} - \alpha_k^{(i)}} - 1 \right) \prod_j \left(\frac{((n_j))^{\beta_j^{(i)}}}{\overline{V}^{\beta_j^{(i)}}} \right) P(\{n_k\};\,t) \right) ,\end{aligned} \tag{3}$$

where $\mathbb{E}$ is the step operator ([18], chapter VI, Eq 3.1), $\sum_i$ represents the sum over all reactions *i*, and rate superscripts (*i*) denote the rates for reaction *i*. The first term in the sum on the right-hand side represents reactions as in Eq (2) (proportional to a constant interaction rate $r_+^{(i)}$) and the second term their inverse reactions (proportional to constant interaction rate $r_-^{(i)}$)). Note that the inverse reaction does not always exist. If it exists, in a MuMoT model definition this would simply be written as an expression like the one in Eq (2), *i.e.* the convention used in MuMoT strictly follows Eq (2). For example, see input cell In[2] in Fig 1; there are also several examples in the online user manual to show how this works [10]. The expression $((n_j))^{\alpha_j} = n_j!/(n_j - \alpha_j)!$ is introduced as an abbreviation. Eq (3) describes the temporal evolution of the joint probability distribution that the system under study is in state $\{n_k\}$ at time *t*. Here, $\{n_k\}$ summarises all agents' individual states as a set. To express changes following interactions we make use of step operators $\mathbb{E}_k$ which increase or decrease the number of agents in state *k* [18]. MuMoT automates the derivation of Eq (3) using the initial model definition according to Eq (2). The Master equation can be accessed as a symbolic equation object for further analysis by expert users, if so desired.

van Kampen Expansion of the Master Equation

In general, there are only very few examples for which Eq (3) can be solved exactly. In what follows we describe an approximation method known as *system size expansion* or *van Kampen expansion* that yields analytical expressions to approximate the solution of a Master equation. However, here we only introduce the main idea of the expansion method and refer to van Kampen's textbook [18] for further details. Let $\Phi_{X_k} = X_k/\overline{V}$ denote the proportion of the population X_k given the system size $\overline{V}$. Note that Φ is a reserved symbol in MuMoT used to express population proportions—the analogue to concentrations of reactants in a chemical reaction. The probability to observe the system in state n_k has a maximum around the macroscopic variable Φ_{X_k} with a deviation around that maximum of order $\sqrt{n_k} \sim \sqrt{\overline{V}}$ [18]. We may now replace the number n_k by a new random variable, say η_{X_k}, according to [18]

$$n_k = \overline{V}\,\Phi_{X_k} + \sqrt{\overline{V}}\,\eta_{X_k}\,. \tag{4}$$

This also means that the probability distribution *P* needs to be rewritten in the new variables, i.e. $P(\{n_k\};\,t) \to P(\{\eta_{X_k}\};\,t)$. Accordingly, the step operators $\mathbb{E}$ in Eq (3) are expanded to yield [18]

$$\mathbb{E} = 1 + \frac{1}{\sqrt{\overline{V}}}\frac{\partial}{\partial \eta_{X_k}} + \frac{1}{2\,\overline{V}}\frac{\partial^2}{\partial \eta_{X_k}^2} + \cdots\,. \tag{5}$$

Calculating the time derivative of $P(\{\eta_{X_k}\};\,t)$ by applying Eqs (4) and (5) to Eq (3) it is possible to get the equation for $P(\{\eta_{X_k}\};\,t)$ expressed in terms of different orders of the systems size $\overline{V}$ (note that the η_{X_k} are time-dependent via Φ_{X_k} in Eq(4)). As a result, there are large terms $\propto \sqrt{\overline{V}}$ which should cancel, yielding the macroscopic equation of motion for Φ_{X_k}. This corresponds to directly deriving the macroscopic ODE for Φ_{X_k} from the underlying reaction by applying the *law of mass action*. The next highest order in this expansion is $\propto \overline{V}^0$. Collecting all terms $\propto \overline{V}^0$ and neglecting all other terms $(\sim \mathcal{O}(\overline{V}^{-1/2}))$ yields a Fokker-Planck equation with terms linear in η_{X_k} (*linear noise approximation*). Although we do not attempt to solve Master equations or their approximations in the form of linear Fokker-Planck equations in MuMoT, we utilise the linear Fokker-Planck equation to compute analytical expressions that represent fluctuations and noise correlations, by deriving equations of motion for first and second order moments of $P(\{\eta_{X_k}\}; t)$ according to

$$\begin{aligned} \frac{\partial}{\partial t}\langle\eta_{X_i}\rangle(t) &= \int d\boldsymbol{\eta}\ \eta_{X_i}\ \frac{\partial}{\partial t}P(\{\eta_{X_k}\}; t), \\ \frac{\partial}{\partial t}\langle\eta_{X_i}\eta_{X_j}\rangle(t) &= \int d\boldsymbol{\eta}\ \eta_{X_i}\eta_{X_j}\ \frac{\partial}{\partial t}P(\{\eta_{X_k}\}; t), \end{aligned} \quad (6)$$

where $d\boldsymbol{\eta} = d\eta_{X_1}\cdots d\eta_{X_K}$, and $\partial P/\partial t$ represents the linear Fokker-Planck equation. Both van Kampen expansion and derivation of the linear Fokker-Planck equation can be readily performed in MuMoT. In addition, in MuMoT explicit expressions for first and second order moments following from Eq (6) may be derived. Furthermore, MuMoT can attempt to obtain analytical solutions for these equations in the stationary state.

All mathematical procedures concerning the Master equation and Fokker-Planck equation make extensive use of Python's SymPy package [19].

Other Methods to Study Noise in MuMoT

Making use of MuMoT's functionality described in the previous paragraph, it is possible to compute and display the temporal evolution of correlation functions $\langle\eta_{X_k}(t)\,\eta_{X_j}(0)\rangle$; examples of how to do this are given in the online user manual [10]. Noise can also be displayed in stream and vector plots; if requested then MuMoT tries to obtain the stationary solutions of the diagonal elements of the second order moments and then project these onto the direction of the eigenvectors of available stable fixed points of the macroscopic ODEs. If the system is too complicated and MuMoT cannot find an analytical solution, noise may be calculated by principled numerical simulation, as described below.

Stochastic Simulation

The Master equation of Eq (3) can be very difficult to solve for even very simple systems, therefore most studies resort to the complementary approach of numerical simulations [20]. Gillespie proposed a probabilistically exact algorithm for simulating chemical reactions called the *stochastic simulation algorithm* (SSA) [21]. Each simulation computes a stochastic temporal trajectory of the state variables from a given user-defined initial condition $\partial P(\{n_k\};\ 0)$ for a user-defined maximum time T. Averaging various trajectories gives an approximation of the solution of Eq (3) (for a given $\partial P(\{n_k\};\ 0)$) that increases in accuracy with the number of simulations. MuMoT implements the SSA via the command SSA(). The user can run a single simulation to generate a single temporal trajectory, or otherwise run several simulations and aggregate the data in a single plot. The user can visualise the entire temporal trajectory (in a plot similar to Fig 3), or the final population distribution $\partial P(\{n_k\};\ T)$ in the form of either a barplot or as points in a 2-dimensional space plane (in which the two axes are state variables). Multiple trajectories can be aggregated in standardised ways of displaying probability distributions, *e.g.*, in the 2-dimensional space plane, simulation aggregates are visualised as ellipses centred on the distribution mean and with 1-σ covariance sizes (*e.g.* see the green ellipse in Fig 4(bottom panels)). This aggregate visualisation can be superimposed on to stream and vector field plots when requested, and if Eq (3) cannot be analytically solved by MuMoT, as discussed above.

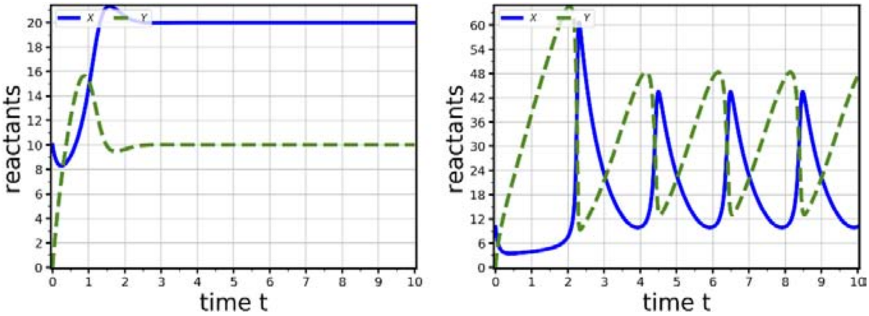


Figure 3. Numerical integration of the Brusselator equations. The Brusselator equations ([3], p.253) exhibit either stable (left) or oscillatory (right) dynamics

according to the parameter values selected. Parameter sets: $\Phi_\alpha = \Phi_\beta = \chi = \delta = \gamma = \xi = 2.0$, $\Phi_{Xt(0)} = 1.0$, system size = 10 (left), $\Phi_\alpha = \chi = \delta = \gamma = \xi = 2.0$, $\Phi_\beta = 5.5$, $\Phi_{Xt(0)} = 1.0$, system size = 10 (right).

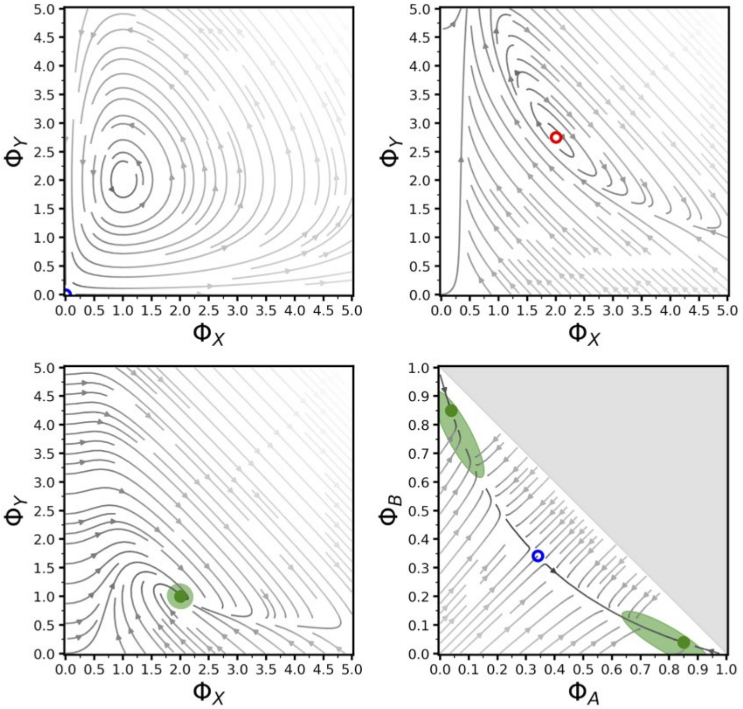


Figure 4. Phase portraits with computed fixed points and noise. Upper-left: oscillatory dynamics in the Lotka-Volterra equations ([3], p.79) (parameters $\Phi_A = \alpha = \beta = \gamma = 2.0$). Upper-right: limit cycle in the Brusellator ([3], p.253) (parameters $\Phi_\alpha = \chi = \delta = \gamma = \xi = 2.0$, $\Phi_\beta = 5.5$). Lower-left: global attractor with isotropic noise in the Brusellator ([3], p.253) (parameters $\Phi_\alpha = \Phi_\beta = \chi = \delta = \gamma = \xi = 2.0$, system size = 10). Lower-right: co-existence of two stable attractors in the honeybee swarming model [2], with anisotropic non-axis-parallel noise (parameters $\Delta = 0.0$, $\mu = 3.0$, $s = 10.0$, system size = 20, runs = 100). Line shading indicates speed of flow, with darker representing faster. Fixed points are denoted as stable (dark solid green circle), saddle (hollow blue circle), or unstable (hollow red circle). Light green ellipses represent 1-σ noise around stable fixed points.

Spatial Noise

MuMoT also enables the study of the effects of spatial noise on a model. Including spatial noise relaxes the sometimes simplistic assumption of a well-mixed system in which interactions between any group of reactants can always happen, at rates proportional to the product of their relative frequencies in the population. Instead, each reactant has a set of *available* reactants with which it can interact at each timestep. The set of possible interactions corresponds to the system's interaction topology, which the user can select among a set of standard graph structures. Graphs are handled by MuMoT through the functionalities offered by the NetworkX library [22] which allows advanced users to easily add new topologies. In the first MuMoT release, the available topologies are the complete graph, the Erdös–Rényi random graph [23], the Barabási–Albert scale-free network [24], and the random geometric graph [25]. The latter is constructed by locating at random uniform locations the reactants in a square environment with edge length 1, and allowing interaction between two reactants when their Euclidean distance is less than or equal to a user-defined distance. The topology of the random geometric graphs can be static or time-varying. The latter is implemented by letting each reactant perform a correlated random walk in the 2-dimensional environment and recomputing the topology each time based on the new distances between reactants.

Spatial noise is difficult to compute analytically in an automatised way, therefore MuMoT computes it numerically via individual-based simulations. Each reactant is simulated as an agent which probabilistically interacts at synchronous discrete timesteps with the available reactants. The agent's behaviour is automatically implemented from the model's reaction kinetics as a probabilistic finite state machine following the technique proposed in [26]. Along with the agents' behaviour, MuMoT automatically sizes the timestep length to match the time-scale with the population-level descriptions (*e.g.* ODEs and Master equation). This feature can be particularly convenient if the user aims at a quantitative comparison between model description levels. Similarly to SSA simulations, the user can select to run individual simulations or to aggregate results from multiple independent simulations to compute statistical distributions.

RESULTS

All results can be reproduced using the MuMoTpaperResults.ipynb Jupyter notebook [10].

Numerical Integration

To illustrate the numerical integration functionality of MuMoT we repeat analyses of the Brusselator equations ([3], p.253) in Fig 3. The equations have two dynamical regimes, one with a single globally stable attractor when $\Phi_\beta \leq \Phi_\alpha^2$ (Fig 3 (left)), and one in which a stable limit cycle exists when $\Phi_\beta > \Phi_\alpha^2$ (Fig 3 (right)).

