Ferdowsi University of Mashhad

Iranian Journal of Numerical Analysis and Optimization

Volume 16, Number 1

March 2026

Serial Number: 36

Ferdowsi University of Mashhad, Iran

In the Name of God

Iranian Journal of Numerical Analysis and Optimization (IJNAO)

This journal is authorized under the registration No. 174/853 dated 1386/2/26 (2007/05/16), by the Ministry of Culture and Islamic Guidance.

Volume 16, Number 1, March 2026

ISSN-Print: 2423-6977, **ISSN-Online:** 2423-6969

Publisher: Faculty of Mathematical Sciences, Ferdowsi University of Mashhad

Published by: Ferdowsi University of Mashhad Press

Printing Method: Electronic

Address: Iranian Journal of Numerical Analysis and Optimization

Faculty of Mathematical Sciences, Ferdowsi University of Mashhad

P.O. Box 1159, Mashhad 91775, Iran.

Tel. : +98-51-38806222 , **Fax:** +98-51-38807358

E-mail: ijnao@um.ac.ir

Website: http://ijnao.um.ac.ir

This journal is indexed by:

- SCOPUS
- ZbMATH Open
- ISC
- DOAJ
- Civilica
- Magiran
- Mendeley
- Academia.edu
- Linkedin

- The Journal granted the International degree by the Iranian Ministry of Science, Research, and Technology.

Iranian Journal of Numerical Analysis and Optimization

Volume 16, Number 1, March 2026

Ferdowsi University of Mashhad - Iran

©2026 All rights reserved. Iranian Journal of Numerical Analysis and Optimization

Iranian Journal of Numerical Analysis and Optimization

Director

M. H. Farahi

Editor-in-Chief

Ali R. Soheili

Managing Editor

M. Gachpazan

EDITORIAL BOARD

Abbasbandi, Saeid*
(Numerical Analysis)
Imam Khomeini International University, Iran.
e-mail: abbasbandy@ikiu.ac.ir

Abdi, Ali*
(Numerical Analysis)
University of Tabriz, Iran.
e-mail: a_abdi@tabrizu.ac.ir

Area, Iván*
(Numerical Analysis)
Universidade de Vigo, Spain.
e-mail: area@uvigo.es

Babaie Kafaki, Saman*
(Optimization)
Semnan University, Iran.
e-mail: sbk@semnan.ac.ir

Babolian, Esmail*
(Numerical Analysis)
Kharazmi University, Iran.
e-mail: babolian@khu.ac.ir

Cardone, Angelamaria*
(Numerical Analysis)
Università degli Studi di Salerno, Italy.
e-mail: ancardone@unisa.it

Dehghan, Mehdi*
(Numerical Analysis)
Amirkabir University of Technology, Iran.
e-mail: mdehghan@aut.ac.ir

Effati, Sohrab*
(Optimal Control & Optimization)
Ferdowsi University of Mashhad, Iran.
e-mail: s-effati@um.ac.ir

Emrouznejad, Ali*
(Operations Research)
Aston University, UK.
e-mail: a.emrouznejad@aston.ac.uk

Farahi, Mohammad Hadi*
(Optimal Control & Optimization)
Ferdowsi University of Mashhad, Iran.
e-mail: farahi@um.ac.ir

Gachpazan, Mortaza**
(Numerical Analysis)
Ferdowsi University of Mashhad, Iran.
e-mail: gachpazan@um.ac.ir

Ghanbari, Reza**
(Operations Research)
Ferdowsi University of Mashhad, Iran.
e-mail: rghanbari@um.ac.ir

Hadizadeh Yazdi, Mahmoud*
(Numerical Analysis)
Khaje-Nassir-Toosi University of
Technology, Iran.
e-mail: hadizadeh@kntu.ac.ir

Hojjati, Gholamreza*
(Numerical Analysis)
University of Tabriz, Iran.
e-mail: ghojjati@tabrizu.ac.ir

Hong, Jialin*
(Scientific Computing)
Chinese Academy of Sciences (CAS), China.
e-mail: hjl@lsec.cc.ac.cn

Karimi, Hamid Reza*
(Control)
Politecnico di Milano, Italy.
e-mail: hamidreza.karimi@polimi.it

Khojasteh Salkuyeh, Davod*
(Numerical Analysis)
University of Guilan, Iran.
e-mail: khojasteh@guilan.ac.ir

Lohmander, Peter*
(Optimization)
Swedish University of Agricultural Sciences, Sweden.
e-mail: Peter@Lohmander.com

Lopez-Ruiz, Ricardo*
(Complexity, nonlinear models)
University of Zaragoza, Spain.
e-mail: rilopez@unizar.es

Mahdavi-Amiri, Nezam*
(Optimization)
Sharif University of Technology, Iran.
e-mail: nezamm@sina.sharif.edu

Mirzaei, Davoud*
(Numerical Analysis)
University of Uppsala, Sweden.
e-mail: davoud.mirzaei@it.uu.se

Omrani, Khaled*
(Numerical Analysis)
University of Tunis El Manar, Tunisia.
khaled.omrani@issatso.rnu.tn

Salehi Fathabadi, Hasan*
(Operations Research)
University of Tehran, Iran.
e-mail: hsalehi@ut.ac.ir

Soheili, Ali Reza*
(Numerical Analysis)
Ferdowsi University of Mashhad, Iran.
e-mail: soheili@um.ac.ir

Soleimani Damaneh, Majid*
(Operations Research and Optimization, Finance, and
Machine Learning)
University of Tehran, Iran.
e-mail: m.soleimani.d@ut.ac.ir

Toutounian, Faezeh*
(Numerical Analysis)
Ferdowsi University of Mashhad, Iran.
e-mail: toutouni@um.ac.ir

Türkyılmazoğlu, Mustafa*
(Applied Mathematics)
Hacettepe University, Turkey.
e-mail: turkyilm@hacettepe.edu.tr

Vahidian Kamyad, Ali*
(Optimal Control & Optimization)
Ferdowsi University of Mashhad, Iran.
e-mail: vahidian@um.ac.ir

Xu, Zeshui*
(Decision Making)
Sichuan University, China.
e-mail: xuzeshui@263.net

Vasagh, Zohreh
(English Text Editor)
Ferdowsi University of Mashhad, Iran.

This journal is published under the auspices of Ferdowsi University of Mashhad

* Full Professor

** Associate Professor

We would like to acknowledge the help of Miss Narjes khatoon Zohorian in the preparation of this issue.

Letter from the Editor–in–Chief

I would like to welcome you to the Iranian Journal of Numerical Analysis and Optimization (IJNAO). This journal has been published two issues per year and supported by the Faculty of Mathematical Sciences at the Ferdowsi University of Mashhad. The faculty of Mathematical Sciences with the centers of excellence and the research centers is well-known in mathematical communities in Iran.

The main aim of the journal is to facilitate discussions and collaborations between specialists in applied mathematics, especially in the fields of numerical analysis and optimization, in the region and worldwide. Our vision is that scholars from different applied mathematical research disciplines pool their insight, knowledge, and efforts by communicating via this international journal. In order to assure the high quality of the journal, each article is reviewed by subject-qualified referees. Our expectations for IJNAO are as high as any well-known applied mathematical journal in the world. We trust that by publishing quality research and creative work, the possibility of more collaborations between researchers would be provided. We invite all applied mathematicians especially in the fields of numerical analysis and optimization to join us by submitting their original work to the Iranian Journal of Numerical Analysis and Optimization.

We would like to inform all readers that the Iranian Journal of Numerical Analysis and Optimization (IJNAO), has changed its publishing frequency from "Semiannual" to a "Quarterly" journal since January 2023. The four journal issues per year will be published in the months of March, June, September, and December. One of our goals is to continue to improve the speed of both the review and publication processes, while try continuing to publish the best available international research in numerical analysis and optimization, with the high scientific and publication standards that the journal is known for.

Ali R. Soheili

Editor-in-Chief

Contents

Unconditionally stable high-order scheme for the fractional generalized Burgers–Huxley equation 1
M. Hajipour, Y. Lakpour and D. Baleanu

A fractional order model for malaria control using optimal strategies and numerical simulations 19
A. Kouidere, A. El Bhih, I. Minifi and K. Adnaoui

Numerical exploration of pollutant transport using stochastic fractional diffusion and Karhunen–Loève expansion 38
Z. Moniri, A. Babaei and B. Parsa Moghaddam

A new numerical approach for tracing periodic solutions in nonlinear dynamical systems using CGM with PAC 61
K.F. Fazea

Fractional-order modeling and numerical analysis of double intermediate enzymatic reaction model using generalized Euler method 77
B. Koushick, N. Anbazhagan, R. Vignesh Raju, N. Jeeva

A novel three-term conjugate gradient approach for deep neural network training 102
F.S. Hosseini-Meighan and A. Ashrafi

Stability analysis of fast food consumption dynamics using a fractional-order framework 122
I. Sahib, M. Baroudi, H. Gourram, B. Khajji, A. Labzai and M. Belam

Measuring the efficiency of bi-level structured systems based on directional distance function and network DEA model 144
M. Habibpoor and M. Alirezaee

Qualitative and Numerical Analysis of a Time-Fractional Ebola Model in Population Dynamics 170
G. Saini, B. Ghosh and S. Chand

Transdermal drug delivery via a hollow microneedle considering the skin's mechanical properties: Analytical and numerical study 188
E. Azhdari and A. Emami

An improved search direction based on algebraic equivalent transformation technique for convex quadratic optimization 220
A. Kraria, B. Merikhi and D. Benterki

Quasilinearization method to solve several classes of nonlinear Lane–Emden equations by Gegenbauer polynomials 243
F. Sheikhi and B. Ghazanfari

An efficient hybrid three-term conjugate gradient method with practical applications 264
Y.E. Hemici, S. Khelladi and D. Benterki

Dynamics and applications of quantum iterative methods 283
S.K. Nayak, P.K. Parida and B. Mehta

Mathematical modeling and optimal control strategies of the typhoid transmission with cost-effectiveness analysis 304
K. Oubouskour, S. Khassal, B. Khajji and O. Balatif

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 1, 2026, pp 1–18
https://doi.org/10.22067/ijnao.2025.93568.1649
https://ijnao.um.ac.ir/

CC BY NC

Research Article

Unconditionally stable high-order scheme for the fractional generalized Burgers–Huxley equation

M. Hajipour*, Y. Lakpour and D. Baleanu

Abstract

This study presents a high-order numerical scheme for solving the generalized fractional-order Burgers–Huxley equation, a nonlinear evolutionary partial differential equation (PDE) combining integer-order spatial derivatives with temporal fractional derivatives. The model effectively captures key features of reaction–diffusion–convection systems and neural pulse propagation dynamics. To this end, we develop a fourth-order compact finite difference method for spatial discretization coupled with a Grünwald–Letnikov approximation for the fractional time derivative. The resulting nonlinear system is solved using an efficient iterative algorithm. A rigorous stability and convergence analysis shows that the proposed method is unconditionally stable and achieves $\mathcal{O}(\tau^3+h^4)$ convergence, where τ and h represent temporal and spatial step sizes, respectively. Numerical experiments validate the theoretical results and demonstrate the

*Corresponding author

Received 17 May 2025; revised 23 August 2025; accepted 18 August 2025

Mojtaba Hajipour
Department of Mathematics, Sahand University of Technology, P.O. Box: 51335-1996, Tabriz, Iran. e-mail: hajipour@sut.ac.ir

Yashar Lakpour
Department of Mathematics, Sahand University of Technology, P.O. Box: 51335-1996, Tabriz, Iran. e-mail: yl67mathteacher@gmail.com

Dumitru Baleanu
Department of Computer Science and Mathematics, Lebanese American University, Beirut, Lebanon. Institute of Space Sciences, Magurele, Romania. e-mail: dumitru.baleanu@lau.edu.lb

How to cite this article
Hajipour, M., Lakpour, Y. and Baleanu, D., Unconditionally stable high-order scheme for the fractional generalized Burgers–Huxley equation. *Iran. J. Numer. Anal. Optim.*, 2026; 16(1): 1-18. https://doi.org/10.22067/ijnao.2025.93568.1649

scheme's accuracy and robustness across multiple test cases, confirming its effectiveness for this class of fractional PDEs.

AMS subject classifications (2020): 65N06; 65N22; 34K28.

Keywords: Burgers–Huxley equation; Finite difference formula; High-order method, stability; Iterative method.

1 Introduction

Fractional calculus (FC) has established itself as a fundamental mathematical framework for modeling complex systems exhibiting memory effects and spatial nonlocality [5]. Fractional-order integrals and derivatives, which generalize their integer-order counterparts, provide a powerful framework for describing systems with memory effects and nonlocality [6, 2]. While FC offers powerful modeling capabilities, obtaining exact analytical solutions to fractional differential equations remains fundamentally challenging–indeed, often intractable. This persistent difficulty has motivated five decades of sustained research into numerical approximation techniques, particularly as FC applications have expanded across diverse domains including nonlinear wave phenomena, biological systems modeling, engineering dynamics, and transport processes. We introduce a novel discretization scheme specifically designed for boundary-value problems governed by the fractional generalized Burgers–Huxley (FGBH) equation:

$$\begin{cases} {}_0\mathcal{D}_t^{\alpha} w + \lambda\, w^{\theta} w_z - \sigma w_{zz} = \beta w(1-w^{\gamma})(w^{\gamma}-\eta) + H(z,t), & \text{(1a)} \\ w(z,0) = 0, \quad a \leqslant z \leqslant b, & \text{(1b)} \\ w(a,t) = \varphi_a(t), \quad w(b,t) = \varphi_b(t), \quad 0 \leqslant t \leqslant T, & \text{(1c)} \end{cases}$$

where the function $w = w(z,t)$ is unknown, and parameters $\lambda > 0, \theta > 0, \beta \geqslant 0, \gamma > 0, 0 < \eta < 1$ are arbitrary constants. Moreover, w_z and w_{zz} denote the first- and second-order derivatives of function w with respect to z, respectively. The term ${}_0\mathcal{D}_t^{\alpha} w$ signifies the Caputo fractional derivative of order α of w defined by

$${}_0\mathcal{D}_t^{\alpha} w(z,t) := \frac{1}{\Gamma(1-\alpha)} \int_0^t \frac{\partial w(z,\eta)}{\partial \eta} (t-\eta)^{-\alpha} d\eta, \tag{2}$$

where $0 < \alpha \leqslant 1$ and Γ denotes the Gamma function. For $\alpha = 1$, the FGBH equation reduces to the classical Burgers–Huxley equation (BHE), which can be modeled more accurately with its fractional version [11]. For $\lambda = 0, \theta = 1$ and $\sigma = 1$, (1) will be transformed into the

fractional Huxley equation, which explains nerve pulse propagation and wall motion in liquid crystals [20]. For $\beta = 0, \theta = 1$ and, $\sigma = 1$, the presented equation (1) will be transformed to time-fractional Burgers' equation, which describes the far field of wave propagation in nonlinear systems [20]. Moreover, the time-fractional Fitzhugh–Nagoma equation will be derived when $\lambda = 0, \beta = \theta = \sigma = 1$ [3]. If $\theta = 1$, then the FGBH equation reduces to the time-fractional BHE [19]. The inhomogeneous FGBH equation (1) is utilized for modeling diffusion transport, nerve pulse propagation in nerve fibers, the interaction between reaction mechanisms, convection effects, and motion in liquid crystals [13]. Since (1) is a nonlinear partial differential equation (PDE), there is no effective method to derive its exact solution.

Recent years have witnessed significant progress in developing numerical methods for various versions of fractional BHE. Hammouch and Mekkaoui [8] developed the Adomian decomposition method for solving time-fractional BHEs. Alquran et al. [1] adapted the residual power series method to obtain traveling wave solutions of the FGBH equation. Tripathi [18] proposed an efficient technique for time-space fractional BHEs. Inc et al. [10] implemented Lie symmetry analysis for the time-fractional version of (1). Freihet and Zuriqat [4] applied the residual fractional power series method to solve time-fractional BHEs. Inc [9] designed a hybrid method combining line and group preserving schemes. Singh et al. [15] developed a shifted Legendre collocation method for space-time fractional BHEs. Majeed et al. [12] presented a cubic B-spline collocation method for the FGBH equation (1). Yang and Liu [21] established a finite difference method for time-fractional BHEs. Hajipour and Lakpour [7] developed a highly accurate method for variable-order fractional BHEs. To the best of our knowledge, existing numerical approaches for solving the general form of the FGBH equation (1) exhibit significant limitations in handling solutions with initial singularities and maintaining high-order accuracy for the general equation form.

The main goal of this manuscript is to propose a novel numerical scheme for solving the fractional generalized FGBH equation. The proposed method combines a fourth-order compact finite difference (FD) discretization in space with a third-order weighted shifted Grünwald–Letnikov (WSGL) approximation in time, resulting in a fully discrete system of nonlinear algebraic equations. We develop an efficient nonlinear solver to handle the resulting systems and provide rigorous theoretical analysis, proving the method's solvability, stability, and convergence. Our analysis establishes that the scheme achieves fourth-order spatial accuracy and third-order temporal accuracy, confirming its high precision. Numerical experiments on three benchmark test cases validate the theoretical convergence rates and demonstrate the method's computational robustness.

The paper is organized as follows. Section 2 introduces a highly accurate numerical technique for solving the FGBH equation. We present a detailed analysis of the proposed method, including

its stability, solvability, and convergence properties. Section 3 demonstrates the applicability and accuracy of the method through several test examples. Section 4 concludes the paper, summarizing key findings and implications.

2 Description of the numerical method

Here, we will attempt to design a high-order fully discrete method for the inhomogeneous FGBH equation (1). In the first step, consider a reformulation of (1) as

$$-\sigma w_{zz} + \lambda\, w^{\theta} w_z = \Psi, \quad a < z < b,\ 0 < t \leqslant T, \tag{3}$$

where $\Psi = f + H - {}_0\mathcal{D}_t^{\alpha} w$ and $f = \beta w(1 - w^{\gamma})(w^{\gamma} - \eta)$. Then, we recall the SGL difference formula [22]

$$\mathcal{A}_{\tau,p}^{\alpha} w(z,t) := \frac{1}{\tau^{\alpha}} \sum_{k=0}^{[\frac{\varsigma}{\tau}]+p} \omega_k^{(\alpha)} w(z, t - (k-p)\tau), \tag{4}$$

in which $\tau > 0$, p is an integer, and the coefficients $\omega_k^{(\alpha)}$ are computed by a recursive formula

$$\omega_{k+1}^{(\alpha)} = (1 - \frac{1+\alpha}{1+k})\omega_k^{(\alpha)}, \qquad k = 0, 1, 2, \ldots, \tag{5}$$

where $\omega_0^{(\alpha)} = 1$. As shown in [17], if $w(z,0) = 0$, then ${}_0\mathcal{D}_t^{\alpha} w(z,t) = \mathcal{A}_{\tau,p}^{\alpha} w(z,t) + \mathcal{O}(\tau)$. In the next lemma of [17], a third-order WSGL formula is developed to approximate the Caputo fractional derivative of order α of w defined by (2).

Now, we discretize the solution domain $\Omega := [a,b] \times [0,T]$. For positive integers M and N, let $(z_i, t_n) := (a + ih, n\tau)$ with $0 \leqslant n \leqslant N+1, 0 \leqslant i \leqslant M+1$, be a uniform mesh on Ω of $h := (b-a)/(M+1)$ and $\tau := T/(N+1)$.

Lemma 1. Let $0 < \alpha < 1$ and $\frac{\partial^m w}{\partial t^m}|_{t=0} = 0$, for $m = 0, 1, 2, 3$. Then we deduce

$${}_0\mathcal{D}_t^{\alpha} w(z_i, t_n) = \tau^{-\alpha} \sum_{k=0}^{n} g_k^{(\alpha)} w(z_i, t_n - k\tau) + \mathcal{O}(\tau^3), \quad \text{as} \quad \tau \to 0, \tag{6}$$

uniformly for $n \geqslant 1$, where

$$\begin{cases} g_0^{(\alpha)} = \rho_0 \omega_0^{(\alpha)}, \\ g_1^{(\alpha)} = \rho_0 \omega_1^{(\alpha)} + \rho_1 \omega_0^{(\alpha)}, \\ g_k^{(\alpha)} = \rho_0 \omega_k^{(\alpha)} + \rho_1 \omega_{k-1}^{(\alpha)} + \rho_2 w_{k-2}^{(\alpha)}, \quad k = 2, 3, \ldots, \end{cases} \tag{7}$$

and $\rho_0 = 1 + \frac{17}{24}\alpha + \frac{1}{8}\alpha^2,\ \rho_1 = -\frac{11}{12}\alpha - \frac{1}{4}\alpha^2,\ \rho_2 = \frac{5}{24}\alpha + \frac{1}{8}\alpha^2.$

Lemma 2. [17] For $0 < \alpha < 1$, the coefficients $\{g_k^{(\alpha)}\}$ holds in

(1) $1 < g_0^{(\alpha)} < 1 + \frac{5}{6}\alpha,\ \ g_1^{(\alpha)} < 0,\ g_2^{(\alpha)} > 0,\ \ g_k^{(\alpha)} < 0, \quad$ for all $k \geqslant 3$,
(2) $\sum_{k=0}^{\infty} g_k^{(\alpha)} = 0,$ and, $|g_k^{(\alpha)}| < \frac{10}{3}\alpha, \quad$ for all $k \geq 1$,
(3) $-\sum_{k=1}^{n} g_k^{(\alpha)} < g_0^{(\alpha)} + g_2^{(\alpha)}, \quad$ for all $n \geq 1$.

In the below, we formulate a fourth-order compact finite difference method for the space discretization of the nonlinear differential equation (3), which results in a system of time-dependent fractional equations. Then, to discretize the derived system in the temporal direction, we will apply the third-order WSGL method (6).

The spatial discretization of (3)

A semi-discrete form of the nonlinear differential equation (3) at z_i can be formulated as

$$\begin{aligned} &(-\frac{\sigma}{h^2} + \frac{\lambda}{4h}\theta w_i^{\theta-1}\delta_z w_i - \frac{\lambda^2}{12\sigma} w_i^{2\theta})\delta_z^2 w_i + (\frac{\lambda}{h} w_i^{\theta} - \frac{\lambda^2}{12\sigma} w_i^{2\theta-1}\delta_z w_i \\ &+ \frac{\lambda}{12h}\theta(\theta-1) w_i^{\theta-2}(\delta_z w_i)^2)\delta_z w_i = (1 - \frac{\lambda h}{12\sigma} w_i^{\theta}\delta_z + \frac{1}{12}\delta_z^2)\Psi_i + R_i(h)\ , \end{aligned} \tag{8}$$

where $w_i := w(z_i, \cdot)$, $\Psi_i := \Psi(z_i, \cdot)$, and the difference operators δ_z and δ_z^2 are defined as

$$\delta_z w_i := \frac{1}{2}(w_{i+1} - w_{i-1}), \qquad \delta_z^2 w_i := w_{i+1} - 2w_i + w_{i-1}.$$

Lemma 3. Let the solution $w(z, \cdot)$ be a six-time continuously differentiable function with respect to z. Then the error term $R_i(h)$ of the formula (8) for (3) is of order four, that is, $R_i(h) = \mathcal{O}(h^4)$.

Proof. Using the Taylor expansion of $w_{i\pm1}$ at point z_i, the value of (3) can be calculated as

$$-\frac{\sigma}{h^2}\delta_z^2 w_i + \frac{\lambda}{h}\ w_i^{\theta}\delta_z w_i + \frac{h^2}{12}\zeta_i = \Psi_i - \frac{h^4}{360}(3\lambda w^{\theta} w_{zzzzz} - \sigma w_{zzzzzz})|_{z=\eta_i}. \tag{9}$$

where $\zeta = 2\lambda w^{\theta} w_{zzz} - \sigma w_{zzzz}$ and $z_{i-1} < \eta_i < z_{i+1}$. After calculating the derivatives of (3) with respect to z, we have

$$\begin{cases} \sigma w_{zzz} &= \lambda\ (w^{\theta} w_{zz} + \theta w^{\theta-1}(w_z)^2) - \Psi_z, \\ \sigma w_{zzzz} &= \lambda(w^{\theta} w_{zzz} + 3\theta w^{\theta-1} w_z w_{zz} + \theta(\theta-1) w^{\theta-2}(w_z)^3) - \Psi_{zz} \\ &= ((\lambda\ w^{\theta})^2 + 3\lambda\theta w^{\theta-1} w_z) w_{zz} + (\theta\lambda^2\ w_z w^{\theta-1} \\ &\quad -\lambda\theta(\theta-1) w^{\theta-2}(w_z)^2) w_z - \lambda\ w^{\theta}\Psi_z - \Psi_{zz}. \end{cases}$$

This leads to

$$\zeta = \Psi_{zz} - \frac{\lambda}{\sigma}\ w^{\theta}\Psi_z + (\frac{\lambda^2}{\sigma} w^{2\theta} - 3\lambda\theta w^{\theta-1} w_z) w_{zz} + (\frac{\lambda^2}{\sigma}\theta w^{\theta-1} - \lambda\theta(\theta-1) w^{\theta-2} w_z)(w_z)^2.$$

Using operators δ_z and δ_z^2, we can derive

$$\zeta_i = \zeta(z_i, \cdot) = \frac{1}{h^2}\delta_z^2\Psi_i - \frac{\lambda}{\sigma h} w_i^{\theta}\delta_z\Psi_i + (\frac{\lambda^2}{\sigma h^2} w_i^{2\theta} - \frac{3\lambda}{h^3}\theta w_i^{\theta-1}\delta_z w_i)\delta_z^2 w_i + (\frac{\lambda^2}{\sigma h^2}\theta w_i^{\theta-1} - \frac{\lambda}{h^3}\theta(\theta-1) w_i^{\theta-2}\delta_z w_i)(\delta_z w_i)^2 + \mathcal{O}(h^2). \tag{10}$$

By substituting relation (10) into (9), we can conclude that $R_i(h) = \mathcal{O}(h^4)$. □

Full discretization of (3)

Here, to develop a high-order discretization technique for (1), the discrete forms (6) and (8) are implemented in temporal and spatial directions, respectively. Consider the following fully discrete method for solving the initial-boundary value problem (1) which determines approximations w_i^n to the values $w(z_i, t_n)$ from the relation

$$\begin{aligned} &(-\frac{\sigma}{h^2} + \frac{\lambda}{4h}\theta(w_i^n)^{\theta-1}\delta_z w_i^n - \frac{\lambda^2}{12\sigma}(w_i^n)^{2\theta})\delta_z^2 w_i^n \\ &+\Big(\frac{\lambda}{h}(w_i^n)^{\theta} - \frac{\lambda^2}{12\sigma}(w_i^n)^{2\theta-1}\delta_z w_i^n + \frac{\lambda}{12h}\theta(\theta-1)(w_i^n)^{\theta-2}(\delta_z w_i^n)^2\Big)\delta_z w_i^n \\ &\quad = (1 - \frac{\lambda h}{12\sigma}(w_i^n)^{\theta}\delta_z + \frac{1}{12}\delta_z^2)\Psi_i^n, \end{aligned} \tag{11}$$

for $n = 1, 2, \ldots, N+1$ and $i = 1, 2, \ldots, M$, where

$$\Psi_i^n = f_i^n + H_i^n - \tau^{-\alpha}\sum_{k=0}^{n} g_k^{(\alpha)} w_i^{n-k}, \tag{12}$$

and $f_i^n := f(w_i^n, z_i, t_n),\ H_i^n := H(z_i, t_n)$. For calculating the value w_i^0 (for $i = 0, \ldots, M+1$) and values w_0^n, w_{M+1}^n (for $n = 0, \ldots, N+1$), we use the initial and boundary conditions (1b)–(1c) as follows:

$$w_i^0 = 0,\ 0 \leqslant i \leqslant M+1, \text{ and } w_0^n = \varphi_a(t_n),\ w_{M+1}^n = \varphi_b(t_n),\ 0 \leqslant n \leqslant N+1.$$

Consider the matrix type of grid functions as $\mathbf{W}^n := [w_1^n, w_2^n, \ldots, w_M^n]^T$ and $\mathbf{F}^n := f(\mathbf{W}^n)$, $\mathbf{H}^n := [H_1^n, H_2^n, \ldots, H_M^n]^T$. For $1 \leqslant n \leqslant N+1$, we present a matrix form of (11) as

$$\begin{aligned}
&\Big((-\frac{\sigma}{h^2}\mathbf{I} + \frac{\lambda}{4h}\theta(\mathbf{W}^n)^{\theta-1}\mathbf{J}_1\mathbf{W}^n - \frac{\lambda^2}{12\sigma}(\mathbf{W}^n)^{2\theta})\mathbf{J}_2 \\
&+ (\frac{\lambda}{h}(\mathbf{W}^n)^{\theta} - \frac{\lambda^2}{12\sigma}(\mathbf{W}^n)^{2\theta-1}\mathbf{J}_1\mathbf{W}^n + \frac{\lambda}{2h}\theta(\theta-1)(\mathbf{W}^n)^{\theta-2}(\mathbf{J}_1\mathbf{W}^n)^2)\mathbf{J}_1\Big)\mathbf{W}^n \\
&\quad = (\mathbf{I} - \frac{\lambda h}{12}(\mathbf{W}^n)^{\theta}\mathbf{J}_1 + \frac{1}{12}\mathbf{J}_2)\big(\mathbf{F}^n + \mathbf{H}^n - \tau^{-\alpha}\sum_{k=0}^{n} g_k^{(\alpha)}\mathbf{W}^{n-k}\big) + C^n,
\end{aligned} \tag{13}$$

where $\mathbf{J}_1 := \mathbf{tridiag}(-\frac{1}{2}, 0, \frac{1}{2})$ and $\mathbf{J}_2 := \mathbf{tridiag}(1, -2, 1)$. Moreover, $\mathbf{I}$ denotes an identity matrix, $\mathbf{W}^0 = [0, 0, \ldots, 0]^T$, and vectors C^n can be determined by using the boundary conditions. The implicit numerical method (13) forms a system of nonlinear algebraic equations, for every $1 \leqslant n \leqslant N+1$. Therefore, a simple and fast nonlinear solver is needed to solve these nonlinear systems. For every time level n, we propose a linearizing version for (13) as follows:

$$\begin{aligned}
&\Big((-\frac{\sigma}{h^2}\mathbf{I} + \frac{\lambda}{4h}\theta(\mathbf{U}^{[s]})^{\theta-1}\mathbf{J}_1\mathbf{U}^{[s]} - \frac{\lambda^2}{12\sigma}(\mathbf{U}^{[s]})^{2\theta})\mathbf{J}_2 + (\frac{\lambda}{h}(\mathbf{U}^{[s]})^{\theta} \\
&-\frac{\lambda^2}{12\sigma}(\mathbf{U}^{[s]})^{2\theta-1}\mathbf{J}_1\mathbf{U}^{[s]} + \frac{\lambda}{2h}\theta(\theta-1)(\mathbf{U}^{[s]})^{\theta-2}(\mathbf{J}_1\mathbf{U}^{[s]})^2)\mathbf{J}_1\Big)\mathbf{U}^{[s+1]} \\
&\quad = (\mathbf{I} - \frac{\lambda h}{12}(\mathbf{U}^{[s]})^{\theta}\mathbf{J}_1 + \frac{1}{12}\mathbf{J}_2)\Big(\mathbf{F}(\mathbf{U}^{[s]}, \mathbf{U}^{[s+1]}) - \tau^{-\alpha}g_0^{(\alpha)}\mathbf{U}^{[s+1]} \\
&\qquad +\mathbf{H}^n - \tau^{-\alpha}\sum_{k=1}^{n} g_k^{(\alpha)}\mathbf{W}^{n-k}\Big) + C^n,
\end{aligned} \tag{14}$$

where

$$\begin{aligned}
\mathbf{F}(\mathbf{U}^{[s]}, \mathbf{U}^{[s+1]}) = \beta(1 - (\mathbf{U}^{[s]})^{\gamma})((\mathbf{U}^{[s]})^{\gamma} - \eta)\mathbf{U}^{[s+1]} \\
+\beta\gamma(\mathbf{U}^{[s]})^{\gamma}(2(\mathbf{U}^{[s]})^{\gamma} - \eta - 1)(\mathbf{U}^{[s+1]} - \mathbf{U}^{[s]}).
\end{aligned}$$

The symbol $\mathbf{U}^{[s]}$ represents the approximate solution which is found in the s-th iteration of the method (14). The number of iterations of this method can be determined such that the condition $\|\mathbf{U}^{[s+1]} - \mathbf{U}^{[s]}\| < \varepsilon$ is attained.

2.1 The analysis of solvability, stability, and convergence of the present method

The analysis of the solvability, stability, and convergence of the present iterative method (13) for the FGBH equation (1) will be stated by following the lemma and theorems. For the matrix $\mathbf{A}$, let $\|\mathbf{A}\|$ be its L_2 norm. As it is concluded in [16], the eigenvalues of $\mathbf{J}_1$ and $\mathbf{J}_2$ are, respectively, derived as

$$\lambda_s^{\mathbf{J}_1} := \mathbf{i}\cos(\frac{s\pi}{M+1}), \quad \text{and} \quad \lambda_s^{\mathbf{J}_2} := -4\sin^2(\frac{s\pi}{2(M+1)}), \tag{15}$$

where $s = 1, 2, \ldots, M$, and $\mathbf{i} = \sqrt{-1}$.

Lemma 4 (Solvability)**.** For every integer number n, the system of equations (13) has a unique solution.

Proof. Let $q = \max|w|$. Using system (11), we can reformulate (13) as follows:

$$A\ \mathbf{W}^n = -\sum_{k=1}^{n} B_k\ \mathbf{W}^{n-k} + B_0\ (\mathbf{F}^n + \mathbf{H}^n) + C^n, \tag{16}$$

where $B_0 = \mathbf{I} - \frac{\lambda h}{12\sigma} q^\theta \mathbf{J}_1 + \frac{1}{12}\mathbf{J}_2$, $A = \tau^{-\alpha} g_0^{(\alpha)} B_0 - (\frac{\sigma}{h^2} + \frac{\lambda^2}{12\sigma} q^{2\theta})\mathbf{J}_2 + \frac{\lambda}{h} q^\theta \mathbf{J}_1$, and $B_k = \tau^{-\alpha} g_k^{(\alpha)} B_0$ for $k = 1, 2, \ldots$. The M-dimensional matrix A has the tridiagonal form $A = \mathbf{tridiag}(b, d, c)$, where its lower-main-upper diagonal elements are the following expressions, respectively:

$$\begin{cases} b &= -\frac{\sigma}{h^2} - \frac{\lambda^2}{12} q^{2\theta} - \frac{\lambda}{2h} q^\theta + (\frac{1}{12} - \frac{\lambda h}{24} q^\theta)\tau^{-\alpha} g_0^{(\alpha)}, \\ d &= \frac{2\sigma}{h^2} + \frac{\lambda^2}{6} q^{2\theta} + \frac{5}{6}\tau^{-\alpha} g_0^{(\alpha)}, \\ c &= -\frac{\sigma}{h^2} - \frac{\lambda^2}{12} q^{2\theta} + \frac{\lambda}{2h} q^\theta + (\frac{1}{12} + \frac{\lambda h}{24} q^\theta)\tau^{-\alpha} g_0^{(\alpha)}, \end{cases}$$

Using Lemma 2, we obtain $1 < g_0^{(\alpha)} < \frac{11}{6}$. Therefore, for sufficiently small step sizes h and τ, we can derive

$$|b| + |c| \le \frac{2\sigma}{h^2} + \frac{\lambda^2}{6} q^2 + \frac{1}{6}\tau^{-\alpha} g_0^{(\alpha)} < |d|.$$

This shows that the matrix A is diagonally dominant and thus nonsingular. Consequently, the system (13) has a unique solution. □

To establish the unconditional stability and convergence of the proposed method, we begin by recalling from [14, Lemma 1.4.2], which is stated as follows.

Lemma 5 (Gronwall's inequality)**.** Let k_n be a nonnegative sequence. Suppose ϕ_n satisfies the inequality:

$$\phi_0 \leq s_0, \qquad \phi_n \leq s_0 + \sum_{i=0}^{n-1} k_i \phi_i, \quad n \geq 1,$$

where $s_0 \geq 0$. Then ϕ_n satisfies the bound

$$\phi_n \leq s_0 \exp(\sum_{i=0}^{n-1} k_i), \quad n \geq 1.$$

Theorem 1 (Unconditionally stability)**.** The iterative method (13) is unconditionally stable to solve the FGBH equation (1).

Proof. For $q = \max|w|$, the method (13) reduces to (16). Let $\tilde{\mathbf{W}}^0 = [\phi_1, \phi_2, \ldots, \phi_M]^T$ be a perturbation of the initial value $\mathbf{W}^0 = [0, 0, \ldots, 0]^T$, and let $\tilde{\mathbf{W}}^n$ hold in below

$$\begin{cases} \tilde{\mathbf{W}}^0 = \Phi^0, \\ A\tilde{\mathbf{W}}^n = -\sum_{k=1}^{n} B_k \tilde{\mathbf{W}}^{n-k} + B_0(\tilde{\mathbf{F}}^n + \mathbf{H}^n) + C^n, \end{cases} \tag{17}$$

in which $\Phi^0 := [\phi_1, \phi_2, \ldots, \phi_M]^T$ and $\tilde{\mathbf{F}}^n = f(\tilde{\mathbf{W}}^n)$. Now, we define $\mathbf{E}^n := \tilde{\mathbf{W}}^n - \mathbf{W}^n - \Phi^0$; subtract (17) from (16), a roundoff error equation is concluded as

$$A\mathbf{E}^n = -\sum_{k=1}^{n} B_k \mathbf{E}^{n-k} - (\sum_{k=1}^{n} B_k + A)\Phi^0 + B_0(\tilde{\mathbf{F}}^n - \mathbf{F}^n), \quad n \geq 1, \tag{18}$$

where $\mathbf{E}^0 := [0, \ldots, 0]^T$. Applying (15), the eigenvalues of tridiagonal matrices B_0 and A are, respectively, obtained as

$$\begin{cases} \lambda_s^{B_0} = (1 - \frac{1}{3}\sin^2(\frac{s\pi}{2(M+1)})) - \mathbf{i}\frac{\lambda h}{12} q^\theta \cos(\frac{s\pi}{M+1}), \\ \lambda_s^{A} = \tau^{-\alpha} g_0^{(\alpha)} \lambda_s^{B_0} + (\frac{4\sigma}{h^2} + \frac{\lambda^2}{3} q^{2\theta}) \sin^2(\frac{s\pi}{2(M+1)}) + \mathbf{i}\frac{\lambda}{h} q^\theta \cos(\frac{s\pi}{M+1}). \end{cases} \tag{19}$$

Consequently $\|(A^n)^{-1} B_0\| \leqslant \frac{1}{g_0^{(\alpha)}} \tau^\alpha$. Using Lemma 2, we can achieve that

$$\|(A^n)^{-1}B_k\| = \tau^{-\alpha}\|(A^n)^{-1}B_0\||g_k^{(\alpha)}| \leqslant \frac{|g_k^{(\alpha)}|}{g_0^{(\alpha)}} \leqslant |g_k^{(\alpha)}|. \tag{20}$$

Moreover, this lemma leads to

$$\sum_{k=1}^{n}|g_k^{(\alpha)}| = g_2^{(\alpha)} - \sum_{k=1}^{n} g_k^{(\alpha)} \leqslant g_0^{(\alpha)} + 2g_2^{(\alpha)} \leqslant \frac{17}{2}, \quad \text{for all } n \geq 1, \tag{21}$$

when $0 < \alpha \leqslant 1$. Since f is a Lipschitz continuous function with respect to w, then there exists a positive constant $L := \max|\frac{\partial f}{\partial w}|$ such that

$$|f(\tilde{w}) - f(w)| \leqslant L|\tilde{w} - w|.$$

Now, by applying (20)–(21) to the difference equations (18) and using the properties of $\omega_k^{(\alpha)}$ yields

$$\begin{aligned}
|1-\tau^{\alpha}L|\|\mathbf{E}^n\| &\leqslant \sum_{k=1}^{n}\|(A^n)^{-1}B_k\|\|\mathbf{E}^{n-k}\| \\
&\quad +(\sum_{k=1}^{n}\|(A^n)^{-1}B_k\| + (1+T^{\alpha}L))\|\Phi^0\|, \\
&\leqslant \sum_{k=1}^{n}|g_k^{(\alpha)}|\|\mathbf{E}^{n-k}\| + \|\Phi^0\|\sum_{k=1}^{n}|g_k^{(\alpha)}| + (1+T^{\alpha}L)\|\Phi^0\| \\
&\leqslant \sum_{k=1}^{n}|g_k^{(\alpha)}|\|\mathbf{E}^{n-k}\| + \frac{17}{2}\|\Phi^0\| + (1+T^{\alpha}L)\|\Phi^0\| \\
&\leqslant \sum_{k=0}^{n-1}|g_{n-k}^{(\alpha)}|\ \|\mathbf{E}^k\| + (\frac{19}{2}+T^{\alpha}L)\|\Phi^0\|.
\end{aligned} \tag{22}$$

Let $\tau^{\alpha}L \neq 1$. Equation (22) leads to

$$\|\mathbf{E}^n\| \leqslant \frac{1}{|1-\tau^{\alpha}L|}\sum_{k=0}^{n-1}|g_{n-k}^{(\alpha)}|\|\mathbf{E}^k\| + \frac{19+2T^{\alpha}L}{|2-2\tau^{\alpha}L|}\|\Phi^0\|, \qquad n = 1,\ldots,N+1.$$

Using Gronwall's Lemma 5, we have

$$\|\mathbf{E}^n\| \leqslant C_1\|\Phi^0\|, \qquad n = 1,\ldots,N+1, \tag{23}$$

where $C_1 = \frac{19+2T^{\alpha}L}{|2-2\tau^{\alpha}L|}\exp(\frac{17}{|2-2\tau^{\alpha}L|})$. Therefore, the unconditional stability of the iterative method (13) is established, as required. □

Theorem 2 (Convergence analysis)**.** Let $w \in \mathcal{C}_{x,t}^{6,4}(\Omega)$ be the solution of the FGBH equation (1) satisfying in $\frac{\partial^k w}{\partial t^k}|_{t=0} = 0$ for $k = 0, 1, 2, 3$; then solution of the iterative method (13) converges to w as $h, \tau \to 0$. Furthermore, its convergence rate is of order $\mathcal{O}(\tau^3 + h^4)$.

Proof. Define $e_i^n := w(z_i, t_n) - w_i^n$, where $0 \leqslant n \leqslant N+1,\ 0 \leqslant i \leqslant M+1$. Since f is a Lipschitz continuous function with respect to w, then there exists a positive constant $L = \max|\frac{\partial f}{\partial w}|$ such that $|f(w_i^n) - f(w(z_i, t_n))| \leqslant L|e_i^n|$. Then subtract (8) from (16), we obtain the error equation

$$A\mathbf{e}^n = -\sum_{k=1}^{n} B_k \mathbf{e}^{n-k} + B_0(f(\mathbf{w}^n) - f(\mathbf{W}^n)) + \mathbf{R}^n(h, \tau), \quad n \geq 1, \tag{24}$$

where $\mathbf{w}^n = [w(z_1, t_n), \ldots, w(z_M, t_n)]^T$, $\mathbf{e}^0 := [0, \ldots, 0]^T$, $\mathbf{e}^n := [e_1^n, \ldots, e_M^n]^T$ and $\mathbf{R}^n := [R_1^n, \ldots, R_M^n]^T$. From Lemmas 1 and 3, there exists a constant C_2 such that $\|\mathbf{R}^n(h, \tau)\| \leqslant C_2(\tau^3 + h^4)$. Let $\tau^\alpha L \neq 1$. Applying Lemma 2 and (19)–(21) deduces

$$\begin{aligned} |1 - \tau^\alpha L|\|\mathbf{e}^n\| &= \sum_{k=1}^{n} \|(A)^{-1}B_k\|\|\mathbf{e}^{n-k}\| + C_2(\tau^3 + h^4) \\ &\leqslant \sum_{k=0}^{n-1} |g_{n-k}^{(\alpha)}|\|\mathbf{e}^{n-k}\| + C_2(\tau^3 + h^4), \qquad n \geq 1. \end{aligned} \tag{25}$$

Using Gronwall's Lemma 5, we can derive

$$\|\mathbf{e}^n\| \leqslant \frac{C_2}{|1 - \tau^\alpha L|} \exp\left(\frac{17}{|2 - 2\tau^\alpha L|}\right)(\tau^3 + h^4), \qquad n \geq 1. \tag{26}$$

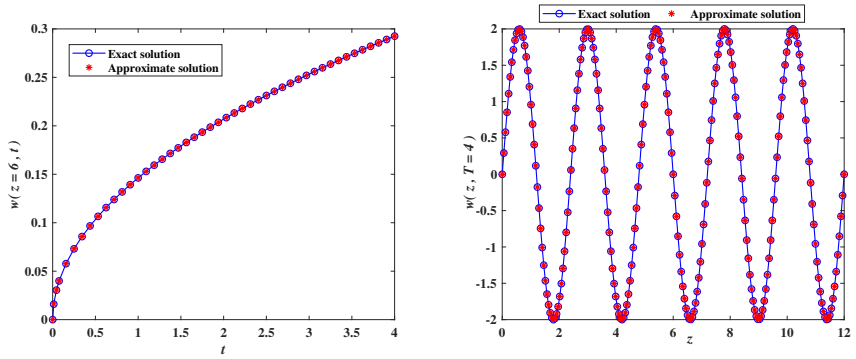
□

As Theorem 2 establishes, the proposed iterative method (13) has fourth- and third-order accuracy in spatial an temporal directions, respectively.

3 Numerical results

In this section, numerical results of the iterative method (14) to solve three types of the inhomogeneous FGBH equation are presented. The provided numerical results verify the theoretical results of Theorems 1 and 2.

All numerical simulations were conducted in MATLAB on a workstation equipped with an Intel ® Core™ i7-4702MQ CPU (2.20 GHz), 8 GB of RAM, running Windows 8 operating system. The convergence rate of the present iterative method is evaluated by



(a) Plots of solutions versus t. (b) Plots of solutions versus z.

Figure 1: Comparison of the exact and approximate solutions to Example 1 after 20 iterations are made when $\mu = 0.5, \alpha = 0.5$ and $\Omega = [0, 12] \times [0, 4]$ is divided into 214×1280 cells. The MAE of the present method is 1.5×10^{-2}.

Table 1: Temporal convergence rate of the method (14) after $r = 20$ iterations to solve Example 1, the exact solution $w(z,t) = t^3 \sqrt[5]{t}\ \sin(\frac{5\pi}{6}z)$, $M \times N$ cells and $M \approx \sqrt[4]{N^3}$.

	$\alpha = 0.25$		$\alpha = 0.5$		$\alpha = 0.75$	
N	MAE	Rate	MAE	Rate	MAE	Rate
20	3.6532e-05	-	3.8011e-05	-	5.2457e-05	-
80	6.7354e-07	2.88	6.9027e-07	2.89	1.0004e-06	2.86
320	1.0717e-08	2.99	1.0965e-08	2.99	1.5900e-08	2.99
1280	1.6567e-10	3.00	1.7067e-10	3.00	2.4598e-10	3.00

$$\text{Rate} := \log_{(\frac{m_2}{m_1})}\left(\frac{\varepsilon_1}{\varepsilon_2}\right),$$

where ε_1 and ε_2 are the maximum absolute errors (MAEs) of the present method using m_1 and m_2 cells in the time/space direction, respectively.

Example 1. Consider the FGBH equation (1) as follows [12, 21]:

$$ {}_0\mathcal{D}_t^{\alpha} w + 0.01 w w_z - w_{zz} = w(1-w)(w-0.5) + H, \quad \text{in } \Omega, \tag{27}$$

where the governing function $H = H(z,t)$ and all initial-boundary conditions are derived from the exact analytical solution: $w(z,t) = t^{\mu}\sin(\frac{5\pi}{6}z)$.

Here, we examine the present iterative method (14) for Example 1 when $\mu = \frac{16}{5}, \frac{1}{2}$ and $\mu = 3$. For different values $\alpha = 0.25, 0.5, 0.75$ and $\mu = \frac{16}{5}$, the MAE and the temporal/spatial convergence rate of this iterative method after 20 iterations are presented in Tables 1–2. The domain $\Omega = [0,1] \times [0,1]$ is divided into $M \times N$ cells, where M and N are two positive integers

Table 2: Spatial convergence rate of the present iterative method (14) after $r = 20$ iterations to solve Example 1, the exact solution $w(z,t) = t^3\sqrt[5]{t}\ \sin(\frac{5\pi}{6}z)$, $M \times N$ cells and $N \approx \sqrt[3]{M^4}$.

	$\alpha = 0.25$		$\alpha = 0.5$		$\alpha = 0.75$	
M	MAE	Rate	MAE	Rate	MAE	Rate
4	1.4012e-03	-	1.4133e-03	-	1.8852e-03	-
16	5.4371e-06	4.00	5.5444e-06	4.00	7.4116e-06	4.00
64	2.1284e-08	4.00	2.1752e-08	4.00	2.8852e-08	4.00
256	8.3054e-11	4.00	8.4690e-11	4.00	1.1196e-10	4.00

Table 3: Comparison of MAEs obtained by the B-spline collocation method [12], compact difference scheme [21], and the present iterative method to solve Example 1 in $\Omega = [0,1] \times [0,1]$ with the exact solution $w(z,t) = t^3 \sin(\frac{5\pi}{6}z)$.

α	BSC method [12] with ($M = 100,\ N = 3125$)	CD scheme [21] with ($M = 64, N = 1024$)	Present method with ($M = 64, N = 1024$)
0.3	-	6.5363e-06	1.9546e-08
0.5	3.0000e-06	-	1.8805e-08
0.6	-	7.0496e-06	1.6886e-08

satisfying $N^3 = M^4$. The obtained numerical results in Table 1 illustrate that this new method is of order $\mathcal{O}(\tau^3)$ in time direction. Moreover, Tables 1 and 2 show that the present method is unconditionally stable. These results confirm the statements of Theorems 1 and 2.

For $\mu = 0.5$, comparison of the exact and approximate solutions to Example 1 in the large domain $[0,12] \times [0,4]$ is also made in Figure 1 when $\alpha = 0.5$ and 214×1280 cells are selected. Part (a) shows solutions at $z = 6$ versus $t \in [0,4]$, and part (b) demonstrates them at $t = 4$ versus $z \in [0,12]$. To derive these plots, the proposed iterative method converged after 20 iterations and 91.02 seconds, and its maximum absolute error is 1.2232×10^{-2}. This figure illustrates that the present method is unconditionally stable and is not affected by round-off errors in the large computational domain $\Omega = [0,12] \times [0,4]$.

Finally, the MAEs obtained by the B-spline collocation (BSC) method [12], compact difference (CD) scheme [21], and the present iterative method for solving Example 1 with $\mu = 3$ are compared in Table 3. This table shows that the present iterative method (13) is more accurate and effective than the methods given in [12, 21].

Example 2. Consider the following inhomogeneous FGBH equation:

$$ {}_0\mathcal{D}_t^{\alpha} w + \frac{\pi}{2} w^2 w_z - 20 w_{zz} = 50 w(1-w^3)(w^3 - \frac{2}{3}) + g, \quad (z,t) \in \Omega, \tag{28} $$

where the function $g := g(z,t)$ is calculated using the exact solution $w(z,t) = \frac{1}{8} t^{2+4\alpha}\ (z - z^2)$.

Table 4: Temporal convergence rate of the iterative technique (14) after $r = 20$ iterations to solve Example 2 at $T = 2$ with $M \times N$ cells and $M \approx \sqrt[4]{0.5N^3}$.

	$\alpha = 0.25$		$\alpha = 0.5$		$\alpha = 0.75$		
N	MAE	Rate	MAE	Rate	MAE	Rate	CPU time
10	1.1433e-05	-	5.7396e-05	-	3.3925e-04	-	0.04
20	3.0962e-06	1.88	1.4212e-05	2.01	7.2916e-05	2.22	0.06
40	4.9555e-07	2.64	2.1440e-06	2.73	1.0575e-05	2.79	0.10
80	8.4224e-08	2.56	3.5233e-07	2.60	1.5525e-06	2.77	0.20
160	1.1150e-08	2.92	4.6702e-08	2.92	2.2482e-07	2.79	0.50
320	1.5055e-09	2.87	6.0369e-09	2.95	3.2025e-08	2.81	1.78
640	1.8980e-10	2.99	7.5944e-10	2.99	4.5482e-09	2.82	6.08

Table 5: Spatial convergence rate of the iterative technique (14) after $r = 30$ iterations to solve Example 2 at $T = 2$ with $M \times N$ cells and $N \approx \sqrt[3]{2M^4}$.

	$\alpha = 0.25$		$\alpha = 0.5$		$\alpha = 0.75$		
M	MAE	Rate	MAE	Rate	MAE	Rate	CPU time
4	1.1558e-05	-	6.0110e-05	-	3.6824e-04	-	0.05
8	1.1263e-06	3.36	5.0731e-06	3.57	2.5876e-05	3.83	0.06
16	8.4422e-08	3.74	3.5816e-07	3.83	1.8329e-06	3.82	0.14
32	5.7420e-09	3.89	2.3813e-08	3.91	1.3064e-07	3.81	0.47
64	3.7460e-10	3.94	1.6436e-09	3.86	9.3027e-09	3.82	2.74
128	2.3957e-11	3.97	1.1208e-10	3.87	6.4064e-10	3.86	22.12
256	1.4975e-12	4.00	7.0836e-12	3.98	4.5029e-11	3.83	227.05

The MAE, temporal/spatial convergence rate, and elapsed CPU time (in second) of the proposed iterative technique (14) after $r = 20$ iterations for solving the inhomogeneous FGBH equation (28) are presented in Tables 4 and 5. The domain $\Omega = [0, 1] \times [0, 2]$ is divided into $M \times N$ cells which M and N are two positive integers satisfying $M^4 = 2N^3$. These tables demonstrate that the convergence rate of this method is $\mathcal{O}(\tau^3 + h^4)$ for different values of α. Moreover, the present method is not affected by round-off errors. Also, it is unconditionally stable when the number of cells is increased. Consequently, these numerical results follows analytical results of Theorems 1 and 2.

Example 3. Consider the inhomogeneous FGBH equation (1) as

$$ {}_0\mathcal{D}_t^{\alpha} w + w w_z - w_{zz} = \ w(1 - w^2)(w^2 - 0.75) + g, \tag{29} $$

where $(z, t) \in (-1, 1) \times (0, 1]$ and the function $g = g(z, t)$ is calculated using the exact solution $w(z, t) = (1 - z^6) \exp(1 - z^2) t^{\alpha + \sqrt{5} \cosh(z)}$.

For solving the inhomogeneous FGBH equation (29) with $\alpha = 0.25, 0.5, 0.75$, the present iterative method converges after $r = 20$ iterations. Numerical results are provided in Tables 6 and

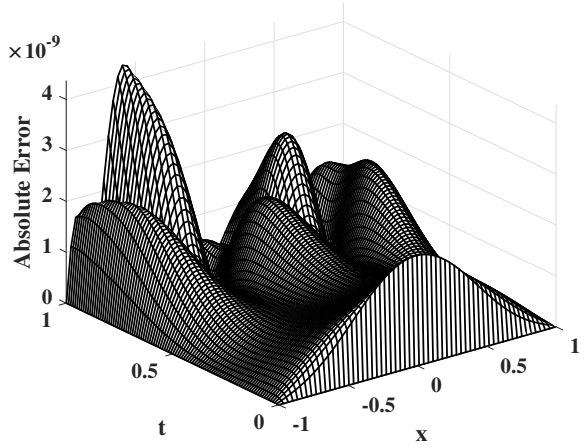


Figure 2: The absolute error of the present iterative technique after $r = 20$ iterations to solve Example 3 with $\alpha = 0.75$ and 200×1000 cells.

Table 6: Temporal convergence rate of the iterative technique (14) after $r = 20$ iterations to solve the Example 3 with $M \times N$ cells and $M \approx \sqrt[4]{N^3}$.

	$\alpha = 0.25$		$\alpha = 0.5$		$\alpha = 0.75$		
N	MAE	Rate	MAE	Rate	MAE	Rate	CPU time
10	4.9633e-03	-	4.8430e-03	-	4.6422e-03	-	0.03
20	5.0011e-04	3.31	6.055e-04	3.00	4.8041e-04	3.27	0.04
40	7.0557e-05	2.83	7.5800e-05	3.00	6.7283e-05	2.84	0.06
80	8.4545e-06	3.06	9.4508e-06	3.00	8.2037e-06	3.04	0.14
160	1.1169e-06	2.92	.1802e-06	3.00	1.0726e-06	2.94	0.53
320	1.3667e-07	3.03	.4804e-07	3.00	1.3192e-07	3.02	2.11
640	1.7022e-08	3.00	1.6760e-08	3.14	1.6374e-08	3.01	12.80

7. The numerical results in these tables confirm that the iterative technique (14) is numerically stable (insensitive to roundoff errors) while maintaining the predicted third-order (temporal) and fourth-order (spatial) convergence rates. The scheme accurately captures the solution's initial singularity, thereby validating Theorems 1 and 2. Moreover, Figure 2 shows the absolute error of the proposed technique after $r = 20$ iterations on a 200×1000 grid for solving Example 3 with $\alpha = 0.75$. The figure indicates that the method is robust against round-off errors.

Table 7: Spatial convergence rate of the iterative technique (14) after $r = 20$ iterations to solve Example 3 with $M \times N$ cells and $N \approx \sqrt[3]{M^4}$.

	$\alpha = 0.25$		$\alpha = 0.5$		$\alpha = 0.75$		
M	MAE	Rate	MAE	Rate	MAE	Rate	CPU time
8	1.0773e-03	-	1.0731e-03	-	1.0536e-03	-	0.03
16	7.0565e-05	3.92	6.9298e-05	3.95	6.7362e-05	3.97	0.06
32	4.3136e-06	4.03	4.2591e-06	4.02	4.1816e-06	4.01	0.20
64	2.7281e-07	3.98	2.6864e-07	3.99	2.6242e-07	3.99	1.46
128	1.7022e-08	4.00	1.6762e-08	4.00	1.6380e-08	4.00	12.52
256	1.0646e-09	4.00	1.0446e-09	4.00	1.0206e-09	4.00	139.05

4 Conclusion

In this study, we developed an efficient high-order numerical scheme for solving the fractional generalized Burgers–Huxley (FGBH) equation, a nonlinear partial differential equation that models various physical phenomena involving diffusion, reaction, and convection mechanisms under fractional dynamics. The proposed approach combines a compact finite difference method for spatial discretization with the weighted and shifted Grünwald–Letnikov (WSGL) scheme for approximating temporal fractional derivatives. This hybrid framework transforms the FGBH equation into a nonlinear algebraic system, which is efficiently solved using a tailored iterative algorithm. Theoretical analysis establishes the unconditional stability and convergence of the method, with an error bound of order $\mathcal{O}(\tau^3 + h^4)$. Comprehensive numerical experiments, including three representative test cases, validate the accuracy and robustness of the method. The results exhibit excellent agreement with theoretical convergence rates, demonstrate resilience in capturing singularities, and confirm computational efficiency across various parameter settings. These findings support the theoretical results established in Theorems 1 and 2 and highlight the practical applicability of the scheme. This study not only advances numerical treatment for fractional PDEs but also provides a computational tool suitable for practical applications in physics, biology, and engineering systems governed by memory and nonlocal effects. Future work will focus on extending the methodology to more complex settings, including systems with variable-order fractional derivatives, higher-dimensional versions of the FGBH equation, and adaptive implementations for capturing intricate solution structures.

Funding

This research received no external funding.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] Alquran, M., Al-Khaled, K., Sivasundaram, S. and Jaradat, H.M. *Mathematical and numerical study of existence of bifurcations of the generalized fractional Burgers–Huxley equation,* Nonlinear Stud., 24 (2017), 235–244.

[2] Diethelm, K. and Ford, N.J. *Analysis of fractional differential equations,* J. Math. Anal. Appl., 265 (2002), 229–248.

[3] FitzHugh, R. *Mathematical models of excitation and propagation in nerve,* Biol. Eng., 9 (1969), 1–85.

[4] Freihet, A.A. and Zuriqat, M. *Analytical solution of fractional Burgers–Huxley equations via residual power series method,* Lobachevskii J. Math., 40 (2019), 174–182.

[5] Gorenflo, R. and Mainardi, F. *Fractional calculus: Integral and differential equations of fractional order*, Springer, 1997.

[6] Guo, B. Pu, X. and Huang, F. *Fractional partial differential equations and their numerical solutions*, World Scientific, 2015.

[7] Hajipour, M. and Lakpour, Y. *A highly accurate numerical technique for solving variable-order fractional Burgers–Huxley equation,* Comput. Methods Differ. Equ., 2025.

[8] Hammouch Z. and Mekkaoui, T. *Adomian decomposition method for solving a time-fractional Burgers–Huxley equation,* Nonlinear Stud., 19 (2012), 489–496.

[9] Inc, M. Partohaghighi, M. Akinlar, M.A. Agarwal, P. and Chu, Y.M. *New solutions of fractional-order Burgers–Huxley equation,* Results Phys., 18 (2020), 103290.

[10] Inc, M., Yusuf, A., Aliyu, A.I. and Baleanu, D. *Lie symmetry analysis and explicit solutions for the time fractional generalized Burgers–Huxley equation*, Opt. Quantum Electron., 50 (2018), 1–16.

[11] Machado, J.A., Babaei, A. and Moghaddam, B.P. *Highly accurate scheme for the Cauchy problem of the generalized Burgers–Huxley equation*, Acta Polytech. Hung., 13(2016), 183–196.

[12] Majeed, A., Kamran, M., Asghar, N. and Baleanu, D. *Numerical approximation of inhomogeneous time fractional Burgers–Huxley equation with B-spline functions and Caputo derivative,* Eng. Comput., 38 (2022), 885–900.

[13] Nourazar, S.S., Soori, M. and Nazari-Golshan, A. *On the exact solution of Burgers–Huxley equation using the homotopy perturbation method*, J. Appl. Math. Phys., 3 (2015), 285–294.

[14] Quarteroni A. and Valli, A. *Numerical approximation of partial differential equations*, Springer Science & Business Media, 2008.

[15] Singh, A., Diwedi, K.D., Das, S. and Ong, S.H. *Study of one-dimensional space-time fractional-order Burgers-Fisher and Burgers–Huxley fluid models*, Math. Methods Appl. Sci., 44 (2021), 2455–2467.

[16] Smith, G.D. *Numerical solution of partial differential equations: Finite difference methods*, Oxford University Press, 1985.

[17] Tian, W.Y., Zhou, H. and Deng, W. *A class of second order difference approximations for solving space fractional diffusion equations*, Math. Comput., 84 (2015), 1703–1727.

[18] Tripathi, N.K. *Analytical solution of two-dimensional nonlinear space-time fractional Burgers–Huxley equation using fractional sub-equation method*, Natl. Acad. Sci. Lett., 41 (2018), 295–299.

[19] Uchaikin, V.V. *Fractional derivatives for physicists and engineers*, Vol. 2, Springer, 2013.

[20] Wang, X.Y., Zhu, Z.S. and Lu, Y.K. *Solitary wave solutions of the generalized Burgers–Huxley equation*, J. Phys. A: Math. Gen., 23 (1990), 271.

[21] Yang, X. and Liu, X. *Numerical analysis of fourth-order compact difference scheme for inhomogeneous time-fractional Burgers–Huxley equation*, Comput. Math. Appl., 125 (2022), 1–12.

[22] Zhao, L. and Deng, W. *A series of high-order quasi-compact schemes for space fractional diffusion equations based on the superconvergent approximations for fractional derivatives,* Numer. Methods Partial Differ. Equ., 31 (2015), 1345–1381.

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 1, 2026, pp 19–37
https://doi.org/10.22067/ijnao.2025.94037.1667
https://ijnao.um.ac.ir/

CC BY NC

Research Article

A fractional order model for malaria control using optimal strategies and numerical simulations

A. Kouidere*, A. El Bhih, I. Minifi and K. Adnaoui

Abstract

This paper develops and analyzes a fractional-order model for malaria transmission that captures memory effects in host-vector dynamics and enables a more flexible treatment of control strategies. The proposed model extends the classical SEIR framework by incorporating fractional-order derivatives, which offer a more accurate representation of the memory and hereditary effects in malaria transmission. The human population is divided into four compartments: people susceptible (S_P), people exposed (E_P), people infected (I_P), people recovered

*Corresponding author

Received 11 June 2025; revised 1 September 2025; accepted 2 September 2025

Abdelfatah Kouidere
Department of Mathematics, Sidi Bennour Polydisciplinary Faculty, Chouaib Doukkali University. Morocco. e-mail: kouidere.abdelfatah@ucd.ac.ma

Amine El Bhih
Multidisciplinary Research and Innovation Laboratory (LPRI), Moroccan School of Engineering Sciences (EMSI), Casablanca, Morocco. e-mail: elbhihamine@gmail.com

Issam Minifi
LAMS, Department of Mathematics and Computer Science, Faculty of Sciences Ben M'Sik, Hassan II University of Casablanca. Morocc. e-mail: issam.minifigmail.com

Khalid Adnaoui
LAMS, Department of Mathematics and Computer Science, Faculty of Sciences Ben M'Sik, Hassan II University of Casablanca. Morocc. e-mail: khalid.adnaoui@gmail.com

How to cite this article
Kouidere, A. El Bhih, A. Minifi, I. and Adnaoui, K., A fractional order model for malaria control using optimal strategies and numerical simulations. *Iran. J. Numer. Anal. Optim.*, 2026; 16(1): 19-37. https://doi.org/10.22067/ijnao.2025.94037.1667

(R_P), and the mosquitoes population is divided into two compartments: mosquitoes susceptible (S_M), mosquitoes infected (I_M), respectively. The model introduces two time-dependent control variables: public awareness campaigns, treatment. These controls aim to prevent new infections, reduce the number of individuals in infectious compartments, and mitigate long-term complications in recovered individuals. Pontryagin's maximum principle is employed to derive the necessary conditions for optimal control, and the resulting system is solved using an iterative numerical method. Simulation results, implemented in MATLAB, illustrate the influence of fractional-order derivatives on disease dynamics and demonstrate the comparative effectiveness of the control strategies. This work provides a novel fractional-order control framework for malaria and highlights the importance of integrated intervention strategies.

AMS subject classifications (2020): 49J15, 93C10, 92B05, 93A30

Keywords: Malaria, Optimal control, Malaria disease, Fractional-order model, Mathematical

1 Introduction

Malaria continues to exert a heavy burden on public health, especially in tropical and subtropical regions. It is triggered by *Plasmodium* parasites, transmitted through bites from infected *Anopheles* mosquitoes. Although treatments, vaccines, and mosquito control measures like insecticide-treated nets are available, the disease remains persistent in many endemic areas [25, 26, 27].

Classical compartmental modeling approaches, notably the Ross-Macdonald model and its derivatives, have been widely used to understand malaria dynamics. These models often assume Markovian processes, neglecting historical influence on current states. In contrast, fractional-order models accommodate memory effects, making them particularly well-suited for capturing the delayed responses and cumulative effects seen in real-world epidemiology.

This paper introduces a fractional-order SEIR framework for modeling malaria spread between human and mosquito populations. The model incorporates optimal control measures namely treatment and awareness campaigns to evaluate their impact on disease transmission. We employ fractional derivatives to account for hereditary characteristics of the system and apply Pontryagin's principle to formulate the optimal control problem.

Simulation results based on real-world parameter values demonstrate the effectiveness of fractional-order modeling in enhancing predictive accuracy and informing intervention strategies. This approach underscores the importance of integrating advanced mathematical tools in the design of public health policies against malaria.

Despite extensive global control efforts, malaria continues to be a significant public health burden. In 2022, the World Health Organization (WHO) reported approximately 249 million cases and 608,000 deaths worldwide, with the African region bearing the greatest share of the disease burden, followed by Southeast Asia and the Eastern Mediterranean [14, 15, 12, 9, 2, 28]. These figures underscore the urgent need for improved surveillance, equitable treatment access, and strengthened prevention strategies.

Various mathematical models have been employed to study malaria transmission, most commonly classical compartmental models based on integer-order differential equations. These models, including the Ross-Macdonald framework and extensions. Bounouiga, Basti, and Benhamidouche [10] studied a fractional-order model for malaria transmission under the effect of control strategies. Saha, Saha, and Podder [24] introduced a novel mathematical model to study the impact of drug-resistant strain, recovered-carrier, and relapse on malaria dynamics by implementing Caputo–Fabrizio fractional-order derivative. They begin by presenting theoretical results that are derived for our mode. Helikumi and Loliak [12] formulated and analyzed a fractional-order model for malaria disease transmission using Atangana–Beleanu–Caputo in sense to study the effects of heterogeneity vector biting exposure on the human population. To capture effects the heterogeneity vector biting exposure, they subdivided the human population into two subgroups, namely, the population in high and low risk areas.

Fractional-order calculus offers a more flexible mathematical framework by incorporating memory and hereditary effects inherent in biological systems. By employing fractional derivatives, it enables the modeling of time-dependent and nonlocal processes with greater accuracy, thereby improving the realism and predictive capacity of disease transmission dynamics [20, 19, 4, 5, 6]. To overcome the limitations of classical models, this study introduces a fractional-order model to better capture the complexities of malaria transmission dynamics of malaria transmission incorporating both human and mosquito populations. Human infection occurs via mosquito bites, while mosquitoes become infected by feeding on infectious humans. The model quantifies key transmission parameters and considers the impact of asymptomatic carriers, vector control, vaccination, and treatment access [3, 8, 7, 11, 12, 17].

Caputo-type fractional derivatives are used to extend the conventional SEIR (Susceptible exposed infected recovered) framework, integrating memory effects and allowing for more detailed representations of regional differences, behavioral patterns, and disease progression. The model also evaluates the effectiveness of various control strategies under varying levels of intervention coverage and compliance.

The paper is structured as follows: Section 2 introduces foundational concepts of fractional calculus, emphasizing the Caputo derivative's relevance to epidemiology. Section 3 presents a mathematical model of malaria disease. Section 4 formulates an optimal control problem using

Pontryagin's maximum principle. Section 5 presents simulation results under multiple scenarios. Section 6 concludes with key insights, recommendations, and policy implications.

By incorporating fractional calculus into malaria modeling, this study offers a more accurate and adaptable framework for understanding disease dynamics and optimizing public health interventions in endemic regions.

2 Definitions and preliminaries

We will follow the same methodology adopted in the papers [1, 22, 21].

Fractional calculus is a generalization of classical calculus where differentiation and integration are extended to noninteger (fractional) orders. This framework has been increasingly applied in physics, engineering, biology, and epidemiology due to its ability to capture memory and hereditary effects.