Phase Portraits with Fixed Point and Noise Calculations

We illustrate the phase portrait functionality of MuMoT in Fig 4 by repeating analyses of a variety of equation systems: the classical Lotka-Volterra equations ([3], p.79), the Brusellator equations ([3], p.253), and a model of collective decision-making by swarming honeybees [2, 14]. These systems can exhibit a variety of dynamics including: oscillations (Fig 4 (upper-left)), unstable fixed points with limit cycles (Fig 4 (upper-right)), globally stable attractors (Fig 4 (bottom-left)), and stable attractors co-existing with saddle points (Fig 4 (bottom-right)). When stable fixed points are present MuMoT can calculate or compute the equilibrium noise around them, dependent on system size (Fig 4 (bottom)); this can be either isotropic (Fig 4 (bottom-left)), or anisoptropic and/or non-axis-parallel (Fig 4 (bottom-right)). This latter case is particularly interesting because the correct noise around the fixed point may differ substantially from simply adding Gaussian noise to the dynamical equations.

Bifurcation Analysis

MuMoT's bifurcation analysis functionality is illustrated through reproducing a number of bifurcation analyses [14] of the honeybee model presented above [2] (Fig 5). These reveal conditions under which the dynamics exhibit: (i) a pitchfork bifurcation (Fig 5 (left)), a sample post-bifurcation phase portrait for which is presented in Fig 4, (ii) an unfolding of the pitchfork bifurcation (*i.e.* saddle-node bifurcation) (Fig 5 (centre)), and (iii) a hysteresis loop (Fig 5 (right)). These can be compared to figures 5(i)-(iii) of [14].

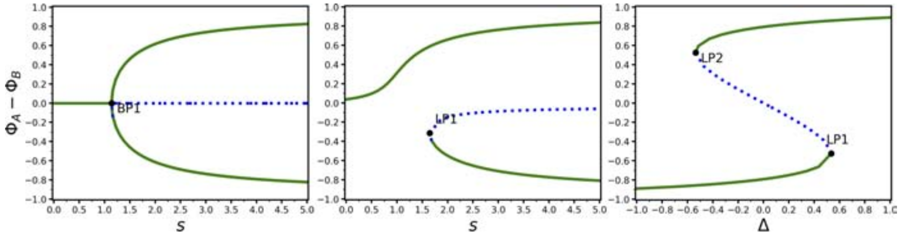


Figure 5. Bifurcation analysis of a nonlinear decision-making model. Bifurcations of the honeybee swarming model [2, 14]. Left: symmetry breaking in the two decision populations through a pitchfork bifurcation, with strength of cross-inhibitory stop-signalling *s* as the bifurcation parameter (cf. [14] Fig 5i) (parameters $\Delta = 0.0$, $\mu = 4.0$). Centre: unfolding of the pitchfork bifurcation into a saddle-node bifurcation (cf. [14] Fig 5ii) (parameters $\Delta = 0.1$, $\mu = 4.0$). Right: hysteresis loop with option quality difference Δ as the bifurcation parameter (cf. [14] Fig 5iii) (parameters $\mu = s = 4.0$). Solid black lines denote stable branches, dashed blue lines denote unstable branches.

Finite Population and Spatial Numerical Simulation

MuMoT can be used to perform a variety of spatial numerical simulations, illustrated in Fig 6 for the honeybee swarming model introduced above [2, 14]. Non-spatial finite-population simulation reproduces the statistics of deadlock breaking observed in [14] (Fig 6 (left)). Spatial noise can also be incorporated either by embedding the model in a network (Fig 6 (centre)) or 2d-plane (Fig 6 (right)).

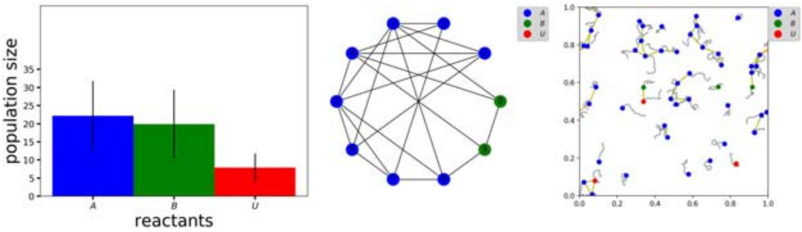


Figure 6. Numerical simulations of a nonlinear decision-making model. Numerical simulations of the honeybee swarming model [2, 14] given various sources of noise. Left: finite-population noise effects during symmetry-breaking in a well-mixed model (parameters $\Delta = 0$, $\mu = 3.0$, $s = 3.0$, $\Phi_{Ut(0)} = 1.0$, system size = 50, time = 10, runs = 10). Centre: finite-population and spatial noise effects due to embedding the model in a random graph. Right: finite-population and spatial noise effects due to embedding the model in a plane, with agents

performing correlated random walks; traces indicate recent agent paths, links indicate current interaction events.

Derivation of the Master Equation and Expansion to Derive the Fokker-Planck Equation

Here we reproduce the analysis presented in [18] (pp. 244-246) to derive the Master Equation and Fokker-Planck equation for the following toy model:

$$\begin{gathered} (A) \xrightarrow{k} X \\ X + X \xrightarrow{h} \emptyset + \emptyset \end{gathered} \tag{7}$$

The automated analysis results in

$$\frac{\partial}{\partial t} P(X,t) \coloneqq \frac{Ak}{\overline{V}} (\mathrm{E}_{\mathrm{op}}(X,-1) - 1)\overline{V} P(X,t) + h(\mathrm{E}_{\mathrm{op}}(X,2) - 1)\frac{X}{\overline{V}}(X-1)P(X,t) \tag{8}$$

And

$$\frac{\partial}{\partial t} P(\eta_X,t) \coloneqq \frac{\Phi_A k}{2} \frac{\partial^2}{\partial \eta_X^2} P(\eta_X,t) + 2\Phi_X^2 h \frac{\partial^2}{\partial \eta_X^2} P(\eta_X,t) + 4\Phi_X \eta_X h \frac{\partial}{\partial \eta_X} P(\eta_X,t) + 4\Phi_X h P(\eta_X,t) \tag{9}$$

as expected.

A substantially more complicated example derivation, for the honeybee swarming model of Eq 1 [2, 14]. This derivation is equivalent to that performed in [2] and results in the same dynamical equations.

AVAILABILITY AND FUTURE DIRECTIONS

MuMoT is available as source code, as a package for Python 3 [27] via PyPI (pypi.python.org), and as a server-based installation currently exemplified by free-to-use access to the interactive user manual and other notebooks using the Binder service [9], which requires only a web browser to use. MuMoT is written in Python 3 and integrates with Jupyter Notebooks [8] and as such is platform-independent. Non-interactive aspects of MuMoT's functionality can also be accessed through using it as a standalone Python package, enabling its modelling and analysis functionality to be used from within third-party code projects. MuMoT is available under the GPL licence version 3.0, and makes use of other software available under the MIT licence. For further details including links to usage information are available at github.com/DiODeProject/MuMoT/.

Numerous software products have been proposed to perform subsets of the analyses offered by MuMoT. For instance, several tools offer the possibility to run the SSA and efficiently analyse reaction kinetics models [28–36]. Similarly, software to analyse mean-field dynamical systems and perform bifurcation analysis is widely available, *e.g*. MATCONT for Matlab [37], or the Dynamica package for Wolfram Mathematica [38]. Linear noise approximations have previously been implemented as well [32]. Several tools offers software to simulate complex systems, dynamical networks, and agent-based models [39–41], some of which run as Jupyter notebooks as MuMoT does [42, 43].

In contrast to the previous solutions, MuMoT combines ease-of-use with a multi-level analysis that spans from ODEs analysis, to statistical physics approximations, bifurcation analysis, and SSA and multiagent simulations, integrated within a simple interactive notebook document interface. This makes MuMoT particularly appropriate for non-experts to learn to build models and apply complex mathematical and computational techniques to them, to communicate research results, and to enable students to interactively explore models, and modelling and analysis techniques.

Future work should focus on integrating MuMoT with other software and standard. For example, the simple textual input method for MuMoT models is very accessible to non-experts, but precludes more sophisticated use cases. Import and export via interchange formats such as Systems Biology Markup Language (SBML) [44] would enable users to connect between MuMoT for general analysis, and external specialist software packages for more detailed analyses; for example StochSS [36] can run the SSA algorithm on cloud infrastructure for larger-scale computations, and perform parameter sweeps and estimation. Embracing data interchange formats will allow MuMoT to take its place as an integral part of the growing ecosystem of open-source modelling software.

ACKNOWLEDGMENTS

We thank Will Furnass with assistance in deployment, documentation, and continuous integration. We thank Renato Pagliara for assistance in testing on the Windows platform.

REFERENCES

1. Tyson JJ, Chen KC, Novak B. Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. Current Opinion in Cell Biology. 2003;15(2):221–231. pmid:12648679
2. Seeley TD, Visscher PK, Schlegel T, Hogan PM, Franks NR, Marshall JAR. Stop signals provide cross inhibition in collective decision-making by honeybee swarms. Science. 2012;335(6064):108–111. pmid:22157081
3. Murray JD. Mathematical Biology I: An Introduction. 3rd ed. Springer-Verlag; 2002.
4. Yildiz E, Ozdaglar A, Acemoglu D, Saberi A, Scaglione A. Binary opinion dynamics with stubborn agents. ACM Transactions on Economics and Computation (TEAC). 2013;1(4):19.
5. Israel G. La Mathématisation du Réel. Seuil; 1996.
6. Marshall JAR, Franks NR. Computer modeling in behavioral and evolutionary ecology: whys and wherefores. In: Modeling Biology: Structures, Behavior, Evolution. The Vienna Series in Theoretical Biology. MIT Press; 2007. p. 335–353.
7. van Rossum G, et al. Python 3; 2008. Available from: https://www.python.org/3/reference/; accessed on 2019-06-12 [cited 2019-03-07].
8. Various. Project Jupyter;. Available from: https://jupyter.org [cited 2019-06-12].
9. Various. Binder;. Available from: https://mybinder.org [cited 2019-06-12].
10. Marshall, James A R and Reina, Andreagiovanni and Bose, Thomas. MuMoT online manual; 2019. Available from: https://mumot.readthedocs.io/en/latest/getting_started.html [cited 2019-06-28].
11. Leff A, Rayfield JT. Web-application development using the model/view/controller design pattern. In: Proceedings of the Fifth IEEE International Enterprise Distributed Object Computing Conference. IEEE; 2001. p. 118–127.
12. van Rossum G, Warsaw B, Coghlan N. PEP 8: style guide for Python code. Python.org; 2001. Available from: https://www.python.org/dev/peps/pep-0008/.
13. Lee BD. Ten simple rules for documenting scientific software. PLoS Computational Biology. 2018;14(12):e1006561. pmid:30571677

14. Pais D, Hogan PM, Schlegel T, Franks NR, Leonard NE, Marshall JAR. A mechanism for value-sensitive decision-making. PLoS one. 2013;8(9):e73216. pmid:24023835
15. Galla T. Independence and interdependence in the nest-site choice by honeybee swarms: agent-based models, analytical approaches and pattern formation. Journal of Theoretical Biology. 2010;262(1):186–196. pmid:19761778
16. Jones E, Oliphant T, Peterson P, et al. SciPy: Open source scientific tools for Python; 2001–. Available from: http://www.scipy.org/.
17. Clewley, R H and Sherwood, W E and LaMar, M D and Guckenheimer, J M. PyDSTool: a software environment for dynamical systems modeling; 2007. Available from: https://pydstool.github.io/PyDSTool/ [cited 2019-06-12].
18. van Kampen NG. Stochastic Processes in Physics and Chemistry: Third Edition. Amsterdam: North-Holland; 2007.
19. Meurer A, Smith CP, Paprocki M, Čertík O, Kirpichev SB, Rocklin M, et al. SymPy: symbolic computing in Python. PeerJ Computer Science. 2017;3:e103.
20. Gillespie DT, Hellander A, Petzold LR. Perspective: Stochastic algorithms for chemical kinetics. The Journal of Chemical Physics. 2013;138(17):170901. pmid:23656106
21. Gillespie DT. A general method for numerically simulating stochastic time evolution of coupled chemical reactions. Journal of Computational Physics. 1976;22:403–434.
22. Hagberg AA, Schult DA, Swart PJ. Exploring network structure, dynamics, and function using NetworkX. In: Proceedings of the 7th Python in Science Conference (SciPy 2008). SciPy; 2008.
23. Erdös P, Rényi A. On random graphs I. Publicationes Mathematicae (Debrecen). 1959;6:290–297.
24. Barabási AL, Albert R. Emergence of scaling in random networks. Science. 1999;286(5439):509–512. pmid:10521342
25. Penrose M. Random Geometric Graphs. Oxford studies in probability. Oxford University Press; 2003.
26. Reina A, Valentini G, Fernández-Oto C, Dorigo M, Trianni V. A design pattern for decentralised decision making. PLoS ONE. 2015;10(10):e0140950. pmid:26496359

27. Marshall, James A R and Reina, Andreagiovanni and Bose, Thomas. MuMoT 1.0.0-release. 2019.
28. Adalsteinsson D, McMillen D, Elston TC. Biochemical Network Stochastic Simulator (BioNetS): software for stochastic modeling of biochemical networks. BMC Bioinformatics. 2004;5(1):24. pmid:15113411
29. Ramsey S, Orrell D, Bolouri H. Dizzy: stochastic simulation of large-scale genetic regulatory networks. Journal of Bioinformatics and Computational Biology. 2005;3(02):415–436. pmid:15852513
30. Mendes P, Hoops S, Sahle S, Gauges R, Dada J, Kummer U. In: Computational Modeling of Biochemical Networks Using COPASI. Totowa, NJ: Humana Press; 2009. p. 17–59.
31. Mauch S, Stalzer M. Efficient formulations for exact stochastic simulation of chemical systems. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2011;8(1):27–35. pmid:21071794
32. Thomas P, Matuschek H, Grima R. Intrinsic Noise Analyzer: A software package for the exploration of stochastic biochemical kinetics using the system size expansion. PLoS ONE. 2012;7(6):e38518. pmid:22723865
33. Abel JH, Drawert B, Hellander A, Petzold LR. GillesPy: A Python package for stochastic model building and simulation. IEEE Life Sciences Letters. 2016;2(3):35–38. pmid:28630888
34. Sanft KR, Wu S, Roh M, Fu J, Lim RK, Petzold LR. StochKit2: software for discrete stochastic simulation of biochemical systems with events. Bioinformatics. 2011;27(17):2457–2458. pmid:21727139
35. Maarleveld TR, Olivier BG, Bruggeman FJ. StochPy: A comprehensive, user-friendly tool for simulating stochastic biological processes. PLoS ONE. 2013;8(11):e79345. pmid:24260203
36. Drawert B, Hellander A, Bales B, Banerjee D, Bellesia G, Daigle BJ, et al. Stochastic Simulation Service: Bridging the gap between the computational expert and the biologist. PLOS Computational Biology. 2016;12(12):e1005220. pmid:27930676
37. Dhooge A, Govaerts W, Kuznetsov YA. MATCONT: A Matlab package for numerical bifurcation analysis of ODEs. SIGSAM Bull. 2004;38(1):21–22.