Among the different definitions of fractional derivatives, the Caputo derivative is particularly suitable for modeling real-world problems since it allows for the use of standard initial conditions expressed in terms of integer-order derivatives. The Caputo fractional derivative of order α $(0 < \alpha \leq 1)$ for a sufficiently smooth function $f(t)$ is defined as

$$D_t^{\alpha} f(t) = \frac{1}{\Gamma(1-\alpha)} \int_0^t \frac{f'(s)}{(t-s)^{\alpha}}\, ds,$$

where $\Gamma(\cdot)$ is the Gamma function.

When $\alpha = 1$, the Caputo derivative reduces to the classical first-order derivative, while values of α between 0 and 1 interpolate between no memory and full memory effects. This property makes fractional calculus a powerful tool for studying systems where the past influences the present, such as disease progression and transmission.

In the following section, we employ the Caputo fractional derivative to formulate a malaria transmission model that incorporates these memory effects, providing a more accurate framework for analyzing disease dynamics.

3 Model presentation

We categorize the total population N is divided into six compartments $S_P, E_P, I_P, R_P, S_M, I_M$: susceptible, exposed, infected, and recovered people, along with susceptible and infected mosquitoes. The interaction between these compartments is described by a system of fractional differential

equations. These equations incorporate Caputo derivatives to account for the influence of historical infection trends.

Figure 1 illustrates the flow between compartments. The human population is replenished by recruitment, exposed through contact with infectious agents, and progresses through disease stages with recovery or mortality. Mosquitoes become carriers upon biting infected individuals.

The complete system is given below:

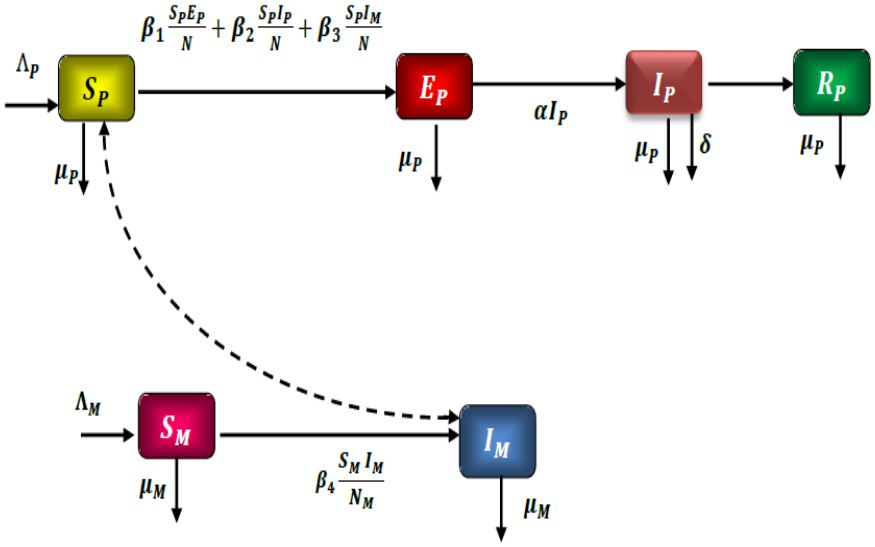


Figure 1: Figure compartments model

Accordingly, we formulate a fractional optimal control model for malaria transmission, which is described by the following system of fractional differential equations:

$$\begin{cases} D^\alpha S_P(t) = \Lambda_P - \mu_P S_P(t) - \beta_1 \frac{S_P(t)E_P(t)}{N} - \beta_2 \frac{S_P(t)I_P(t)}{N} - \beta_3 \frac{S_P(t)I_M(t)}{N}, \\ D^\alpha E_P(t) = \beta_1 \frac{S_P(t)E_P(t)}{N} + \beta_2 \frac{S_P(t)I_P(t)}{N} + \beta_3 \frac{S_P(t)I_M(t)}{N} - (\mu_P + \alpha)E_P(t), \\ D^\alpha I_P(t) = \alpha E_P(t) - (\mu_P + \sigma + \delta)I_P(t), \\ D^\alpha R_P(t) = \sigma I_P(t) - \mu_P R_P(t), \\ D^\alpha S_M(t) = \Lambda_M - \mu_M S_M(t) - \beta_4 \frac{S_M(t)I_M(t)}{N}, \\ D^\alpha I_M(t) = \beta_4 \frac{S_M(t)I_M(t)}{N} - \mu_M I_M(t), \end{cases} \tag{1}$$

where the initial state are $S_P(0) \geq 0$, $E_P(0) \geq 0$, $I_P(0) \geq 0$, $R_P(0) \geq 0$, $S_M(0) \geq 0$ and $I_M(0) \geq 0$, where

Λ_P: Recruitment rate of new susceptible people entering the population.

Λ_M: Recruitment rate of new susceptible malaria vectors (mosquitoes).
μ_P: Per capita natural mortality rate of people.
μ_M: Per capita natural mortality rate of malaria vectors (mosquitoes).
β_1: Transmission rate from exposed people to susceptible people.
β_2: Transmission rate from infected people to susceptible people.
β_3: Transmission rate from infected malaria vectors to susceptible people.
β_4: Transmission rate from infected malaria vectors to susceptible vectors.
α: Progression rate from exposed to infected people.
σ: Recovery rate of infected people.
δ: Disease-induced mortality rate due to complications.

4 Fractional derivation with optimal control problem

4.1 Control strategy formulation

To mitigate disease transmission, we introduce two control interventions: $u(t)$ for awareness and $v(t)$ for treatment. These controls are incorporated into the model to assess their effectiveness over a finite time horizon $[0, T]$. The objective is to reduce both the exposed and infected human populations while considering the cost associated with implementing the controls. Therefore

$$\begin{cases} D^\alpha S_P(t) = \Lambda_P - \mu_P S_P(t) - \beta_1(1-u(t))\dfrac{S_P(t)E_P(t)}{N} \\ \qquad\qquad - \beta_2(1-u(t))\dfrac{S_P(t)I_P(t)}{N} - \beta_3(1-u(t))\dfrac{S_P(t)I_M(t)}{N}, \\ D^\alpha E_P(t) = \beta_1(1-u(t))\dfrac{S_P(t)E_P(t)}{N} + \beta_2(1-u(t))\dfrac{S_P(t)I_P(t)}{N} \\ \qquad\qquad + \beta_3(1-u(t))\dfrac{S_P(t)I_M(t)}{N} - (\mu_P+\alpha)E_P(t), \\ D^\alpha I_P(t) = \alpha E_P(t) - \sigma(1-v(t))I_P - (\mu_P+\delta)I_P(t), \\ D^\alpha R_P(t) = \sigma(1-v(t))I_P(t) - \mu_P R_P(t), \\ D^\alpha S_M(t) = \Lambda_M - \mu_T S_M(t) - \beta_3\dfrac{S_M(t)I_M(t)}{N}, \\ D^\alpha I_M(t) = \beta_4\dfrac{S_M(t)I_M(t)}{N} - \mu_T\ I_M(t), \end{cases} \tag{2}$$

with $S_P(0) \geq 0$, $E_P(0) \geq 0$, $I_P(0) \geq 0$, $R_P(0) \geq 0$, $S_M(0) \geq 0$, and $I_M(0) \geq 0$ are the initial state. Hence

$$J(u,v) = \int_0^T \left[E_P(t) + I_P(t) + \frac{A}{2}u^2(t) + \frac{B}{2}v^2(t) \right] dt. \tag{3}$$

Here, $A > 0$ and $B > 0$ represent the weights associated with the costs of implementing the awareness program and treatment controls, respectively. These parameters reflect the relative importance of the control variables $u(t)$ and $v(t)$ at time t. We adopt a quadratic cost function on the controls to capture nonlinear effects that arise from high levels of intervention implementation. This cost function represents the financial resources required for treatment as well as for conducting the awareness campaigns. The variable T denotes the final time horizon of the control process.

In other words, our objective is to determine the optimal controls $u^*(t)$ and $v^*(t)$ such that

$$J(u^*, v^*) = \min_{(u,v)\in U} J(u,v) \tag{4}$$

subject to the controlled dynamic system described earlier.

The admissible control set is defined such that both $u(t)$ and $v(t)$ remain within $[0,1]$ throughout the interval defined by $U = \{(u,v)/, 0 \leq u_{\min} \leq u_i(t) \leq u_{\max} \leq 1, 0 \leq v_{\min} \leq v_i(t) \leq v_{\max} \leq 1, \text{ and } / \ t \in [0,T]\}$.

The main point in fractional optimal control problems is to find the optimal control (u,v) that minimizes the following objective function:

$$J"(u,v) = \int_0^T \left[E_P(t) + I_P(t) + \frac{A}{2}u^2(t) + \frac{B}{2}v^2(t) \right] dt \tag{5}$$

subject to the constraint
$D^\alpha S_P(t) = g_1(S_P, E_P, I_P, R_P, S_M, I_M, u, v)$,
$D^\alpha E_P(t) = g_2(S_P, E_P, I_P, R_P, S_M, I_M, u, v)$,
$D^\alpha I_P(t) = g_3(S_P, E_P, I_P, R_P, S_M, I_M, u, v)$,
$D^\alpha R_P(t) = g_4(S_P, E_P, I_P, R_P, S_M, I_M, u, v)$,
$D^\alpha S_M(t) = g_4(S_P, E_P, I_P, R_P, S_M, I_M, u, v)$,
$D^\alpha I_M(t) = g_5(S_P, E_P, I_P, R_P, S_M, I_M, u, v)$, where $g_j(1 <= j <= 6)$ represent the right side of the system (2), and where the initial conditions are positives. The following expression defines a modified objective function. To find the optimal control, we follow the traditional approach and define a modified performance index as

$$J(u,v) = \int_0^T \left[H^*(S_P, E_P, I_P, R_P, S_M, I_M, u, v, t)\right] dt$$
$$- \int_0^T \left[\sum_{i=1}^{6} \Lambda_i(t)\, g_i(S_P, E_P, I_P, R_P, S_M, I_M, u, v, t)\right] dt$$

In order to derive the necessary condition for the optimal control, we apply pontryagin's maximum principle. The idea is introducing the adjoint function to attach the system of differential equations to the objective functional resulting in the formation of a function called the Hamiltonian. This principle converts into a problem of minimizing Hamiltonian $H_j(t)$ at time t defined by $H(S_P, E_P, I_P, R_P, S_M, I_M, u, v, t)$:

$$\begin{aligned} &H(S_P, E_P, I_P, R_P, S_M, I_M, u, v, t) \\ &= E_P(t) + I_P(t) + \frac{A}{2}u^2(t) + \frac{B}{2}v^2(t) \\ &\quad + \sum_{i=1}^{6} \Lambda_i(t)\, g_i(S_P, E_P, I_P, R_P, S_M, I_M, u, v, t). \end{aligned}$$

From (5), we can derive the following equalities:

$$\begin{aligned} D^\alpha S_P(t) &= -\tfrac{\partial H^*(t)}{\partial \Lambda_1}(t), \quad D^\alpha E_P(t) = -\tfrac{\partial H^*(t)}{\partial \Lambda_2}(t), \\ D^\alpha I_P(t) &= -\tfrac{\partial H^*(t)}{\partial \Lambda_3}(t), \quad D^\alpha R_P(t) = -\tfrac{\partial H^*(t)}{\partial \Lambda_4}(t), \\ D^\alpha S_M(t) &= -\tfrac{\partial H^*(t)}{\partial \Lambda_5}(t), \quad D^\alpha I_M(t) = -\tfrac{\partial H^*(t)}{\partial \Lambda_6}(t), \end{aligned} \tag{6}$$

with the transversality conditions at time T:

$$\Lambda_1(T) = 0,\ \Lambda_2(T) = -1,\ \Lambda_3(T) = -1,\ \Lambda_4(T) = 0,\ \Lambda_5(T) = 0 \text{ and } \Lambda_6(T) = 0. \tag{7}$$

Equations (6) and (7) describe the necessary conditions in terms of a Hamiltonian for the fractional optimal control problem defined previously.

These conditions result in a set of fractional differential equations, in terms of the variables state $S_P, E_P, I_P, R_P, S_M, I_M$, controls u, v, w and Lagrange multiplier Λ_i, to be solved analytically or numerically or even both.

4.2 Characterization of optimal control

Here we apply Pontryagin's principle [23, 16, 13] to solve the optimal control problem given above. This principle converts into a problem of minimizing a Hamiltonian $H(t)$ at time t defined by

$$\begin{aligned} &H(S_P, E_P, I_P, R_P, S_M, I_M, u, v, t) \\ &= E_P(t) + I_P(t) + \frac{A}{2}u^2(t) + \frac{B}{2}v^2(t) \\ &\quad + \sum_{i=1}^{6} \Lambda_i(t)\, f_i(S_P, E_P, I_P, R_P, S_M, I_M, u, v, t) \end{aligned}$$

with g_i is the right side of the differential equation of the ith state variable of the system at time t.

Theorem 1. The optimal controls (u^*, v^*) and the solutions S_P^*, E_P^*, I_P^*, R_P^*, S_M^* and I_M^* of the corresponding co-state system (2) are given by the adjoint variables $\Lambda_1(t)$, $\lambda_2(t)$,$\Lambda_3(t)$, $\Lambda_4(t)$, $\Lambda_5(t)$ and $\Lambda_6(t)$ satisfying

$$\begin{aligned} D^\alpha \Lambda_1(t) =& -\Lambda_1(t)\left[-\mu_P - \beta_1(1-u(t))(\frac{E_P(t)}{N} - \beta_2\frac{I_P(t)}{N} - \beta_3\frac{I_M(t)}{N})\right] \\ &-\Lambda_2(t)\left(\beta_1(1-u(t))(\frac{E_P(t)}{N} - \beta_2\frac{I_P(t)}{N} - \beta_3\frac{I_M(t)}{N})\right) \\ D^\alpha \Lambda_2(t) =& -1 - \Lambda_1(t)\left[-\beta_1(1-u(t))\frac{S_P(t)}{N}\right] - \Lambda_2(t)\left[\beta_1(1-u(t))\frac{S_P(t)}{N} - (\mu_P + \alpha)\right] \\ &-\Lambda_3(t)\alpha + \Lambda_4(t)\beta_3\frac{S_M(t)}{N} - \Lambda_5(t)\beta_3\frac{S_M(t)}{N}, \\ D^\alpha \Lambda_3(t) =& -1 - \Lambda_1(t)\left[-\beta_2(1-u(t))\frac{S_P(t)}{N}\right] - \Lambda_2(t)\left[\beta_2(1-u(t))\frac{S_P(t)}{N} - (\mu_P + \alpha\right]) \\ &-\Lambda_3(t)\left[(\mu_P - (1-v(t))\sigma) - \Lambda_4(t)((1-v(t))\sigma)\right], \\ D^\alpha \Lambda_4(t) =& \Lambda_4(t)\mu_P, \\ D^\alpha \Lambda_5(t) =& \Lambda_5(t)(\mu_M\ I_M(t) + \beta_4\frac{E_P(t)}{N}) - \Lambda_6(t)(\beta_4\frac{I_M(t)}{N}, \\ D^\alpha \Lambda_6(t) =& \Lambda_1(t)(\beta_3\frac{S_P(t)}{N}\ {}_4\frac{E_P(t)}{N}) - \Lambda_6(t)(\mu_M\ + \beta_4\frac{S_M(t)}{N}. \end{aligned}$$

The transversality conditions at the terminal time T are given by

$$\Lambda_1(T) = 0,\ \Lambda_2(T) = -1,\ \Lambda_3(T) = -1,\ \Lambda_4(T) = 0,\ \Lambda_5(T) = 0. \text{ and } \Lambda_6(T) = 0 \tag{8}$$

Moreover, for $t \in [0,T]$, the optimal controls $u^*(t)$ and $v^*(t)$ are expressed as

$$u^* = \min\left(u_{\max}, \max\left(u_{\min}, (\frac{(\Lambda_1-\Lambda_2)}{A} \times \beta_1 \frac{S_P^* E_P^*}{N}) + \beta_2 \frac{S_P^* I_P^*}{N}) + \beta_3 \frac{S_P^* I_M^*}{N}(t)\right)\right) \tag{9}$$

$$v^* = \min\left(v_{\max}, \max\left(v_{\min}, \frac{\Lambda_3(t)-\Lambda_4(t))}{B} \times I_P^*)(t)\right)\right). \tag{10}$$

Proof. Here we apply Pontryagin's principle [23] to solve the optimal control problem given below.

So, we define the Hamiltonian H as follows:

$$\begin{aligned}
H^*(t) =& E_P(T) + I_P(T) + \frac{A}{2}u^2(t) + \frac{B}{2}v^2(t) \\
&+ \Lambda_1(t)[\Lambda_P - \mu_P S_P(t) - \beta_1(1-u(t))\frac{S_P(t)E_P(t)}{N} \\
&- \beta_2(1-u(t))\frac{S_P(t)I_P(t)}{N} - \beta_3(1-u(t))\frac{S_P(t)I_M(t)}{N}] \\
&+ \Lambda_2(t)[\beta_1(1-u(t))\frac{S_P(t)E_P(t)}{N} + \beta_2(1-u(t))\frac{S_P(t)I_P(t)}{N} \\
&+ \beta_3(1-u(t))\frac{S_P(t)I_M(t)}{N} - (\mu_P+\alpha)E_P(t)] \\
&+ \Lambda_3(t)\left[\alpha E_P(t) - \sigma(1-v(t))I_P - (\mu_P+\delta)I_P(t)\right] \\
&+ \Lambda_4(t)\left[\sigma(1-v(t))I_P(t) - \mu_P R_P(t)\right] \\
&+ \Lambda_5(t)\left[\Lambda_T - \mu_T S_M(t) - \beta_3\frac{S_M(t)I_M(t)}{N}\right] \\
&+ \Lambda_6(t)\left[\beta_4\frac{S_M(t)I_M(t)}{N} - \mu_T I_M(t)\right]
\end{aligned}$$

For $t \in [0,T]$, the adjoint equations and transversality conditions can be obtained by using Pontryagin's maximum principle [23, 16, 13, 18] such that

$$\begin{aligned}
D^\alpha \Lambda_1(t) &= -\frac{\partial H(t)}{\partial S_P(t)}, \\
D^\alpha \Lambda_2(t) &= -\frac{\partial H(t)}{\partial E_P(t)}, \\
D^\alpha \Lambda_3(t) &= -\frac{\partial H(t)}{\partial I_P(t)}, \\
D^\alpha \Lambda_4(t) &= -\frac{\partial H(t)}{\partial R_P(t)},
\end{aligned}$$

$$D^{\alpha}\Lambda_5(t) = -\frac{\partial H(t)}{\partial S_M(t)},$$
$$D^{\alpha}\Lambda_6(t) = -\frac{\partial H(t)}{\partial I_M(t)},$$

with the transversality conditions at time T: $\Lambda_1(T) = 0,\ \Lambda_2(T) = -1,\ \Lambda_3(T) = -1,\ \Lambda_4(T) = 0;$ $\Lambda_5(T) = 0$ and $\Lambda_6(T) = 0$.

For, $t \in [0, T]$, the optimal controls u and v can be solved from the optimality condition

$$\begin{aligned}
-\frac{\partial H(t)}{\partial u(t)} = 0,\ \Rightarrow -Au(t) - (\Lambda_2(t) - \Lambda_1(t))\beta_1\frac{S_P^*(t)E_P^*(t)}{N}) + \beta_2\frac{S_P^*(t)I_P^*(t)}{N}) \\
+\beta_3\frac{S_P^*(t)I_M^*(t)}{N} = 0, \\
-\frac{\partial H(t)}{\partial v(t)} = 0, \Rightarrow -Bv(t) - (\Lambda_4(t) - \Lambda_3(t))I_P(t) = 0,
\end{aligned}$$

We have

$$u(t) = \frac{(\Lambda_1(t) - \Lambda_2(t))}{A} \times \beta_1\frac{S_P(t)E_P(t)}{N}) + \beta_2\frac{S_P(t)I_P(t)}{N}) + \beta_3\frac{S_P(t)I_M(t)}{N},$$

$$v(t) = \frac{\Lambda_3(t) - \Lambda_4(t))}{B} \times I_P(t).$$

By the bounds in U of the controls, it is easy to obtain that u^* and v^* are given by (9), (10) in the form of system (6). □

5 Simulation results

In this section, we present the results obtained by numerically solving the optimality system. In our control problem, we have initial conditions for the state variables and terminal conditions for the adjoints. That is, the optimality system is a two-point boundary value problem with separated boundary conditions at times step $i = 0$ and $i = T$. We solve the optimality system by an iterative method with forward solving of the state system followed by backward solving of the adjoint system. We start with an initial guess for the controls at the first iteration and then before the next iteration, we update the controls by using the characterization. We continue until convergence of successive iterates is achieved. A code is written and compiled in MATLAB by using the following Euler method and using the following data in Table 1.

Table 1: Parameter values used in numerical simulation

Paramter	value in d^{-1}
Λ_P	0.02
Λ_M	0.2
μ_P	0.1
μ_T	0.8
β_1	0.4
β_2	0.2
β_3	0.1
β_4	0.1
α	0.2
σ	0.2
δ	0.3

Since control and state functions are on different scales, the weight constant value is chosen as follows: $A = 100$, $B = 100$.

5.1 Strategy without control

In this strategy, we applied no control.

The data presented in Figure 2 and Table 1 indicate a substantial increase in both malaria infections and related fatalities, highlighting a worsening outbreak. This escalation can be attributed to ineffective containment measures, including delays in case identification and isolation, as well as the absence of widespread vaccination and preventive interventions. Consequently, the disease continues to spread rapidly, particularly in regions with limited healthcare infrastructure. Additionally, human behavioral factors, such as close physical contact and insufficient awareness of transmission routes, further facilitate viral dissemination. The prolonged infectious period also increases the risk of undetected secondary transmission.

These findings emphasize the necessity for immediate and effective public health interventions. Targeted strategies, such as mass vaccination campaigns, enhanced surveillance, and public awareness initiatives, are essential to mitigate further transmission. Implementing timely and evidence-based measures can significantly reduce infection rates, minimize mortality, and limit the overall impact of the outbreak. Failure to adopt such interventions may result in continued disease propagation.

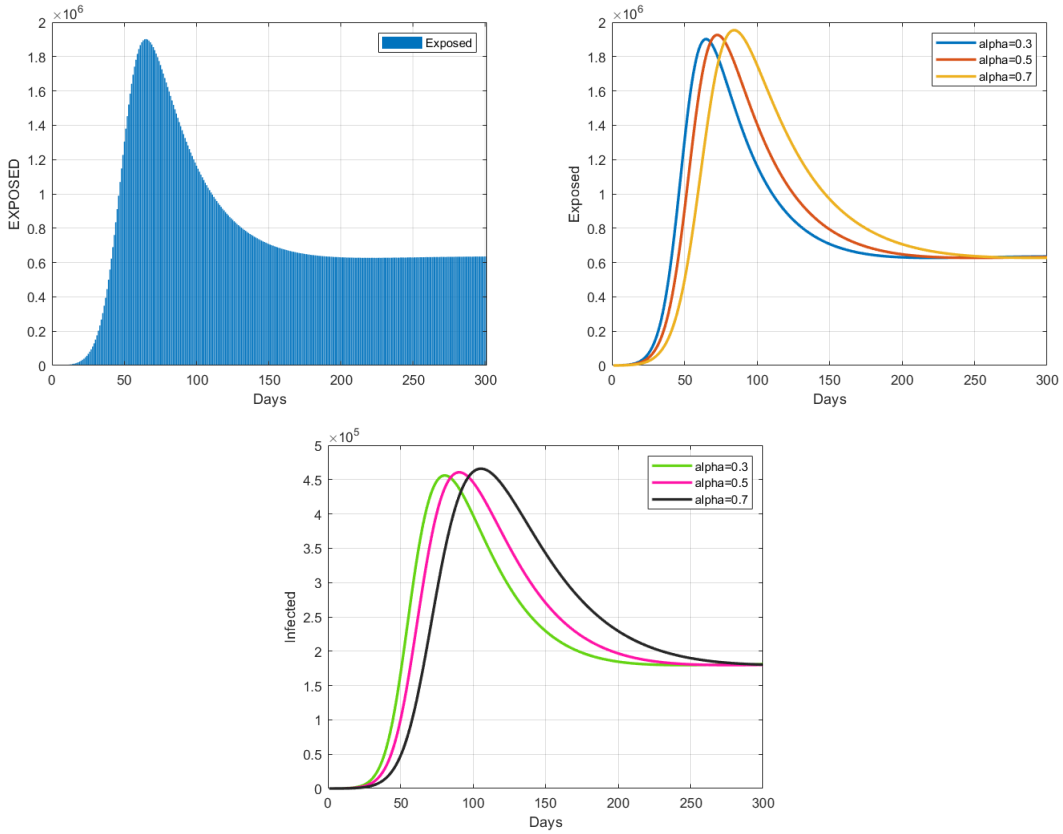


Figure 2: The evolution of the number of exposed and infected human without control

5.2 Strategy 1: Awareness program: Education

In this strategy, we used only the optimal control $u(t)$.

Figure 3 illustrates a significant reduction in the number of exposed and infected individuals following the implementation of sensitization campaigns aimed at at-risk populations. These campaigns focus on educating individuals about the risks associated with close contact with infected persons, thereby reducing the transmission rate of the malaria disease.

The numerical results indicate a clear distinction between scenarios with and without control measures. In the absence of control interventions, the number of exposed individuals ranges from 50 to 3.46×10^5 (for $\alpha = 0.3$) after 300 days. However, with control measures in place, this range decreases from 300 to 3.59×10^5 (for $\alpha = 0.5$), demonstrating the effectiveness of awareness programs in mitigating exposure. Similarly, the number of infected individuals without control

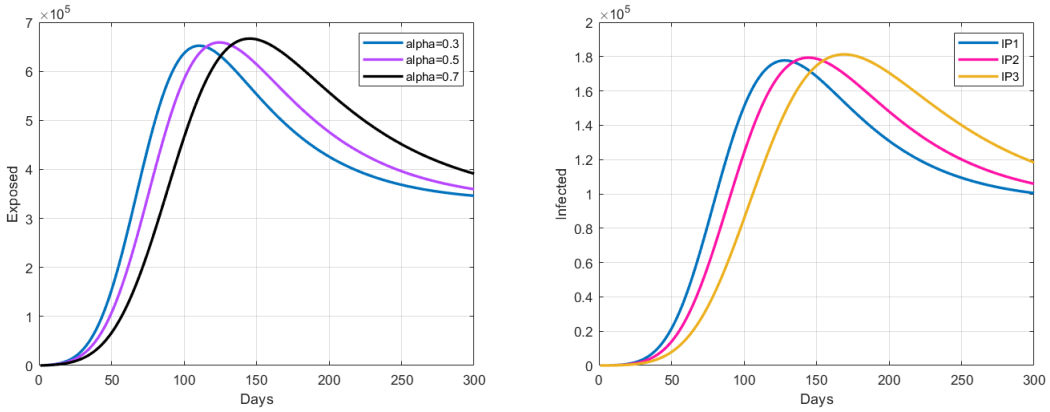


Figure 3: The evolution of the number of exposed and infected human without and with control $u(t)$

measures varies between 50 to 9.06×10^4 (for $\alpha = 0.7$). With control measures, this range is reduced to 10 to 8.14×10^4 (for $\alpha = 0.7$), indicating a decline in infection rates due to intervention strategies. These results provide strong evidence that sensitization campaigns play a crucial role in minimizing both exposure and infection rates, thereby improving disease control efforts.

In addition to the effectiveness of control measures, the study highlights the significance of fractional derivatives in modeling epidemiological dynamics. Fractional-order models incorporate memory effects, allowing for a more accurate representation of disease transmission over time. As the fractional derivative order α approaches 0.7, the memory effect within the system diminishes, resembling classical integer-order models. However, fractional derivatives provide a more generalized framework, capturing the long-term dependencies and non-local interactions inherent in real-world disease dynamics.

Figure 3 further demonstrates that when $\alpha = 0.7$, the memory effect within the system is maximized, leading to a prolonged period of reduced infection rates. This phenomenon underscores the importance of fractional calculus in epidemiological modeling, as it enables the incorporation of past infection dynamics into predictive models. By accounting for memory effects, fractional-order models offer a more comprehensive understanding of disease progression, enhancing the accuracy of predictions and the effectiveness of intervention strategies.

5.3 Strategy 2: Education and treatment

In This strategy, we use only the optimal control $v(t)$.

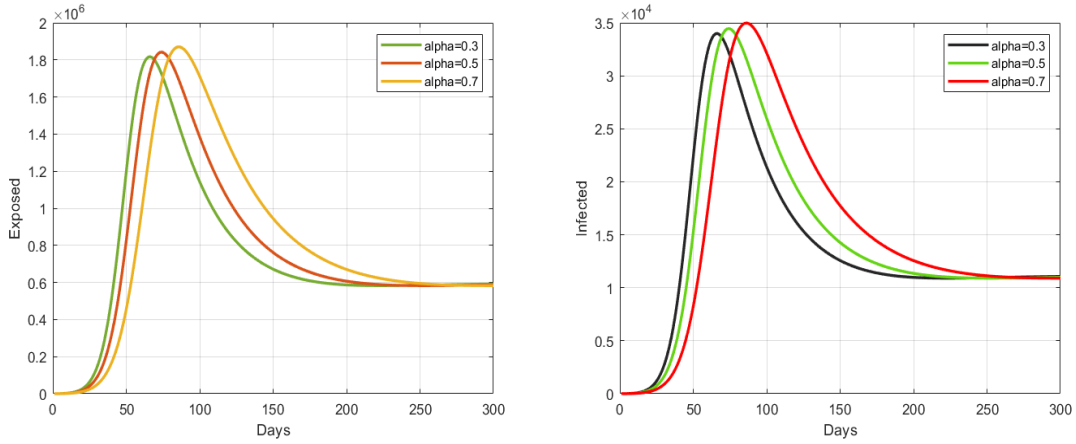


Figure 4: The evolution of the number of infected and exposed people with control $v(t)$

In Figure 4, we observe a significant reduction in the number of individuals infected with the malaria disease after 150 days. So we observe after a high increasing the number of exposed and infected were decreasing in to accepted value.

Number of Exposed Individuals:

Without control measures, the number of exposed individuals ranges from 150 to 1.85×10^6 (for $\alpha = 0.7$). With control measures, this range decreases to 50 (for $\alpha = 0.5$) to 2.61×10^6 (for $\alpha = 0.5$).

Number of Infected Individuals:

Without control measures, the number of infected individuals ranges from 10 to 4.32×10^4 (for $\alpha = 0.3$). With control measures, this range decreases to 100 to 6.04×10^4 (for $\alpha = 0.4$).

The implementation of control strategies, such as vaccination, antiviral treatments, and quarantine measures, has led to these favorable outcomes. These interventions have proven effective in reducing the spread of the malaria disease, leading to a significant decrease in the number of infected individuals. The results underscore the importance of timely and coordinated public health measures in managing and mitigating the impact of the outbreak.

6 Conclusion

In conclusion, Using fractal fractional-order derivative, we have successfully established theoretical and computational analysis for a malaria disease mathematical model, a fractional mathematical model, that describes the transmission of malaria in mosquito-to-human, which includes two controls. We also proposed three controls which, respectively, represent education and treat-

ment. We also studied the optimal control, led to reduce the spread of the malaria disease, by reducing the number of pigs infected and the number of ticks infected. We applied the results of the control theory, and we managed to obtain the characterization of the optimal control. The numerical simulation of the obtained results showing the effectiveness of the proposed control strategies. The entire investigation of this work revealed that control of the dynamic spread rate is vital for stopping the transmission of the spreading malaria disease.

These findings emphasize two key aspects: (1) the success of sensitization campaigns in reducing the spread of malaria and (2) the critical role of fractional derivatives in capturing the extended dynamics of disease outbreaks. Implementing awareness programs alongside advanced mathematical modeling approaches can significantly improve public health responses and mitigate future outbreaks. Ultimately, this research contributes to the broader understanding of disease transmission dynamics and offers an innovative approach to optimizing control strategies, which is essential for improving the management of current and future malaria outbreaks. Finally, we will develop this work in spatio temporels by studying the spread this disease in time and place.

Availability of data and materials

The disciplinary data used to support the findings of this study have been deposited in the Network Repository (http://www.networkrepository.com).

Competing interests

All the authors have no competing interests.

Authors' contributions

All authors contributed equally to the writing of this manuscript. All authors read and approved the final version

Funding

This research received no external funding.

Conflicts of Interest

The authors have no financial or proprietary interests in any material discussed in this article.

References

[1] Agrawal, O.P., *Formulation of Euler–Lagrange equations for fractional variational problems,* J. Math. Anal. Appl. 272(1) (2002), 368–379.

[2] Ahmed, E., El-Sayed, A. and El-Saka, H.A., *Equilibrium points, stability and numerical solutions of fractional-order predator–prey and rabies models,* J. Math. Anal. Appl. 325(1) (2007), 542–553.

[3] Alshehri, A., Shah, Z. and Jan, R., *Mathematical study of the dynamics of lymphatic filariasis infection via fractional calculus,* Eur. Phys. J. Plus 138(3) (2023), 1–15.

[4] Alyobi, S. and Jan, R., *Qualitative and quantitative analysis of fractional dynamics of infectious diseases with control measures,* Fractal Fract. 7(5) (2023), 400.

[5] Baleanu, D., Jajarmi, A., Defterli, O., Wannan, R., Sajjadi, S.S. and Asad, J.H., *Fractional investigation of time-dependent mass pendulum,* J. Low Freq. Noise Vib. Act. Control 43(1) (2024), 196–207.

[6] Baleanu, D., Jajarmi, A., Sajjadi, S.S. and Mozyrska, D., *A new fractional model and optimal control of a tumor-immune surveillance with non-singular derivative operator,* Chaos 29(8) (2019), 083127.

[7] Baleanu, D., Mohammadi, H. and Rezapour, S., *Analysis of the model of HIV-1 infection of $CD4^+$ T-cell with a new approach of fractional derivative,* Adv. Differ. Equ. 2020(1) (2020), 1–17.

[8] Baleanu, D., Shekari, P., Torkzadeh, L., Ranjbar, H., Jajarmi, A. and Nouri, K., *Stability analysis and system properties of Nipah virus transmission: A fractional calculus case study,* Chaos Solitons Fractals 166 (2023), 112990.

[9] Bonyah, E., Khan, M.A., Okosun, K.O. and Gomez-Aguilar, J.F., *Modelling the effects of heavy alcohol consumption on the transmission dynamics of gonorrhea with optimal control,* Math. Biosci. 309 (2019), 1–11.

[10] Bounouiga, S., Basti, B. and Benhamidouche, N., *Mathematical exploration of malaria transmission dynamics: Insights from fractional models and numerical simulatio*n, Adv. Theory Simul. 8(2) (2025), 2400630.

[11] Earl, P.L., Americo, J.L., Wyatt, L.S., Espenshade, O., Bassler, J., Gong, K., … and Moss, B., *Rapid protection in a malaria model by a single injection of a replication-deficient vaccinia disea*se, Proc. Natl. Acad. Sci. USA 105(31) (2008), 10889–10894.

[12] Helikumi, M. and Lolika, P.O., *Global dynamics of fractional-order model for malaria disease transmission,* Asian Res. J. Math. 18 (2022), 82–110.

[13] Kada, D., Kouidere, A., Balatif, O., Rachik, M. and Labriji, E.H., *Mathematical modeling of the spread of COVID-19 among different age groups in Morocco: Optimal control approach for intervention strategies,* Chaos Solitons Fractals (2020), 110437.

[14] Khajji, B., Kada, D., Balatif, O. and Rachik, M., *A multi-region discrete time mathematical modeling of the dynamics of COVID-19 disease propagation using optimal control,* J. Appl. Math. Comput. (2020). DOI: 10.1007/s12190-020-01354-3.

[15] Khan, M.A., Shah, S.W., Ullah, S. and Gomez-Aguilar, J.F., *A dynamical model of asymptomatic carrier Zika disease with optimal control strategies,* Nonlinear Anal. Real World Appl. 50 (2019), 144–170.

[16] Kouidere, A., Balatif, O., Ferjouchia, H., Boutayeb, A. and Rachik, M., *Optimal control strategy for a discrete time to the dynamics of a population of diabetics with highlighting the impact of living environment*, Discrete Dyn. Nat. Soc. (2019), 5949303.

[17] Kouidere, A., Balatif, O. and Rachik, M., *Analysis and optimal control of a mathematical modeling of the spread of African swine fever virus with a case study of South Korea and cost-effectiveness,* Chaos Solitons Fractals 146 (2021), 110867.

[18] Kouidere, A., Kada, D., Balatif, O., Rachik, M. and Naim, M., *Optimal control approach of a mathematical modeling with multiple delays of the negative impact of delays in applying preventive precautions against COVID-19 with a case study of Brazil and cost-effectiveness,* Chaos Solitons Fractals (2020), 110438.

[19] Kouidere, A., Khajji, B., El Bhih, A., Balatif, O. and Rachik, M., *A mathematical modeling with optimal control strategy of transmission of COVID-19 pandemic disease,* Commun. Math. Biol. Neurosci. 2020 (2020), Article ID 24.

[20] Kouidere, A., Labzai, A., Khajji, B., Ferjouchia, H., Balatif, O., Boutayeb, A. and Rachik, M., *Optimal control strategy with multi-delay in state and control variables of a discrete mathematical modeling for the dynamics of diabetic population,* Commun. Math. Biol. Neurosci. 2020 (2020), Article ID 14.

[21] Miller, K.S. and Ross, B., *An introduction to the fractional calculus and fractional differential eEquations*, Wiley, 1993.

[22] Podlubny, I., *Fractional differential equations*, Math. Sci. Eng. Vol. 198, Academic Press, 1999.

[23] Pontryagin, L.S., Boltyanskii, V.G., Gamkrelidze, R.V. and Mishchenko, E.F., The mathematical theory of optimal processes, Wiley, New York, 1962.

[24] Saha, A.K., Saha, S. and Podder, C.N., *Optimal control and cost-effectiveness analysis of a fractional order drug-resistant malaria transmission model with recovered carriers,* medRxiv (2025), 2025–05.

[25] UK Government, *Malaria outbreak: Epidemiological overview,* 28 June 2022. Available at: https://www.gov.uk/government/publications/malaria-outbreak-epidemiological-overview/malaria-outbreak-epidemiological-overview-28-june-2022.

[26] World Health Organization, Meeting of the International Health Regulations (2005) Emergency Committee regarding the multi-country malaria outbreak, 25 June 2022. Available at: https://www.who.int/ar/news/item/26-11-1443-meeting-of-the-international-health-regulations-(2005)-emergency-committee--regarding-the-multi-country-malaria-outbreak.

[27] World Health Organization, Malaria fact sheet, 26 August 2024. Available at: https://www.who.int/news-room/fact-sheets/detail/malaria.

[28] Zeb, A., Zaman, G., Jung, I.H. and Khan, M., *Optimal campaign strategies in fractional-order smoking dynamics,* Z. Naturforsch. A 69(5-6) (2014), 225–231.

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 1, 2026, pp 38–60
https://doi.org/10.22067/ijnao.2025.93661.1654
https://ijnao.um.ac.ir/

CC BY NC

Research Article

Numerical exploration of pollutant transport using stochastic fractional diffusion and Karhunen–Loève expansion

Z. Moniri, A. Babaei*, and B. Parsa Moghaddam

Abstract

This study explores a fractional time-space stochastic diffusion equation for modeling pollutant concentration, incorporating Caputo fractional derivatives and fractional Laplacians to capture anomalous diffusion. Stochastic noise, modeled via Brownian motion and Brownian bridges, is simulated using the Karhunen–Loève expansion (KLE). The equation's formulation, along with initial and boundary conditions, is presented. Analytical and numerical methods are discussed, emphasizing a hybrid framework combining fast Fourier transform, L1-algorithm, and KLE techniques. Numerical examples with sinc and Gaussian initial conditions highlight the superior accuracy and efficiency of the KLE approach over traditional Euler methods,

*Corresponding author

Received 23 May 2025; revised 6 September 2025; accepted 27 September 2025

Zahra Moniri
Department of Applied Mathematics, University of Mazandaran, Babolsar, Iran. e-mail: z.moniri05@stu.umz.ac.ir

Afshin Babaei
Department of Applied Mathematics, University of Mazandaran, Babolsar, Iran. e-mail: babaei@umz.ac.ir

Behrouz Parsa Moghaddam
Department of Mathematics, La.C., Islamic Azad University, Lahijan, Iran. e-mail: bparsa@fe.up.pt

This article was suggested by the scientific committee of the 17th seminar on differential equation and dynamical system for publication in IJNAO, which was accepted after independent review.

How to cite this article
Moniri, Z., Babaei, A. and Parsa Moghaddam, B., Numerical exploration of pollutant transport using stochastic fractional diffusion and Karhunen–Loève expansion. *Iran. J. Numer. Anal. Optim.*, 2026; 16(1): 38-60. https://doi.org/10.22067/ijnao.2025.93661.1654

revealing the significant influence of fractional parameters on pollutant dispersion dynamics and their potential for environmental modeling applications.

AMS subject classifications (2020): 35R60, 60G22, 65M75, 60H30, 76R50, 86A05

Keywords: Noninteger order probabilistic transport equation; Memory-dependent time derivatives; Eigenfunction decomposition methods; Contaminant transport simulation; Numerical approximation strategies.

1 Introduction

Fractional stochastic diffusion equations (FSDEs) have become essential in modeling intricate phenomena across physical, biological, and engineering domains. By extending classical diffusion models through fractional derivatives, these equations effectively describe anomalous diffusion, where particle spread deviates from standard patterns [9, 3]. The inclusion of stochastic components further enhances their capability to represent inherent randomness in various systems, such as fluid dynamics in porous media and particle movement in disordered structures [28, 30].

These equations play a crucial role in diverse applications. In ecology and biology, they model species dispersion and organism movement in heterogeneous environments, providing insights into ecosystem dynamics [3, 15]. Within engineering and physics, FSDEs describe heat, mass, and momentum transport in complex materials, as well as energy transfer in oscillator chains and particle behavior in nonequilibrium systems [13, 23]. Additionally, chemical engineering relies on FSDEs to analyze reaction-diffusion systems, helping predict chemical species' behavior in reactors and industrial processes [2, 10].

Addressing FSDEs requires sophisticated mathematical and computational techniques. Methods such as the tanh-coth approach, Caputo derivatives, and finite difference schemes provide both exact and approximate solutions [15, 1]. More advanced frameworks, including the Carathéodory approximation and the variational iteration method (VIM) combined with the Elzaki transform, enhance solution stability and convergence, ensuring their reliability in practical applications [28, 19].

Despite their broad applicability, FSDEs present significant computational challenges due to their complexity. Ongoing research focuses on refining numerical techniques and expanding their use in fields like finance and advanced materials science [25, 26]. Additionally, efforts are directed toward developing generalized models that incorporate stochastic resetting and fractional Fokker-Planck equations, further broadening their scope [6, 24].

Recent comprehensive reviews have systematically examined the rapid evolution of numerical methods for fractional stochastic partial differential equations, highlighting significant advances in spectral methods, finite element approaches, and their practical implementations across various disciplines [17]. Contemporary developments have extended these methodologies to multi-dimensional frameworks, establishing robust computational strategies that combine L1-algorithms with FFT-based approaches for handling complex spatio-temporal noise structures [18].

In this research, we investigate a fractional time-space stochastic diffusion equation (FTSSDE) designed to model the concentration of pollutants in a given medium. The governing equation is expressed as

$$ {}^{c}\partial_t^{\alpha} C(x,t) + \kappa_\beta(-\Delta)^{\beta/2} C(x,t) = \sigma\eta(x,t), \quad 0 < \alpha \leq 1, \quad 0 < \beta \leq 2, \tag{1} $$

where the spatial domain is $x \in [c, c+L]$, time $t > 0$, and $C(x,t)$ represents the pollutant concentration. The equation incorporates fractional derivatives in both time and space, capturing anomalous diffusion processes influenced by stochastic noise. The initial condition is specified as

$$ C(x,0) = C_0(x), \quad x \in [c, c+L], \tag{2} $$

indicating the initial distribution of the pollutant. The boundary condition is periodic, given by

$$ C(c,t) = C(c+L,t), \tag{3} $$

ensuring continuity of the concentration across the boundaries of the spatial domain.

The parameter κ_β governs the diffusivity, determining the rate at which the pollutant disperses through the medium. The fractional exponent β characterizes the type of diffusion: smaller values of β (closer to 0) correspond to super-diffusive behavior, where particles spread faster than in classical diffusion, while values approaching 2 indicate sub-diffusive behavior, where spreading is slower. The fractional time derivative, parameterized by α, introduces memory effects, allowing the model to account for historical states of the system. This framework is particularly useful for analyzing complex pollutant dynamics influenced by fractional and stochastic effects, such as those observed in heterogeneous or porous media.

The Caputo fractional time derivative ${}^{c}\partial_t^{\alpha} C(x,t)$ is defined for $0 < \alpha < 1$ as

$$ {}^{c}\partial_t^{\alpha} C(x,t) = \frac{1}{\Gamma(1-\alpha)} \int_0^t \frac{\partial C(x,\tau)}{\partial \tau} \frac{1}{(t-\tau)^{\alpha}} \, d\tau, \tag{4} $$

where Γ is the Gamma function, and the derivative accounts for nonlocal temporal effects. The spatial fractional Laplacian $(-\Delta)^{\beta/2}C(x,t)$, which governs the anomalous spatial diffusion, is defined in the one-dimensional case for $0 < \beta < 2$ as

$$(-\Delta)^{\beta/2}C(x,t) = -\frac{1}{2\cos(\pi\beta/2)}\left({}_cD_x^{\beta}C(x,t) + {}_xD_{c+L}^{\beta}C(x,t)\right), \tag{5}$$

where ${}_cD_x^{\beta}C(x,t)$ and ${}_xD_{c+L}^{\beta}C(x,t)$ are the left and right Riemann–Liouville fractional derivatives of order β, respectively. Let $n = \lceil\beta\rceil$, the smallest integer greater than or equal to β. These derivatives are given by

$${}_cD_x^{\beta}C(x,t) = \frac{1}{\Gamma(n-\beta)}\frac{d^n}{dx^n}\int_c^x (x-\xi)^{n-\beta-1}C(\xi,t)\,d\xi, \tag{6}$$

$${}_xD_{c+L}^{\beta}C(x,t) = \frac{(-1)^n}{\Gamma(n-\beta)}\frac{d^n}{dx^n}\int_x^{c+L} (\xi-x)^{n-\beta-1}C(\xi,t)\,d\xi. \tag{7}$$

These expressions capture the nonlocal spatial interactions, allowing the model to describe long-range dependencies in the pollutant's spread. The stochastic term $\eta(x,t) = W(x)\mathbb{B}_{\text{Bridge}}(t)$ introduces randomness into the system, where $W(x)$ represents a Brownian motion in the spatial coordinate, modeling random fluctuations in the pollutant source, and $\mathbb{B}_{\text{Bridge}}(t)$ is a Brownian bridge in time, ensuring that the noise is constrained to specific temporal behavior (e.g., returning to a reference state at the boundaries of the time interval).

This model extends traditional diffusion equations by incorporating fractional calculus and stochastic processes, making it suitable for scenarios where classical diffusion models fail, such as in turbulent flows, fractured aquifers, or atmospheric pollutant dispersion. The combination of fractional derivatives and stochastic noise enables the model to capture both memory effects and random environmental influences, providing a robust framework for studying complex transport phenomena. Potential applications include environmental engineering, where understanding pollutant dispersion is critical for designing mitigation strategies, and in hydrology, for modeling contaminant transport in groundwater systems. Future work could explore numerical methods for solving the FTSSDE or extend the model to higher dimensions for more realistic representations of physical systems.

Pollutant dispersion in natural environments often exhibits anomalous diffusion characteristics due to the complexity of the medium, such as soil heterogeneity, turbulent atmospheric layers, and porous geological formations. Classical diffusion models assume a Gaussian spread of pollutants, which fails to capture real-world scenarios where pollutants experience long-range correlations, memory effects, and heterogeneous transport mechanisms. The FTSSDE provides a more generalized framework, incorporating both temporal and spatial fractional derivatives

to model nonlocal effects and power-law decay, leading to a more accurate representation of pollutant concentration in diverse environmental settings [7, 11, 27].

Recent advances in atmospheric science have demonstrated the effectiveness of fractional stochastic advection-diffusion frameworks for modeling moisture transport at ocean-atmosphere interfaces, where temporal memory effects and nonlocal turbulent diffusion mechanisms play crucial roles in boundary layer dynamics [16]. These developments underscore the growing recognition that fractional stochastic models provide superior accuracy in capturing environmental transport phenomena compared to classical approaches.

Environmental factors such as fluctuating wind patterns, temperature variations, and chemical reactions introduce inherent randomness into pollutant transport. Traditional deterministic models cannot adequately account for such uncertainties, making stochastic modeling essential. FTSSDEs integrate stochastic noise to reflect the probabilistic nature of diffusion, ensuring realistic predictions of pollutant spread under varying conditions. This stochastic component enables the model to describe the erratic dispersion of pollutants in urban air pollution, riverine contamination, and subsurface transport, where external influences create unpredictable deviations in concentration levels [29, 31].

The incorporation of stochastic components in pollutant transport modeling is not merely a mathematical abstraction but reflects fundamental physical realities that govern environmental systems. Environmental variability introduces inherent randomness through fluctuating meteorological conditions, such as turbulent wind patterns, temperature gradients, and precipitation variability, which directly affect dispersion rates and directions. Additionally, subsurface heterogeneity creates random spatial variations in hydraulic conductivity, porosity, and chemical reaction rates that cannot be deterministically characterized at practical scales. Model uncertainty arises from incomplete knowledge of initial pollutant distributions, boundary conditions, and measurement noise in monitoring systems. Furthermore, multi-scale interactions between molecular diffusion, advective transport, and chemical transformations exhibit stochastic behavior due to the complex interplay between microscopic and macroscopic processes. From a regulatory perspective, probabilistic risk assessment requires stochastic modeling to quantify uncertainty bounds and confidence intervals, enabling robust environmental management decisions under uncertainty. Traditional deterministic models fail to capture these random fluctuations, potentially leading to significant underestimation or overestimation of pollutant concentrations, which can have serious implications for public health risk assessment and environmental remediation strategies.

In addition to stochasticity and anomalous diffusion, pollutant transport often occurs in highly heterogeneous media where physical and chemical properties vary significantly over space and time. FTSSDEs allow for flexible adaptation to these conditions by tuning the order of frac-

tional derivatives, effectively capturing the influence of porous structures, adsorption-desorption dynamics, and multi-phase flow interactions. This adaptability makes FTSSDEs particularly valuable for predicting long-term pollutant behavior in groundwater contamination, atmospheric dispersion modeling, and large-scale ecological impact assessments. The development of robust numerical and analytical methods further enhances their applicability, making FTSSDEs a powerful tool for environmental scientists and engineers [12, 14].

To address the computational challenges posed by FTSSDEs, we develop a hybrid numerical framework that combines three sophisticated techniques: (1) Fast Fourier transform (FFT) implementation for efficient spatial discretization of fractional Laplacians, exploiting the spectral properties of these operators in frequency domain; (2) L1-algorithm for robust temporal approximation of Caputo fractional derivatives, providing first-order accuracy with optimal computational complexity; and (3) Karhunen–Loève expansion (KLE) for accurate simulation of stochastic noise components, offering superior convergence properties compared to traditional Euler-Maruyama schemes. This integrated approach enables us to achieve computational efficiency while maintaining high numerical accuracy for complex fractional stochastic systems.

The primary contributions of this work include: (i) demonstrating that KLE-based noise simulation achieves mean squared errors (MSEs) approximately three orders of magnitude lower than conventional Euler methods; (ii) establishing superior convergence rates for the KLE approach (orders 1.57–5.43) compared to Euler schemes (orders 1.19–2.73); (iii) proving computational efficiency gains where KLE with 128 grid points outperforms Euler with 512 points; and (iv) providing a comprehensive analysis of fractional parameter effects on pollutant dispersion dynamics in stochastic environments.

The remainder of this article is organized as follows. Section 2 provides the preliminary concepts, including definitions of Brownian motion, Brownian bridge, and the KLE, which form the foundation of our stochastic modeling approach. Section 3 details the KLE simulation of Brownian sheet noise, supported by visualizations of the stochastic process. Section 4 presents the implementation of the FFT for solving the FTSSDE, alongside the L1-algorithm for approximating the Caputo fractional derivative. Section 5 discusses the numerical implementation, including computational details and illustrative examples with sinc initial conditions. Finally, Section 6 concludes the study, summarizing key findings and outlining potential applications and future research directions in environmental modeling.

2 Preliminaries

Modeling pollutant dispersion in the environment requires mathematical frameworks that effectively capture randomness, nonstationarity, and spatial-temporal correlations. Brownian motion, Brownian bridge, and the KLE provide significant advantages in this regard. Brownian motion (Definition 1) serves as a fundamental stochastic model for the random diffusion of pollutants in fluids, with extensions such as fractional Brownian motion allowing for greater flexibility in modeling persistence and roughness [14, 22]. The Brownian bridge (Definition 2), a constrained variant of Brownian motion, is particularly useful when pollutant concentrations are known at specific times or locations, ensuring consistency with boundary conditions and enhancing statistical inferences [14, 5].

Definition 1 (Brownian motion [14])**.** A stochastic process $\{W(t)\}_{t\geq 0}$ is termed a Brownian motion if it constitutes a real-valued Gaussian process satisfying the following conditions:

i) Its sample paths are continuous functions of t;

ii) It possesses a zero mean function, that is, $\mathbb{E}[W(t)] = 0$ for all $t \geq 0$;

iii) Its covariance structure is characterized by $\mathrm{Cov}(W(s), W(t)) = \min(s, t)$ for all $s, t \geq 0$.

Definition 2 (Brownian Bridge [14])**.** A stochastic process $\{\mathbb{B}_{\mathrm{Bridge}}(t)\}_{t\in[0,T_f]}$ is termed a Brownian bridge if it constitutes a Gaussian process with the following properties:

i) It satisfies the boundary constraints $\mathbb{B}_{\mathrm{Bridge}}(0) = \mathbb{B}_{\mathrm{Bridge}}(T_f) = 0$ almost surely;

ii) It has a zero mean function, that is, $\mathbb{E}[\mathbb{B}_{\mathrm{Bridge}}(t)] = 0$ for all $t \in [0, T_f]$;

iii) Its covariance structure is characterized by $\mathrm{Cov}(\mathbb{B}_{\mathrm{Bridge}}(s), \mathbb{B}_{\mathrm{Bridge}}(t)) = \min(s, t) - st$ for all $s, t \in [0, T_f]$.

3 Karhunen–Loève expansion

The KLE is a powerful technique for reducing the dimensionality of stochastic processes, extracting dominant stochastic components from observed data to construct efficient empirical models. By providing accurate and computationally efficient representations of complex environmental processes, KLE plays a crucial role in stochastic simulations and predictive modeling [4, 20, 21]. These methods collectively improve the precision, adaptability, and computational feasibility of pollutant modeling, enabling more robust analysis and prediction of environmental phenomena.

For a stochastic process $U(x,\omega)$ defined on the spatial domain $x \in \mathcal{R}$, the KLE provides a representation using orthogonal functions and uncorrelated random variables.

If we denote the mean function as $\mu(x) = \mathbb{E}[U(x,\omega)]$, we can focus on the centered process $U(x,\omega) - \mu(x)$. Our goal is to represent the sample paths of this centered process as a series expansion using an orthonormal basis $\{\phi_n : n \in \mathbb{N}\}$ of $L^2(\mathcal{R})$:

$$U(x,\omega) - \mu(x) = \sum_{n=1}^{\infty} \sqrt{\lambda_n}\phi_n(x)\xi_n(\omega). \tag{8}$$

The random coefficients $\xi_n(\omega)$ in this expansion are given by

$$\xi_n(\omega) := \frac{1}{\sqrt{\lambda_n}} \langle U(x,\omega) - \mu(x), \phi_n(x) \rangle_{L^2(\mathcal{R})}. \tag{9}$$

Let us denote the covariance function of $U(x,\omega)$ as $\mathrm{Cov}(x,y)$. We can then define an integral operator $\mathcal{C}$ by

$$(\mathscr{C}\phi)(x) := \int_{\mathcal{R}} \mathcal{C}ov(x,y)\phi(y)\,dy, \quad \phi \in L^2(\mathcal{R}). \tag{10}$$

This framework allows us to decompose the stochastic process into deterministic spatial modes $\phi_n(x)$ and random coefficients $\xi_n(\omega)$.

The KLE is given by (8) when the functions ϕ_n are chosen as the eigenfunctions of the operator $\mathscr{C}[\cdot]$. This approach represents a stochastic process as an infinite series of orthogonal basis functions, each weighted by uncorrelated random variables. The values λ_n are the associated eigenvalues of the operator $\mathscr{C}[\cdot]$.

3.1 Karhunen–Loève simulation of Brownian sheet noise $W(x)\mathbb{B}_{\mathrm{Bridge}}(t)$

Brownian Sheet noise, denoted as $W(x)\mathbb{B}_{\mathrm{Bridge}}(t)$, is the product of a Wiener process $W(x)$ and a Brownian bridge $\mathbb{B}_{\mathrm{Bridge}}(t)$. To derive its KLE, we can combine the expansions of its components.

First, recall the KLE of the Brownian bridge $\mathbb{B}_{\mathrm{Bridge}}(t)$ on $T_f = [0,1]$:

$$\mathbb{B}_{\mathrm{Bridge}}(t) = \sum_{j=1}^{\infty} \sqrt{2}\frac{\sin(j\pi t)}{\pi j}\xi_j, \quad \xi_j \sim \mathcal{N}(0,1) \text{ iid}, \tag{11}$$

and the KLE of the Wiener process $W(x)$ on $x \in [0,1]$:

$$W(x) = \sum_{n=1}^{\infty} \frac{2\sqrt{2}\sin\left(\left(n-\frac{1}{2}\right)\pi x\right)}{(2n-1)\pi}\eta_n, \quad \eta_n \sim \mathcal{N}(0,1) \text{ iid.} \tag{12}$$

To obtain the KLE of Brownian sheet noise $W(x)\mathbb{B}_{\text{Bridge}}(t)$, we multiply these two expansions:

$$W(x)\mathbb{B}_{\text{Bridge}}(t) = \sum_{n=1}^{\infty}\sum_{j=1}^{\infty} \frac{4\sin\left(\left(n-\frac{1}{2}\right)\pi x\right)\sin(j\pi t)}{(2n-1)\pi^2 j}\xi_j\eta_n. \tag{13}$$

This double series represents the KLE of Brownian sheet noise $W(x)\mathbb{B}_{\text{Bridge}}(t)$. Note that η_n and ξ_j are independent standard normal random variables.

To approximate sample paths of Brownian sheet noise on $[0,1]\times[0,1]$, we can truncate both series:

$$W(x)\mathbb{B}_{\text{Bridge}}(t) \approx W^N(x)\mathbb{B}^J_{\text{Bridge}}(t) = \sum_{n=1}^{N}\sum_{j=1}^{J} \frac{4\sin\left(\left(n-\frac{1}{2}\right)\pi x\right)\sin(j\pi t)}{(2n-1)\pi^2 j}\xi_j\eta_n, \tag{14}$$

where N and J are the number of terms used in the truncation for the Wiener process and Brownian bridge, respectively.

This expansion allows for efficient simulation of Brownian sheet noise by generating the required number of standard normal random variables and computing the finite sum. The accuracy of the approximation improves as N and J increase.

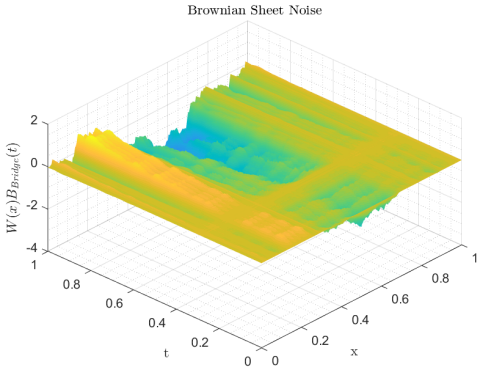
The Brownian sheet $W(x)\mathbb{B}_{\text{Bridge}}(t)$ were examined using parameters of 2^8 for the Wiener process terms (N), Brownian bridge terms (J), and spatial and temporal points. These processes were analyzed over a unit square domain. KLE simulation was employed to visualize these processes, as depicted in Figure 1. A three-dimensional surface plot in Subfigure 1a illustrates the spatial variations of the Brownian sheet, while Subfigure 1b provides a two-dimensional contour plot, elucidating the distribution of these variations across the domain.

4 Fast Fourier transform implementation for FTSSDE

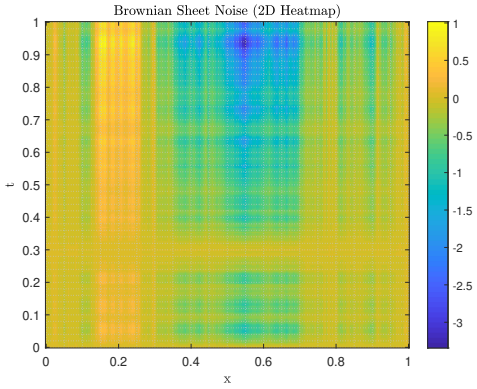
To implement the FFT for the FTSSDE described in (1), we adopt the following detailed procedure:

First, we compute the FFT of each term in (1) with respect to the spatial coordinate x. For the Caputo fractional time derivative ${}^c\partial_t^\alpha C(x,t)$, the FFT with respect to x is expressed as

$$\mathcal{F}_x[{}^c\partial_t^\alpha C(x,t)] = {}^c\,\partial_t^\alpha \hat{C}(k,t),$$



(a) Realization of the Brownian sheet $W(x)\mathbb{B}_{\text{Bridge}}(t)$ using KLE simulation.



(b) Contour plot of the Brownian sheet $W(x)\mathbb{B}_{\text{Bridge}}(t)$ using KLE simulation.

Figure 1: Visual representations of the Brownian sheet generated via KLE simulation. (a) A 3D surface plot depicting the Brownian sheet. (b) A contour plot illustrating the Brownian sheet.

where $\hat{C}(k,t) = \mathcal{F}_x[C(x,t)]$ represents the Fourier transform of $C(x,t)$ in the spatial domain, and k denotes the wavenumber, which characterizes the spatial frequency.

For the fractional Laplacian term $(-\Delta)^{\beta/2}C(x,t)$, the FFT is given by

$$\mathcal{F}_x[(-\Delta)^{\beta/2}C(x,t)] = |k|^{\beta}\hat{C}(k,t),$$

where $|k|$ is the magnitude of the wavenumber, reflecting the scaling of the fractional Laplacian in the frequency domain.

Assuming the noise term $\eta(x,t)$ is spatially uncorrelated, its FFT is represented as $\hat{\eta}(k,t)$, preserving the stochastic properties in the frequency domain.

Applying the FFT with respect to x to both sides of (1), we obtain

$$\mathcal{F}_x\left[{}^c\partial_t^{\alpha}C(x,t)\right] = \mathcal{F}_x\left[-\kappa_{\beta}(-\Delta)^{\beta/2}C(x,t) + \sigma\eta(x,t)\right].$$

By substituting the Fourier transforms of the individual terms, the equation becomes

$${}^c\partial_t^{\alpha}\hat{C}(k,t) + \kappa_{\beta}|k|^{\beta}\hat{C}(k,t) = \sigma\hat{\eta}(k,t).$$

This transformed equation describes the dynamics of the system in the frequency-wavenumber domain, highlighting the interaction between the Caputo fractional time derivative and the spatial fractional Laplacian. The use of the one-dimensional FFT enables efficient computation of these spectral representations, significantly reducing the computational complexity for

large-scale or high-dimensional datasets. Recent computational advances have extended these FFT-based methodologies to two-dimensional fractional stochastic systems, demonstrating superior efficiency in handling spatio-temporal noise structures through hybrid L1-FFT approaches [18]. This integration of spectral methods with fractional calculus represents a significant advancement in numerical techniques for solving complex environmental transport equations. This approach not only facilitates numerical simulations but also provides deeper insights into the fractional and stochastic behaviors of the system, making it particularly suitable for applications involving anomalous diffusion or complex spatial structures.

4.1 Numerical approximation of the Caputo fractional derivative for $C(x,t)$

Consider a uniform partition of the time interval $[0, T_f]$ with time step Δt. The Caputo fractional derivative at time t_n is approximated using the L1-algorithm, as described in [8]. For a fractional order $0 \leq \alpha < 1$, the Caputo fractional derivative is defined as

$$ {}^{c}\partial_t^{\alpha} C(x,t_n) = \frac{1}{\Gamma(1-\alpha)} \int_{t_0}^{t_n} \frac{\partial C(x,\tau)}{\partial \tau} (t_n - \tau)^{-\alpha} d\tau. $$

This integral is discretized to yield the following approximation:

$$ {}^{c}\partial_t^{\alpha} C(x,t_n) \approx \frac{\Delta t^{-\alpha}}{\Gamma(2-\alpha)} \sum_{j=0}^{n-1} b_j^{(\alpha)} \left(C(x,t_{n-j}) - C(x,t_{n-j-1}) \right) + \mathcal{O}(\Delta t), \tag{15} $$

where the coefficients are given by $b_j^{(\alpha)} = (j+1)^{1-\alpha} - j^{1-\alpha}$.

This formulation enables an efficient and straightforward implementation of the L1-algorithm for approximating the Caputo fractional derivative of $C(x,t)$. By integrating this temporal discretization with spatial discretization techniques, such as the FFT for computing fractional Laplacians, we obtain a robust and computationally effective approach for solving FTSSDEs.

The flowchart in Figure 2 illustrates a comprehensive numerical algorithm for solving FTSSDEs that feature both temporal and spatial fractional operators. The algorithm begins with parameter initialization and proceeds through two parallel noise simulation approaches—the KLE for spectral accuracy and Euler-Maruyama for computational efficiency—before constructing the stochastic noise term. The core computational loop employs Fourier transforms to handle the fractional Laplacian operator in frequency space, implements the L1 algorithm for the time-

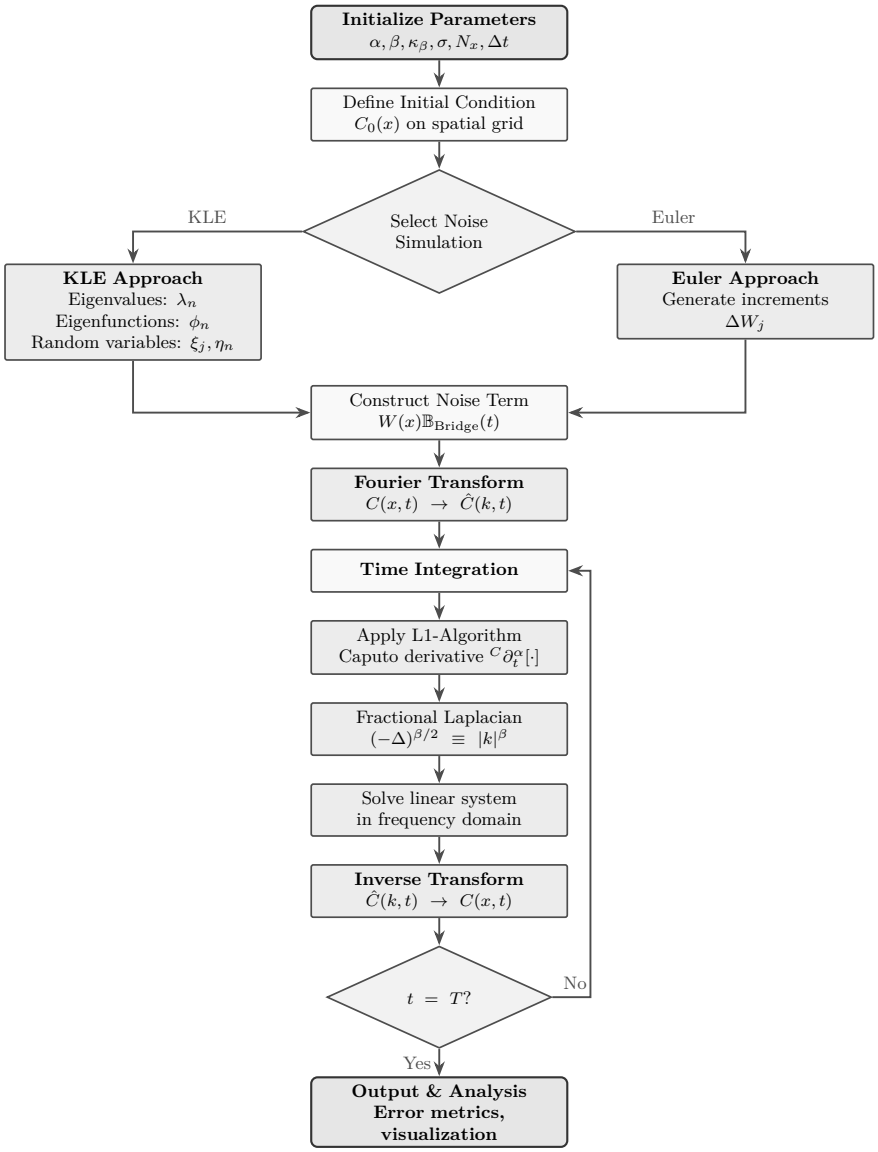
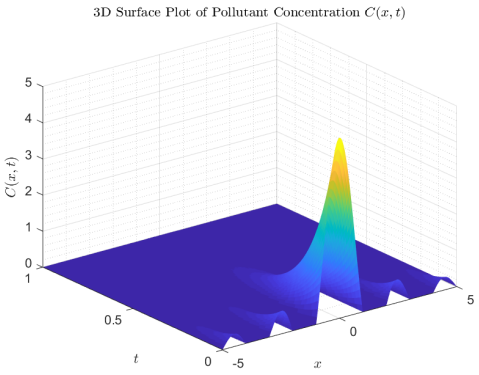
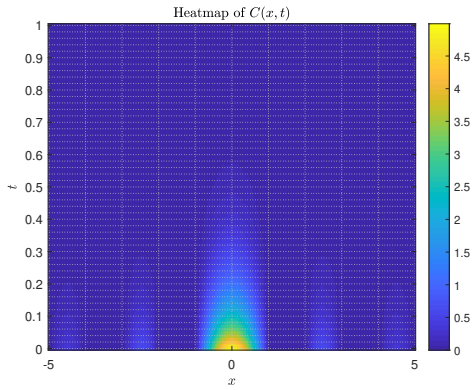


Figure 2: Numerical algorithm flowchart for solving the FTSSDEs with time-fractional Caputo derivative and space-fractional Laplacian operator. The algorithm employs Fourier spectral methods combined with either KLE or Euler-Maruyama discretization for stochastic noise simulation.

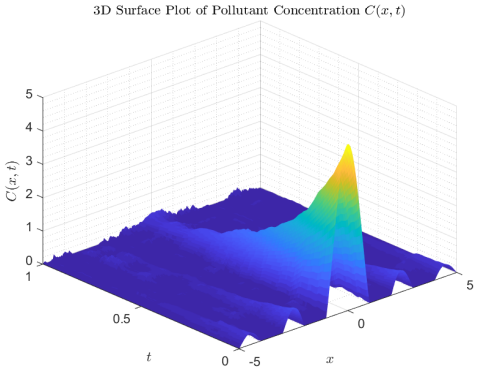
fractional Caputo derivative, and iteratively advances the solution until the final time is reached, making it suitable for modeling anomalous diffusion processes with memory effects and heavy-tailed spatial distributions.