38. Beer RD. Dynamica: a Mathematica package for the analysis of smooth dynamical systems; 2018. Available from: http://mypage.iu.edu/~rdbeer/.

39. Wilensky U. NetLogo. Northwestern University, Evanston, IL: Center for Connected Learning and Computer-Based Modeling; 1999. Available from: http://ccl.northwestern.edu/netlogo/.

40. Kiran M, Richmond P, Holcombe M, Chin LS, Worth D, Greenough C. FLAME: Simulating large populations of agents on parallel hardware architectures. In: Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1. AAMAS'10. Richland, SC: IFAAMAS; 2010. p. 1633–1636.

41. Luke S, Cioffi-Revilla C, Panait L, Sullivan K, Balan G. MASON: A multiagent simulation environment. SIMULATION. 2005;81(7):517–527.

42. Sayama H. PyCX: a Python-based simulation code repository for complex systems education. Complex Adaptive Systems Modeling. 2013;1(2).

43. Medley JK, Choi K, König M, Smith L, Gu S, Hellerstein J, et al. Tellurium notebooks–An environment for reproducible dynamical modeling in systems biology. PLOS Computational Biology. 2018;14(6):e1006220. pmid:29906293

44. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics. 2003;19(4):524–531. pmid:12611808

Chapter

13

EFFECTS OF GREENHOUSE GASES AND HYPOXIA ON THE POPULATION OF AQUATIC SPECIES: A FRACTIONAL MATHEMATICAL MODEL

Pushpendra Kumar[1], V. Govindaraj[1], Vedat Suat Erturk[2], and Mohamed S. Mohamed[3]

[1]Department of Mathematics, National Institute of Technology Puducherry, Karaikal, 609609, India.
[2]Department of Mathematics, Faculty of Arts and Sciences, Ondokuz Mayis University, Atakum, 55200, Samsun, Turkey.
[3]Department of Mathematics and Statistics, College of Science, Taif University, Taif, 21944, Saudi Arabia.

ABSTRACT

Study of ecosystems has always been an interesting topic in the view of real-world dynamics. In this paper, we propose a fractional-order nonlinear mathematical model to describe the prelude of deteriorating quality of water cause of greenhouse gases on the population of aquatic animals. In the

Citation: (APA): Kumar, P., Govindaraj, V., Erturk, V. S., & Mohamed, M. S. (2022). Effects of greenhouse gases and hypoxia on the population of aquatic species: a fractional mathematical model. Advances in Continuous and Discrete Models, 2022(1), 1-19. (19 pages).

proposed system, we recall that greenhouse gases raise the temperature of water, and because of this reason, the dissolved oxygen level goes down, and also the rate of circulation of disintegrated oxygen by the aquatic animals rises, which causes a decrement in the density of aquatic species. We use a generalized form of the Caputo fractional derivative to describe the dynamics of the proposed problem. We also investigate equilibrium points of the given fractional-order model and discuss the asymptotic stability of the equilibria of the proposed autonomous model. We recall some important results to prove the existence of a unique solution of the model. For finding the numerical solution of the established fractional-order system, we apply a generalized predictor–corrector technique in the sense of proposed derivative and also justify the stability of the method. To express the novelty of the simulated results, we perform a number of graphs at various fractional-order cases. The given study is fully novel and useful for understanding the proposed real-world phenomena.

MSC: 26A33; 65D05; 65D30; 65L07; 92B05

Keywords: Dissolved oxygen, Temperature, Aquatic species, Greenhouse gases, Fractional mathematical model, Numerical method, Modified Caputo fractional Derivative

INTRODUCTION

In the study of greenhouse effects, we know that in the day the sun warms up the atmosphere of earth. But when the Earth supercools at the night, then the presented heat is radiated again into the environment. In the duration of this process, the heat is exploited by the greenhouse gases in the environment of earth. This process makes the layer of the earth thermal, which causes the possibility of living being's survival on earth. However, because of the increment in the level of greenhouse gases, the earth's temperature has raised simultaneously. This has caused a number of drastic impacts. In the list of reasons of greenhouse effect, deforestation, burning of fossil fuels, farming, industrial waste, and landfills play a major role. The major effects of increased greenhouse gases are depletion of ozone layer, global warming, air and smog pollution, water bodies acidification, etc. Since the starting of the industrial revolution, the concentration of carbon dioxide, chlorofluorocarbon (CFC), nitrous oxide, and methane have enhanced in the environment, and there is firm witness that the venomous impacts of greenhouse gases on our ecological systems have been taken account as a outcome of human bustles.

Aquatic life simply means to stay in surface water, and water in this paragon is specified as a marine habitat. Living beings that live in the water either permanently or momentarily are called aquatic animals and plants, and these compose the beings in water aquatic life. It is well known that the increment in the temperature of water causes the reduction in the concentration of mixed oxygen of the aquatic environment and also rises the requirement of mixed oxygen for the aquatic animals. Invertebrates, fish, and other aquatic species rely upon the amount of oxygen decomposed in the water, and in the absence of it, they may not live. A small changes in concentration of mixed oxygen can effect the conformation of aquatic society [1]. So the rate of survival of the aquatic density (Fig. 1) goes down under hypoxia, and the oxygen necessary for their living raises with growth in temperature [2, 3]. Hence, because of the combined influences of reduced concentration of mixed oxygen and enhanced demand of oxygen by the animals, the warming of species bodies rises the death rate of species population [3, 4]. To define these dynamics, a number of models have been proposed, but only few models [5, 6] have been given to simulate the effects of dissolved oxygen and temperature on the population of aquatic species.

Figure 1. Some aquatic animals.

In the matter of the above discussion, in this study, we prepare a fractional-order mathematical system to simulate the joint influences of low mixed oxygen density, exalted water temperature, and raised oxygen demand on the extinction or survival of aquatic population. Fractional derivatives are one of the most effective tools for simulations and have been proposed in many different ways (for instance, see [7–9]). Fractional-order models have been widely used to define a number of real-world problems because their memory effects make these models more visible in the literature. Recently, a number of fractional-order models have been prepared by researchers. In this regard, in [10–18] the authors have proposed a number of fractional-order mathematical models to describe the dynamics of Covid-19 epidemic. In [19, 20] the authors have simulated the fractional-order dynamics of well-known lassa hemorrhagic fever. The applications of fractional derivatives in ecology can be seen in [21]. Regarding some more specific areas, nonclassical derivatives have been successfully used to derive the structure of tuberculosis [22], malaria [23], mosaic disease [24], Nipah epidemic [25], canine distemper virus [26], and huanglongbing transmission [27]. In [28] the authors used a fractional-order time-delay mathematical model to describe the process of oncolytic virotherapy. A study on analytic solution for oxygen diffusion from capillary to tissues via fractional derivatives is proposed in [29]. Also, an application of a new generalized Caputo derivative to define the famous love story of Layla and Majnun is given in [30]. So the literature of fractional-order calculus is increasing exponentially day by day. Also, a number of true and false results come on the various fractional derivatives. Recently, in [31] the authors have proved that in the case of evolution equations in terms of the Caputo–Fabrizio and Atangana–Baleanu fractional derivatives, intrinsic discontinuities occur. The geometry of fractional-order derivatives is still not well-defined, but their applications in different scientific fields make them more visible to the literature. Some important studies related to the properties of fractional derivatives, special functions, and different types of inequalities can be learned from [32–35]. Nonstandard Chebyshev collocation and finite difference schemes for solving fractional diffusion equations are proposed in [36]. Some novel analysis on the fractional differential equations for the generalized Mittag-Leffler function are discussed in [37]. A study on the analytical solutions of the fractional-order equations with uncertainty is proposed in [38]. Alderremy et al. [39] have discussed some novel models of the multispace-fractional Gardner equation. A study on spectral collocation method for solving smoking model is proposed in [25]. In [40] the authors have proposed a study on Darcy–

Brinkman–Forchheimer model for nanobioconvection stratified MHD flow through an elastic surface. In [41] a reduced differential transform scheme for simulating nonlinear biomathematics models is given. In [42] a study on dynamical features and signal flow graph of nonlinear noninteger order smoking mathematical model is explored. In [43], some numerical methods for a model of relativistic electrons arising in the laser thermonuclear fusion are investigated. The manuscript is designed as follows: In Sect. 2, firstly, we remind some important definitions and results. In Sect. 3, we give a complete description of the proposed fractional-order nonlinear model, where we define the significance and importance of every small part of the model. Then in Sect. 4, we give a complete mathematical analysis related to the solution existence, derivation, and stability. To show the correctness of our results, in Sect. 5, we present the necessary graphs at various fractional-order values and parameter weights. At the end, a conclusion gives a comfortable end to the paper.

PRELIMINARIES

Firstly, we remind some important definitions and results.

Definition 1 ([44])

The new definition of the Caputo-type fractional derivative $D_{d_+}^{\sigma,\varkappa}$ of order $\sigma > 0$ (called a new generalized Caputo) for the function $\Psi \in C^1([d, T]))$ is given by

$$\left(D_{d_+}^{\sigma,\varkappa}\Psi\right)(\xi) = \frac{\varkappa^{\sigma-n+1}}{\Gamma(n-\sigma)}\int_d^{\xi} s^{\varkappa-1}\left(\xi^{\varkappa}-s^{\varkappa}\right)^{n-\sigma-1}\left(s^{1-\varkappa}\frac{d}{ds}\right)^{n}\Psi(s)\,ds, \quad \xi > d, \tag{1}$$

where $\rho > 0$, $d \geq 0$, *and* $n - 1 < \sigma \leq n$.

Lemma 1 ([45])

For $0 < b < 1$ *and a nonnegative integer* ϱ, *there exist positive constants* $\mathcal{C}_{b,1}$ *and* $\mathcal{C}_{b,2}$, *dependent only on* b, *such that*

$$(\varrho + 1)^{b} - \varrho^{b} \leq \mathcal{C}_{b,1}(\varrho + 1)^{b-1}$$

and

$$(\varrho + 2)^{b+1} - 2(\varrho + 1)^{b+1} + \varrho^{b+1} \leq \mathcal{C}_{b,2}(\varrho + 1)^{b-1}.$$

Lemma 2 ([45])

Let $d_{q,s} = (s - q)^{b-1}$ *for* $q = 1, 2, \ldots, s-1$ *and* $d_{q,s} = 0$ *for* $q \geq s$, *let* M, b, h, T > 0, *and* $rh \leq T$, *where r is a positive integer. Let* $\sum_{q=r}^{q=s} d_{q,s}|e_q| = 0$ *for* r>s≥1. *If*

$$|e_s| \leq Mh^b \sum_{q=1}^{s-1} d_{q,s}|e_q| + |\beta_0|, \quad s = 1, 2, \ldots, r,$$

then

$$|e_r| \leq \mathcal{C}|\beta_0|, \quad r = 1, 2, \ldots,$$

where $\mathcal{C}$ *is a positive constant not dependent on r and h.*

MODEL DYNAMICS

Now we propose a fractional-order mathematical model to study the proposed dynamics. In [1] the authors have already given an idea on the proposed topic by using an integer-order model. We propose a fractional-order model because it is well known that the memory effects, which cannot be studied in the classical case, can be easily observed by fractional-order derivatives. It is very important that when we propose a fractional-order model, it should have the same time dimension on both sides of the system. Taking care of all these aspects, we define the novel fractional-order model as follows:

$$\begin{cases} {}^{C}D_t^{\sigma,\varkappa}N(t) = G(U,T)N - \frac{g_0^{\sigma}N^2}{\gamma_0}, \\ {}^{C}D_t^{\sigma,\varkappa}T(t) = w^{\sigma}C - \zeta_1(T - T_{10}) + \gamma^{\sigma}(Z_0 - Z), \\ {}^{C}D_t^{\sigma,\varkappa}C(t) = A_0 - \delta_1^{\sigma}C, \\ {}^{C}D_t^{\sigma,\varkappa}Z(t) = O_c^{\sigma} - \Lambda_1^{\sigma}Z - \Lambda^{\sigma}ZC, \\ {}^{C}D_t^{\sigma,\varkappa}U(t) = \gamma_1\beta^{(T-T_0)}(D_s(T) - U) - \delta_2^{\sigma}UN - \zeta^{\sigma}(T - T_{\text{opt}}), \end{cases} \tag{2}$$

where ${}^{C}D_t^{\sigma,\varkappa}$ is the new generalized Caputo-type fractional-order operator of order *σ*. In this model,

$$G(U,T) = g_0^{\sigma}\left(\exp\left(-b\left(\frac{T - T_{\text{opt}}}{T_{\max} - T_{\text{opt}}}\right)\right) + \left(\frac{U - U_0(T)}{U + 1}\right)\right),$$

$$U_0(T) = \beta_{10}^{\sigma} + \beta_{11}^{\sigma}(T - T_{\text{opt}}),$$

and

$$D_s(T) = \frac{D_{s0}}{1 + T - T_{\text{opt}}}.$$

In this model, we have five different classes, in which class *N* shows the logistically crescent aquatic species density whose rate of growth is taken as a function of temperature and mixed oxygen, *U* justifies the dissolved oxygen concentrations, *T* defines the water temperature average of the species, *C* expresses the greenhouse gases accumulative concentrations, and *Z* justifies the concentration of ozone. Also, the term G(U, T) expresses the specific rate of growth of the species, which is in fact an exponentially decreasing function of *T* for $T > T_{opt}$ and increasing function of *U*. The function $U_0(T)$ denotes the quantity of dissolved oxygen demanded by species population, which rises with temperature increase.