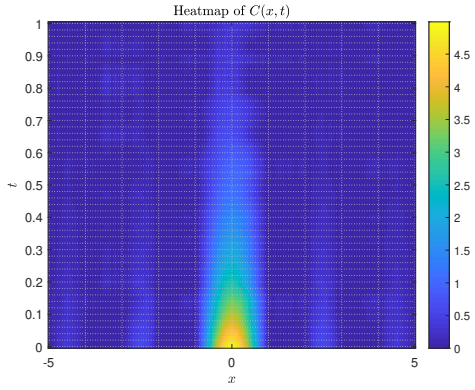
(a) Three-dimensional surface plot of pollutant concentration $C(x,t)$ using Euler Brownian noise simulation.



(b) Contour plot of pollutant concentration $C(x,t)$ with Euler simulation of Brownian noise.



(c) Three-dimensional surface plot of pollutant concentration $C(x,t)$ using KLE Brownian noise simulation.



(d) Contour plot of pollutant concentration $C(x,t)$ with KLE Brownian noise simulation.

Figure 3: Example 1: Three-dimensional surface representations and contour plots displaying pollutant concentration $C(x,t)$ over space and time, using the sinc function as the initial condition.

5 Numerical implementation

To solve the FTSSDEs efficiently, we employed MATLAB's `fft` and `ifft` functions to handle the spatial discretization of partial differential equations via the FFT. This approach leverages the computational efficiency of FFT to approximate fractional Laplacians, significantly reducing the computational cost for large-scale simulations. To ensure the reproducibility of our results,

we initialized the random number generator with a fixed seed of 123 using MATLAB's Mersenne Twister algorithm. This configuration ensures consistent generation of pseudo-random numbers across multiple runs, allowing for reliable replication of results under identical conditions. By fixing the seed, we mitigate variability in stochastic simulations, which is critical for validating the robustness of our proposed scheme and for computing stable statistical metrics, such as confidence intervals. This methodology strikes an optimal balance between numerical accuracy and computational efficiency, enabling precise statistical analysis without excessive resource demands. All simulations were conducted on a system equipped with an Intel(R) Core(TM) i7-7500U CPU (2.70–2.90 GHz), 12.0 GB RAM, and a 64-bit operating system, ensuring sufficient computational power for the tasks.

Example 1. To illustrate the performance of our numerical scheme, we consider the FTSSDE given by (1) in a one-dimensional spatial domain. The equation is parameterized as follows: the fractional derivative order is set to $\alpha = 0.95$, reflecting a near-integer temporal memory effect; the fractional Laplacian order is $\beta = 1.35$, which governs the spatial nonlocality; the diffusion coefficient is $\kappa_\beta = 0.01$, controlling the strength of diffusion; and the noise standard deviation is $\sigma = 0.3$, representing the intensity of stochastic forcing. The spatial domain is defined as $[c, c+L]$ with a length of L units, and the simulation runs over a total time of T_f. The FTSSDE is expressed as

$$ {}^{c}\partial_t^{\alpha} C(x,t) + \kappa_\beta(-\Delta)^{\beta/2} C(x,t) = \sigma W(x)\mathbb{B}_{\text{Bridge}}(t), \quad x \in [c, c+L],\, t \in [0, T_f], \tag{16} $$

where $W(x)$ denotes a spatial white noise process, and $\mathbb{B}_{\text{Bridge}}(t)$ represents a Brownian bridge process, introducing temporal stochasticity. The initial condition for the pollutant concentration $C(x,t)$ is specified as

$$ C(x,0) = C_0(x) = 5\text{sinc}(x), \quad x \in [c, c+L], \tag{17} $$

where the $\text{sinc}(x) = \frac{\sin(x)}{x}$ function provides a smooth, oscillatory profile shifted by 0.5 to ensure nonnegativity, mimicking a realistic initial distribution of a pollutant. The boundary conditions are as defined in (3), typically enforcing periodicity or zero-flux conditions to suit the physical context of the problem. This setup allows us to evaluate the numerical scheme's ability to capture the complex interplay between fractional derivatives, diffusion, and stochastic noise in a controlled yet challenging scenario.

Employing the sinc function as the initial condition provides a novel framework for studying pollutant dispersion dynamics. The $\text{sinc}(x) = \frac{\sin(x)}{x}$ function, characterized by a pronounced central peak and decaying oscillatory tails, serves as an effective model for capturing the spatial heterogeneity of pollutant concentrations, particularly in scenarios involving localized sources or

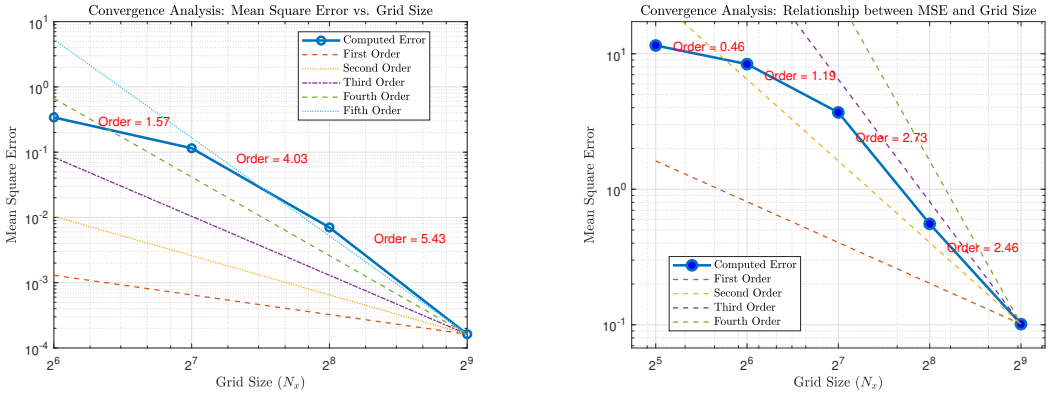
asymmetric distributions. This choice facilitates a detailed exploration of how pollutant sources interact with environmental factors and dispersion mechanisms, offering valuable insights into the complex behavior of environmental pollution. By leveraging the sinc function's unique properties, our approach enhances the ability to analyze the influence of fractional derivatives, stochastic noise, and diffusion processes, contributing to a deeper understanding of pollutant transport in diverse ecological settings.

In Example 1, Figure 3 illustrates the temporal evolution of the pollutant concentration $C(x,t)$ within a one-dimensional spatial domain $[-5,5]$, using the sinc function as the initial condition. The visualizations include three-dimensional surface plots and contour plots, generated through simulations employing both the Euler method and the KLE for Brownian noise. The domain, with a length of $L = 10$ units, is discretized into $N_x = 256$ spatial points to ensure high resolution in capturing spatial variations. The simulation spans a total time of $T_f = 1$ unit, divided into time steps of $\Delta t = 0.01$, providing fine temporal granularity for accurate tracking of the solution's evolution. For the KLE-based simulation, 512 terms are utilized to represent the stochastic noise, ensuring a robust approximation of the Brownian bridge process. These plots effectively demonstrate the numerical scheme's capability to model the intricate dynamics of pollutant dispersion under the influence of fractional operators and stochastic forcing, highlighting differences in the behavior of $C(x,t)$ between the Euler and KLE approaches.

Table 1: Convergence analysis for Euler and KLE simulations

	Euler Simulation		**KLE Simulation**	
Grid Size	**MSE**	**Convergence Order**	**MSE**	**Convergence Order**
512	0.1011	–	1.6259×10^{-4}	–
256	0.55561	2.4583	7.0254×10^{-3}	5.4333
128	3.6878	2.7306	0.11504	4.0335
64	8.3852	1.1851	0.34073	1.5665

Table 1 and Figure 4 present a comparative convergence analysis between KLE and Euler simulations for Brownian noise across uniform grid sizes from 64 to 512 points. The KLE simulation demonstrates markedly superior performance, achieving MSE values approximately three orders of magnitude lower than the Euler method at equivalent grid resolutions (e.g., 1.6×10^{-4} versus 0.1011 at 512 grid points). Moreover, the KLE approach exhibits exceptional convergence properties, with orders ranging from 1.5665 to 5.4333, substantially outperforming the Euler method's more modest convergence rates of 1.1851 to 2.7306. This remarkable difference in convergence behavior is clearly visualized in the log-log plots of Figure 4, where the steeper slope of the KLE simulation curve indicates its superior convergence rate.



(a) Convergence analysis for KLE simulation

(b) Convergence analysis for Euler simulation

Figure 4: Convergence analysis showing the relationship between MSE and uniform grid resolution N_x. The analysis compares performance of the proposed method using both KLE and Euler approaches for Brownian noise simulation. Reference lines indicating different rates of convergence (ROCs) are included for comparison.

These findings have significant computational implications, as the KLE method can achieve comparable or superior accuracy with considerably coarser grids compared to the Euler approach. For instance, a KLE simulation with just 128 grid points ($MSE \approx 0.115$) delivers better accuracy than an Euler simulation with 512 points ($MSE \approx 0.101$), potentially yielding substantial computational savings. The superior performance of the KLE method can be attributed to its spectral decomposition approach, which effectively captures the essential stochastic features of Brownian motion, in contrast to the Euler method's incremental approximation strategy. This fundamental difference in methodology results in both enhanced accuracy and improved convergence properties, making the KLE approach particularly advantageous for applications requiring high precision numerical solutions of stochastic processes.

Example 2. To further validate the numerical scheme and explore the influence of different initial conditions on pollutant dispersion, we consider the FTSSDE given by (1) with a Gaussian initial distribution. The equation parameters remain consistent with Example 1: the fractional derivative order is $\alpha = 0.95$, the fractional Laplacian order is $\beta = 1.35$, the diffusion coefficient is $\kappa_\beta = 0.01$, and the noise standard deviation is $\sigma = 0.2$.

The initial condition for the pollutant concentration $C(x,t)$ is specified as a Gaussian distribution:

$$C(x,0) = C_0(x) = A \exp\left(-\frac{(x-x_0)^2}{2\sigma_{\text{gauss}}^2}\right) + C_{\text{medium}}, \quad x \in [c, c+L], \tag{18}$$

where $A = 5.0$ represents the peak concentration at the center of the distribution, $x_0 = 0$ is the center coordinate of the Gaussian profile, $\sigma_{\text{gauss}} = 1.5$ controls the spread of the initial

pollutant distribution, and $C_{\text{medium}} = 0.01$ denotes the background concentration level in the medium. This Gaussian profile models a realistic scenario where pollutants are initially concentrated around a point source with smooth decay away from the center, commonly observed in environmental contamination from localized industrial discharge or chemical spills.

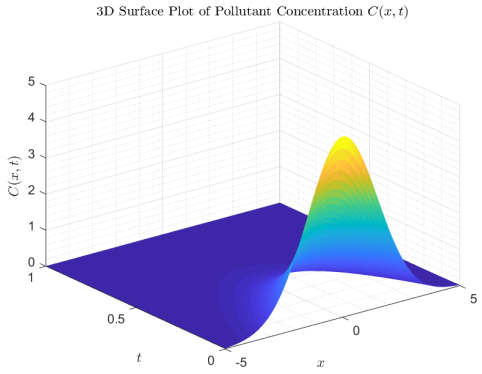
The Gaussian initial condition provides an ideal test case for evaluating the numerical scheme's performance in handling smooth, bell-shaped distributions that are prevalent in environmental modeling. Unlike the oscillatory nature of the sinc function used in Example 1, the Gaussian profile offers a monotonic decay from the central peak, allowing for a clearer observation of how fractional derivatives and stochastic noise influence the spreading and evolution of pollutant plumes over time. The boundary conditions follow (3), maintaining consistency with the periodic domain setup employed throughout this study.

In Example 2, Figure 5 presents the temporal evolution of the pollutant concentration $C(x,t)$ within the same one-dimensional spatial domain $[-5,5]$ as in Example 1, but using the Gaussian distribution as the initial condition. The domain maintains a length of $L = 10$ units, discretized into $N_x = 256$ spatial points, with the simulation spanning a total time of $T_f = 1$ unit using time steps of $\Delta t = 0.01$. The visualizations include both three-dimensional surface plots and contour plots, comparing simulations employing the Euler method and the KLE for Brownian noise representation. For the KLE-based simulation, 512 terms are utilized to ensure robust approximation of the stochastic processes. These plots demonstrate the smooth evolution of the initially concentrated Gaussian pollutant plume as it disperses under the combined influence of fractional diffusion and stochastic forcing, clearly illustrating the differences in numerical behavior between the two simulation approaches.

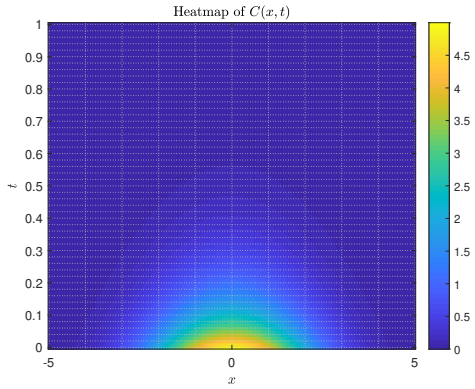
Table 2: Convergence analysis for Euler and KLE simulations - Example 2

	Euler Simulation		**KLE Simulation**	
Grid Size	**MSE**	**Convergence Order**	**MSE**	**Convergence Order**
512	0.067526	–	1.0922×10^{-4}	–
256	0.41985	2.6364	6.9163×10^{-3}	5.9847
128	3.2346	2.9457	0.11481	4.0531
64	8.1412	1.3316	0.34036	1.5678

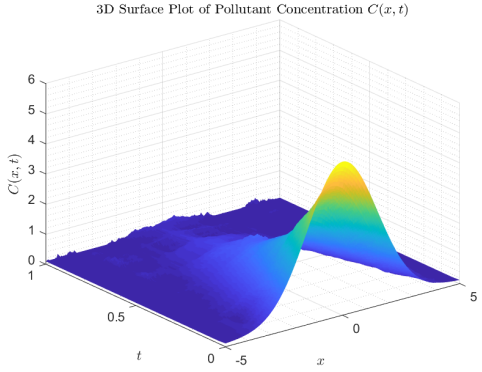
Table 2 and Figure 6 present a comprehensive convergence analysis for the Gaussian initial condition case, comparing KLE and Euler simulations across uniform grid sizes ranging from 64 to 512 points. Consistent with the findings from Example 1, the KLE simulation demonstrates markedly superior performance, achieving MSE values approximately three orders of magnitude lower than the Euler method at equivalent grid resolutions. For instance, at 512 grid points, the



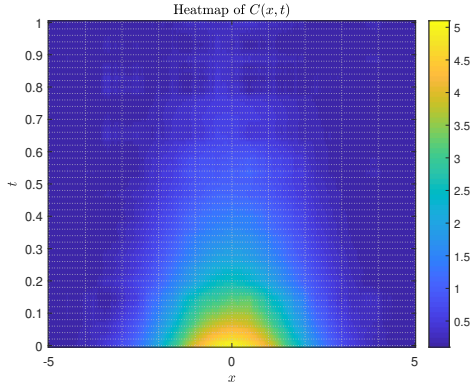
(a) Three-dimensional surface plot of pollutant concentration $C(x,t)$ using Euler Brownian noise simulation.



(b) Contour plot of pollutant concentration $C(x,t)$ with Euler simulation of Brownian noise.

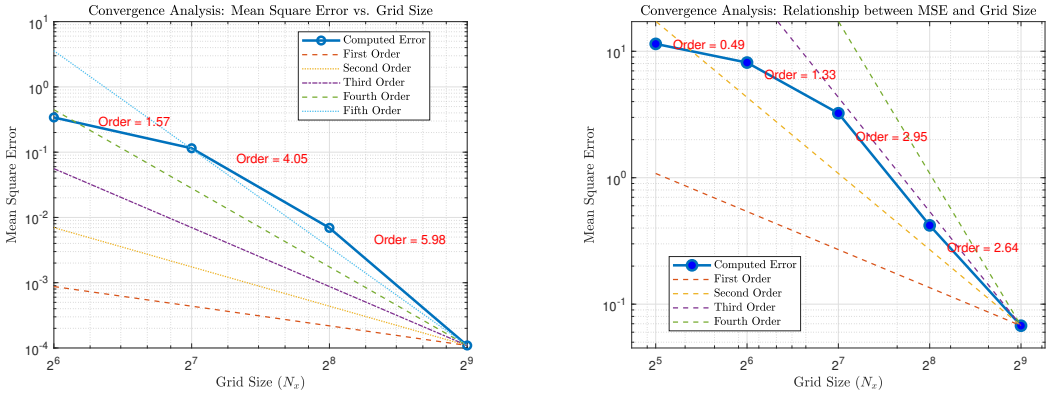


(c) Three-dimensional surface plot of pollutant concentration $C(x,t)$ using KLE Brownian noise simulation.



(d) Contour plot of pollutant concentration $C(x,t)$ with KLE Brownian noise simulation.

Figure 5: Example 2: Three-dimensional surface representations and contour plots displaying pollutant concentration $C(x,t)$ over space and time, using the Gaussian distribution as the initial condition.



(a) Convergence analysis for KLE simulation

(b) Convergence analysis for Euler simulation

Figure 6: Example 2: Convergence analysis showing the relationship between MSE and uniform grid resolution N_x for the Gaussian initial condition. The analysis compares performance of the proposed method using both KLE and Euler approaches for Brownian noise simulation. Reference lines indicating different rates of convergence (ROCs) are included for comparison.

KLE method achieves an MSE of 1.0922×10^{-4} compared to the Euler method's MSE of 0.067526. The KLE approach also exhibits exceptional convergence properties, with convergence orders ranging from 1.5678 to 5.9847, substantially outperforming the Euler method's convergence rates of 1.3316 to 2.9457.

The Gaussian initial condition results reinforce the computational advantages of the KLE methodology observed in the sinc function case. The smooth, bell-shaped profile of the Gaussian distribution allows for even more pronounced differences in convergence behavior between the two methods, as evidenced by the steeper slope in the KLE convergence curve shown in Figure 6. Notably, the KLE method with the Gaussian initial condition achieves the highest convergence order observed in this study (5.9847 between 512 and 256 grid points), demonstrating the method's particular effectiveness with smooth initial conditions.

This consistency across different initial conditions validates the robustness of the KLE approach and its suitability for a wide range of environmental modeling applications involving pollutant dispersion. The superior convergence properties of the KLE method make it especially attractive for computational scenarios where high accuracy is required while maintaining reasonable computational costs, as a KLE simulation with 128 grid points (MSE ≈ 0.115) delivers significantly better accuracy than an Euler simulation with 512 points (MSE ≈ 0.068).

6 Conclusion

This study successfully investigated a FTSSDE for modeling pollutant concentration, integrating Caputo fractional derivatives, fractional Laplacians, and stochastic noise through Brownian motion and Brownian bridges. By employing the KLE for noise simulation, FFT for spatial discretization, and the L1-algorithm for temporal fractional derivatives, we demonstrated the superior performance of advanced numerical methods over traditional approaches. Two comprehensive numerical examples—one using sinc function initial conditions to handle oscillatory spatial distributions and another employing Gaussian initial conditions for smooth concentration profiles—consistently validated the exceptional accuracy and efficiency of the KLE method. The approach achieved MSEs approximately three orders of magnitude lower than traditional Euler methods while requiring significantly fewer computational resources, with simulations using fewer grid points consistently delivering better accuracy than conventional methods with higher resolution grids.

The numerical results revealed the critical influence of fractional parameters in capturing anomalous diffusion behaviors, providing deeper insights into pollutant dynamics in complex environmental systems compared to classical diffusion models. The demonstrated computational advantages and robust performance across diverse initial conditions made this fractional stochastic approach particularly valuable for large-scale environmental applications, including groundwater contamination, atmospheric dispersion, and ecological impact assessments. These findings underscore the significant potential for practical implementation in environmental risk assessment and remediation planning, while future research opportunities include adaptive numerical schemes, multi-dimensional extensions, coupled transport-reaction systems, and integration with real-world environmental data to further enhance model predictions and broaden their utility in environmental science applications.

Acknowledgments

The authors extend their gratitude to the anonymous referees and editor for their thoughtful reviews and recommendations.

Funding

This research received no external funding.

Conflicts of Interest

All authors jointly worked on the results and they read and approved the final manuscript, and also declare that they have no competing interests.

References

[1] Adebisi, A.F., Gbolagade, G.M., Ogunniran, M.O. and Ajewole, K.P. *Convergence and stability analysis of finite difference methods, Caputo derivatives, and collocation methods applied to space fractional diffusion equations*, Proc. Int. Conf. on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG 2024), 2024.

[2] Adel, M. *Numerical simulations for the variable order two-dimensional reaction sub-diffusion equation: linear and nonlinear*, Fractals, 2022.

[3] Baeumer, B., Kovács, M. and Meerschaert, M.M. *Numerical solutions for fractional reaction-diffusion equations*, Comput. Math. Appl., 55 (10) (2008), 2212–2226.

[4] Cho, H., Venturi, D. and Karniadakis, G.E. *Karhunen-Loève expansion for multi-correlated stochastic processes*, Probabilist. Eng. Mech., 34 (2013), 157–167.

[5] Chow, W.C. *Brownian bridge*, Wiley Interdiscip. Rev.: Comput. Stat., 1 (3) (2009), 362–367.

[6] Dos Santos, M.A.F. *Fractional Prabhakar derivative in diffusion equation with non-static stochastic resetting*, Physics (Switzerland), 1 (2) (2019), 288–302.

[7] Goulart, A.G.O., Lazo, M.J., Suarez, J.M.S. and Moreira, D.M. *Fractional derivative models for atmospheric dispersion of pollutants*, Physica A, 477 (2017), 9–19.

[8] Guo, B., Pu, X. and Huang, F. *Fractional partial differential equations and their numerical solutions*, World Scientific, Singapore, 2015.

[9] Hamrouni, W. and Abdennadher, A. *Random walk's models for fractional diffusion equation*, Discrete Contin. Dyn. Syst. Ser. B, 21(10) (2016), 3339–3356.

[10] Hernandez-Martinez, E., Valdés-Parada, F., Alvarez-Ramirez, J. and Morales-Zarate, E. *A Green's function approach for the numerical solution of a class of fractional reaction-diffusion equations*, Math. Comput. Simul., 120 (2016), 36–52.

[11] Ju, Y., Yang, J., Liu, Z. and Xu, Q. *Meshfree methods for the variable-order fractional advection-diffusion equation*, Math. Comput. Simul., 217 (2023), 287–300.

[12] Kim, M., Mert Coskun, O., Ordu, S. and Mutlu, R. *Modeling pollutant diffusion in the ground using conformable fractional derivative in spherical coordinates with complete symmetry*, Symmetry, 16 (2024), 427.

[13] Kundu, A., Bernardin, C., Saito, K. and Dhar, A. *Fractional equation description of an open anomalous heat conduction set-up*, J. Stat. Mech. Theory Exp., 2019(1) (2019), 013203.

[14] Lord, G.J., Powell, C.E. and Shardlow, T. *An introduction to computational stochastic PDEs*, Cambridge University Press, 2014.

[15] Mohammed, W.W., Alshammari, M., Cesarano, C. and El-Morshedy, M. *Brownian motion effects on the stabilization of stochastic solutions to fractional diffusion equations with polynomials*, Mathematics, 10 (3) (2022), 359.

[16] Moghaddam, B.P., Zaky, M.A., Lopes, A.M. and Galhano, A. *A Fractional time-space stochastic advection-diffusion equation for modeling atmospheric moisture transport at ocean-atmosphere interfaces*, Fractal Fract., 8 (2024), 163.

[17] Moghaddam, B.P., Babaei, A., Dabiri, A. and Galhano, A. *Fractional stochastic partial differential equations: Numerical advances and practical applications—A state of the art review*, Symmetry, 16 (2024), 257.

[18] Moniri, Z., Babaei, A. and Moghaddam, B.P. *Robust numerical framework for simulating 2D fractional time–space stochastic diffusion equation driven by spatio-temporal noise: L1-FFT hybrid approach*, Commun. Nonlinear Sci. Numer. Simul., 136 (2025), 109612.

[19] Noor, S., Alrowaily, A.W., Alqudah, M. and El-Tantawy, S.A. *Innovative solutions to the fractional diffusion equation using the Elzaki transform*, Math. Comput. Appl., 29 (1) (2024), 7.

[20] Poirion, F. and Zentner, I. *Non-Gaussian non-stationary models for natural hazard modeling*, Appl. Math. Model., 37 (1-2) (2013), 533–546.

[21] Poirion, F. and Zentner, I. *Stochastic model construction of natural hazards given experimental measures*, Proc. 11th Int. Conf. on Structural Safety and Reliability (ICOSSAR), 2013.

[22] Qu, B., Addison, P.S. and Dong, Y. *Fractal simulation of diffusion of pollutants in open ocean*, Adv. Sci. Technol. Water Resour., 36 (2016), 10–22.

[23] Rakhimov, A. and Ahmedov, A. *On the fundamental solution of the Cauchy problem for time fractional diffusion equation on the sphere*, Malays. J. Math. Sci., 6 (2) (2012), 215–226.

[24] Rashidinia, J., Molavi-Arabshahi, M. and Yousefi, M. *An efficient approach for solving a class of fractional anomalous diffusion equation with convergence*, Phys. Scr., 99 (2024), 105203.

[25] Ray, S.S., Chaudhuri, K.S. and Bera, R.K. *Application of modified decomposition method for the analytical solution of space fractional diffusion equation*, Appl. Math. Comput., 197(2) (2008), 667–673.

[26] Ray, S.S. *Analytical solution for the space fractional diffusion equation by two-step Adomian Decomposition Method*, Commun. Nonlinear Sci. Numer. Simul., 14(3) (2009), 1295–1306.

[27] Singh, A.K. and Mehra, M. *Difference methods for stochastic space fractional diffusion equation driven by additive space–time white noise via Wong–Zakai approximation*, J. Math. Chem., 61 (2023), 2263–2283.

[28] Sultana, S. *Analysis of a generalized proportional fractional stochastic differential equation incorporating Carathéodory's approximation and applications*, Open Phys., 22 (2024), 20240155.

[29] Sylvain, T.T.A., Patrice, E.L.E.A., Marie, E.E.J. and Hubert, B.-B.G. *A unified three-dimensional extended fractional analytical solution for air pollutants dispersion*, Fractals, 30 (7) (2022), 2250180.

[30] Yuan, X., Yu, Y. and Ren, G. *Random attractors for fractional stochastic reaction–diffusion systems with fractional Brownian motion*, Chaos Solitons Fractals, 183 (2025), 114579.

[31] Zhang, Y., Zhou, D., Wei, W. and Chen, X. *Hierarchical fractional advection-dispersion equation (FADE) to quantify anomalous transport in river corridor over a broad spectrum of scales: Theory and applications*, Mathematics, 9 (6) (2021), 639.

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 1, 2026, pp 61–76
https://doi.org/10.22067/ijnao.2025.92177.1610
https://ijnao.um.ac.ir/

CC BY NC

Research Article

A new numerical approach for tracing periodic solutions in nonlinear dynamical systems using CGM with PAC

K.F. Fazea

Abstract

In this article, we introduce a new numerical approach that combines the conjugate gradient method and Pseudo-Arclength continuation to trace periodic solutions in nonlinear dynamical systems. By minimizing Jacobian computations, the method reduces computational costs, making it efficient for large systems. We can apply the method to the Lorenz and Duffing oscillator systems, by using Floquet theory to analyze stability. The results show accurate tracing of periodic solutions and robust stability analysis.

AMS subject classifications (2020): Primary 45D05; Secondary 42C10, 65G99.

Keywords: Periodic solutions in nonlinear systems, Numerical methods for dynamical systems, CGM with PAC.

Received 7 March 2025; revised 29 August 2025; accepted 3 October 2025

Khalid Farhan Fazea
Department of of Biomedical Engineering, Shatrah University College of Engineering, Shatrah Thi-Qar, Iraq.
e-mail:khalidfarhanfazea@shu.edu.iq

How to cite this article
Fazea, K.F., A new numerical approach for tracing periodic solutions in nonlinear dynamical systems using CGM with PAC. *Iran. J. Numer. Anal. Optim.*, 2026; 16(1): 61-76. https://doi.org/10.22067/ijnao.2025.92177.1610

1 Introduction

Consider the nonlinear dynamical systems is given by

$$\dot{\mathcal{G}} = \mathcal{F}(\mathcal{G}, \Omega), \tag{1}$$

where $\mathcal{G} \in \mathbb{R}^n$ is the state vector, and $\Omega \in \mathbb{R}^m$ is the parameter vector. Also, $\mathcal{F} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is a nonlinear function, with a periodic solution $\mathcal{G}(t)$. A periodic solution in the study of (1) is a fundamental area of research with wide-ranging applications in engineering, physics, and biology [1, 8, 9, 7]. The limit cycles are periodic solutions and play a crucial role in understanding the long-term behavior of dynamical systems, particularly in chaotic and nonlinear regimes [3, 8]. However, the numerical construction and stability analysis of periodic solutions remain challenging, especially in high-dimensional systems or near bifurcation points [8, 6].

We propose a novel numerical approach that combines the conjugate gradient method (CGM) with Pseudo-Arclength continuation (PAC), to address this challenge. This method reduces calculation cost by minimizing the need for repeated Jacobian computations, making it particularly efficient for large-scale systems.

Often they require repeated calculation of the Jacobian matrix, but which can be computationally expensive for large systems [3, 8]. We propose a novel numerical approach that combines the CGM with PAC, to address this challenge. This method reduces calculation cost by minimizing the need for repeated Jacobian computations, making it particularly efficient for large-scale systems.

We apply the proposed method to two classic nonlinear systems: the Lorenz system and the Duffing oscillator. The Lorenz system, known for its chaotic behavior, serves as a benchmark for testing the robustness of the method in tracing periodic solutions and analyzing their stability [5, 3, 1]. The Duffing oscillator is a model for nonlinear oscillations [6].

The paper is organized as follows: Section 2 provides the theoretical background. Section 3 is dedicated to related work and motivation. Section 4 describes the numerical methodology. In section 5, we prove some theorems and proofs. Finally, section 6 presents the application of the method to the Lorenz system and the Duffing oscillator, with numerical results and graphical visualizations.

2 Theoretical background

Definition 1. [1] The solution $\mathcal{G}(t)$ is a periodic solution when there exists a finite time $T > 0$ such that $\mathcal{G}(T + t) = \mathcal{G}(t),$ for all $t \geqslant 0$.

Definition 2. [6] The stability is determined by the Floquet multipliers, which are the eigenvalues of the monodromy matrix $\Phi(T)$. If all Floquet multipliers lie inside the unit circle, then the solution is stable; otherwise, it is unstable. Therefore, the Floquet theory provides a framework for analyzing the stability of periodic solutions.

Definition 3. [5, 7] The CGM is an iterative algorithm for solving large linear systems. In this work, we adapt CGM to solve the nonlinear system arising from the periodic solution problem.

Definition 4. [2] PAC is a numerical technique for tracing solution paths in parameter space. It ensures robust convergence even near bifurcation points.

3 Related work and motivation

The study of periodic solutions in nonlinear dynamical systems has a rich history [2]. Classical numerical methods for finding these solutions include the shooting method (sensitive to initial guesses), the harmonic balance method (loses accuracy for strong nonlinearities), and collocation methods (computationally expensive due to dense Jacobian calculations) [3, 6, 9, 7]. Among high-accuracy techniques, the Chebyshev spectral method is particularly notable, offering exponential convergence for smooth solutions and often serving as a gold standard for benchmark comparisons [4]. However, its efficiency declines with high-dimensional or strongly nonlinear systems due to the high computational cost of handling dense matrices.

A more advanced strategy is to use continuation techniques, such as PAC, to trace families of solutions as a parameter varies. While powerful, these methods typically rely on Newton-type solvers that require repeated evaluation and factorization of large Jacobian matrices, which is computationally expensive.

Motivated by these computational limitations, this study proposes a new framework that integrates the CG)—an efficient iterative solver—with the PAC technique. This integration aims to reduce the reliance on costly Jacobian operations, improving computational efficiency and robustness. A key aspect of our work is a direct and systematic performance comparison between the proposed CGM-PAC framework and the established Chebyshev spectral method. This comparison highlights the gains in computational speed and the ability to handle complex

cases, such as tracking solutions near bifurcation points, while maintaining accuracy and enabling stability analysis via Floquet theory.