The term $D_s(T)$ is defined for the consideration that if the water temperature level is high, close to the optimum temperature, then the natural loaded dissolved oxygen concentration reduces. The significance of all other parameters is completely given in Table 1. The more deep texture of the given model in classical sense can be learned from [1].

Table 1. Description of model parameters

g_0	Intrinsic growth rate
β_{10}	Dissolved oxygen's minimum natural concentration needed by the aquatic species
β_{11}	Increment rate in the mass of mixed oxygen demanded for the species per unit rise in the level of temperature above the suitable temperature
T_{opt}	Optimal water temperature for the aquatic species maximum rate of growth
γ_0	Carrying capacity of the environment
D_{s0}	Dissolved oxygen's natural saturated concentration at $T = T_{opt}$
A_0	Ejection rate of greenhouse gases cause of anthropogenic bustles
w	Increment rate in the temperature of water cause of greenhouse gases
ζ_1	Coefficient of heat transfer of surface

O_c	Physic manufacture of concentration of ozone per unit time in the environment
Λ_1	Natural deterioration rate of concentration of ozone
Λ	Deterioration rate of concentration of ozone cause of greenhouse gases
γ_1	Coefficient of reaeration at the reference temperature
δ_2	Deterioration rate of mixed oxygen because of breathing by the species
ζ	Deterioration rate of mixed oxygen because of a rise in the temperature above the suitable temperature
β	A constant that succumbs upon the tincture state of the water body
γ	Variations rate in the water temperature because of changes in the ozone concentration level associated with its threshold value
Z_0	Threshold of concentration of ozone below which temperature will rise
T_{10}	Temperature of the environment
δ_1	Depletion rate of greenhouse gases
T_0	Context temperature (associated with the turbulence degree in the water, in which turn succumbs on the depth and speed of the river)
b	Constant that incarnates the toxic influence of divergence of T from T_{opt} and divergence of T from T_{max}
T_{max}	Maximum temperature of water at which growth can occur
$N(0)$	Initial population of N
$T(0)$	Initial population of T
$C(0)$	Initial population of C
$Z(0)$	Initial population of Z
$U(0)$	Initial population of U

The equilibria of the given fractional-order mathematical model can be obtained by solving the following system:

$$G(U,T)N - \frac{g_0^\sigma N^2}{\gamma_0} = 0, \tag{3}$$

$$w^\sigma C - \zeta_1(T - T_{10}) + \gamma^\sigma(Z_0 - Z) = 0, \tag{4}$$

$$A_0 - \delta_1^\sigma C = 0, \tag{5}$$

$$O_c^\sigma - \Lambda_1^\sigma Z - \Lambda^\sigma ZC = 0, \tag{6}$$

$$\gamma_1 \beta^{(T-T_0)}\big(D_s(T) - U\big) - \delta_2^\sigma UN - \zeta^\sigma (T - T_{\text{opt}}) = 0. \tag{7}$$

Equation (5) gives

$$C = \frac{A_0}{\delta_1^\sigma}. \tag{8}$$

Equation (6) gives

$$Z = \frac{O_c^\sigma}{\Lambda_1^\sigma + \Lambda^\sigma C}. \tag{9}$$

Equation (4) gives

$$T = \frac{w^\sigma C + \zeta_1 T_{10} + \gamma^\sigma (Z_0 - Z)}{\zeta_1}. \tag{10}$$

Here we have two different types of equilibrium points.

1. Boundary equilibrium point $\bar{E} = (\bar{U}, \bar{Z}, \bar{C}, \bar{T}, \bar{N})$:

$\bar{N} = 0$ (no species population), $\bar{U} = (\frac{D_{s0}}{1+\bar{T}-T_{\text{opt}}} - \frac{\zeta^\sigma (\bar{T}-T_{\text{opt}})}{\gamma_1 \beta^{(\bar{T}-T_0)}})$. Here $\bar{C}, \bar{Z}, \bar{T}$ are given by (8), (9), (10), respectively.

A boundary equilibrium point $\bar{E}$ exists if $D_{s0}\gamma_1\beta^{(\bar{T}-T_0)} - \zeta^\sigma(\bar{T} - T_{\text{opt}})(1 + \bar{T} - T_{\text{opt}}) > 0$ and $Z_0 > \bar{Z}$.

2. Interior equilibrium point E*(U*, Z*, C*, T*, N*), where $N^* = \gamma_0(\exp(-b(\frac{T^*-T_{\text{opt}}}{T_{\max}-T_{\text{opt}}})) + \frac{U^*-(\beta_{10}^\sigma+\beta_{11}^\sigma(T^*-T_{\text{opt}}))}{1+U^*})$ (species population exists) and N* > 0, provided that $U^* - (\beta_{10}^\sigma + \beta_{11}^\sigma(T^* - T_{\text{opt}})) > 0$. Here C*, Z*, T* are given by Equations (8), (9), (10), respectively, and U* is the positive root of the quadratic equation

$$a_1 {U^*}^2 + b_1 U^* + c_1 = 0, \tag{11}$$

where $a_1 = \gamma_1 \beta^{ó(T^*-T_0)}(1+T^*-T_{\text{opt}}) + \delta_2^{ó}\gamma_0 \exp(-b(\frac{T^*-T_{\text{opt}}}{T_{\max}-T_{\text{opt}}}))(1+T^*-T_{\text{opt}}) + \delta_2^{ó}\gamma_0(1+T^*-T_{\text{opt}})$, $b_1 = \gamma_1\beta^{(T^*-T_0)}(1 + T^* - T_{\text{opt}}) - \gamma_1\beta^{(T^*-T_0)}D_{s0} + \delta_2^\sigma \gamma_0 \exp(-b(\frac{T^*-T_{\text{opt}}}{T_{\max}-T_{\text{opt}}}))(1 + T^* - T_{\text{opt}}) + \delta_2\gamma_0(1 + T^* - T_{\text{opt}})(\beta_{10}^\sigma + \beta_{11}^\sigma(T^* - T_{\text{opt}})) + \zeta^\sigma(T^* - T_{\text{opt}})$

$(1 + T^* - T_{\text{opt}}), c_1 = \zeta^\sigma (T^* - T_{\text{opt}})(1 + T^* - T_{\text{opt}}) - \gamma_1 \beta^{(T^*-T_0)} D_{s_0}$. When the given conditions are taken account, the quadratic equation (11) has at least one positive root if $a_1 > 0$, $b_1 > 0$, and $c_1 < 0$. Now we derive the following nonautonomous system after solving the given model (2) for *C*:

$$^{C}D_t^{\sigma,\varkappa} N(t) = G(U,T)N - \frac{g_0^\sigma N^2}{\gamma_0}, \tag{12}$$

$$^{C}D_t^{\sigma,\varkappa} U(t) = \gamma_1 \beta^{(T-T_0)} \big(D_s(T) - U\big) - \delta_2^\sigma UN - \zeta^\sigma (T - T_{\text{opt}}), \tag{13}$$

since $Z^* \le \limsup_{t\to\infty} Z(t), C^* \le \limsup_{t\to\infty} C(t), T^* \le \limsup_{t\to\infty} T(t)$.

Hence the fractional-order nonautonomous model (12)–(13) can be specified in the following equivalent fractional-order autonomous model:

$$\begin{aligned} ^{C}D_t^{\sigma,\varkappa} N(t) = g_0^\sigma \left(\exp\left(-b\left(\frac{T^* - T_{\text{opt}}}{T_{\max} - T_{\text{opt}}} \right)\right) + \frac{U - (\beta_{10}^\sigma + \beta_{11}^\sigma (T^* - T_{\text{opt}}))}{1 + U} \right) \\ - N - \frac{g_0^\sigma N^2}{\gamma_0}, \end{aligned} \tag{14}$$

$$^{C}D_t^{\sigma,\varkappa} U(t) = \gamma_1 \beta^{(T^*-T_0)} \left(\frac{D_{s_0}}{1 + T^* - T_{\text{opt}}} - U \right) - \delta_2^\sigma UN - \zeta^\sigma (T^* - T_{\text{opt}}). \tag{15}$$

The equilibrium points of the dynamic system (14)–(15) are calculated by the following group of equations:

$$g_0^\sigma \left(\exp\left(-b\left(\frac{T^* - T_{\text{opt}}}{T_{\max} - T_{\text{opt}}} \right)\right) + \frac{U - (\beta_{10}^\sigma + \beta_{11}^\sigma (T^* - T_{\text{opt}}))}{1 + U} \right) - N - \frac{g_0^\sigma N^2}{\gamma_0} = 0, \tag{16}$$

$$\gamma_1 \beta^{(T^*-T_0)} \left(\frac{D_{s_0}}{1 + T^* - T_{\text{opt}}} - U \right) - \delta_2^\sigma UN - \zeta^\sigma (T^* - T_{\text{opt}}) = 0. \tag{17}$$

1. Boundary equilibrium point $\bar{\bar{E}}(\bar{\bar{U}},\bar{\bar{N}})$:

$$\bar{\bar{N}} = 0 \quad \text{(no species population)},$$

$$\bar{\bar{U}} = \left(\frac{D_{s_0}}{1 + T^* - T_{\text{opt}}} - \frac{\zeta^\sigma (T^* - T_{\text{opt}})}{\gamma_1 \beta^{(T^*-T_{\text{opt}})}} \right).$$

The existence of boundary equilibrium point $\bar{\bar{E}}$ provides

$$D_{s_0} \gamma_1 \beta^{(T^*-T_{\text{opt}})} - \zeta^\sigma \big(1 + T^* - T_{\text{opt}}\big)\big(T^* - T_{\text{opt}}\big) > 0.$$

2. Interior equilibrium point $E^{**}(U^{**}, N^{**})$:

$N^{**} = \gamma_0 (\exp(-b(\frac{T^* - T_{\text{opt}}}{T_{\max} - T_{\text{opt}}})) + \frac{U^{**} - (\beta_{10}^\sigma + \beta_{11}^\sigma (T^* - T_{\text{opt}}))}{1 + U^{**}})$ (aquatic population exists) and $N^{**} > 0$, provided that

$$U^{**} - (\beta_{10}^{\sigma} + \beta_{11}^{\sigma}(T^* - T_{\text{opt}})) > 0,$$

where U** is a positive root of the quadratic equation

$$A_1 U^{*^2} + B_1 U^* + C_1 = 0, \tag{18}$$

where $A_1 = \gamma_1 \beta^{(T^*-T_0)}(1+T^*-T_{\text{opt}}) + \delta_2^{\acute{o}}\gamma_0 \exp(-b(\frac{T^*-T_{\text{opt}}}{T_{\max}-T_{\text{opt}}}))(1+T^*-T_{\text{opt}}) + \delta_2^{\acute{o}}\gamma_0(1+T^*-T_{\text{opt}}),$
$B_1 = \gamma_1 \beta^{(T^*-T_0)}(1+T^*-T_{\text{opt}}) - \gamma_1 \beta^{(T^*-T_0)} D_{s_0} + \delta_2^{\sigma}\gamma_0 \quad \exp(-b(\frac{T^*-T_{\text{opt}}}{T_{\max}-T_{\text{opt}}}))(1+T^* - T_{\text{opt}}) - \delta_2^{\sigma}\gamma_0(1+T^*-T_{\text{opt}})(\beta_{10}^{\sigma}+\beta_{11}^{\sigma}(T^*-T_{\text{opt}})) + \zeta^{\sigma}(T^*-T_{\text{opt}}) \; (1+T^*-T_{\text{opt}})$, $C_1 = \zeta^{\sigma}(T^* - T_{\text{opt}})(1+T^*-T_{\text{opt}}) - \gamma_1\beta^{(T^*-T_0)}D_{s_0}$.

If the given constraints are satisfied, then the quadratic equation specified by (18) has at least one positive root if

$$A_1 > 0, \qquad B_1 > 0, \qquad C_1 < 0.$$

Lemma 3

For the fractional-order mathematical system (14)–(15), $\bar{\bar{E}}$ *is locally asymptotically stable if* $g_0^{\acute{o}}(\exp(-b(\frac{T^*-T_{\text{opt}}}{T_{\max}-T_{\text{opt}}})) + \frac{\bar{\bar{U}}-(\beta_{10}^{\acute{o}}+\beta_{11}^{\acute{o}}(T^*-T_{\text{opt}}))}{1+\bar{\bar{U}}}) < 0$ *and is an unstable saddle point if* $g_0^{\acute{o}}(\exp(-b(\frac{T^*-T_{\text{opt}}}{T_{\max}-T_{\text{opt}}})) + \frac{\bar{\bar{U}}-(\beta_{10}^{\acute{o}}+\beta_{11}^{\acute{o}}(T^*-T_{\text{opt}}))}{1+\bar{\bar{U}}}) > 0$.

Proof

After the linearization, taking the Laplace transform of both sides of system (14)–(15), the Jacobian matrix for system (14)–(15) simulated at $\bar{\bar{E}}$ is given by

$$M_{11} = \begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{bmatrix},$$

where

$$a_{11} = g_0^{\sigma}\left(\exp\left(-b\left(\frac{T^*-T_{\text{opt}}}{T_{\max}-T_{\text{opt}}}\right)\right) + \frac{\bar{\bar{U}}-(\beta_{10}^{\sigma}+\beta_{11}^{\sigma}(T^*-T_{\text{opt}}))}{1+\bar{\bar{U}}}\right),$$

$a_{21} = -\delta_2\bar{\bar{U}}, a_{22} = -\gamma_1\beta^{(T^*-T_0)}$. The eigenvalues associated with the matrix M_{11} are $\lambda_1 = g_0^{\acute{o}}(\exp(-b(\frac{T^*-T_{\text{opt}}}{T_{\max}-T_{\text{opt}}})) + \frac{\bar{\bar{U}}-(\beta_{10}^{\acute{o}}+\beta_{11}^{\acute{o}}(T^*-T_{\text{opt}}))}{1+\bar{\bar{U}}}), \lambda_2 = -\gamma_1\beta^{(T^*-T_0)}$.

The eigenvalue λ_1 is negative if $g_0^{\acute{o}}(\exp(-b(\frac{T^* - T_{\text{opt}}}{T_{\max} - T_{\text{opt}}})) + \frac{\overline{\overline{U}} - (\beta_{10}^{\acute{o}} + \beta_{11}^{\acute{o}}(T^* - T_{\text{opt}}))}{1 + \overline{\overline{U}}}) < 0$ and positive if $g_0^{\acute{o}}(\exp(-b(\frac{T^* - T_{\text{opt}}}{T_{\max} - T_{\text{opt}}})) + \frac{\overline{\overline{U}} - (\beta_{10}^{\acute{o}} + \beta_{11}^{\acute{o}}(T^* - T_{\text{opt}}))}{1 + \overline{\overline{U}}}) > 0$.

The other eigenvalue λ_2 is negative. Hence the required results are obtained.

Lemma 4

The given equilibrium point E** *of the fractional-order system* (14)–(15) *is always locally asymptotically stable.*

Proof

The Jacobian matrix of system (14)–(15) with respect to E** is

$$M_{22} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix},$$

where $b_{11} = \frac{-g_0^{\acute{o}} N^{**}}{\gamma_0}$, $b_{12} = g_0^{\acute{o}} N^{**}(\frac{1 + \beta_{10}^{\acute{o}} + \beta_{11}^{\acute{o}}(T^* - T_{\text{opt}})}{(1 + U^{**})^2}), b_{21} = -\delta_2^{\acute{o}} U^{**}, b_{22} = -\gamma_1 \beta^{(T^* - T_0)} - \delta_2^{\acute{o}} N^{**}$.

The behavior of the eigenvalues is estimated by using Hurwitz's criteria in the quadratic equation

$$\begin{aligned} &\lambda^2 + \lambda\left(\frac{-g_0^\sigma N^{**}}{\gamma_0} + \gamma_1 \beta^{(T^* - T_0)} - \delta_2^\sigma N^{**}\right) \\ &\quad + \left(\delta_2^\sigma U^{**} g_0^\sigma N^{**}\left(\frac{1 + \beta_{10}^\sigma + \beta_{11}^\sigma (T^* - T_{\text{opt}})}{(1 + U^{**})^2}\right) + \frac{g_0^\sigma N^{**}}{\gamma_0} \gamma_1 \beta^{(T^{**} - T_0)} + \frac{g_0^\sigma \delta_2^\sigma N^{**^2}}{\gamma_0}\right) \\ &= 0. \end{aligned} \tag{19}$$

Using Hurwitz's criteria, we observe that the eigenvalues λ_1, λ_2 of the matrix M_{22} are negative if $T^* > T_{\text{opt}}$. Thus we get that E** is locally asymptotically stable under the restriction $T^* > T_{\text{opt}}$.

Now for the deformation of fractional-order system (2), we convert it to an equivalent compact form in the case of singular kernels as follows:

$$\begin{cases} {}^{C}D_t^{\sigma,\varkappa}N(t) = \mathcal{Q}_1(t,N), \\ {}^{C}D_t^{\sigma,\varkappa}T(t) = \mathcal{Q}_2(t,T), \\ {}^{C}D_t^{\sigma,\varkappa}C(t) = \mathcal{Q}_3(t,C), \\ {}^{C}D_t^{\sigma,\varkappa}Z(t) = \mathcal{Q}_4(t,Z), \\ {}^{C}D_t^{\sigma,\varkappa}U(t) = \mathcal{Q}_5(t,U). \end{cases} \tag{20}$$

Here $\mathcal{Q}_1, \mathcal{Q}_2, \mathcal{Q}_3, \mathcal{Q}_4, \mathcal{Q}_5$ are the proposed kernels with respect to the given classes N, T, C, Z, U, respectively.

FRACTIONAL-ORDER ANALYSIS ON THE PROPOSED MODEL

Analysis of the Existence and Uniqueness of the Solution

Proving the existence of the solution for fractional-order systems is always a sensitive part because not all fractional differential equations have their proof of the existence of a solution. In this area a number of works have been done, and lots of researchers work. Here, before deriving the solution of the proposed model, we first prove that the given fractional-order model has a unique solution. We give the results only for the class N(ζ), and the results are as for the other model classes. So we recall the model equation for *N*,

$${}^{C}D_\zeta^{\sigma,\varkappa}N(\zeta) = \mathcal{Q}_1(\zeta,N), \tag{21a}$$

$$N(0) = N_0, \tag{21b}$$

and the relative Volterra integral equation

$$N(\zeta) = N(0) + \frac{\varkappa^{1-\sigma}}{\Gamma(\sigma)} \int_0^\zeta \xi^{\varkappa-1}\left(\zeta^\varkappa - \xi^\varkappa\right)^{\sigma-1} \mathcal{Q}_1(\xi,N)\, d\xi. \tag{22}$$

Theorem 1 ([46] (Existence))

Let $0 < \sigma \le 1, N_0 \in \mathbb{R}, K > 0,$ *and* $T^* > 0$. *Define* $\mathcal{Q} := \{(\zeta,N) : \zeta \in [0,T^*], |N - N_0| \le K\},$ *and let the mapping* $\mathcal{Q}_1 : \mathcal{Q} \to \mathbb{R}$ *be continuous. Further, define* $M := \sup_{(\zeta,N)\in\mathcal{Q}} |\mathcal{Q}_1(\zeta,N)|$ *and*

$$T = \begin{cases} T^* & \text{if } M = 0, \\ \min\{T^*, (\frac{K\Gamma(\sigma+1)\rho^\sigma}{M})^{\frac{1}{\sigma}}\} & \text{otherwise.} \end{cases} \tag{23}$$

Then the IVP (21a)–(21b) *has a solution* N ∈ C[0, T]].

Lemma 5 ([46])

By considering the result of Theorem 1*a function* N ∈ C[0, T] *solves the IVP* (21a)–(21b) *if and only if it solves the Volterra integral equation* (22).

Theorem 2 ([46] (Uniqueness))

Let $N(0) \in \mathbb{R}, K > 0,$ *and* $T^* > 0$. *Further, let* $0 < \sigma \leq 1$ *and* $m = \lceil \sigma \rceil$. *For the set* $\mathcal{Q}$ *given in Theorem* 1, *let* $\mathcal{Q}_1 : \mathcal{Q} \to \mathbb{R}$ *be a continuous function that satisfies the Lipschitz condition with respect to the second variable*, *that is*,

$$\left|\mathcal{Q}_1(\zeta, N_1) - \mathcal{Q}_1(\zeta, N_2)\right| \leq V|N_1 - N_2|$$

with a constant V > 0 *independent of* ζ, N_1, *and* N_2. *Then the IVP* (21a)–(21b) *has a unique solution* N ∈ C[0, T].

Numerical Solution of the Proposed Model with Application of the Generalized Predictor–Corrector Technique

In the last few years, a number of fractional-order numerical schemes have been proposed by the scientists to solve various types of dynamical models. Very recently, the authors of [47] have proposed a new numerical method in the generalized Caputo derivative sense. Here we solve the proposed model with the help of generalized P-C scheme for the solution of the IVP (21a)–(21b) by following the methodology proposed in [44]. Also, we will analyze the stability of the given scheme. In that way, we first recall the above given Volterra integral equation (22), which gives

$$N(\zeta) = N(0) + \frac{\varkappa^{1-\sigma}}{\Gamma(\sigma)} \int_0^\zeta \xi^{\varkappa-1}\left(\zeta^\varkappa - \xi^\varkappa\right)^{\sigma-1} \mathcal{Q}_1(\xi, N)\, d\xi. \tag{24}$$

Now with supposing that a unique solution exists for the function $\mathcal{Q}_1$ on the interval [0, T], we divide the adopted interval [0, T] into *N* unequal subparts $\{[\zeta_k, \zeta_{k+1}], k = 0, 1, \ldots, N - 1\}$ using the mesh points

$$\begin{cases} \zeta_0 = 0, \\ \zeta_{k+1} = (\zeta_k^\varkappa + h)^{1/\varkappa}, \quad k = 0, 1, \ldots, \mathbb{N} - 1, \end{cases} \tag{25}$$

where $h = \frac{T^\varkappa}{\mathbb{N}}$. Now let us try to analyze the approximations $S_k, k = 0, 1, \ldots, \mathbb{N}$, to get a numerical solution of the given IVP. Suppose that we have already derived the approximations $N_j \approx N(\zeta_j)\ (j = 1, 2, \ldots, k)$ and want to derive approximations $N_{k+1} \approx N(\zeta_{k+1})$ by means of the integral equation

$$N(\zeta_{k+1}) = N(0) + \frac{\varkappa^{1-\sigma}}{\Gamma(\sigma)} \int_0^{\zeta_{k+1}} \xi^{\varkappa-1}\left(\zeta_{k+1}^{\varkappa} - \xi^{\varkappa}\right)^{\sigma-1} \mathcal{Q}_1(\xi, N)\, d\xi. \tag{26}$$

By substitution z = $\xi^{\varkappa}$ we get

$$N(\zeta_{k+1}) = N(0) + \frac{\varkappa^{-\sigma}}{\Gamma(\sigma)} \int_0^{\zeta_{k+1}^{\varkappa}} \left(\zeta_{k+1}^{\varkappa} - z\right)^{\sigma-1} \mathcal{Q}_1\left(z^{1/\varkappa}, N\left(z^{1/\varkappa}\right)\right) dz, \tag{27}$$

that is,

$$N(\zeta_{k+1}) = N(0) + \frac{\varkappa^{-\sigma}}{\Gamma(\sigma)} \sum_{j=0}^{k} \int_{\zeta_j^{\varkappa}}^{\zeta_{k+1}^{\varkappa}} \left(\zeta_{k+1}^{\varkappa} - z\right)^{\sigma-1} \mathcal{Q}_1\left(z^{1/\varkappa}, N\left(z^{1/\varkappa}\right)\right) dz. \tag{28}$$

Now, to simulate the right-side of Eq. (28), applying the trapezoidal quadrature rule with respect to the weight function $(\zeta_{k+1}^{\varkappa} - z)^{\sigma-1}$ and shifting the function $G_1(z^{1/\varkappa}, N(z^{1/\varkappa}))$ by its piecewise linear interpolant with nodes $\zeta_j^{\varkappa}$ $(j = 0, 1, \ldots, k+1),$ we get

$$\begin{aligned}
&\int_{\zeta_j^{\varkappa}}^{\zeta_{k+1}^{\varkappa}} (\zeta_{k+1}^{\varkappa} - z)^{\sigma-1} \mathcal{Q}_1(z^{1/\varkappa}, N(z^{1/\varkappa}))\, dz \\
&\quad \approx \frac{h^{\sigma}}{\sigma(\sigma+1)} [((k-j)^{\sigma+1} - (k-j-\sigma)(k-j+1)^{\sigma}) \\
&\quad \times G_1(\zeta_j, N(\zeta_j)) + ((k-j+1)^{\sigma+1} - (k-j+\sigma+1)(k-j)^{\sigma}) \mathcal{Q}_1(\zeta_{j+1}, N(\zeta_{j+1}))].
\end{aligned} \tag{29}$$

So, fitting the above-proposed approximations in Eq. (28), we establish the corrector term for $N(\zeta_{k+1}), k = 0, 1, \ldots, \mathbb{N} - 1$:

$$N(\zeta_{k+1}) \approx N(0) + \frac{\varkappa^{-\sigma} h^{\sigma}}{\Gamma(\sigma+2)} \sum_{j=0}^{k} a_{j,k+1} \mathcal{Q}_1(\zeta_j, N(\zeta_j)) + \frac{\varkappa^{-\sigma} h^{\sigma}}{\Gamma(\sigma+2)} \mathcal{Q}_1(\zeta_{k+1}, N(\zeta_{k+1})), \tag{30}$$

where

$$a_{j,k+1} = \begin{cases} k^{\sigma+1} - (k-\sigma)(k+1)^{\sigma} & \text{if } j = 0, \\ (k-j+2)^{\sigma+1} + (k-j)^{\sigma+1} - 2(k-j+1)^{\sigma+1} & \text{if } 1 \le j \le k. \end{cases} \tag{31}$$

The final task for our solution is changing the quantity $N(\zeta_{k+1})$ on the right-hand side of formula (30) with the predictor value $N^{P}(\zeta_{k+1}),$ which can be calculated by applying the one-step Adams–Bashforth technique to the integral equation (27). In this case, by changing the mapping $\mathcal{Q}_1(z^{1/\varkappa}, N(z^{1/\varkappa}))$ by the quantity $\mathcal{Q}_1(\zeta_j, N(\zeta_j))$ at each integral in Eq. (28) we get

$$
\begin{aligned}
N^P(\zeta_{k+1}) &\approx N(0) + \frac{\varkappa^{-\sigma}}{\Gamma(\sigma)} \sum_{j=0}^{k} \int_{\zeta_j^{\varkappa}}^{\zeta_{j+1}^{\varkappa}} (\zeta_{k+1}^{\varkappa} - z)^{\sigma-1} \mathcal{Q}_1(\zeta_j, N(\zeta_j))\, dz \\
&= N(0) + \frac{\rho^{-\sigma} h^{\sigma}}{\Gamma(\sigma+1)} \sum_{j=0}^{k} [(k+1-j)^{\sigma} - (k-j)^{\sigma}] \mathcal{Q}_1(\zeta_j, N(\zeta_j)).
\end{aligned}
\tag{32}
$$

So our P-C method for deriving the approximations $N_{k+1} \approx N(\zeta_{k+1})$ is totally evaluated by the formula

$$
N_{k+1} \approx N(0) + \frac{\varkappa^{-\sigma} h^{\sigma}}{\Gamma(\sigma+2)} \sum_{j=0}^{k} a_{j,k+1} \mathcal{Q}_1(\zeta_j, N_j) + \frac{\varkappa^{-\sigma} h^{\sigma}}{\Gamma(\sigma+2)} \mathcal{Q}_1(\zeta_{k+1}, N_{k+1}^P),
\tag{33}
$$

where $N_j \approx N(\zeta_j), j = 0, 1, \ldots, k,$ and the predicted value $N_{k+1}^P \approx N^P(\zeta_{k+1})$ can be simulated as mentioned in Eq. (32) with the terms $a_{j,k+1}$ estimated according to (31).