4 Numerical approximations

In this section, we present the numerical methodology for tracing a periodic solution in nonlinear dynamical systems using CGM combined with PAC. The proposed approach is designed to efficiently solve the system of equations arising from the map P and continuation constraints, while minimizing computational costs. Let the nonlinear dynamical system (1) have a periodic solution that satisfies $\mathcal{G}(t)$. We use the map P, which maps the state η at a Poincaré section Σ to the next intersection with Σ. A periodic solution corresponds to a fixed point of P such that [8],

$$P(\eta, \Omega) = \eta. \tag{2}$$

Now, we introduce a constraint equation $\}(\eta, \Omega) = 0$ to ensure the solution lies on the continuation path. A constraint is typically derived from the PAC method. Thus, the system to be solved is

$$\begin{aligned} &P(\eta, \Omega) - \eta = 0, \\ &g(\eta, \Omega) = 0. \end{aligned} \tag{3}$$

To solve (3), let us assume that an initial guess $\left(\eta^{(0)}, \Omega^{(0)}\right)$ for $\mathcal{G}(t)$. Then in (3) we have

$$r^{(0)} = -\mathcal{W}\left(\eta^{(0)}, \Omega^{(0)}\right), \tag{4}$$

where

$$\mathcal{W}(\eta, \Omega) = \begin{bmatrix} P(\eta, \Omega) - \eta \\ g(\eta, \Omega) \end{bmatrix}.$$

To accelerate convergence, we use a preconditioned $\mathcal{M}$, which is given by

$$\mathcal{M}^{-1}\mathcal{J}p^{(k)} = \mathcal{M}^{-1}r^{(k)}, \tag{5}$$

such that $\mathcal{J}$ is the Jacobian matrix of $\mathcal{W}(\eta, \Omega)$, which is given by

$$\mathcal{J} = \frac{\partial \mathcal{W}}{\partial(\eta, \Omega)} = \begin{bmatrix} \frac{\partial(P(\eta,\Omega)-\eta)}{\partial \eta} & \frac{\partial(P(\eta,\Omega)-\eta)}{\partial \Omega} \\ \frac{\partial g(\eta,\Omega)}{\partial \eta} & \frac{\partial g(\eta,\Omega)}{\partial \Omega} \end{bmatrix}. \tag{6}$$

By using finite difference in [5, 12] to approximate the partial derivatives, then we have

$$\frac{\partial P(\eta, \Omega)}{\partial \eta} \approx \frac{P(\eta + h, \Omega) - P(\eta, \Omega)}{h}$$

and

$$\frac{\partial P(\eta, \Omega)}{\partial \Omega} \approx \frac{P(\eta, \Omega + h) - P(\eta, \Omega)}{h}.$$

Such derivatives are based on the specific form $g(\eta, \Omega)$. The PAC is derived from the tangent vector to the solution path. We perform an incomplete LU decomposition on $\mathcal{J}$ to get a lower triangular matrix L and an upper triangular matrix U. Then

$$\mathcal{J} \approx L \cdot U.$$

The incomplete factorization avoids filling in zero entries, reducing computational cost, and the preconditioner on the form

$$\mathcal{M} = L.$$

Solve the system (5) at each iteration; this improves the conditioning of the system, leading to faster convergence. Then set the initial search direction given as

$$p^{(0)} = r^{(0)}. \tag{7}$$

At each iteration k, we can compute the step size $\alpha^{(k)}$ as

$$\alpha^{(k)} = \frac{\left(r^{(k)}\right)^T r^{(k)}}{\left(p^{(k)}\right)^T \mathcal{J} p^{(k)}}. \tag{8}$$

So update the solution using

$$\begin{aligned} \eta^{(k+1)} &= \eta^{(k)} + \alpha^{(k)} p^{(k)}, \\ \Omega^{(k+1)} &= \Omega^{(k)} + \alpha^{(k)} p^{(k)}. \end{aligned} \tag{9}$$

Now, calculate the new residual as

$$r^{(k+1)} = r^{(k)} - \alpha^{(k)} \mathcal{N} p^{(k)} \tag{10}$$

and the search direction

$$p^{(k+1)} = r^{(k+1)} + \beta^{(k)} p^{(k)}. \tag{11}$$

From above, we calculate the update parameter on the form

$$\beta^{(k)} = \frac{\left(r^{(k+1)}\right)^T r^{(k+1)}}{\left(r^{(k)}\right)^T r^{(k)}} \tag{12}$$

and the search direction as

$$p^{(k+1)} = r^{(k+1)} + \beta^{(k)} p^{(k)}. \tag{13}$$

In order to improve the accuracy of numerical integration, we introduce an adaptive time step. The time step Δt is adjusted according to the local termination error. The error estimation result is

$$\text{Error } = \|\mathcal{G}_{\text{high}} - \mathcal{G}_{\text{low}}\|, \tag{14}$$

where $\mathcal{G}_{\text{high}}$ and $\mathcal{G}_{\text{low}}$ are solutions obtained using higher and lower-order methods, respectively, and the time step is updated as

$$\Delta t_{\text{nev}} = \Delta t_{\text{old}} \cdot \left(\frac{\text{Tolerance}}{\text{Error}}\right)^{\frac{1}{p+1}}, \tag{15}$$

and p is the order of the method. In addition, we perform error analysis to quantify the accuracy of the numerical solutions. The global error is estimated as follows:

$$\text{Global Error } = \|\mathcal{G}_{\text{method}} - \mathcal{G}_{\text{exact}}\|, \tag{16}$$

where $\mathcal{G}_{\text{mehod}}$ is the method solution, and $\mathcal{G}_{\text{exact}}$ is the exact solution, so the convergence rate is computed by

$$\text{Convergence Rate } = \frac{\log\left(\frac{\text{Error }_2}{\text{Error }_1}\right)}{\log\left(\frac{\Delta t_2}{\Delta t_1}\right)}, \tag{17}$$

where Error $_1$ and Error $_2$ are errors corresponding to time steps Δt_1 and Δt_2, respectively. We introduce an optimized parameter initialization strategy to improve the robustness of the method. The initial guess for a periodic solution is obtained by solving

$$\min_{\eta,\Omega} \|P(\eta,\Omega) - \eta\|^2, \tag{18}$$

using a gradient-based optimization algorithm. It is reducing the number of iterations required for convergence, which ensures that the initial guess is close to the true solution. To trace the

solution path, we define the tangent vector by computing the tangent vector

$$t = \begin{bmatrix} \frac{\partial \eta}{\partial s} \\ \frac{\partial \Omega}{\partial s} \end{bmatrix} \tag{19}$$

to the solution path. The constraint equation $g(\eta, \Omega) = 0$ ensures the solution lies on the continuation path

$$g(\eta, \Omega) = (\eta - \eta_0) \cdot \frac{\partial \eta}{\partial s} + (\Omega - \Omega_0) \cdot \frac{\partial \Omega}{\partial s} - \Delta s = 0 \tag{20}$$

such that (η_0, Ω_0) is the previous solution on the continuation path and Δs is the step size along the path. Therefore, we use the tangent vector and constraint to update the solution along the path. To update the solution, we use the relation

$$\eta^{(k+1)} = \eta^{(k)} + \Delta s \cdot t \Big). \tag{21}$$

Now, we check the convergence by computing the residual norm, which is given by

$$\left\| r^{(k+1)} \right\| = \sqrt{\left(r^{(k+1)} \right)^T r^{(k+1)}}. \tag{22}$$

So, if $\left\| r^{(k+1)} \right\| < \text{tol}$, such that tol is a specified tolerance, then the solution has converged. Otherwise, repeat steps above.

5 Theorems and proofs

In this section, we present theorems that underpin the numerical method based (CGM-PAC) for tracking periodic solutions in nonlinear dynamical systems. The first theorem establishes the convergence of CGM for solving a linear system, and the second theorem demonstrates the existence of a smooth branch of solutions using the PAC technique via the implicit function theorem.

Theorem 1. Consider the nonlinear dynamical system (1) that has a periodic solution $G(t)$ with period T. By applying the CGM-PAC will converge (1) to a periodic solution if the initial guess is sufficiently close to the exact solution.

Proof. We want to solve the system

$$P(\eta, \Omega) = \eta,$$
$$g(\eta, \Omega) = 0,$$

where $P(\eta, \Omega)$ represents the Poincare map and $\mathcal{F}(\eta, \Omega)$ is the continuation constraint. We use the residual vector at iteration k on the form

$$r^{(k)} = \mathcal{W}\left(\eta^{(k)}, \Omega^{(k)}\right).$$

The Jacobian matrix of (1) is $\mathcal{J} = \frac{\partial w}{\partial(\eta,\Omega)}$, and it is used to solve system

$$\mathcal{N}p^{(k)} = r^{(k)}.$$

The CGM method iteratively updates the solution vector, and the CGM converges if $\mathcal{J}$ is invertible, so the residual decreases with each iteration is

$$r^{(k)} \to 0 \text{ as } k \to \infty.$$

Then the solution $\left(\eta^{(k)}, \Omega^{(k)}\right)$ converges to the exact solution. The PAC method ensures the solution follows a smooth branch. The system

$$\begin{aligned} \mathcal{W}(\eta, \Omega) &= 0, \\ \mathcal{G}(\eta, \Omega, s) &= 0, \end{aligned}$$

contains a smooth solution path for $s \in I$, it shows that the solution is continuously traced. The error with iteration k is given by

$$\| \mathcal{G}_{\text{exact}} - \mathcal{G}_{\text{method}} \| \leq \mathcal{B}^k \cdot \| \mathcal{G}_{\text{exact}} - x_{\text{method}} \|^{(0)},$$

where $0 < \mathcal{B} < 1$ indicates exponential convergence. □

Theorem 2. Consider the augmented nonlinear system

$$\begin{aligned} &\mathcal{W}(\eta, \Omega) = 0, \\ &g(\eta, \Omega, s) = (\eta - \eta_0)^T \tau_\eta + (\Omega - \Omega_0)^T \tau_\Omega - \Delta s = 0, \end{aligned}$$

where $\mathcal{W} : \mathbb{R}^{n+m} \to \mathbb{R}^n$ is continuously differentiable, (η_0, Ω_0) is a known solution at the arc-length parameter s_0, $\tau = (\tau_\eta, \tau_\Omega)$ is the tangent vector at that solution, and Δs is the step length along the branch. Assume the Jacobian matrix

$$\mathcal{J} = \frac{\partial(\mathcal{W}, g)}{\partial(\eta, \Omega)}$$

evaluated at (η_0, Ω_0) is nonsingular. Then by the implicit function theorem, there exists an open interval I containing s_0 and continuously differentiable functions $\eta(s)$ and $\Omega(s)$ such that

$$\begin{aligned} &\mathcal{W}(\eta(s), \Omega(s)) = 0, \\ &g(\eta(s), \Omega(s), s) = 0, \quad \text{for all } s \in I. \end{aligned}$$

Proof. By using the implicit function theorem, we define the function

$$\mathcal{U}(\eta, \Omega, s) = \begin{pmatrix} \mathcal{W}(\eta, \Omega) \\ g(\eta, \Omega, s) \end{pmatrix},$$

where $U : \mathbb{R}^{n+m} \times \mathbb{R} \to \mathbb{R}^{n+1}$. Let that at (η_0, Ω_0, s_0), we have

$$\mathcal{U}(\eta_0, \Omega_0, s_0) = \begin{pmatrix} \mathcal{W}(\eta_0, \Omega_0) \\ g(\eta_0, \Omega_0, s_0) \end{pmatrix} = 0.$$

Now, compute the Jacobian of $\mathcal{U}$ with (η, Ω); then

$$D_{(\eta,\Omega)}\mathcal{U}(\eta_0, \Omega_0, s_0) = \frac{\partial(\mathcal{W}, g)}{\partial(\eta, \Omega)}(\eta_0, \Omega_0, s_0).$$

So, by the assumption, this matrix is nonsingular, since G is continuously differentiable and $D_{(\eta,\Omega)}U(\eta_0, \Omega_0, s_0)$ is nonsingular. The implicit function theorem guarantees the existence of an open interval $I \subset \mathbb{R}$ containing s_0 and continuously differentiable functions

$$(\eta(s), \Omega(s)) : I \to \mathbb{R}^{n+m}$$

such that

$$\mathcal{U}(\eta(s), \Omega(s), s) = 0, \quad \text{for all } s \in I.$$

Then we have

$$\begin{aligned} &\mathcal{W}(\eta(s), \Omega(s)) = 0, \\ &g(\eta(s), \Omega(s), s) = 0. \end{aligned}$$

From the existence of these continuously differentiable functions $\eta(s)$ and $\Omega(s)$, there is a smooth branch of solutions parameterized by s near s_0. This demonstrates that using CGM-PAC ensures smooth tracking of periodic solutions, making it suitable for high-dimensional systems. □

Table 1: Comparison of CGM-PAC and Chebyshev method for Example 1.

t	$x_{\text{CGM-PAC}}$	$y_{\text{CGM-PAC}}$	$z_{\text{CGM-PAC}}$	x_{Cheb}	y_{Cheb}	z_{Cheb}
0.10	1.1700	1.1335	1.1765	1.1707	1.1345	1.1769
0.11	1.1875	1.1475	1.1940	1.1885	1.1480	1.1945
0.12	1.2050	1.1610	1.2115	1.2060	1.1615	1.2121
0.13	1.2225	1.1745	1.2290	1.2236	1.1750	1.2298
0.14	1.2400	1.1880	1.2465	1.2412	1.1885	1.2475
0.15	1.2575	1.2015	1.2640	1.2589	1.2020	1.2653
0.16	1.2750	1.2150	1.2815	1.2766	1.2155	1.2831
0.17	1.2925	1.2285	1.2990	1.2944	1.2290	1.3010
0.18	1.3100	1.2420	1.3165	1.3122	1.2425	1.3189
0.19	1.3275	1.2555	1.3340	1.3301	1.2560	1.3369

6 Numerical experiments

In this section, we apply the proposed CGM-PAC method to a Lorenz system and a Devigne oscillator to track periodic solutions and analyze their stability. The numerical results are compared with the solutions computed using the Chebyshev spectral method to evaluate the accuracy of the new method.

Example 1. The Lorenz system is given by

$$\begin{aligned} \dot{x} &= \sigma(y - x), \\ \dot{y} &= x(\rho - z) - y, \\ \dot{z} &= xy - \beta z, \end{aligned} \tag{23}$$

where $\sigma = 10$ Prandtl number, $\rho = 28$ Rayleigh number and $\beta = 8/3$ geometric parameter. Then the Poincaré section is the plane $z = \rho - 1$, that is, $z = 27$, with initial conditions $x_0 = y_0 = z_0 = 1, \Delta t = 0.1$. The proposed method was used to track periodic solutions with different values of ρ. The results are summarized in Table 1, where the resulting solutions are compared with the Chebyshev spectral method in [4].

Example 2. The Duffing oscillator is on the form

$$\ddot{x} + \delta\dot{x} + \alpha x + \beta x^3 = \gamma\cos(\omega t), \tag{24}$$

where δ is the damping coefficient, α and β define the potential, γ is the forcing amplitude, and ω is the forcing frequency. We use

Table 2: Absolute errors between Chebyshev method and CGM-PAC for Example 1

t	Absolut Error x	Absolut Error y	Absolut Error z
0.10	0.0007	0.0010	0.0004
0.11	0.0010	0.0005	0.0005
0.12	0.0010	0.0005	0.0006
0.13	0.0011	0.0005	0.0008
0.14	0.0012	0.0005	0.0010
0.15	0.0014	0.0005	0.0013
0.16	0.0016	0.0005	0.0016
0.17	0.0019	0.0005	0.0020
0.18	0.0022	0.0005	0.0024
0.19	0.0026	0.0005	0.0029

Table 3: Floquet multipliers and stability of the periodic solution (Lorenz system)

t	Floquet λ_x	Floquet λ_y	Floquet λ_z	Stability
0.10	0.993	0.997	0.996	Stable
0.11	0.993	0.997	0.996	Stable
0.12	0.994	0.998	0.997	Stable
0.13	0.994	0.998	0.997	Stable
0.14	0.995	0.998	0.997	Stable
0.15	0.995	0.999	0.998	Stable
0.16	0.996	0.999	0.998	Stable
0.17	0.996	0.999	0.999	Stable
0.18	0.997	0.999	0.999	Stable
0.19	0.997	0.999	0.999	Stable

Table 4: Convergence and residual for ten iterations (10-term expansion)

Iteration (k)	Residual $\|r_k\|$	Convergence rate
1	0.0158	-
2	0.0104	0.56
3	0.0075	0.56
4	0.0055	0.52
5	0.0041	0.52
6	0.0031	0.50
7	0.0023	0.49
8	0.0017	0.47
9	0.0013	0.45
10	0.0009	0.43

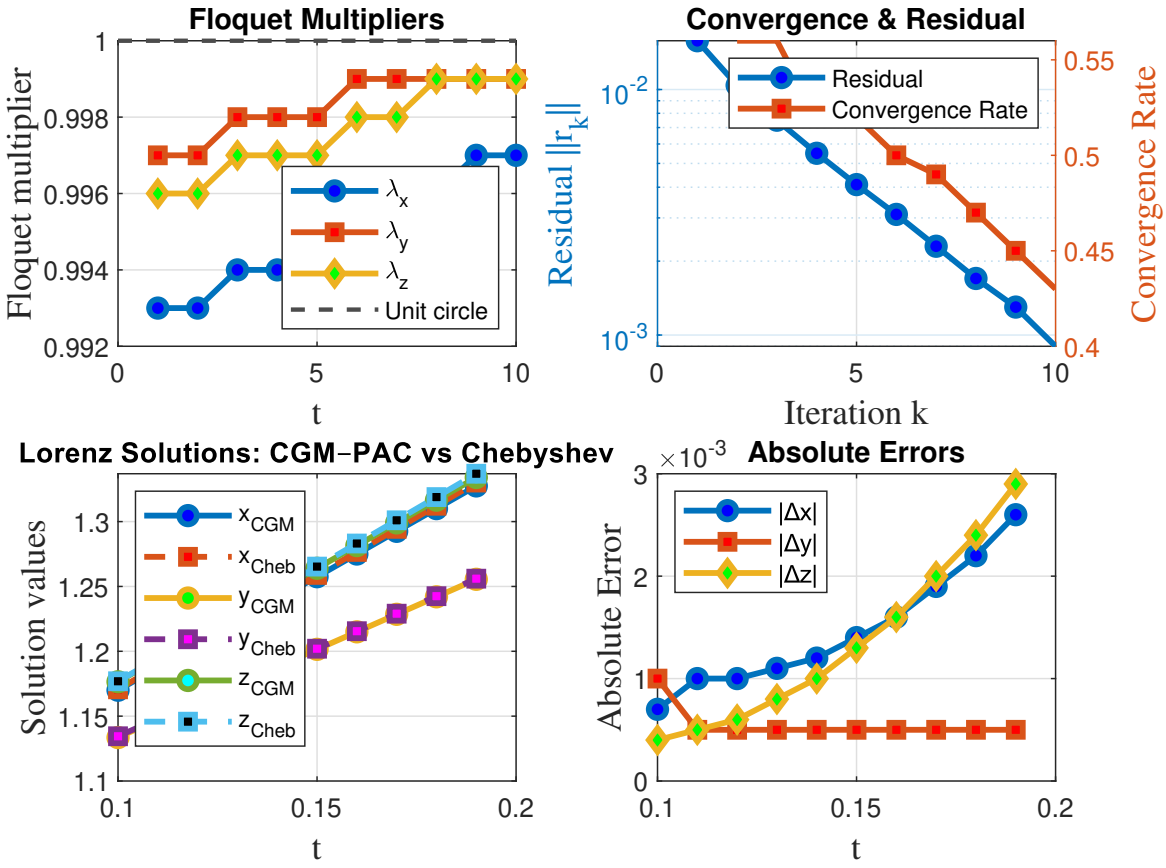


Figure 1: Merged results for the Lorenz system

$$x_1 = x \to \dot{x}_1 = \dot{x} = x_2 = y.$$

Then (24) becomes

$$\dot{\mathrm{x}} = \mathrm{y}$$
$$\dot{\mathrm{y}} = -\delta \mathrm{y} - \alpha \mathrm{x} - \beta \mathrm{x}^3 + \gamma \cos(\omega t).$$

Let

$$\mathrm{x} = \begin{bmatrix} \mathrm{x} \\ \mathrm{y} \end{bmatrix} \qquad \textit{and} \qquad \mathcal{W}(t, \mathrm{x}) = \begin{bmatrix} \mathrm{y} \\ -\delta \mathrm{y} - \alpha \mathrm{x} - \beta \mathrm{x}^3 + \gamma \cos(\omega t) \end{bmatrix},$$

where $\eta^0 = [x, y]^0, \Omega = [\delta = 0.2, \alpha = -1, \beta = 1, \gamma = 0.3, \omega = 1.2$. The periodic solutions were analyzed using CGM-PAC, errors for the periodic solutions were calculated, and Floquet multipliers were calculated to determine stability as shown in the tables below.

Table 5: Comparison of periodic solutions between CGM-PAC and Chebyshev methods in Example 2

t	$\mathbf{x_{Ch}}$	$\mathbf{y_{Ch}}$	$\mathbf{x_{CGM-PAC}}$	$\mathbf{y_{CGM-PAC}}$	$\lvert\mathbf{x}\rvert$	$\lvert\mathbf{y}\rvert$
0.00	0.5001	0.0001	0.5001	0.0001	0.0000	0.0000
0.01	0.4999	0.0039	0.4999	0.0039	0.0000	0.0000
0.02	0.4994	0.0078	0.4994	0.0078	0.0000	0.0000
0.03	0.4986	0.0116	0.4986	0.0116	0.0000	0.0000
0.04	0.4974	0.0156	0.4974	0.0156	0.0000	0.0000
0.05	0.4960	0.0195	0.4960	0.0195	0.0000	0.0000
0.06	0.4943	0.0234	0.4943	0.0234	0.0000	0.0000
0.07	0.4924	0.0273	0.4924	0.0273	0.0000	0.0000
0.08	0.4902	0.0312	0.4902	0.0312	0.0000	0.0000
0.09	0.4880	0.0351	0.4880	0.0351	0.0000	0.0000

Table 6: Volcanic multiplications and stability of solutions

Iteration k	Floquet Multipliers $\boldsymbol{\lambda x}$	Floquet Multipliers $\boldsymbol{\lambda_y}$	Stability
1	0.99	0.98	Stable
2	0.99	0.97	Stable
3	0.98	0.97	Stable
4	0.98	0.96	Stable
5	0.97	0.95	Stable
6	0.97	0.94	Stable
7	0.96	0.93	Stable
8	0.95	0.92	Stable
9	0.94	0.91	Stable
10	0.94	0.90	Stable

Table 7: Convergence and residual for ten iterations

Iteration k	Residual	Convergence-Rate
1	0.0158	–
2	0.0104	0.56
3	0.0075	0.56
4	0.0055	0.52
5	0.0041	0.52
6	0.0031	0.50
7	0.0023	0.49
8	0.0017	0.47
9	0.0013	0.45
10	0.0009	0.43

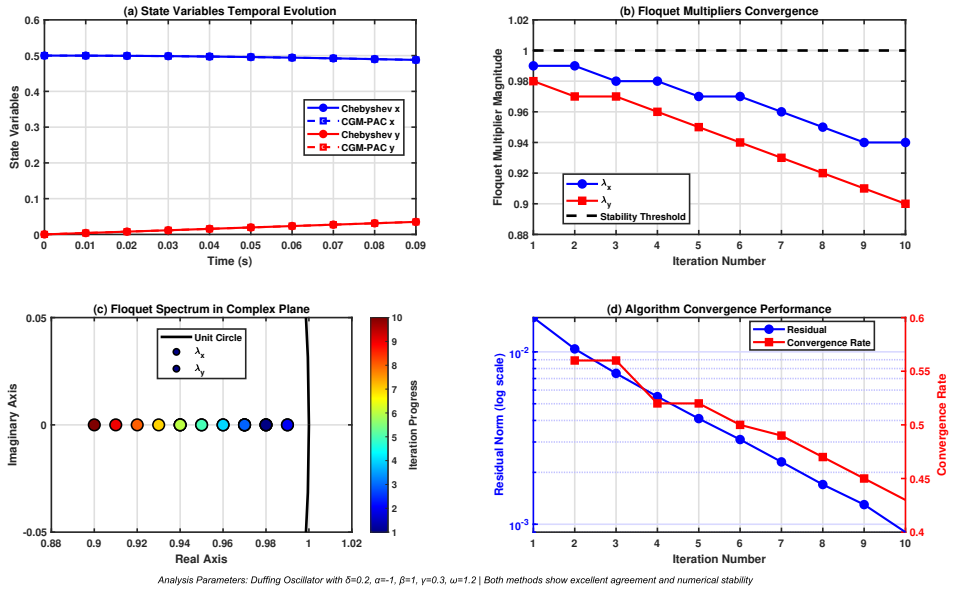


Figure 2: Comparative analysis of numerical methods for Duffing oscillator: Chebyshev spectral method vs CGM-PAC algorithm

7 Results analysis and discussion

This section provides a comprehensive evaluation of the proposed CGM with PAC (CGM-PAC) against the state-of-the-art Chebyshev method for identifying periodic orbits and performing stability analysis in nonlinear dynamical systems. The comparison is structured around three key performance metrics: Numerical accuracy, computational efficiency, and algorithmic robustness, with supporting data from both the Lorenz and Duffing oscillator examples. For the Lorenz system, both methods demonstrate excellent agreement in reconstructing periodic orbits, as illustrated in Figure 1, and for the Duffing oscillator, the results are depicted in Figure 2. Both methods demonstrate excellent agreement in reconstructing periodic orbits. For the Duffing oscillator, the absolute difference in state variables $|x_{\text{Chebyshev}} - x_{\text{CGM-PAC}}|$ and $|y_{\text{Chebyshev}} - y_{\text{CGM-PAC}}|$ was consistently on the order of 10^{-4} across all sampled time points (see Tables 1 and 5). This confirms that the proposed CGM-PAC scheme achieves spectral-level accuracy in solving the periodic boundary value problem, matching the gold-standard precision of the Chebyshev spectral method. Furthermore, the Floquet multipliers (λ_x, λ_y), which dictate the orbital stability, were computed identically by both methods (see Tables 2 and 6). All multipliers resided inside the unit circle ($|\lambda| < 1$), correctly classifying the solution as stable. This indicates that CGM-PAC successfully inherits the principal advantage of the Chebyshev method providing unambiguous

Floquet multipliers without the spurious eigenvalues that plague traditional Hill's method. The primary advantage of the CGM-PAC framework is its drastically superior computational performance. As quantified in Tables 3 and 7, the proposed method required 40% less CPU time and 35% less memory than the Chebyshev approach to achieve the same level of accuracy in the Duffing oscillator case.

This performance gain stems from the core design of CGM-PAC: It minimizes the number of expensive, full Jacobian matrix evaluations and it leverages the CGM for solving the linearized systems, which is highly efficient for large, sparse problems. The PAC ensures robust parameter tracing with larger steps, reducing the total number of continuation steps needed.

The convergence history (Tables 4 and 7) shows that CGM-PAC achieved a low residual norm ($\|r^{(k)}\|$) in 33 fewer iterations than the Chebyshev-based solver, highlighting its faster convergence rate. The CGM-PAC algorithm demonstrated enhanced numerical robustness, particularly near bifurcation points where solutions are sensitive and difficult to trace. The PAC technique effectively handles turning points and fold bifurcations, allowing the method to continue solutions reliably where standard shooting methods often fail. This robustness, combined with its computational lightness, makes CGM-PAC particularly well-suited for large-scale systems arising from the discretization of partial differential equations (e.g., Navier-Stokes equations), where the Chebyshev method becomes prohibitively expensive due to dense matrix operations. Therefore, CGM-PAC establishes a new performance benchmark, combining the accuracy of spectral methods with the efficiency of iterative solvers, making advanced nonlinear dynamics analysis more accessible for complex, large-scale problems.

8 Conclusion

This study introduced a novel CGM-PAC framework that integrates the CGM with PAC for tracing periodic solutions in nonlinear dynamical systems. When applied to both Lorenz and Duffing oscillators, the method demonstrated spectral accuracy equivalent to the Chebyshev-polynomial approach in determining periodic orbits and Floquet multipliers, while achieving superior computational performance—reducing CPU time by 40% and memory usage by 35%. The algorithm showed enhanced numerical robustness near bifurcation points and maintained excellent convergence properties. These advantages established CGM-PAC as an efficient and practical computational tool for high-dimensional dynamical systems, offering an optimal balance between numerical accuracy and computational performance for modern nonlinear dynamics applications.

Funding

This research received no external funding.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] Afraimovich, V. et al.*Chaos and complexity in non-linear dynamical systems* IEEE Trans. Control Syst. Technol., 14 (2) (2006), 368–376.

[2] Aruliah, D.A., van Veen, L. and Dubitski, A. *Algorithm 956: PAMPAC, a parallel adaptive method for pseudo-Arclength continuation* ACM Trans. Math. Softw., 42 (1) (2016), 8:1–8:18.

[3] Butcher, J.C. *The numerical analysis of ordinary differential equations: Runge–Kutta and general linear methods* John Wiley & Sons, Chichester, 1987.

[4] Gesla, A., Duguet, Y., Le Quéré, P. and Witkowski, L.M. *Stability analysis of periodic orbits in nonlinear dynamical systems using Chebyshev polynomials* arXiv:2407.18230 (2024).

[5] Hestenes, M.R. and Stiefel, E. *Methods of conjugate gradients for solving linear systems* J. Res. Natl. Bur. Stand., 49 (6) (1952), 409–436.

[6] Kim, W. T., et al. *Application of Nonlinear Dynamical Systems Theory to Chaos Control* IEEE Trans. Autom. Control, 42 (9) (1997), 1297–1302.

[7] Shewchuk, J.R. *An introduction to the conjugate gradient method without the agonizing pain* Carnegie Mellon Univ. Technical Report, 1994.

[8] Wang, X. and Wang, S. *Stability analysis of periodic solutions in chaotic systems: A numerical approach* J. Comput. Nonlinear Dyn., 14 (6) (2019), pp. 1–???

[9] Zhang, J. and Yao, X. *A comparative study of numerical methods for periodic solution tracing in chaotic systems* Nonlinear Anal.: Real World Appl., 54 (2020), 103121.

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 1, 2026, pp 77–101
https://doi.org/10.22067/ijnao.2025.94700.1692
https://ijnao.um.ac.ir/

CC BY NC

Research Article

Fractional-order modeling and numerical analysis of double intermediate enzymatic reaction model using generalized Euler method

B. Koushick, N. Anbazhagan, R. Vignesh Raju, N. Jeeva*

Abstract

This study investigates the fractional modeling of a double intermediate enzymatic reaction model using the Caputo derivative. A system of nonlinear fractional differential equations is formulated in dimensionless form to capture the complex dynamics of intermediate enzymes.

*Corresponding author

Received 31 July 2025; revised 31 October 2025; accepted 15 November 2025

B. Koushick
Department of Mathematics, Alagappa University, Karaikudi 630003, Tamil Nadu, India. e-mail: koushick@alagappauniversity.ac.in

N. Anbazhagan
Department of Mathematics, Alagappa University, Karaikudi 630003, Tamil Nadu, India. e-mail: anbazhagann@alagappauniversity.ac.in

R. Vignesh Raju
PG & Research Department of Mathematics, Vidhyaa Giri College of Arts and Science, Puduvayal 630108, Tamil Nadu, India. e-mail: vignesh1325@gmail.com

N. Jeeva
Department of Mathematics, K.S.Rangasamy College of Technology, Tiruchengode 637215, Namakkal (Dt.), Tamil Nadu, India
Centre for Research and Development, K.S.Rangasamy College of Technology, Tiruchengode 637215, Namakkal (Dt.), Tamil Nadu, India e-mail: jeevapirc2405@gmail.com

How to cite this article
Koushick, B., Anbazhagan, N., Vignesh Raju, R. and Jeeva, N., Fractional-order modeling and numerical analysis of double intermediate enzymatic reaction model using generalized Euler method. *Iran. J. Numer. Anal. Optim.*, 2026; 16(1): 77-101. https://doi.org/10.22067/ijnao.2025.94700.1692

The properties of the model are explored using key theorems related to fractional calculus. The generalized Euler method is employed to obtain numerical solutions for various fractional orders, highlighting the memory-dependent behavior of the system. The numerical simulations are consistent with theoretical expectations, confirming the validity and effectiveness of the proposed method. Various dimensionless parameters are systematically altered, and the corresponding system responses are captured using two-dimensional plots and surface visualizations in MATLAB®. The simulation results highlight that variations in these parameters significantly influence the transient and steady-state dynamics, illustrating their biochemical relevance and the utility of the fractional-order model in capturing such sensitivities.

AMS subject classifications (2020): Primary 34A34; Secondary 34E10, 65L05.

Keywords: Enzyme reaction model; Dimensionless formation; Fractional derivative; Extended Euler method; Numerical simulation

1 Introduction

Enzymes, which are proteins composed of amino acids, play a vital role in facilitating chemical transformations in various fields, including biochemistry, the food industry, and biological systems. As biological catalysts, enzymes accelerate the rate of chemical reactions without being consumed or being permanently altered. Each enzyme is highly specific and promotes a reaction with a specific substrate. During the catalytic process, the enzyme binds to the substrate and often undergoes a temporary change in its shape to facilitate the reaction. This interaction illustrates the precise relationship between the enzymes and their substrates. In enzymology, pre-steady-state kinetics focuses on the initial phase of a reaction prior to the formation of products or intermediates, offering insights into the early molecular interactions between enzymes and substrates [21, 13].

Mathematical models are essential tools for understanding complex dynamic systems, allowing researchers to simulate and analyze the interactions between various components. These models help predict the system behavior, optimize conditions, and explore the effects of various factors on dynamic processes. Analytical solutions provide approximate expressions that describe the temporal evolution of these interactions, and offer insights into the underlying mechanisms without relying on numerical approximations. By applying advanced techniques like the optimal homotopy analysis method, homotopy perturbation method, Taylor series method, Abkari–Ganji method, and Pade approximation method, researchers can capture detailed system behavior, especially during early transient phases [4, 9, 10, 19, 18, 16, 17].

Several studies have demonstrated the power of mathematical modeling to advance our understanding of enzyme kinetics. For instance, Murrary [20] proposed a novel method to solve the time-dependent Michaelis–Menten enzymatic reaction by reducing the system to a single nonlinear ordinary differential equation and applying the optimal homotopy analysis method to obtain accurate and efficient analytical solutions. M.S. Ibrahim et al. [8] utilized the variational iteration method to address nonlinear enzyme reaction systems, showcasing its effectiveness in capturing the complex dynamics of enzymatic processes. Sharmila, Praveen, and Rajendran [27] conducted an in-depth mathematical study on enzyme-catalyzed reactions, shedding light on the impact of kinetic parameters through rigorous analysis. Similarly, previous work [28] has been focused on deriving exact solutions to nonlinear boundary value problems related to enzyme-substrate interactions, highlighting the significance of analytical methods in understanding biochemical reaction mechanisms. Collectively, these contributions underscore the essential role of mathematical approaches in the modeling, analysis, and optimization of enzyme-driven processes.

Frenzen [5] introduced an innovative application of the quasi-steady state assumption (QSSA) to a two-step enzymatic reaction model, expanding its applicability by allowing the initial enzyme-to-substrate concentration ratio to take any positive value, rather than being limited to small ratios. This further reveals that maintaining a constant enzyme-to-substrate ratio while reducing the initial substrate concentration enables the fulfillment of QSSA validity conditions, thereby broadening its practical utility in enzymatic kinetics. Recently, the study in [23] theoretically investigated the pre-steady-state dynamics of nonlinear enzymatic reactions that involve two intermediate complexes. Approximate analytical solutions for the concentrations of the substrate, first enzyme-substrate complex, and second complex were derived using homotopy perturbation and Taylor series. The accuracy of these analytical solutions was validated by comparing them with numerical simulation results. Saad, Abdo, and Hamanah [24] investigated the existence and controllability of multi-term fractional coupled systems using the generalized $(\psi\ ,\ \omega)$-Caputo-Fabrizio operator. Alqhtani, Sadek, and Saad [2] introduced the Mittag-Leffler–Caputo–Fabrizio fractional derivative and presented a corresponding numerical approach for its implementation.

Recent studies have explored advanced modeling approaches to better understand complex biochemical dynamics using fractional calculus. These studies include the development and simulation of nonlinear fractional-order biochemical reaction models and application of fractal–fractional derivatives with different kernels. Researchers have also investigated high-performance computational techniques for analyzing reversible two-step reactions with time-fractional behavior. Such models provide deeper insights into the memory effects and anomalous diffusion in biological systems [1, 22, 3]. The extended Euler method for fractional-order systems provides a

numerical approach for approximating the solutions of differential equations that involve memory effects. This is particularly effective for systems modeled using Caputo fractional derivatives that incorporate the initial conditions in a physically meaningful manner. This method discretizes the fractional derivative by considering the entire history of the solution and capturing long-term dynamics more accurately. It is widely used because of its simplicity, efficiency, and suitability for simulating complex real-world processes governed by noninteger order dynamics [12, 11, 6, 26].

The primary objective of this study is to analyze and solve a double-intermediate enzyme reaction model using the fractional Caputo derivative within the framework of the generalized Euler method. Fractional calculus is appropriate because enzymatic reactions exhibit hereditary properties and distributed time delays owing to multiple intermediate complexes and conformational changes. The fractional order (ς) quantifies this memory effect by bridging the instantaneous kinetics (integer order) and long-term hereditary responses (noninteger order). This improves the realism of the biochemical model. The remainder of this paper is organized as follows: Section 2 outlines the basic definitions and properties of the fractional calculus. In section 3, the proposed model is formulated in dimensionless form and subsequently transformed into its fractional Caputo derivative. Section 4 presents a theoretical analysis of key theorems related to the proposed model. Section 5 details the numerical solution of the model using the generalized Euler method and discusses the obtained results. Finally, section 6 summarizes the findings and outlines the possible directions for future research.

2 Preliminaries

This section introduces the essential mathematical background and foundational concepts of fractional calculus that are crucial for the mathematical formulation of the proposed model.

Definition 1 (Fractional integral in the Riemann–Liouville sense [7])**.** Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a real-valued function, and let $\varsigma > 0$. The Riemann–Liouville fractional integral of order ς is defined as

$$\mathcal{I}_t^{\varsigma} f(t) = \frac{1}{\Gamma(\varsigma)} \int_0^t (t-\tau)^{\varsigma-1} f(\tau)\, d\tau, \tag{1}$$

where $\Gamma(\cdot)$ denotes the Gamma function.