Therefore the derivation for the approximate solution of the proposed system (2) is derived successfully and defined by the following equations:

$$
\begin{aligned}
N_{k+1} &\approx N(0) + \frac{\varkappa^{-\sigma} h^{\sigma}}{\Gamma(\sigma+2)} \sum_{j=0}^{k} a_{j,k+1} \mathcal{Q}_1(\zeta_j, N_j) + \frac{\varkappa^{-\sigma} h^{\sigma}}{\Gamma(\sigma+2)} \mathcal{Q}_1(\zeta_{k+1}, N_{k+1}^P), \\
T_{k+1} &\approx T(0) + \frac{\varkappa^{-\sigma} h^{\sigma}}{\Gamma(\sigma+2)} \sum_{j=0}^{k} a_{j,k+1} \mathcal{Q}_2(\zeta_j, T_j) + \frac{\varkappa^{-\sigma} h^{\sigma}}{\Gamma(\sigma+2)} \mathcal{Q}_2(\zeta_{k+1}, T_{k+1}^P), \\
C_{k+1} &\approx C(0) + \frac{\varkappa^{-\sigma} h^{\sigma}}{\Gamma(\sigma+2)} \sum_{j=0}^{k} a_{j,k+1} \mathcal{Q}_3(\zeta_j, C_j) + \frac{\varkappa^{-\sigma} h^{\sigma}}{\Gamma(\sigma+2)} \mathcal{Q}_3(\zeta_{k+1}, C_{k+1}^P), \\
Z_{k+1} &\approx Z(0) + \frac{\varkappa^{-\sigma} h^{\sigma}}{\Gamma(\sigma+2)} \sum_{j=0}^{k} a_{j,k+1} \mathcal{Q}_4(\zeta_j, Z_j) + \frac{\varkappa^{-\sigma} h^{\sigma}}{\Gamma(\sigma+2)} \mathcal{Q}_4(\zeta_{k+1}, Z_{k+1}^P), \\
U_{k+1} &\approx U(0) + \frac{\varkappa^{-\sigma} h^{\sigma}}{\Gamma(\sigma+2)} \sum_{j=0}^{k} a_{j,k+1} \mathcal{Q}_5(\zeta_j, U_j) + \frac{\varkappa^{-\sigma} h^{\sigma}}{\Gamma(\sigma+2)} \mathcal{Q}_5(\zeta_{k+1}, U_{k+1}^P),
\end{aligned}
\tag{34}
$$

where

$$
\begin{aligned}
N^P(\zeta_{k+1}) &\approx N(0) + \frac{\varkappa^{-\sigma} h^{\sigma}}{\Gamma(\sigma+1)} \sum_{j=0}^{k} [(k+1-j)^{\sigma} - (k-j)^{\sigma}] \mathcal{Q}_1(\zeta_j, N(\zeta_j)), \\
T^P(\zeta_{k+1}) &\approx T(0) + \frac{\varkappa^{-\sigma} h^{\sigma}}{\Gamma(\sigma+1)} \sum_{j=0}^{k} [(k+1-j)^{\sigma} - (k-j)^{\sigma}] \mathcal{Q}_2(\zeta_j, T(\zeta_j)), \\
C^P(\zeta_{k+1}) &\approx C(0) + \frac{\varkappa^{-\sigma} h^{\sigma}}{\Gamma(\sigma+1)} \sum_{j=0}^{k} [(k+1-j)^{\sigma} - (k-j)^{\sigma}] \mathcal{Q}_3(\zeta_j, C(\zeta_j)), \\
Z^P(\zeta_{k+1}) &\approx Z(0) + \frac{\varkappa^{-\sigma} h^{\sigma}}{\Gamma(\sigma+1)} \sum_{j=0}^{k} [(k+1-j)^{\sigma} - (k-j)^{\sigma}] \mathcal{Q}_4(\zeta_j, Z(\zeta_j)), \\
U^P(\zeta_{k+1}) &\approx U(0) + \frac{\varkappa^{-\sigma} h^{\sigma}}{\Gamma(\sigma+1)} \sum_{j=0}^{k} [(k+1-j)^{\sigma} - (k-j)^{\sigma}] \mathcal{Q}_5(\zeta_j, U(\zeta_j)).
\end{aligned}
\tag{35}
$$

Method Stability

Theorem 3

Let $\mathcal{Q}_1(\zeta, N), \mathcal{Q}_2(\zeta, T), \mathcal{Q}_3(\zeta, C), \mathcal{Q}_4(\zeta, Z), \mathcal{Q}_5(\zeta, U)$ *satisfy the Lipschitz condition, and let* $N_j, T_j, C_j, Z_j, U_j \; (j = 1, \ldots, k+1)$ *be approximate solutions of the derived P-C method* (34) *and* (35), *respectively. Then the proposed numerical algorithm* (34)–(35) *is conditionally stable.*

Proof

Let $\tilde{N}_0, \tilde{N}_j (j = 0, \ldots, k+1),$ and $\widetilde{N^{P}_{k+1}} \; (k = 0, \ldots, \mathcal{N} - 1)$ be perturbations of N_0, N_j, and N^P_{k+1}, respectively. Then the given below perturbation equations are estimated with the help of Eqs. (34) and (35).

$$\widetilde{N^{P}_{k+1}} = \tilde{N}_0 + \frac{\varkappa^{-\sigma} h^{\sigma}}{\Gamma(\sigma+1)} \sum_{j=0}^{k} b_{j,k+1} \big(\mathcal{Q}_1(\zeta_j, N_j + \tilde{N}_j) - \mathcal{Q}_1(\zeta_j, N_j)\big), \tag{36}$$

where $b_{j,k+1} = [(k+1-j)^{\sigma} - (k-j)^{\sigma}],$

$$\widetilde{N_{k+1}} = \tilde{N}_0 + \frac{\varkappa^{-\sigma} h^{\sigma}}{\Gamma(\sigma+2)} \big(\mathcal{Q}_1(\zeta_{k+1}, N^P_{k+1} + \widetilde{N^{P}_{k+1}}) - \mathcal{Q}_1(\zeta_{k+1}, N^P_{k+1})\big) + \frac{\varkappa^{-\sigma} h^{\sigma}}{\Gamma(\sigma+2)} \times \sum_{j=0}^{k} a_{j,k+1} \big(\mathcal{Q}_1(\zeta_j, N_j + \tilde{N}_j) - \mathcal{Q}_1(\zeta_j, N_j)\big). \tag{37}$$

Using the Lipschitz condition, we obtain

$$|\widetilde{N_{k+1}}| \leq \zeta_0 + \frac{\varkappa^{-\sigma} h^{\sigma} m_1}{\Gamma(\sigma+2)} \left(|\widetilde{N^{P}_{k+1}}| + \sum_{j=1}^{k} a_{j,k+1} |\tilde{N}_j| \right), \tag{38}$$

where $\zeta_0 = \max_{0 \leq k \leq \mathcal{N}} \{ |\tilde{N}_0| + \frac{\varkappa^{-\sigma} h^{\sigma} m_1 a_{k,0}}{\Gamma(\sigma+2)} |\tilde{N}_0| \}$. Also, from Eq. (3.18) in [45] we derive

$$\left|\widetilde{N^{P}_{k+1}}\right| \leq \eta_0 + \frac{\varkappa^{-\sigma} h^{\sigma} m_1}{\Gamma(\sigma+1)} \sum_{j=1}^{k} b_{j,k+1} |\tilde{N}_j|, \tag{39}$$

where $\eta_0 = \max_{0 \leq k \leq N} \{ |\tilde{N}_0| + \frac{\varkappa^{-\sigma} h^{\sigma} m_1 b_{k,0}}{\Gamma(\sigma+1)} |\tilde{N}_0| \}$. Substituting $|\widetilde{N^{P}_{k+1}}|$ from Eq. (39) into Eq. (38) results in

$$|\tilde{N_{k+1}}| \le \sigma_0 + \frac{\varkappa^{-\sigma} h^{\sigma} m_1}{\Gamma(\sigma+2)}\left(\frac{\varkappa^{-\sigma} h^{\sigma} m_1}{\Gamma(\sigma+1)}\sum_{j=1}^{k} b_{j,k+1}|\tilde{N_j}| + \sum_{j=1}^{k} a_{j,k+1}|\tilde{N_j}|\right), \tag{40}$$

$$\le \sigma_0 + \frac{\varkappa^{-\sigma} h^{\sigma} m_1}{\Gamma(\sigma+2)}\sum_{j=1}^{k}\left(\frac{\varkappa^{-\sigma} h^{\sigma} m_1}{\Gamma(\sigma+1)} b_{j,k+1} + a_{j,k+1}\right)|\tilde{N_j}|, \tag{41}$$

$$\le \sigma_0 + \frac{\varkappa^{-\sigma} h^{\sigma} m_1 C_{\sigma,2}}{\Gamma(\sigma+2)}\sum_{j=1}^{k}(k+1-j)^{\sigma-1}|\tilde{N_j}|, \tag{42}$$

where $\sigma_0 = \max\{\zeta_0 + \frac{\varkappa^{-\sigma} h^{\sigma} m_1 a_{k+1,k+1}}{\Gamma(\sigma+2)}\eta_0\},\ C_{\sigma,2}$ is a positive constant depending on σ (by Lemma 1), and h is assumed to be small enough. Using Lemma 2, we have $|\tilde{N_{k+1}}| \le C\sigma_0$. which concludes the proof.

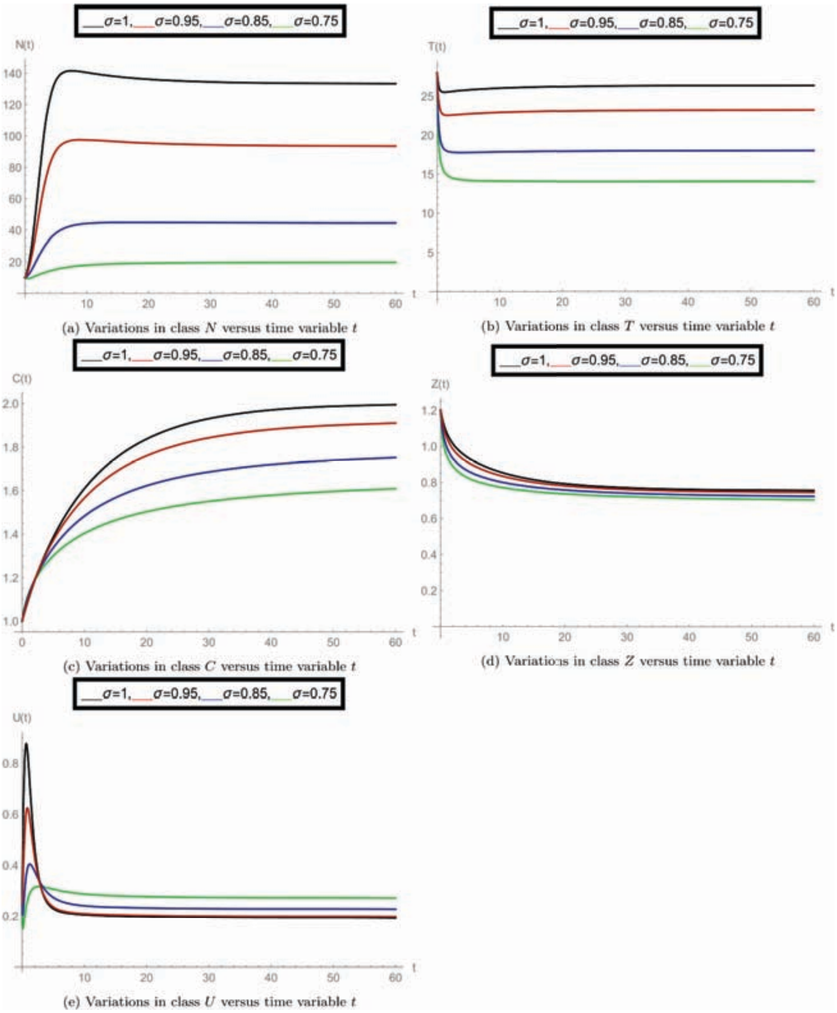
EXPERIMENTAL SIMULATIONS

After finishing all necessary theoretical analysis, we start to perform some experimental calculations to show the correctness of our results. We use *Mathematica* software for performing the number of graphs. For the case of interior equilibrium point $E^*(U^*, Z^*, C^*, T^*, N^*)$, we use the following set of parameter values:

$g_0 = 0.9$, $\beta_{10} = 0.03$, $\beta_{11} = 0.001$, $T_{opt} = 24$, $\gamma_0 = 150$, $D_{s0} = 4$, $A_0 = 0.2$, $w = 2.10$, $\zeta_1 = 3.5$, $O_c = 1.10$, $\Lambda_1 = 0.66$, $\Lambda = 0.4$, $\gamma_1 = 2$, $\delta_2 = 0.2$, $\zeta = 0.0019$, $\beta = 1.024$, $\gamma = 4$, $z_0 = 10.10$, $T_{10} = 14.50$, $\delta_1 = 0.1$, $T_0 = 20$, $b = 1.30$, $T_{max} = 35$, $N(0) = 10$, $T(0) = 28$, $C(0) = 1$, $Z(0) = 1.2$, $U(0) = 0.25$. Here we observe that for the case of fractional order $\sigma = 1$ (when the model behaves like an integer-order system), the authors of [1] have calculated the value of the interior equilibrium point $E^*(0.1023, 0.7534, 2.0000, 26.3818, 122.7088)$ and then, in this case, have specified the constraints for the solution boundedness, equilibrium point E^* stability, and positivity of the solution. Our target is to explore the dynamics of all model classes with respect to the interior equilibrium points at different fractional-order values σ.

4In the set of Fig. 2, we observed the nature of all model classes separately at different fractional-order values σ. In subfigure 2(a) the dynamics of density of aquatic population N is plotted at $\sigma = 1, 0.95, 0.85, 0.75$. Here we observed that at $\sigma = 1$ the numerically calculated equilibrium point is satisfied for class N and also at other values of order σ, it changed simultaneously.

Similarly, subfigure 2(b) shows the average water temperature of the species (class *T*), subfigure 2(c) shows the concentration of greenhouse gases (class *C*), subfigure 2(d) shows the ozone concentration (class *Z*), and subfigure 2(e) shows the dynamics of dissolved oxygen concentration (class *U*). The simultaneous changes in the given model classes at particular values of σ can be seen from the set of Fig. 3. Overall, we observed that when the fractional-order σ changes, the dynamics of the model, along with interior equilibrium point changes, justifies the importance of the fractional-order model.



(a) Variations in class *N* versus time variable *t*

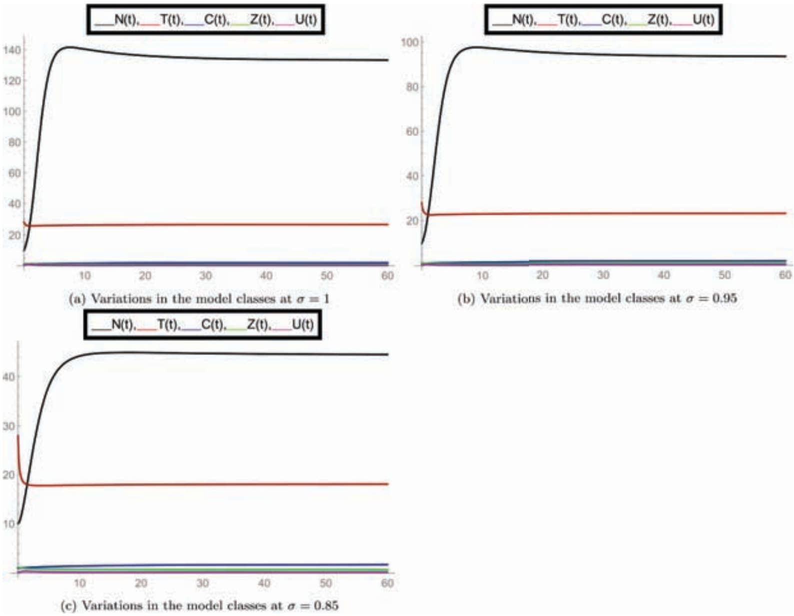
(b) Variations in class *T* versus time variable *t*

(c) Variations in class *C* versus time variable *t*

(d) Variations in class *Z* versus time variable *t*

(e) Variations in class *U* versus time variable *t*

Figure 2. Separate plots of all model classes at various fractional-order values σ for the case of interior equilibrium point $E^*(U^*, Z^*, C^*, T^*, N^*)$.



(a) Variations in the model classes at $\sigma = 1$

(b) Variations in the model classes at $\sigma = 0.95$

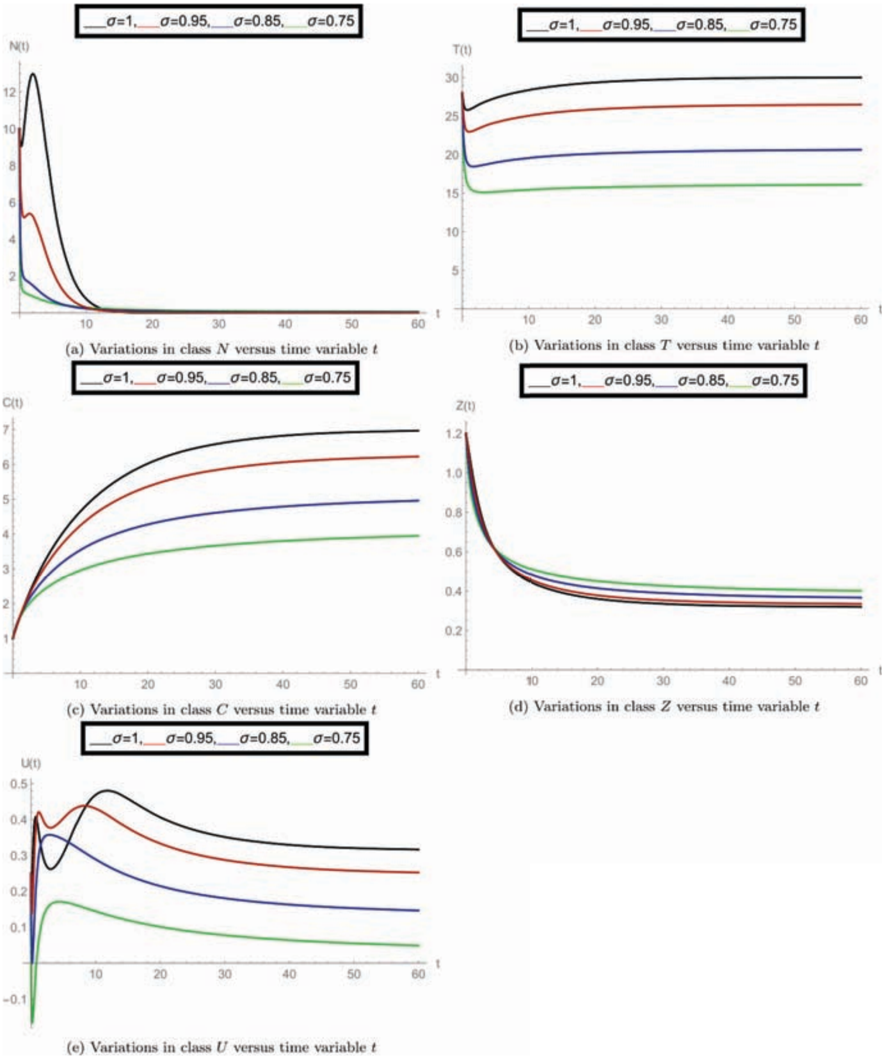
(c) Variations in the model classes at $\sigma = 0.85$

Figure 3. Mixed plots of all model classes at fractional-order values σ = 1, 0.95, 0.85 for the case of interior equilibrium point $E^*(U^*, Z^*, C^*, T^*, N^*)$.

As investigated above, now we consider the case of boundary equilibrium point $\bar{E} = (\bar{U}, \bar{Z}, \bar{C}, \bar{T}, \bar{N})$. In this case, we consider the following parameter values:

$g_0 = 0.9$, $\beta_{10} = 0.03$, $\beta_{11} = 0.50$, $T_{opt} = 24$, $\gamma_0 = 150$, $D_{s0} = 4$, $A_0 = 0.7$, w = 2.1710, $\zeta_1 = 3.5$, $O_c = 1.10$, $\Lambda_1 = 0.66$, $\Lambda = 0.4$, $\gamma_1 = 1$, $\delta_2 = 0.6$, $\zeta = 0.11$, $\beta = 1.024$, $\gamma = 4$, $z_0 = 10.10$, $T_{10} = 14.50$, $\delta_1 = 0.1$, $T_0 = 20$, b = 1.30, $T_{max} = 35$, N(0) = 10, T(0) = 28, C(0) = 1, Z(0) = 1.2, U(0) = 0.25. For the given parameter weights, the value of boundary equilibrium point $\bar{E} = (\bar{U}, \bar{Z}, \bar{C}, \bar{T}, \bar{N})$ at fractional-order σ = 1 (when the model behaves like an integer-order system given in [1]) is $\bar{E}(0.0474, 0.3179, 7.0, 30.0216, 0)$. In that integer-order case, the boundary equilibrium point is linearly asymptotically stable.

For the noninteger-order observations, in the set of Fig. 4, we analyzed the nature of proposed model classes separately at various fractional-order values *σ*. In subfigure 4(a), the dynamics of density of aquatic population *N* is plotted at σ = 1, 0.95, 0.85, 0.75. Here we can see that for σ = 1, the numerically calculated equilibrium point is satisfied for population *N* and that at other values of order *σ*, it changes simultaneously. Following the same way, subfigure 4(b) specifies the average water temperature of the species (class *T*), subfigure 4(c) demonstrates the concentration of greenhouse gases (class *C*), subfigure 4(d) shows the ozone concentration (class *Z*), and subfigure 4(e) shows the dynamics of dissolved oxygen concentration (class *U*).



(a) Variations in class *N* versus time variable *t*

(b) Variations in class *T* versus time variable *t*

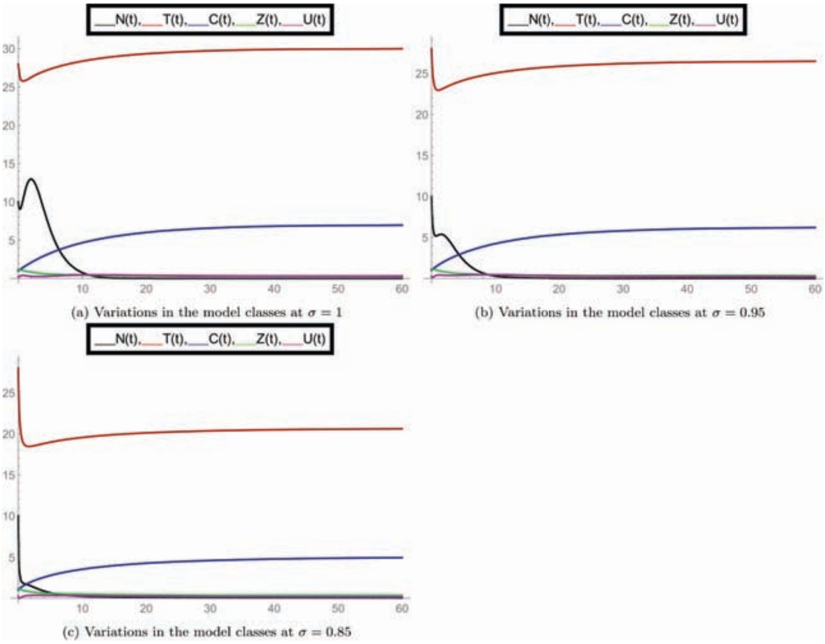
(c) Variations in class *C* versus time variable *t*

(d) Variations in class *Z* versus time variable *t*

(e) Variations in class *U* versus time variable *t*

Figure 4. Separate plots of all model classes at various fractional-order values σ for the case of boundary equilibrium point $\bar{E}=(\bar{U},\bar{Z},\bar{C},\bar{T},\bar{N})$.

The simultaneous changes in the given model classes at particular value of σ can be analyzed from the set of Fig. 5. Overall, we can see that when the fractional-order σ changes, the dynamics of the model changes along with boundary equilibrium point, which satisfies the role of fractional-order operator.



(a) Variations in the model classes at $\sigma = 1$

(b) Variations in the model classes at $\sigma = 0.95$

(c) Variations in the model classes at $\sigma = 0.85$

Figure 5. Mixed plots of all model classes at fractional-order values σ = 1, 0.95, 0.85 for the case of boundary equilibrium point $\bar{E} = (\bar{U}, \bar{Z}, \bar{C}, \bar{T}, \bar{N})$.

From the above given experimental analysis we see that the fractional-order dynamics with memory effects is much stronger than the integer-order dynamics. Here we have more varieties to understand the structure of the proposed ecosystem dynamics at various fractional-order values along with different values of equilibrium points. The modified Caputo fractional derivative is fully suitable to simulate the novel results with the help of given fractional-order model.

CONCLUSION

In our study, we have simulated a novel fractional-order mathematical system to study the prelude of deteriorating quality of water because of greenhouse gases on the population of aquatic animals. It has been shown in the given system that greenhouse gases raise the temperature of water, and because of this reason, the dissolved oxygen level goes down, and also the rate of circulation of disintegrated oxygen by the species rises, which causes a decrement in the density of aquatic species. We have used a new generalized Caputo-type fractional-order derivative to simulate the given dynamics. Equilibrium points for the given fractional model have been calculated, and important discussion on the asymptotic stability of the equilibria of a new autonomous system has been evaluated. We have reminded some important results to prove the existence of unique solution for the fractional-order cases. For finding the numerical solution of the given system, we used a generalized predictor–corrector algorithm in the sense of the new generalized Caputo derivative and also justified the stability of the technique. To prove the importance and correctness of the numerically simulated results, we have performed a number of graphs at different fractional-order values. The given derivative and algorithm work very well to understand the dynamics of the given model. From this study the effects of greenhouse gases and hypoxia on the population of aquatic species can be clearly understood with memory effects. For the future scope, the given ecosystem can be further solved by any other fractional-order derivatives. Also, some new mathematical models can be proposed to simulate the structure of given real-world problems.

ACKNOWLEDGEMENTS

The authors are thankful to the Taif University (supporting project number TURSP-2020/160), Taif, Saudi Arabia.

FUNDING

The research was funded by Taif University TURSP-2020/160.

AUTHORS’ CONTRIBUTIONS

PK: Investigation, conceptualization, formal analysis, methodology, resources, visualization, writing original draft. VG: Investigation, supervision, software, visualization, writing review and editing. VSE: Conceptualization, investigation, supervision, software, visualization, writing review and editing. MSM: Conceptualization, investigation, supervision, funding acquirements, software, visualization, writing review and editing. All authors read and approved the final manuscript.