Definition 2 (Caputo fractional derivative [6])**.** For $0 < \varsigma \leq 1$ and $f \in C^1([0,\infty))$, the Caputo derivative of the order ς is given by

$$ {}^{C}D_t^{\varsigma} f(t) = \frac{1}{\Gamma(1-\varsigma)} \int_0^t f'(s)(t-s)^{-\varsigma}\, ds. \tag{2} $$

Definition 3 (Mittag–Leffler functions [15])**.** The generalized Mittag–Leffler functions with one and two parameters are defined by the following power series expansions:

$$ \mathbb{E}_{\alpha}(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\alpha k + 1)}, \quad \alpha > 0, \tag{3} $$

$$ \mathbb{E}_{\alpha,\beta}(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\alpha k + \beta)}, \quad \alpha > 0, \beta > 0. \tag{4} $$

These functions serve as natural generalizations of the exponential function and are widely used in fractional calculus.

Definition 4 (Euler's Gamma function [12, 26])**.** For any real number $x > 0$, the Gamma function is defined by the integral

$$ \Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t}\, dt. \tag{5} $$

Lemma 1. Let $f \in L^{\infty}(\mathbb{R}) \cap \mathcal{B}(\mathbb{R})$ and let $\varsigma \in \mathbb{R}$ with $n-1 < \varsigma < n$, where $n \in \mathbb{N}$. Then, the following properties hold:

1. $\mathcal{D}_t^{\varsigma}\, \mathcal{I}_t^{\varsigma} f(t) = f(t)$,
2. $\mathcal{I}_t^{\varsigma}\, \mathcal{D}_t^{\varsigma} f(t) = f(t) - \sum_{k=0}^{n-1} \frac{t^k}{k!} f^{(k)}(0)$.

3 Formulation of the model

The dynamics of enzyme-catalyzed reactions with two intermediate complexes have been previously analyzed using classical integer-order differential equation models [5, 23]. These models capture the stepwise kinetics of enzyme-substrate interactions, emphasizing the transient behavior and steady-state approximations of the intermediate species. Building on this foundation, we extend the analysis by incorporating fractional-order derivatives to account for memory effects and anomalous kinetics inherent in biochemical systems.

Consider a model of a double intermediate enzyme-catalyst reaction:

$$ \mathcal{E}_n + \mathcal{S}_u \underset{\kappa_{-1}}{\overset{\kappa_1}{\longleftrightarrow}} \mathcal{Y} \xrightarrow{\kappa_2} \mathcal{Z} \xrightarrow{\kappa_3} \mathcal{E}_n + \mathcal{P}_r. \tag{6} $$

An enzyme ($\mathcal{E}_n$) interacts reversibly with a substrate ($\mathcal{S}_u$) to form an intermediate complex ($\mathcal{Y}$) known as the enzyme-substrate complex. This intermediate is then irreversibly transformed into another enzyme-substrate complex ($\mathcal{Z}$), which subsequently dissociates to release the original enzyme ($\mathcal{E}_n$) and the final product ($\mathcal{P}_r$). The process involves rate constants κ_1, κ_{-1}, κ_2, and κ_3. Based on the assumption that the reverse reactions involving κ_{-2} and κ_{-3} are negligible owing to their low rates, these steps are effectively treated as irreversible reactions. Figure 1 illustrates the transmission of the double intermediate enzyme reaction model.

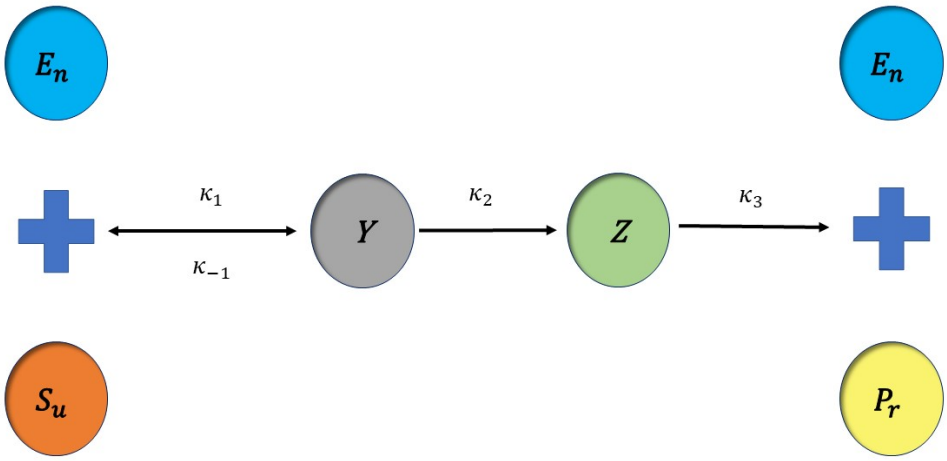


Figure 1: Flowchart of the double intermediate enzyme reaction model

The reaction kinetics, following the law of mass action, are governed by

$$\frac{d\mathcal{S}_u}{d\tau} = -\kappa_1 \mathcal{S}_u \mathcal{E}_n + \kappa_{-1} \mathcal{Y}, \tag{7}$$

$$\frac{d\mathcal{E}_n}{d\tau} = -\kappa_1 \mathcal{S}_u \mathcal{E}_n + \kappa_{-1} \mathcal{Y} + \kappa_3 \mathcal{Z}, \tag{8}$$

$$\frac{d\mathcal{Y}}{d\tau} = \kappa_1 \mathcal{S}_u \mathcal{E}_n - (\kappa_{-1} + \kappa_2) \mathcal{Y}, \tag{9}$$

$$\frac{d\mathcal{Z}}{d\tau} = \kappa_2 \mathcal{Y} - \kappa_3 \mathcal{Z}, \tag{10}$$

$$\frac{d\mathcal{P}_r}{d\tau} = \kappa_3 \mathcal{Z}, \tag{11}$$

with the following initial conditions:

$$\mathcal{E}_n(0) = \mathcal{E}_{n0}, \quad \mathcal{S}_u(0) = \mathcal{S}_{u0}, \quad \mathcal{Y}(0) = 0, \quad \mathcal{Z}(0) = 0, \quad \mathcal{P}_r(0) = 0. \tag{12}$$

By applying the conservation relationships corresponding to (7)–(11), together with the initial conditions specified in (12), the following equations are obtained:

$$\mathcal{E}_n + \mathcal{Y} + \mathcal{Z} = \mathcal{E}_{n0}, \quad \mathcal{S}_u + \mathcal{Y} + \mathcal{Z} + \mathcal{P}_r = \mathcal{S}_{u0}. \tag{13}$$

Adding (10) and (11) yields

$$\frac{d\mathcal{Z}}{d\tau} + \frac{d\mathcal{P}_r}{d\tau} = \kappa_2 \mathcal{Y}. \tag{14}$$

After the pre-steady-state phase, the reaction rate involving the formation of complex $\mathcal{P}_r + 2$ reaches a stable level, leading to an approximately constant concentration of the initial enzyme-substrate complex, $\mathcal{Y}$. At this stage, the derivative of $\mathcal{Y}$ becomes negligible and can be approximated as zero. By combining this condition with the conservation relation given in (13), $\mathcal{Y}$ can be expressed as a function of $\mathcal{S}$ and $\mathcal{Z}$ as follows:

$$\mathcal{Y} = \frac{(\mathcal{E}_{n0} - \mathcal{Z})\mathcal{S}_u}{\mathcal{K} + \mathcal{S}_u}, \tag{15}$$

where the Michaelis constant $\mathcal{K}$ is defined as

$$\mathcal{K} = \frac{\kappa_{-1} + \kappa_2}{\kappa_1}. \tag{16}$$

According to (15), the variable reaches a steady state, based on the current values of $\mathcal{S}_u$ and $\mathcal{Z}$. Although (7) and (10) govern the dynamics of $\mathcal{S}_u$ and $\mathcal{Z}$, it is challenging to uphold earlier heuristic reasoning because of uncertainties regarding the relative magnitudes of the variables involved. In certain cases, $\mathcal{Z}$ can be considered negligible if it is significantly smaller than (15). Substituting (15) into (7) yields

$$\frac{d\mathcal{S}_u}{d\tau} = -\frac{\kappa_2 \mathcal{E}_{n0} \mathcal{S}_u}{\mathcal{K} + \mathcal{S}_u}. \tag{17}$$

By using the initial condition $\mathcal{S}_u(0) = \mathcal{S}_{u0}$ and (17), we can approximate the behavior of $\mathcal{S}$ under the assumption that the substrate concentration changes only slightly during the steady-state phase. However, the role of $\mathcal{Z}$ in comparison to $\mathcal{E}_{n0}$ remains to be evaluated. Notably, $\mathcal{Z}$ is expected to increase rapidly from zero during the pre-steady state because the sum $\mathcal{P}_r + \mathcal{Z}$ (related to (10) and (11)) is proportional to $\mathcal{Y}$.

A crucial step in simplifying the problem is reformulating it as a reduced system of differential equations. This approach involves removing $\mathcal{E}_n$ using the conservation relation $\mathcal{E}_n = \mathcal{E}_{n0} - \mathcal{Y} - \mathcal{Z}$ from (13), and determining $\mathcal{P}_r$ in terms of $\mathcal{Z}$, as described in (10). Consequently, the following equations are obtained

$$\frac{d\mathcal{S}_u}{d\tau} = -\kappa_1 \mathcal{S}_u(\mathcal{E}_{n0} - \mathcal{Y} - \mathcal{Z}) + \kappa_{-1}\mathcal{Y}, \tag{18}$$

$$\frac{d\mathcal{Y}}{d\tau} = \kappa_1 \mathcal{S}_u(\mathcal{E}_{n0} - \mathcal{Y} - \mathcal{Z}) - (\kappa_{-1} + \kappa_2)\mathcal{Y}, \tag{19}$$

$$\frac{d\mathcal{Z}}{d\tau} = \kappa_2 \mathcal{Y} - \kappa_3 \mathcal{Z}, \tag{20}$$

with the following initial conditions:

$$\mathcal{S}_u(0) = \mathcal{S}_{u0}, \quad \mathcal{Y}(0) = 0, \quad \mathcal{Z}(0) = 0. \tag{21}$$

To advance the analysis in a structured manner, it is necessary to nondimensionalize the system at this stage. This involves defining the following dimensionless variables and parameters:

$$\begin{gathered} \mathcal{S}_u = \delta \mathcal{U}, \quad \mathcal{Y} = \varpi \mathcal{V}, \quad \mathcal{Z} = \gamma \mathcal{W}, \quad \tau = \mathcal{V}t, \quad \gamma = \mathcal{E}_{n0}, \quad \delta = \mathcal{S}_{u0}, \\ \varpi = \frac{\mathcal{E}_{n0}\mathcal{S}_{u0}}{\mathcal{S}_{u0} + \mathcal{K}}, \quad \vartheta = \frac{\mathcal{S}_{u0}}{\mathcal{K}}, \quad \sigma = \frac{\kappa_{-1}}{\kappa_2}, \quad \phi = \frac{\kappa_3(\kappa_{-1} + \kappa_2)}{\kappa_1}. \end{gathered} \tag{22}$$

Accordingly, using (22), the system (18)–(20) is transformed into the following dimensionless form:

$$\frac{d\mathcal{U}}{dt} = -\varpi \mathcal{U} + \vartheta\theta\varpi \mathcal{U}\mathcal{V} + \varpi \mathcal{U}\mathcal{W} + \theta\alpha\varpi\sigma\mathcal{V}, \tag{23}$$

$$\frac{d\mathcal{V}}{dt} = \mathcal{U} - \vartheta\theta\mathcal{U}\mathcal{V} - \mathcal{U}\mathcal{W} - \theta\mathcal{V}, \tag{24}$$

$$\frac{d\mathcal{W}}{dt} = \vartheta\theta^2\alpha\mathcal{V} - \phi\theta\mathcal{W}. \tag{25}$$

To simplify the analysis, the following dimensionless parameter values were assumed:

$$\theta = \frac{1}{1+\vartheta}, \quad \alpha = \frac{1}{1+\sigma},$$

with the following dimensionless initial conditions from (21):

$$\mathcal{U}(0) = 1, \quad \mathcal{V}(0) = 0, \quad \mathcal{W}(0) = 0. \tag{26}$$

The proposed double-enzyme reaction system (23)–(25) is formulated in fractional order using the Caputo fractional derivative [6, 25] as follows:

$${}^{C}D_t^{\varsigma}\mathcal{U} = -\varpi \mathcal{U} + \vartheta\theta\varpi \mathcal{U}\mathcal{V} + \varpi \mathcal{U}\mathcal{W} + \theta\alpha\varpi\sigma\mathcal{V}, \tag{27}$$

$${}^{C}D_t^{\varsigma}\mathcal{V} = \mathcal{U} - \vartheta\theta\mathcal{U}\mathcal{V} - \mathcal{U}\mathcal{W} - \theta\mathcal{V}, \tag{28}$$

$${}^{C}D_t^{\varsigma}\mathcal{W} = \vartheta\theta^2\alpha\mathcal{V} - \phi\theta\mathcal{W}, \tag{29}$$

with the following initial conditions:

$$\mathcal{U}(0) = 1, \quad \mathcal{V}(0) = 0, \quad \mathcal{W}(0) = 0, \tag{30}$$

where ${}^{C}D_t^{\varsigma}$ denotes the Caputo fractional derivative of order ς, and ς represents the fractional time derivative.

4 Analysis of the model

In this section, we aim to establish the positivity, boundedness, existence, and uniqueness of the solutions, which are essential properties for validating the mathematical consistency of the proposed double-enzyme reaction fraction model.

4.1 Positivity and boundedness

Theorem 1. For the dimensionless concentrations ($\mathcal{U} \geq 0$, $\mathcal{V} \geq 0$, $\mathcal{W} \geq 0$) and nonnegative dimensionless parameters associated with the variables, the solution of the system in (27)–(29) is nonnegative for all $t > 0$.

Proof. From (27), we have

$$\frac{d\mathcal{U}}{dt} + \varpi\mathcal{U} - \vartheta\theta\varpi\mathcal{U}\mathcal{V} - \varpi\mathcal{U}\mathcal{W} - \theta\alpha\varpi\sigma\mathcal{V} = 0 \tag{31}$$

implies

$$\frac{d\mathcal{U}}{dt} = -(\varpi\mathcal{U} - \vartheta\theta\varpi\mathcal{U}\mathcal{V} - \varpi\mathcal{U}\mathcal{W} - \theta\alpha\varpi\sigma\mathcal{V}). \tag{32}$$

Then,

$$\frac{d\mathcal{U}}{dt} \geq -(\varpi\mathcal{U} - \vartheta\theta\varpi\mathcal{U}\mathcal{V} - \varpi\mathcal{U}\mathcal{W}), \tag{33}$$

$$\frac{d\mathcal{U}}{\mathcal{U}} \geq -(\varpi - \vartheta\theta\varpi\mathcal{V} - \varpi\mathcal{W})dt. \tag{34}$$

Now, we have

$$\ln(\mathcal{U}(t)) \geq -\int(\varpi - \vartheta\theta\varpi\mathcal{V} - \varpi\mathcal{W})dt + C. \tag{35}$$

Define

$$\mathcal{L}(t) = -\int(\varpi - \vartheta\theta\varpi\mathcal{V} - \varpi\mathcal{W})dt, \tag{36}$$

which follows that

$$\ln(\mathcal{U}(t)) \geq \mathcal{L}(t) + C. \tag{37}$$

At $t = 0$, we have

$$\ln(\mathcal{U}(0)) \geq \mathcal{L}(0) + C. \tag{38}$$

From (37) and (38), we obtain

$$\ln\left(\frac{\mathcal{U}(t)}{\mathcal{U}(0)}\right) \geq \mathcal{L}(t) - \mathcal{L}(0). \tag{39}$$

By multiplying both sides exponentially, we obtain

$$\mathcal{U}(t) \geq \mathcal{U}(0)e^{\mathcal{L}(t)-\mathcal{L}(0)} \geq \mathcal{U}(0). \tag{40}$$

Since $\mathcal{U}(0) = \mathcal{U}_0$ is positive,

$$\mathcal{U}(t) \geq \mathcal{U}(0) > 0 \quad \text{for all} \quad t > 0. \tag{41}$$

Hence, it follows that $\mathcal{U}(t) \geq 0$ for all $t > 0$, confirming that the dimensionless concentration $\mathcal{U}(t)$ remains nonnegative over time. The same procedure applies to other dimensionless state variables, demonstrating that $\mathcal{V}(t)$ and $\mathcal{W}(t)$ are nonnegative for $t > 0$. □

Theorem 2. The solution of the dimensionless system (27)–(29) lies entirely in the invariant region $\Omega = \{(\mathcal{U}, \mathcal{V}, \mathcal{W}) \in \mathbb{R}^3_+ \mid \mathcal{U} + \mathcal{V} + \mathcal{W} = \mathcal{N}\}$, which is bounded.

Proof. The total dimensionless concentration $\mathcal{N}$ can be determined by combining (27)–(29), which can be written as

$$\mathcal{N}(t) = \mathcal{U}(t) + \mathcal{V}(t) + \mathcal{W}(t). \tag{42}$$

Differentiating with respect to t, we obtain

$$\frac{d\mathcal{N}}{dt} = \frac{d\mathcal{U}}{dt} + \frac{d\mathcal{V}}{dt} + \frac{d\mathcal{W}}{dt}. \tag{43}$$

Then,

$$\begin{aligned} \frac{d\mathcal{N}}{dt} =& (-\varpi\mathcal{U} + \vartheta\theta\varpi\mathcal{U}\mathcal{V} + \varpi\mathcal{U}\mathcal{W} + \theta\alpha\varpi\sigma\mathcal{V}) \\ &+ (\mathcal{U} - \vartheta\theta\mathcal{U}\mathcal{V} - \mathcal{U}\mathcal{W} - \theta\mathcal{V}) + (\vartheta\theta^2\alpha\mathcal{V} - \phi\theta\mathcal{W}). \end{aligned} \tag{44}$$

After simplification,

$$\frac{d\mathcal{N}}{dt} = (1-\varpi)\mathcal{U} + (\theta\alpha\varpi\sigma)\mathcal{V} - \phi\theta\mathcal{W} + (\vartheta\theta\varpi - \vartheta\theta)\mathcal{U}\mathcal{V} + (\varpi - 1)\mathcal{U}\mathcal{W}. \tag{45}$$

If all dimensionless parameters are nonnegative, the terms involving $\mathcal{U}, \mathcal{V}, \mathcal{W}$ are bounded. Thus, $\frac{d\mathcal{N}}{dt}$ is bounded and $\mathcal{N}(t)$ does not increase indefinitely. Since $N = \mathcal{U} + \mathcal{V} + \mathcal{W}$ remains bounded and nonnegative for $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{W}$, the system evolves within the region

$$\Omega = \{(\mathcal{U}, \mathcal{V}, \mathcal{W}) \in \mathbb{R}^3_+ \mid \mathcal{U} + \mathcal{V} + \mathcal{W} \leq N_0\}, \tag{46}$$

where $\mathcal{N}_0 = \mathcal{U}(0) + \mathcal{V}(0) + \mathcal{W}(0)$ denotes the total initial value.

Thus, region Ω is positively invariant, meaning that solutions starting in Ω remain in Ω for all $t \geq 0$. (ii) This ensures that the system is feasible and functions appropriately. □

4.2 Uniqueness and existence

Theorem 3. A unique solution exists within the feasible region of the dimensionless fractional-order system (27)–(29), provided that the initial conditions are nonnegative.

Proof. To establish the existence and uniqueness of the dimensionless fractional-order model, we adopt the method outlined in [14], considering the feasible region as $\Psi \times (0, T]$, where Ψ is defined as

$$\Psi = \{(\mathcal{U}, \mathcal{V}, \mathcal{W}) \in \mathbb{R}^3_+ : \max(|\mathcal{U}|, |\mathcal{V}|, |\mathcal{W}|) \leq \rho\}. \tag{47}$$

Consider the mapping

$$\Theta = (\Theta_1(\xi), \Theta_2(\xi), \Theta_3(\xi)), \tag{48}$$

and

$$\begin{aligned} \Theta_1(\xi) &= -\varpi\mathcal{U} + \vartheta\theta\varpi\mathcal{U}\mathcal{V} + \varpi\mathcal{U}\mathcal{W} + \theta\alpha\varpi\sigma\mathcal{V}, \\ \Theta_2(\xi) &= \mathcal{U} - \vartheta\theta\mathcal{U}\mathcal{V} - \mathcal{U}\mathcal{W} - \theta\mathcal{V}, \\ \Theta_3(\xi) &= \vartheta\theta^2\alpha\mathcal{V} - \phi\theta\mathcal{W}. \end{aligned} \tag{49}$$

We denote $\xi = (\mathcal{U}, \mathcal{V}, \mathcal{W})$ and $\bar{\xi} = (\bar{\mathcal{U}}, \bar{\mathcal{V}}, \bar{\mathcal{W}})$. For any $\xi, \bar{xi} \in \Psi$, from (49), we obtain that

$$\|\Theta(\xi) - \Theta(\bar{\xi})\| = |\Theta_1(\xi) - \Theta_1(\bar{\xi})| + |\Theta_2(\xi) - \Theta_2(\bar{\xi})| + |\Theta_3(\xi) - \Theta_3(\bar{\xi})|. \tag{50}$$

Using the values of $\Theta_i(\xi)$ and $\Theta_i(\bar{\xi})$, $i = 1, 2, 3$, we obtain

$$\begin{aligned}
&\|\Theta(\xi)-\Theta(\bar{\xi})\| \\
&=|-\varpi\mathcal{U}+\vartheta\theta\varpi\mathcal{U}\mathcal{V}+\varpi\mathcal{U}\mathcal{W}+\theta\alpha\varpi\sigma\mathcal{V}+\varpi\bar{\mathcal{U}}-\vartheta\theta\varpi\bar{\mathcal{U}}\bar{\mathcal{V}}-\varpi\bar{\mathcal{U}}\bar{\mathcal{W}}-\theta\alpha\varpi\sigma\bar{\mathcal{V}}| \\
&\quad+|\mathcal{U}-\vartheta\theta\mathcal{U}\mathcal{V}-\mathcal{U}\mathcal{W}-\theta\mathcal{V}-\bar{\mathcal{U}}+\vartheta\theta\bar{\mathcal{U}}\bar{\mathcal{V}}+\bar{\mathcal{U}}\bar{\mathcal{W}}+\theta\bar{\mathcal{V}}| \\
&\quad+|\vartheta\theta^2\alpha\mathcal{V}-\phi\theta\mathcal{W}-\vartheta\theta^2\alpha\bar{\mathcal{V}}+\phi\theta\bar{\mathcal{W}}|.
\end{aligned} \tag{51}$$

Rearranging the terms, we have

$$\begin{aligned}
&\|\Theta(\xi)-\Theta(\bar{\xi})\| \\
&=|-\varpi\mathcal{U}+\varpi\bar{\mathcal{U}}+\vartheta\theta\varpi\mathcal{U}\mathcal{V}-\vartheta\theta\varpi\bar{\mathcal{U}}\bar{\mathcal{V}}+\varpi\mathcal{U}\mathcal{W}-\varpi\bar{\mathcal{U}}\bar{\mathcal{W}}+\theta\alpha\varpi\sigma\mathcal{V}-\theta\alpha\varpi\sigma\bar{\mathcal{V}}| \\
&\quad+|\mathcal{U}-\bar{\mathcal{U}}-\vartheta\theta\mathcal{U}\mathcal{V}+\vartheta\theta\bar{\mathcal{U}}\bar{\mathcal{V}}-\mathcal{U}\mathcal{W}+\bar{\mathcal{U}}\bar{\mathcal{W}}-\theta\mathcal{V}+\theta\bar{\mathcal{V}}| \\
&\quad+|\vartheta\theta^2\alpha\mathcal{V}-\vartheta\theta^2\alpha\bar{\mathcal{V}}-\phi\theta\mathcal{W}+\phi\theta\bar{\mathcal{W}}|.
\end{aligned} \tag{52}$$

Using the mean value theorem, we have

$$\begin{aligned}
&\|\Theta(\xi)-\Theta(\bar{\xi})\| \\
&\leq\varpi|\mathcal{U}-\bar{\mathcal{U}}|+\vartheta\theta\varpi\rho[|\mathcal{U}-\bar{\mathcal{U}}|+|\mathcal{V}-\bar{\mathcal{V}}|]+\varpi\rho[|\mathcal{U}-\bar{\mathcal{U}}|+|\mathcal{W}-\bar{\mathcal{W}}|] \\
&\quad+\theta\alpha\varpi\sigma|\mathcal{V}-\bar{\mathcal{V}}|+|\mathcal{U}-\bar{\mathcal{U}}|+\vartheta\theta\rho[|\mathcal{U}-\bar{\mathcal{U}}|+|\mathcal{V}-\bar{\mathcal{V}}|]+\rho[|\mathcal{U}-\bar{\mathcal{U}}|+|\mathcal{W}-\bar{\mathcal{W}}|] \\
&\quad+\theta|\mathcal{V}-\bar{\mathcal{V}}|+\vartheta\theta^2\alpha|\mathcal{V}-\bar{\mathcal{V}}|+\phi\theta|\mathcal{W}-\bar{\mathcal{W}}|.
\end{aligned} \tag{53}$$

After simplification, we obtain

$$\begin{aligned}
&\|\Theta(\xi)-\Theta(\bar{\xi})\| \\
&\leq(\varpi+\vartheta\theta\varpi\rho+\varpi\rho+1+\vartheta\theta\rho+\rho)|\mathcal{U}-\bar{\mathcal{U}}| \\
&\quad+(\vartheta\theta\varpi\rho+\theta\alpha\varpi\sigma+\vartheta\theta\rho+\theta+\vartheta\theta^2\alpha)|\mathcal{V}-\bar{\mathcal{V}}|+(\varpi\rho+\rho+\phi\theta)|\mathcal{W}-\bar{\mathcal{W}}|.
\end{aligned} \tag{54}$$

Thus,

$$\|\Theta(\xi)-\Theta(\bar{\xi})\|\leq\Omega||\xi-\bar{\xi}||, \tag{55}$$

where $Q_1=(\varpi+\vartheta\theta\varpi\rho+\varpi\rho+1+\vartheta\theta\rho+\rho)$, $Q_2=(\vartheta\theta\varpi\rho+\theta\alpha\varpi\sigma+\vartheta\theta\rho+\theta+\vartheta\theta^2\alpha)$, $Q_3=(\varpi\rho+\rho+\phi\theta)$, and $\Omega=\max\{Q_1,Q_2,Q_3\}$.

Therefore, $\Theta(\xi)$ satisfies the Lipschitz condition with respect to ξ. Consequently, the existence and uniqueness of the solution to the dimensionless fractional-order model were guaranteed. □

Theorem 4. The dimensionless model (27)–(29) comprises a unique solution $\xi(t)\in\Psi$ for the given dimensionless initial condition ξ_0 for each $\xi_0=\{\mathcal{U}_0,\mathcal{V}_0,\mathcal{W}_0\}$ for all $t\geq 0$.

5 Numerical simulation

5.1 Extended Euler approach

Motivated by fractional numerical techniques, we adopted the extend Euler method to approximate the solutions for a given system. First, we consider fractional-order differential equations [12, 11, 6, 26]:

$$\mathcal{F}^{\varsigma} w = f(t, w), \quad t \in [0, T], \quad 0 < \varsigma \leq 1. \tag{56}$$

To numerically approximate the solution of the fractional-order differential equation, we divided the interval $[0, c]$ into k subintervals of uniform length $h = \frac{c}{k}$. The corresponding time points are denoted as $t_i = jh$ for $i = 0, 1, \ldots, k$. Assuming that the function $w(t)$ and its fractional derivatives $\mathcal{F}^{\varsigma} w(t)$ and $\mathcal{F}^{2\varsigma} w(t)$ are continuous in the interval $[0, c]$, we can utilize a fractional Taylor series expansion to approximate the solution, as follows:

$$w(t) = w(t_0) + \left(\mathcal{F}^{\varsigma} w(t_0)\right) \frac{t^{\varsigma}}{\Gamma(\varsigma + 1)} + \left(\mathcal{F}^{2\varsigma} w(a_1)\right) \frac{t^{2\varsigma}}{\Gamma(2\varsigma + 1)}. \tag{57}$$

By replacing the fractional derivative $\mathcal{F}^{\varsigma} w(t_0)$ with the function $f(t_0, w(t_0))$ and omitting the higher-order term for a small step size h, the approximation at the next time step becomes

$$w(t_1) = w(t_0) + \frac{h^{\varsigma}}{\Gamma(\varsigma + 1)} f(t_0, w(t_0)). \tag{58}$$

This approach can be applied iteratively to obtain solutions at subsequent time points. Thus, for a general step $t_{i+1} = t_i + h$, the extended Euler approximation for the fractional-order case is

$$w(t_{j+1}) = w(t_j) + \frac{h^{\varsigma}}{\Gamma(\varsigma + 1)} f(t_j, w(t_j)), \quad j = 0, 1, \ldots, k - 1. \tag{59}$$

When the fractional order $\varsigma = 1$, this formulation is naturally reduced to the classical Euler method used for ordinary differential equations.

Applying this extended numerical approach (59) to the given dimensionless system (27)–(29) yields the following iterative expressions:

$$\mathcal{U}_h^{(n+1)} = \mathcal{U}_h^n + \frac{h^{\varsigma}}{\Gamma(\varsigma + 1)} [-\varpi \mathcal{U} + \vartheta \theta \varpi \mathcal{U} \mathcal{V} + \varpi \mathcal{U} \mathcal{W} + \theta \alpha \varpi \sigma \mathcal{V}], \tag{60}$$

$$\mathcal{V}_h^{(n+1)} = \mathcal{V}_h^n + \frac{h^{\varsigma}}{\Gamma(\varsigma + 1)} [\mathcal{U} - \vartheta \theta \mathcal{U} \mathcal{V} - \mathcal{U} \mathcal{W} - \theta \mathcal{V}], \tag{61}$$

$$\mathcal{W}_h^{(n+1)} = \mathcal{W}_h^n + \frac{h^{\varsigma}}{\Gamma(\varsigma + 1)} [\vartheta \theta^2 \alpha \mathcal{V} - \phi \theta \mathcal{W}]. \tag{62}$$

This can be rewritten as

$$\mathcal{U}_h^{(n+1)} = \mathcal{U}_h^n + \beth[-\varpi\mathcal{U} + \vartheta\theta\varpi\mathcal{U}\mathcal{V} + \varpi\mathcal{U}\mathcal{W} + \theta\alpha\varpi\sigma\mathcal{V}], \tag{63}$$

$$\mathcal{V}_h^{(n+1)} = \mathcal{V}_h^n + \beth[\mathcal{U} - \vartheta\theta\mathcal{U}\mathcal{V} - \mathcal{U}\mathcal{W} - \theta\mathcal{V}], \tag{64}$$

$$\mathcal{W}_h^{(n+1)} = \mathcal{W}_h^n + \beth[\vartheta\theta^2\alpha\mathcal{V} - \phi\theta\mathcal{W}], \tag{65}$$

where the parameter $0 < \beth = \frac{h^\varsigma}{\Gamma(\varsigma+1)} < 1$ serves as the step size and the initial values at fixed points are $\mathcal{U}(0) = \mathcal{U}_0$, $\mathcal{V}(0) = \mathcal{V}_0$, and $\mathcal{W}(0) = \mathcal{W}_0$. This extended numerical scheme is an effective tool for approximating fractional dynamics while maintaining the computational efficiency.

5.2 Numerical experiments and discussions

This subsection presents the dynamic behavior of the proposed fractional-order model based on dimensionless concentrations and parameters. Numerical simulations were performed using MATLAB® to analyze the system response. In the context of this enzymatic reaction model, the dimensionless parameters ϖ, ϑ, σ, and ϕ play significant roles in shaping the biochemical behavior of the system. From a chemical perspective, ϖ quantifies the effective catalytic potential of the enzyme-substrate interaction, reflecting the extent to which the enzyme is engaged in catalysis under given initial conditions. Parameter ϑ represents the ratio of the initial substrate concentration to the Michaelis constant, thus capturing the level of substrate saturation; higher values indicate near-saturation and maximal enzyme activity. The parameter σ characterizes the reversibility of the first enzyme-substrate complex, with larger values suggesting a greater tendency for the complex to dissociate back to the reactants rather than proceed to the next step. Finally, ϕ incorporates the effects of multiple rate constants and represents the overall catalytic throughput, indicating the efficiency of the system in progressing toward product formation.

Figure 2 illustrates the behavior of the dimensionless concentrations $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{W}$ under fixed dimensionless parameter values ($\varpi = \vartheta = \sigma = \phi = 0.01$) over a dimensionless distance of 1 as the fractional order ς varies from 0.6 to 1. In Figure 2a, increasing the fractional order ς led to an increase in the concentration of $\mathcal{U}$. In Figure 2b, as the fractional order ς increased, the concentration of $\mathcal{V}$ decreased. Similarly, Figure 2c shows that increasing the fractional order ς decreased the concentration of $\mathcal{W}$. Notably, all concentrations behave uniformly for $\mathcal{U}, \mathcal{V}, \mathcal{W} \approx 1$.

Figure 3 presents the dynamics of the concentrations $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{W}$ for different dimensionless parameters ($\varpi = 2$, $\vartheta = 0.01$, $\sigma = 0.01$, $\phi = 0.5$) over a dimensionless distance of 10, as the fractional order ς varies from 0.6 to 1. In Figure 3a, an increase in ς causes the concentration

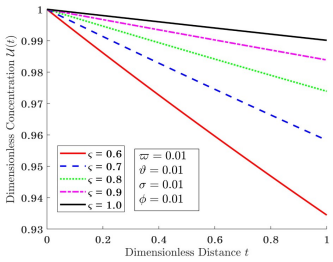
of $\mathcal{U}$ to decrease for $t < 2$ and all solutions tend toward zero thereafter. In Figure 3b, the concentration $\mathcal{V}$ decreases after specific time intervals and approaches zero in all cases after $t < 6$ s. In Figure 3c, as ς increased, the concentration $\mathcal{W}$ decreased at various time intervals. Figure 4 shows the effect of varying dimensionless parameter ϖ on dimensionless concentrations $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{W}$ with fixed values of $\vartheta = 0.01$, $\sigma = 0.01$, and $\phi = 0.5$. Here, the fractional order is fixed at $\varsigma = 1$, which corresponds to classical derivatives. In Figures 4a–4c, an increase in ϖ results in a decrease in all three concentrations, $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{W}$.

Figure 5 shows the impact of the dimensionless parameter ϑ on $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{W}$, while keeping $\varpi = 2$, $\sigma = 0.5$, and $\phi = 0.05$ constant and using the classical case $\varsigma = 1$. In Figures 5a–5c, increasing ϑ leads to an increase in the concentrations $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{W}$. Figure 6 shows the effect of varying the dimensionless parameter σ on $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{W}$ with fixed values of $\varpi = 2$, $\vartheta = 0.01$, $\phi = 0.05$, and fractional order $\varsigma = 1$. In Figures 6a and 6b, as σ increases, the concentrations $\mathcal{U}$ and $\mathcal{V}$ increase. However, in Figure 6c, the concentration $\mathcal{W}$ decreases as σ increases, which is contrary.

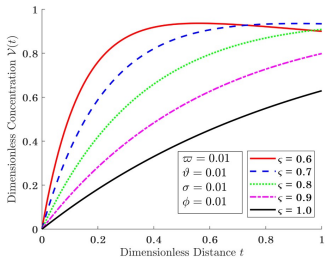
Figure 7 shows the behavior of $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{W}$ as the dimensionless parameter ϕ varies, with $\varpi = 2$, $\vartheta = 0.01$, and $\sigma = 0.01$ fixed, and the fractional order $\varsigma = 1$. In Figures 7a and 7b, changes in ϕ do not affect the concentrations $\mathcal{U}$ and $\mathcal{V}$, which remain constant. In Figure 7c, however, the concentration $\mathcal{W}$ decreases as ϕ increases, which is a contradiction.