REFERENCES

1. Chaturvedi, D., Misra, O.P.: Simultaneous effects of the rise in temperature due to greenhouse gases and hypoxia on the dynamics of the aquatic population: a mathematical model. J. Appl. Math. Comput. 63(1), 59–85 (2020)
2. Misra, O.P., Chaturvedi, D.: Fate of dissolved oxygen and survival of fish population in aquatic ecosystem with nutrient loading: a model. Model. Earth Syst. Environ. 2(3), 1–14 (2016)
3. Vaquer-Sunyer, R., Duarte, C.M.: Thresholds of hypoxia for marine biodiversity. Proc. Natl. Acad. Sci. 105(40), 15452–15457 (2008)
4. Vaquer-Sunyer, R., Duarte, C.M.: Temperature effects on oxygen thresholds for hypoxia in marine benthic organisms. Glob. Change Biol. 17(5), 1788–1797 (2011)
5. Sekerci, Y., Petrovskii, S.: Mathematical modelling of plankton–oxygen dynamics under the climate change. Bull. Math. Biol. 77(12), 2325–2353 (2015)
6. Sekerci, Y., Petrovskii, S.: Global warming can lead to depletion of oxygen by disrupting phytoplankton photosynthesis: a mathematical modelling approach. Geosciences 8(6), 201 (2018)
7. Caputo, M., Fabrizio, M.: A new definition of fractional derivative without singular kernel. Prog. Fract. Differ. Appl. 1(2), 1–13 (2015)
8. Kilbas, A., Srivastava, H.M., Trujillo, J.J.: Theory and Applications of Fractional Differential Equations. Elsevier, Amsterdam (2006)
9. Podlubny, I.: Fractional Differential Equations: An Introduction to Fractional Derivatives, Fractional Differential Equations, to Methods of Their Solution and Some of Their Applications. Elsevier, Amsterdam (1998)
10. Kumar, P., Suat Erturk, V.: The analysis of a time delay fractional COVID-19 model via Caputo type fractional derivative. Math. Methods Appl. Sci. (2020). https://doi.org/10.1002/mma.6935
11. Kumar, P., Erturk, V.S., Abboubakar, H., Nisar, K.S.: Prediction studies of the epidemic peak of coronavirus disease in Brazil via new generalised Caputo type fractional derivatives. Alex. Eng. J. 60(3), 3189–3204 (2021)
12. Kumar, P., Erturk, V.S., Murillo-Arcila, M., Banerjee, R., Manickam, A.: A case study of 2019-nCOV cases in Argentina with the real data based on daily cases from March 03, 2020 to March 29, 2021 using

classical and fractional derivatives. Adv. Differ. Equ. 2021(1), 1 (2021)

13. Gao, W., Veeresha, P., Baskonus, H.M., Prakasha, D.G., Kumar, P.: A new study of unreported cases of 2019-nCOV epidemic outbreaks. Chaos Solitons Fractals 138, 109929 (2020)
14. Nabi, K.N., Abboubakar, H., Kumar, P.: Forecasting of COVID-19 pandemic: from integer derivatives to fractional derivatives. Chaos Solitons Fractals 141, 110283 (2020)
15. Kumar, P., Erturk, V.S., Nisar, K.S., Jamshed, W., Mohamed, M.S.: Fractional dynamics of 2019-nCOV in Spain at different transmission rate with an idea of optimal control problem formulation. Alex. Eng. J. 61, 2204–2219 (2021)
16. Nabi, K.N., Kumar, P., Erturk, V.S.: Projections and fractional dynamics of COVID-19 with optimal control strategies. Chaos Solitons Fractals 145, 110689 (2021)
17. Kumar, P., Erturk, V.S.: A case study of Covid-19 epidemic in India via new generalised Caputo type fractional derivatives. Math. Methods Appl. Sci. (2021). https://doi.org/10.1002/mma.7284
18. Kumar, P., Erturk, V.S., Murillo-Arcila, M.: A new fractional mathematical modelling of COVID-19 with the availability of vaccine. Results Phys. 24, 104213 (2021)
19. Atangana, A.: A novel model for the lassa hemorrhagic fever: deathly disease for pregnant women. Neural Comput. Appl. 26(8), 1895–1903 (2015)
20. Kumar, P., Erturk, V.S., Yusuf, A., Sulaiman, T.A.: Lassa hemorrhagic fever model using new generalized Caputo-type fractional derivative operator. Int. J. Model. Simul. Sci. Comput. 12, 2150055 (2021)
21. Kumar, P., Erturk, V.S.: Environmental persistence influences infection dynamics for a butterfly pathogen via new generalised Caputo type fractional derivative. Chaos Solitons Fractals 144, 110672 (2021)
22. Abboubakar, H., Kumar, P., Erturk, V.S., Kumar, A.: A mathematical study of a tuberculosis model with fractional derivatives. Int. J. Model. Simul. Sci. Comput. 12, 2150037 (2021)
23. Abboubakar, H., Kumar, P., Rangaig, N.A., Kumar, S.: A malaria model with Caputo–Fabrizio and Atangana–Baleanu derivatives. Int. J. Model. Simul. Sci. Comput. 12, 2150013 (2020)
24. Kumar, P., Erturk, V.S., Almusawa, H.: Mathematical structure of mosaic disease using microbial biostimulants via Caputo and

Atangana–Baleanu derivatives. Results Phys. 24, 104186 (2021)

25. Agarwal, P., Singh, R.: Modelling of transmission dynamics of Nipah virus (Niv): a fractional order approach. Phys. A, Stat. Mech. Appl. 547, 124243 (2020)
26. Kumar, P., Erturk, V.S., Yusuf, A., Nisar, K.S., Abdelwahab, S.F.: A study on canine distemper virus (CDV) and rabies epidemics in the red fox population via fractional derivatives. Results Phys. 25, 104281 (2021)
27. Kumar, P., Erturk, V.S., Nisar, K.S.: Fractional dynamics of huanglongbing transmission within a citrus tree. Math. Methods Appl. Sci. 44, 11404–11424 (2021)
28. Kumar, P., Erturk, V.S., Yusuf, A., Kumar, S.: Fractional time-delay mathematical modeling of oncolytic virotherapy. Chaos Solitons Fractals 150, 111123 (2021)
29. Morales-Delgado, V.F., Gómez-Aguilar, J.F., Saad, K.M., Khan, M.A., Agarwal, P.: Analytic solution for oxygen diffusion from capillary to tissues involving external force effects: a fractional calculus approach. Phys. A, Stat. Mech. Appl. 523, 48–65 (2019)
30. Kumar, P., Erturk, V.S., Murillo-Arcila, M.: A complex fractional mathematical modeling for the love story of Layla and Majnun. Chaos Solitons Fractals 150, 111091 (2021)
31. Angstmann, C.N., Jacobs, B.A., Henry, B.I., Xu, Z.: Intrinsic discontinuities in solutions of evolution equations involving fractional Caputo–Fabrizio and Atangana–Baleanu operators. Mathematics 8(11), Article ID 2023 (2020)
32. Agarwal, P., Baleanu, D., Chen, Y., Momani, S., Machado, J.A.T.: Fractional calculus. In: ICFDA: International Workshop on Advanced Theory and Applications of Fractional Calculus. Amman (2019)
33. Agarwal, P., Agarwal, R.P., Ruzhansky, M. (eds.): Special Functions and Analysis of Differential Equations 1st edn. Chapman and Hall/CRC, London (2020)
34. Agarwal, P., Dragomir, S.S., Jleli, M., Samet, B. (eds.): Advances in Mathematical Inequalities and Applications Springer, Singapore (2018)
35. Ruzhansky, M., Cho, Y.J., Agarwal, P., Area, I. (eds.): Advances in Real and Complex Analysis with Applications Springer, Singapore (2017)

36. Agarwal, P., El-Sayed, A.A.: Non-standard finite difference and Chebyshev collocation methods for solving fractional diffusion equation. Phys. A, Stat. Mech. Appl. 500, 40–49 (2018)
37. Agarwal, P., Al-Mdallal, Q., Cho, Y.J., Jain, S.: Fractional differential equations for the generalized Mittag-Leffler function. Adv. Differ. Equ. 2018, Article ID 58 (2018)
38. Salahshour, S., Ahmadian, A., Senu, N., Baleanu, D., Agarwal, P.: On analytical solutions of the fractional differential equation with uncertainty: application to the Basset problem. Entropy 17(2), 885–902 (2015)
39. Alderremy, A.A., Saad, K.M., Agarwal, P., Aly, S., Jain, S.: Certain new models of the multi space-fractional Gardner equation. Phys. A, Stat. Mech. Appl. 545, 123806 (2020)
40. Shahid, A., Mohamed, M.S., Bhatti, M.M., Doranehgard, M.H.: Darcy–Brinkman–Forchheimer model for nano-bioconvection stratified MHD flow through an elastic surface: a successive relaxation approach. Mathematics 9(19), 2514 (2021)
41. Gepreel, K.A., Mahdy, A.M.S., Mohamed, M.S., Al-Amiri, A.: Reduced differential transform method for solving nonlinear biomathematics models. Comput. Mater. Continua 61(3), 979–994 (2019)
42. Mahdy, A.M.S., Mohamed, M.S., Gepreel, K.A., AL-Amiri, A., Higazy, M.: Dynamical characteristics and signal flow graph of nonlinear fractional smoking mathematical model. Chaos Solitons Fractals 141, 110308 (2020)
43. Khater, M.M., Mohamed, M.S., Park, C., Attia, R.A.: Effective computational schemes for a mathematical model of relativistic electrons arising in the laser thermonuclear fusion. Results Phys. 19, 103701 (2020)
44. Odibat, Z., Baleanu, D.: Numerical simulation of initial value problems with generalized Caputo-type fractional derivatives. Appl. Numer. Math. 156, 94–105 (2020)
45. Li, C., Zeng, F.: The finite difference methods for fractional ordinary differential equations. Numer. Funct. Anal. Optim. 34(2), 149–179 (2013)
46. Erturk, V.S., Kumar, P.: Solution of a COVID-19 model via new generalized Caputo-type fractional derivatives. Chaos Solitons Fractals 139, 110280 (2020)

47. Kumar, P., Erturk, V.S., Kumar, A.: A new technique to solve generalized Caputo type fractional differential equations with the example of computer virus model. J. Math. Ext. 15, 1–23 (2021)

INDEX

The Use of Mathematical Structures: Modelling Real Phenomena

The process of describing real-world problems as mathematical structures and abstract objects is often referred to as mathematical modelling. A mathematical model of a real-world problem consists of an approximate description in the form of a differential equations' system. Modern scientists and engineers strive to solve these equations to help them understand the origin and discover new features concerning a real-world problem. Then, once the physical interpretation of the mathematical solution is clearly captured, they often attempt to improve or extend these mathematical modelling approximations to more general situations by increasing the complexity of mathematical models. This book includes several articles devoted to the mathematical modelling of real-world problems in physics, mechanical engineering, biology, and biochemistry. It is divided into four thematic sections. Each section covers a different topic in mathematical modelling for describing and understanding physical phenomena.

The first part of this book (chapters 1 to 3) reflects on mathematical modelling from a more philosophical and generic point of view. Chapter 1 inquiries about the explanatory role of mathematics in empirical science. The author attempts to answer the question: "Are there genuine mathematical explanations of physical phenomena, and if so, how can mathematical theories, which are typically thought to concern abstract mathematical objects, explain contingent empirical matters?". Chapter 2 reflects on the application of quantum mathematical modelling in the fields of psychology, economics, and decision science. The author discusses whether quantum mathematical models are necessary for dealing with specific phenomena in the aforementioned fields, or whether the classical (probabilistic or statistical) mathematical models will suffice. Chapter 3 is focused on the application of stability theory (a fundamental part of mathematical modelling of natural phenomena) in the fields of chemistry and biology. The authors propose that both fields can be conceptually connected through the concept of stability.

The second part of this book (chapters 4 to 9) is devoted to the application of mathematical modelling to specific real-world problems in fluid dynamics and mechanical engineering. It is focused on solving ordinary differential equations (ODEs) used for describing a wide variety of phenomena in physics, biology, chemistry, and several other fields. Chapter 4 describes the application of the Riccati-Bernoulli sub-ODE method for finding exact travelling wave solutions, solitary wave solutions and peaked wave solutions of nonlinear partial differential equations, which play an important role in the study of nonlinear physical phenomena. Chapter 5 describes the mathematical modelling of mantle convection at a high Rayleigh number with variable viscosity and viscous dissipation. Mantle convection is responsible for numerous physical and chemical phenomena occurring on the surface and in the interior of the Earth. The study is important for shading light on the mechanism behind this type of convection, which remains an unsolved problem since the rheology of mantle rocks. Chapter 6 describes an application of a multi-component multiphase reactive transport model for geothermal reservoir simulation. Chapter 7 describes the modelling and dynamic characteristics of a non-metal pressurized reservoir with variable volume. A closed reservoir may provide an advantage of having a smaller volume when compared to open reservoirs which are large, heavy, polluted, and threaten the operation of hydraulic systems. Chapter 8 describes the modelling and natural characteristic analysis of cycloid ball transmission using lumped stiffness method. The study is motivated by the possibility of improving the dynamic precision of robot systems. Chapter 9 describes the modelling of flowslides and debris avalanches in natural and engineered slopes. The study aims to provide a better understanding of how slope instability is affected by the rainfall from the ground surface and water springs from a bedrock.

The third part of this book (chapters 10 to 13) is devoted to the application of mathematical modelling to the fields of biology and biochemistry. Chapter 10 is focused on the Lane-Emden boundary value problem, arising in numerous real-life chemical and biochemical phenomena. Chapter 11 aims to provide a mathematical analysis and modelling of a prey-predator system to describe the effect of predation between prey and predator with a nonlinear functional response. Chapter 12 describes the mathematical modelling of collective behaviour appearing at several levels of biological complexity, from single cells to super-organisms. Chapter 13 describes a fractional mathematical model for studying the effects of greenhouse gases and hypoxia on the population of aquatic species.

Olga Moreira is a Ph.D. and M.Sc. in Astrophysics and B.Sc. in Physics/Applied Mathematics (Astronomy). She is an experienced technical writer and data analyst. As a graduate student, she held two research grants to carry out her work in Astrophysics at two of the most renowned European institutions in the fields of Astrophysics and Space Science (the European Space Agency, and the European Southern Observatory). She is currently an independent scientist, peer-reviewer and editor. Her research interest is solar physics, machine learning and artificial neural networks.

AP | ARCLER
PRESS

ISBN 978-1-77469-697-2
00000
9 781774 696972