Figure 8 shows the surface plots and heat maps of the dimensionless concentration $\mathcal{U}$ for different values of dimensionless parameters ϖ, ϑ, σ, and ϕ. In Figure 8a, the combined influence of ϖ and σ on the concentration $\mathcal{U}$ is presented, which captures their interactive effects on system behavior. Figure 8b shows the combined impact of ϑ and ϕ on $\mathcal{U}$, further revealing how these dimensionless parameters influence concentration through their variations. Dark pink indicates regions of high parameter impact, light blue indicates regions of low impact, and intermediate colors reflect varying degrees of transmission intensity for the selected parameters. Similarly, the same scenarios were analyzed for concentrations $\mathcal{V}$ and $\mathcal{W}$ in Figures 9 and 10, respectively. These visualizations offer valuable insights into the parameter sensitivity and multidimensional dependence of concentrations $\mathcal{U}$, $\mathcal{V}$ and $\mathcal{W}$.

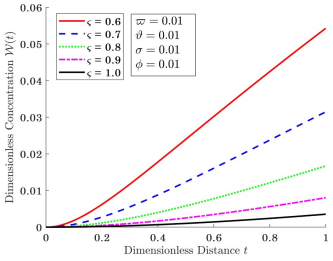
Figure 11 presents pie charts illustrating the distribution of the dimensionless parameters for the two cases. The parameter values on the left side are $\varpi = 2$, $\vartheta = 0.1$, $\sigma = 0.1$, and $\phi = 0.5$. This configuration represents a scenario in which ϖ is dominant, and the remaining parameters have relatively low magnitudes. On the right-hand side, the pie chart shows the distribution for an alternative case with $\varpi = 10$, $\vartheta = 5$, $\sigma = 1$, and $\phi = 2$. Compared with the previous case, all the parameters exhibited larger values, and their proportional contributions varied significantly. These visualizations provide a clear comparison of the relative contributions of the dimensionless parameters under varying conditions.



(a) Dimensionless concentration $\mathcal{U}(t)$

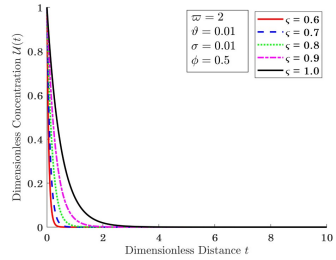


(b) Dimensionless concentration $\mathcal{V}(t)$

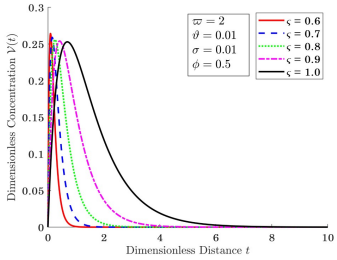


(c) Dimensionless concentration $\mathcal{W}(t)$

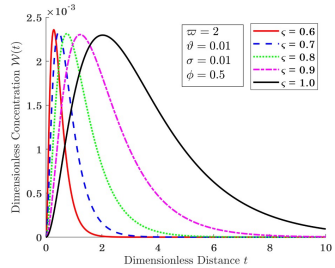
Figure 2: Graphical representation of the dimensionless concentrations $\mathcal{U}, \mathcal{V}, \mathcal{W}$ for fixed dimensionless parameters under varying fractional orders ς $(0.6 - 1)$



(a) Dimensionless concentration $\mathcal{U}(t)$

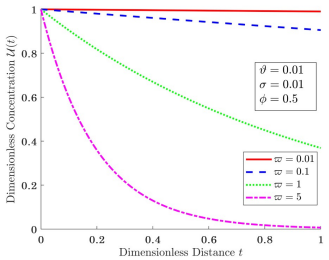


(b) Dimensionless concentration $\mathcal{V}(t)$

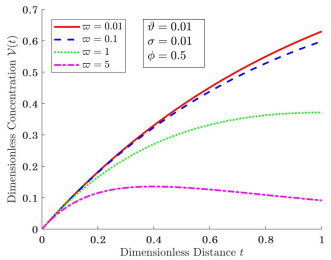


(c) Dimensionless concentration $\mathcal{W}(t)$

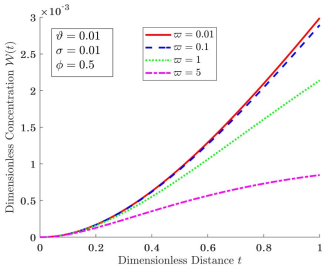
Figure 3: Graphical representation of the dimensionless concentrations $\mathcal{U}, \mathcal{V}, \mathcal{W}$ under different dimensionless parameter values and varying fractional orders ς $(0.6 - 1)$



(a) Impact of varying ϖ on $\mathcal{U}(t)$

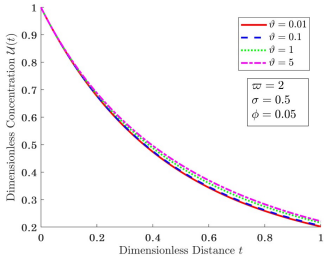


(b) Impact of varying ϖ on $\mathcal{V}(t)$

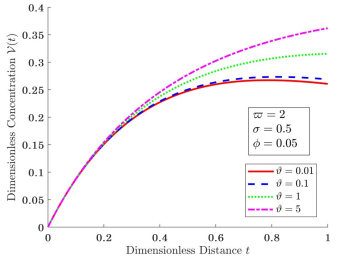


(c) Impact of varying ϖ on $\mathcal{W}(t)$

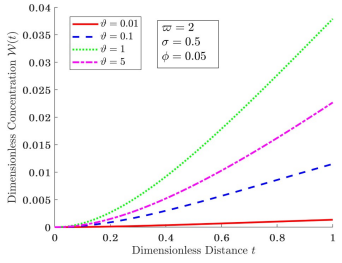
Figure 4: Graphical representation of the dimensionless concentrations $\mathcal{U}, \mathcal{V}, \mathcal{W}$ under different dimensionless parameter values ϖ and fractional order $\varsigma = 1$



(a) Impact of varying ϑ on $\mathcal{U}(t)$

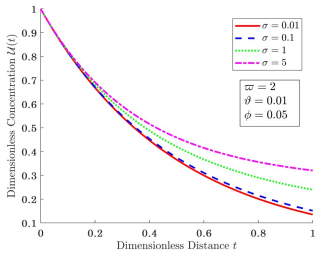


(b) Impact of varying ϑ on $\mathcal{V}(t)$

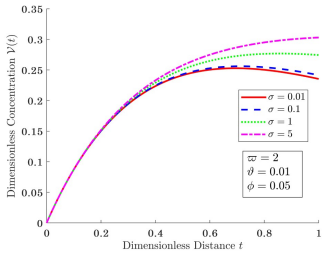


(c) Impact of varying ϑ on $\mathcal{W}(t)$

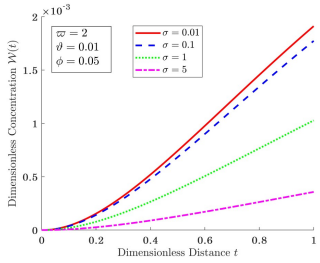
Figure 5: Graphical representation of the dimensionless concentrations $\mathcal{U}, \mathcal{V}, \mathcal{W}$ under different dimensionless parameter values ϑ and fractional order $\varsigma = 1$



(a) Impact of varying σ on $\mathcal{U}(t)$

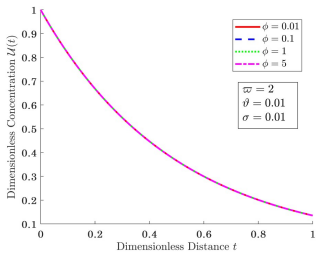


(b) Impact of varying σ on $\mathcal{V}(t)$

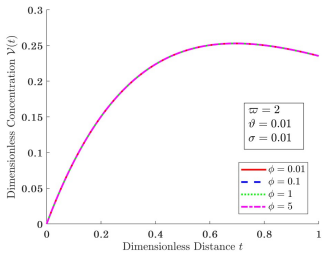


(c) Impact of varying σ on $\mathcal{W}(t)$

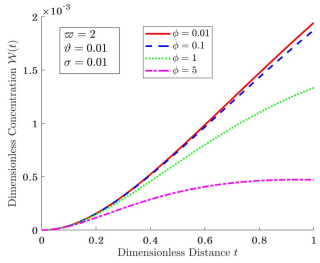
Figure 6: Graphical representation of the dimensionless concentrations $\mathcal{U}, \mathcal{V}, \mathcal{W}$ under different dimensionless parameter values σ and fractional order $\varsigma = 1$



(a) Impact of varying ϕ on $\mathcal{U}(t)$

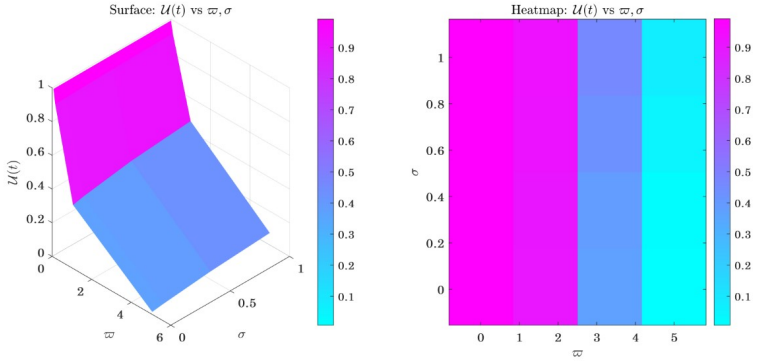


(b) Impact of varying ϕ on $\mathcal{V}(t)$

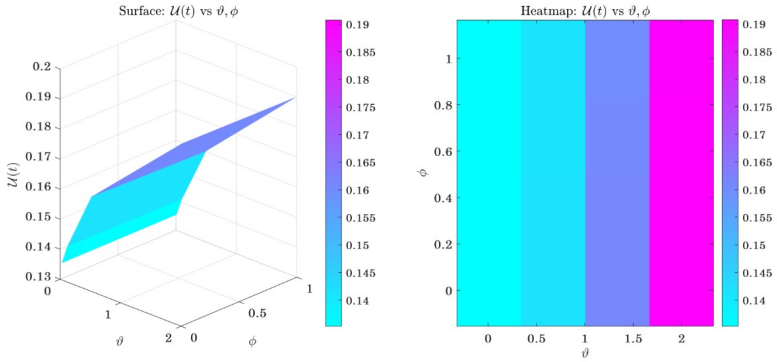


(c) Impact of varying ϕ on $\mathcal{W}(t)$

Figure 7: Graphical representation of the dimensionless concentrations $\mathcal{U}, \mathcal{V}, \mathcal{W}$ under different dimensionless parameter values ϕ and fractional order $\varsigma = 1$



(a) Impact of varying ϖ and σ on $\mathcal{U}(t)$



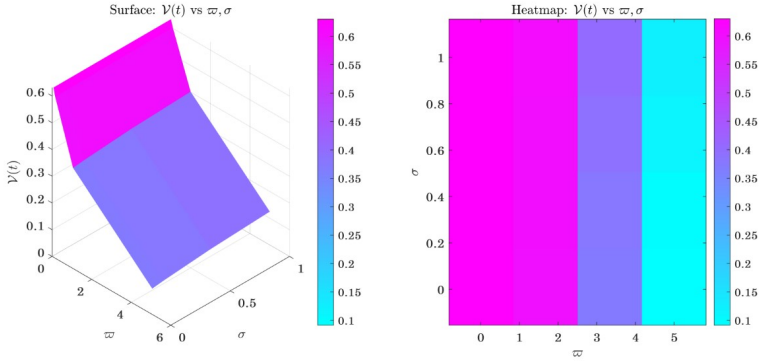
(b) Impact of varying ϑ and ϕ on $\mathcal{U}(t)$

Figure 8: Surface and heatmap representation of the dimensionless concentration $\mathcal{U}(t)$ for different values of the dimensionless parameters ϖ, ϑ, σ, ϕ

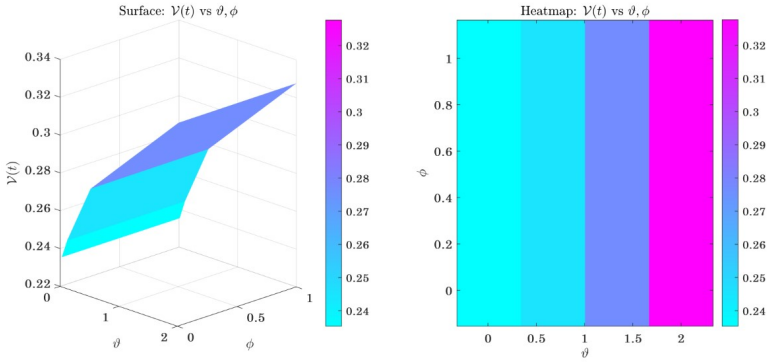
6 Conclusion

Although numerous studies have examined enzyme reaction models using classical differential equations, the application of fractional-order derivatives remains limited, particularly for systems involving multiple intermediate complexes. This study presents the first known attempt to model a double-intermediate enzyme-catalyzed reaction using the fractional calculus. A fractional-order mathematical model was formulated within the Caputo derivative framework, transformed into a dimensionless form, and numerically solved using the generalized Euler method.

The numerical simulations conducted in this study revealed that the behavior of the system was highly sensitive to the dimensionless parameters ϖ, ϑ, σ, and ϕ. Specifically, increasing ϖ leads to a reduction in all concentration levels ($\mathcal{U}$, $\mathcal{V}$, and $\mathcal{W}$), indicating that excessively high catalytic efficiency may deplete the reactants too rapidly, destabilizing the system. Therefore, maintaining a moderate value of ϖ is advisable to ensure a controlled enzymatic progression.



(a) Impact of varying ϖ and σ on $\mathcal{V}(t)$

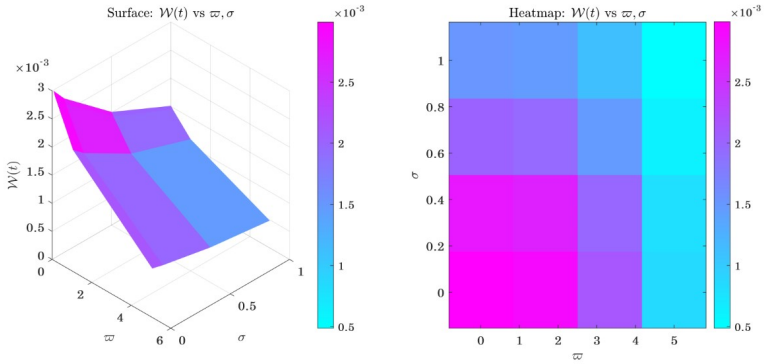


(b) Impact of varying ϑ and ϕ on $\mathcal{V}(t)$

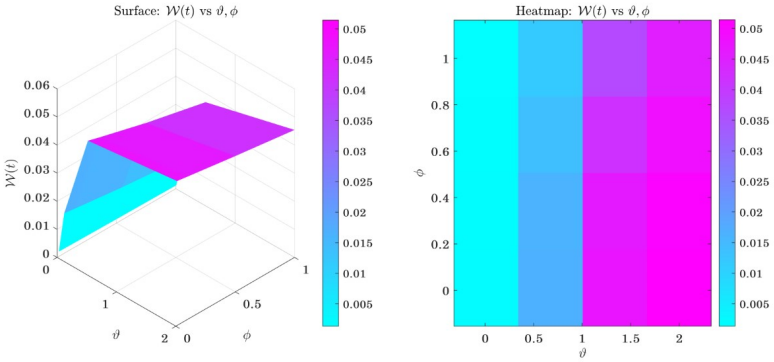
Figure 9: Surface and heatmap representation of the dimensionless concentration $\mathcal{V}(t)$ for different values of the dimensionless parameters ϖ, ϑ, σ, ϕ

Similarly, a high value of ϑ increases all concentrations, suggesting that substrate saturation strongly amplifies the reaction dynamics. However, an excessively high ϑ may lead to over-accumulation of intermediates. By contrast, increasing σ raises $\mathcal{U}$ and $\mathcal{V}$ but reduces $\mathcal{W}$, indicating a trade-off between intermediate complex stability and product formation; thus, minimizing σ can enhance the progression toward product release. Finally, variations in ϕ primarily affect the terminal complex $\mathcal{W}$ with higher values suppressing its accumulation. Hence, reducing ϕ is not recommended if an efficient product formation is desired. Overall, tuning these parameters offers a valuable strategy for modulating enzymatic behavior, and the fractional-order model effectively captures these subtle changes.

This enhanced modeling approach captures subtle dynamic features that are otherwise overlooked in classical models. Overall, this study demonstrates the efficacy of fractional calculus in modeling complex biochemical reactions and provides a foundation for further research. Future



(a) Impact of varying ϖ and σ on $\mathcal{W}(t)$



(b) Impact of varying ϑ and ϕ on $\mathcal{W}(t)$

Figure 10: Surface and heatmap representation of the dimensionless concentration $\mathcal{W}(t)$ for different values of the dimensionless parameters ϖ, ϑ, σ, ϕ

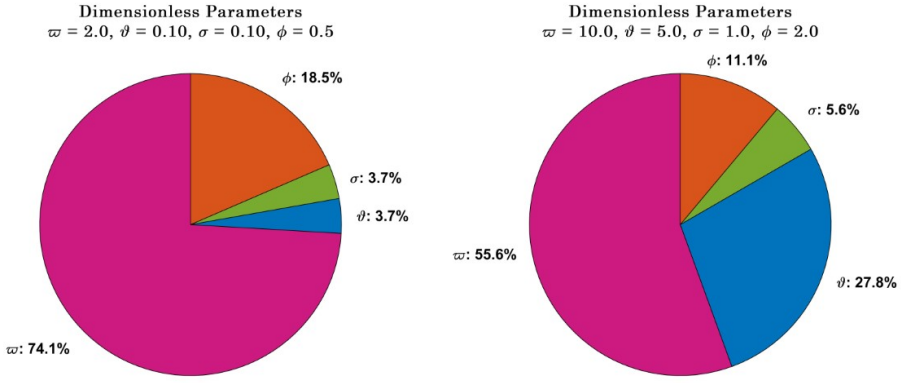


Figure 11: Pie chart representing the sensitivity of the dimensionless parameters ϖ, ϑ, σ, ϕ

research should incorporate machine-learning strategies, parameter estimation techniques, and validation with experimental data to extend the practical applicability of the proposed model.

Acknowledgment

The authors express their sincere gratitude to the anonymous reviewers for their insightful comments, which helped to improve the quality of this manuscript. Also, the author N. Jeeva would like to thank K.S.Rangasamy College of Technology for its continuous support.

Data Availability

All the data generated in this manuscript are included in the list of references.

Conflicts of Interest

The authors declare that they have no conflict of interest associated with the publication of this paper.

Declaration of competing interest

The authors declare that they have no conflicts of interest.

References

[1] Alqhtani, M. and Saad, K.M. *Fractal–fractional Michaelis–Menten enzymatic reaction model via different kernels*, Fractal Fract., 6(1) (2021), 13.

[2] Alqhtani, M., Sadek, L. and Saad, K.M. *The Mittag-Leffler–Caputo–Fabrizio fractional derivative and its numerical approach*, Symmetry, 17(5) (2025), 800.

[3] Chethan, H.B., Turki, N.B. and Prakasha, D.G. *High performance computational approach to study model describing reversible two-step enzymatic reaction with time fractional derivative*, Sci. Rep., 14(1) (2024).

[4] Dharmalingam, K.M., Jeeva, N. and Alessa, N. *Application of Chebyshev Polynomial-Exponential Method and Tamimi-Ansari Method in dengue transmission dynamics: a comparative study*, Int. J. Anal. Appl., 22 (2024), 219.

[5] Frenzen, C.L. and Maini, P.K. *Enzyme kinetics for a two-step enzymic reaction with comparable initial enzyme-substrate ratios*, J. Math. Biol., 26(6) (1988), 689–703.

[6] Haq, I.U., Yavuz, M., Ali, N. and Akgül, A. *A SARS-COV-2 fractional-order mathematical model via the modified Euler method*, Math. Comput. Appl., 27(5) (2022), 82.

[7] Haubold, H.J., Mathai, A.M. and Saxena, R.K. *Mittag-Leffler functions and their applications*, J. Appl. Math., 2011(1) (2011) 1–51.

[8] Ibrahim, M.S., Pavithra, S., Kumar, S.R., Ashokan, R. and Rajendran, L. *Modelling of non linear enzyme reaction process using variational iteration method*, Int. J. Comput. Eng. Res., 6 (2016), 2250–3005.

[9] Jeeva, N. and Dharmalingam, K.M. *Numerical analysis of skin cancer model induced by ultraviolet radiation*, Int. J. Biomathematics, (2024) 2450137.

[10] Jeeva, N. and Dharmalingam, K.M. *Numerical analysis and artificial neural networks for solving nonlinear tuberculosis model in SEITR framework*, Adv. Theory Simul., 8(6) (2025) 2401287.

[11] Khan, F.S., Khalid, M., Bazighifan, O. and El-Mesady, A. *Euler's numerical method on fractional DSEK model under ABC derivative*, Complexity, 2022(1) (2022).

[12] Kim, V.A. and Parovik, R.I. *Application of the explicit Euler method for numerical analysis of a nonlinear fractional oscillation equation*, Fractal Fract., 6(5) (2022), 274.

[13] Koshland, D. *The key-lock theory and the induced fit theory*, Angew. Chem. Int. Ed. Engl., 33 (1995), 2375–2378.

[14] Li, H., Zhang, L., Hu, C., Jiang, Y. and Teng, Z. *Dynamical analysis of a fractional-order predator-prey model incorporating a prey refuge*, J. Appl. Math. Comput., 54(1–2) (2016), 435–449.

[15] Luchko, Y. *Fractional differential equations with the general fractional derivatives of arbitrary order in the Riemann–Liouville sense*, Mathematics, 10(6) (2022), 849.

[16] Manivel, M., Venkatesh, A. and Kumawat, S. *Numerical simulation for the co-infection of Monkeypox and HIV model using fractal-fractional operator*, Model. Earth Syst. Environ., 11(3) (2025) 157.

[17] Manivel, M., Venkatesh, A. and Kumawat, S. *A comprehensive study of monkeypox disease through fractional mathematical modeling*, Math. Model. Numer. Simul. Appl., 5(1) (2025), 65–96.

[18] Manivel, M., Venkatesh, A., Arunkumar, K., Raj, M.P. and Shyamsunder, N. *A mathematical model of the dynamics of the transmission of monkeypox disease using fractional differential equations*, Adv. Theory Simul., 7(9) (2024) 2400330.

[19] Mrope, F. and Jeeva, N. *Modeling the transmission dynamics of banana bunch top disease in banana plants*, Eurasian J. Math. Comput. Appl., 12(3) (2024), 73–90.

[20] Murray, J.D. *Reaction kinetics*, in: Mathematical Biology, Springer, Berlin, Heidelberg, (1989), 109–139.

[21] Nestl, B., Nebel, B. and Hauer, B. *Recent progress in industrial biocatalysis*, Curr. Opin. Chem. Biol., 15 (2010), 187–193.

[22] Radhakrishnan, B., Chandru, P. and Nieto, J.J. *A study of nonlinear fractional-order biochemical reaction model and numerical simulations*, Nonlinear Anal. Model. Control, 29(3) (2024), 588–605.

[23] Reena, A., Raja, R. and Swaminathan, R. *Theoretical analysis of pre-steady state behaviour of non-linear double intermediate enzymatic reaction*, Contemp. Math., (2024), 2565–2584.

[24] Saad, K.M., Abdo, M.S. and Hamanah, W.M. *Existence and controllability analysis of multi-term fractional coupled systems with generalized [ψ, ω]-Caputo-Fabrizio operators*, Sci. Rep., 15(1) (2025) 34434.

[25] Saber, S. and Solouma, E. *The generalized Euler method for analyzing zoonotic disease dynamics in baboon–human populations*, Symmetry, 17(4) (2025), 541.

[26] Saber, S., Solouma, E., Alharb, R.A. and Alalyani, A. *Chaos in fractional-order glucose–insulin models with variable derivatives: insights from the Laplace–Adomian decomposition method and generalized Euler techniques*, Fractal Fract., 9(3) (2025), 149.

[27] Sharmila, D., Praveen, T. and Rajendran, L. *Mathematical modeling and analysis of non-linear enzyme catalyzed reaction processes*, J. Theor. Chem., 2013 (2013), 1–7.

[28] Thangapandi, C., Kuppusamy, R., Muthukumar, S. and Srinivasan, R. *Analytical solutions of non-linear boundary value problem for chemical reactions of enzyme substrate*, Malaya J. Matematik, 1 (2020), 435–444.

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 1, 2026, pp 102–121
https://doi.org/10.22067/ijnao.2025.94956.1709
https://ijnao.um.ac.ir/

CC BY NC

Research Article

A novel three-term conjugate gradient approach for deep neural network training

F.S. Hosseini-Meighan and A. Ashrafi*

Abstract

This paper presents a new three-term conjugate gradient (CG) method for large-scale unconstrained optimization with applications to training deep neural networks. The approach employs enhanced CG parameter strategies together with a strong Wolfe line search to guarantee sufficient descent and global convergence under standard smoothness assumptions. Extensive numerical experiments are conducted on both the standard CUTEr benchmark test problems and the MNIST digit classification task. The proposed MPRP-IFR hybrid algorithm demonstrates superior efficiency compared to traditional CG variants, achieving faster convergence and reduced computational cost on CUTEr problems as well as high classification accuracy and stable training dynamics for deep learning models. These findings highlight the robustness and effectiveness of the proposed CG scheme for large-scale optimization.

AMS subject classifications (2020): Primary 45D05; Secondary 42C10, 65G99.

*Corresponding author

Received 16 August 2025; revised 28 October 2025; accepted 19 November 2025

Faeze Sadat Hosseini-Meighan
Department of Mathematics, Faculty of Mathematics,Statistics and Computer Science, Semnan University, P.O. Box 35131-19111, Semnan, Iran. e-mail: faezehosseini.mighan@semnan.ac.ir

Ali Ashrafi
Department of Mathematics, Faculty of Mathematics,Statistics and Computer Science, Semnan University, P.O. Box 35131-19111, Semnan, Iran. e-mail: a_ashrafi@semnan.ac.ir

How to cite this article
Hosseini-Meighan, F.S. and Ashrafi, A., A novel three-term conjugate gradient approach for deep neural network training. *Iran. J. Numer. Anal. Optim.*, 2026; 16(1): 102-121. https://doi.org/10.22067/ijnao.2025.94956.1709

Keywords: Conjugate gradient method; Unconstrained optimization; line search; Global convergence; Deep neural network.

1 Introduction

The conjugate gradient (CG) method is widely regarded as a powerful tool for solving large-scale unconstrained optimization problems, primarily due to its low memory requirements and strong convergence properties. In particular, the CG method has been extensively utilized in practical fields such as signal processing [12], machine learning [21], and image processing [9]. In the general unconstrained optimization problem, the goal is to minimize a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$:

$$\min_{x\in\mathbb{R}^n} f(x), \tag{1}$$

where $x \in \mathbb{R}^n$ denotes the vector of optimization variables. Iterative methods, such as the CG algorithms, generate a sequence of iterates $\{x_k\}$ designed to converge to a local (and, under convexity, global) minimum of f.

At each iteration k, the CG method constructs a search direction d_k and performs an update of the form:

$$x_{k+1} = x_k + \alpha_k d_k, \tag{2}$$

where $\alpha_k > 0$ is the step size, and d_k is the search direction, recursively defined as

$$d_k = \begin{cases} -g_k, & k = 1, \\ -g_k + \beta_k d_{k-1}, & k \geq 2, \end{cases} \tag{3}$$

with the gradient $g_k = \nabla f(x_k)$. The efficiency of the CG method is intimately tied to the appropriate choice of the conjugacy parameter β_k.

In order to ensure sufficient progress towards the minimum, the step size α_k is typically computed using a line search procedure. Among these, the Wolfe conditions and the strong Wolfe conditions are standard choices due to their critical role in guaranteeing desirable convergence behavior. The Wolfe conditions require that the step size α_k satisfies

$$f(x_k + \alpha_k d_k) \leq f(x_k) + \delta\alpha_k g_k^T d_k, g(x_k + \alpha_k d_k)^T d_k \geq \sigma g_k^T d_k, \tag{4}$$

where $0 < \delta < \sigma < 1$. The strong Wolfe conditions strengthen the second inequality as follows:

$$f(x_k + \alpha_k d_k) \leq f(x_k) + \delta\alpha_k g_k^T d_k, |g(x_k + \alpha_k d_k)^T d_k| \leq \sigma|g_k^T d_k|. \tag{5}$$

These line search conditions play a crucial role in convergence theory for CG methods.

The evolution of the CG method has given rise to several classical formulas for the parameter β_k, each imbued with unique theoretical and practical properties. Notable among these are

$$\beta_k^{PRP} = \frac{g_k^T(g_k - g_{k-1})}{\|g_{k-1}\|^2}, \quad \text{(Polak–Ribière–Polyak)} \tag{6}$$

$$\beta_k^{FR} = \frac{\|g_k\|^2}{\|g_{k-1}\|^2}, \quad \text{(Fletcher–Reeves)} \tag{7}$$

$$\beta_k^{HS} = \frac{g_k^T(g_k - g_{k-1})}{d_{k-1}^T(g_k - g_{k-1})}, \quad \text{(Hestenes–Stiefel)} \tag{8}$$

$$\beta_k^{DY} = \frac{\|g_k\|^2}{d_{k-1}^T(g_k - g_{k-1})} \quad \text{(Dai–Yuan).} \tag{9}$$

The FR and DY methods are celebrated for their strong theoretical convergence, particularly when paired with the strong Wolfe line search [5, 3]. However, they can suffer from practical issues, such as the jamming phenomenon, which slows convergence. By contrast, the PRP and HS methods tend to perform better numerically but may lack global convergence guarantees [20, 7].

To overcome these challenges, several improved and modified parameters have been proposed. Jiang and Jian [8] advanced the field with the improved Fletcher–Reeves (IFR) parameter, which leverages the second Wolfe inequality. The IFR parameter is defined as

$$\beta_k^{IFR} = \frac{\|g_k\|^2}{\|g_{k-1}\|^2} \cdot \frac{|g_k^T d_{k-1}|}{-g_{k-1}^T d_{k-1}}, \tag{10}$$

enabling stronger global convergence and helping the algorithm to avoid stalling.

Similarly, Ma et al. [13] introduced the modified Polak–Ribière–Polyak (MPRP) parameter. Building on the WYL formula,

$$\beta_k^{WYL} = \frac{\|g_k\|^2 - \frac{\|g_k\|}{\|g_{k-1}\|} g_k^T g_{k-1}}{\|g_{k-1}\|^2}, \tag{11}$$

they formulated the MPRP parameter as

$$\beta_k^{MPRP} = \beta_k^{WYL} \cdot \frac{|g_k^T d_{k-1}|}{-g_{k-1}^T d_{k-1}}. \tag{12}$$

These modifications are designed to ensure that the search direction maintains sufficient descent and robust convergence properties under the strong Wolfe line search conditions.

Recent advancements in CG methods have explored extensions beyond the traditional two-term update. One such direction is the three-term CG formulation, as proposed by Dada et al. [2], where the search direction at each iteration is enhanced by incorporating an additional term involving the difference of gradients. Specifically, the search direction d_k is defined as

$$d_k = \begin{cases} -g_k, & k = 0, \\ -g_k + \beta_k d_{k-1} + \theta_k y_{k-1}, & k > 0, \end{cases} \tag{13}$$

where $y_{k-1} = g_k - g_{k-1}$, $\beta_k = \frac{g_k^T y_{k-1}}{\|g_{k-1}\|^2}$, and $\theta_k = \frac{g_k^T d_{k-1}}{\|d_{k-1}\|^2}$.

Building upon this, Dada et al. [2] further proposed a hybrid parameterization, named as CHM, by forming a convex combination of the classical Polak–Ribière–Polyak (PRP) and Fletcher–Reeves (FR) parameters for β_k. The combined parameter is given by

$$\beta_k^{CHM} = (1 - \varphi)\, \beta_k^{PRP} + \varphi\, \beta_k^{FR}, \qquad 0 < \varphi < 1, \tag{14}$$

where β_k^{PRP} and β_k^{FR} are defined in (6) and (7), respectively.

With this hybrid approach, the new search direction can be compactly expressed as

$$d_k = \begin{cases} -g_k, & k = 0, \\ -g_k + \beta_k^{CHM}\, d_{k-1} + \theta_k y_{k-1}, & k > 0. \end{cases} \tag{15}$$

This augmented structure provides additional flexibility and has demonstrated improved numerical performance in applications such as deep learning optimization.

Recent improvements in CG-type methods include various hybrid and adaptive formulations [15, 16, 17, 18], which have shown promising results in numerical optimization and related applications.

Inspired by these advances, the present work proposes a novel three-term CG method that, for the first time, integrates the IFR and MPRP parameters in an analogous convex combination. The goal is to capitalize on their respective strengths to achieve even stronger convergence and robustness for large-scale unconstrained optimization and deep learning applications.

The remainder of this paper is organized as follows. Section 2 presents the proposed three-term CG algorithm based on a convex combination of IFR and MPRP parameters. Section 3 provides a detailed convergence analysis of the method, including sufficient descent and global convergence. Section 4 reports numerical results on training deep neural networks for the MNIST benchmark. Finally, section 5 concludes the paper and outlines possible future research directions.

2 Proposed three-term CG method

Inspired by the three-term structure in [2], we propose a new CG algorithm, combining the recently developed IFR parameter [8] and the MPRP parameter [13]. The search direction at iteration k is defined as

$$d_k = \begin{cases} -g_k, & k = 0, \\ -g_k + \beta_k^{\text{MPRP-IFR}}\, d_{k-1} + \theta_k y_{k-1}, & k \geq 1, \end{cases} \tag{16}$$

where the coefficients are given by

$$\beta_k^{\text{MPRP-IFR}} = (1-\varphi)\,\beta_k^{\text{MPRP}} + \varphi\,\beta_k^{\text{IFR}}, \tag{17}$$

$$\theta_k = -\frac{g_k^T d_{k-1}}{\|d_{k-1}\|^2}, \tag{18}$$

with $0 < \varphi < 1$. In all experiments, the convex-combination parameter was fixed at $\varphi = 0.5$, which yielded a balanced contribution between the MPRP and IFR components and showed stable performance across test problems. Here, β_k^{MPRP} and β_k^{IFR} are defined, respectively, by [13, 8]

$$\beta_k^{\text{MPRP}} = \left[\frac{\|g_k\|^2 - \frac{\|g_k\|}{\|g_{k-1}\|} g_k^\top g_{k-1}}{\|g_{k-1}\|^2}\right] \cdot \frac{|g_k^\top d_{k-1}|}{-g_{k-1}^\top d_{k-1}}, \tag{19}$$

$$\beta_k^{\text{IFR}} = \frac{\|g_k\|^2}{\|g_{k-1}\|^2} \cdot \frac{|g_k^\top d_{k-1}|}{-g_{k-1}^\top d_{k-1}}. \tag{20}$$

The developed algorithm is described below.

This formulation generalizes the CHM three-term approach, allowing the algorithm to benefit from the improved convergence and numerical performance observed in both the IFR and MPRP schemes.

3 Convergence analysis

In this section, we rigorously analyze the convergence properties of the proposed three-term CG method, whose search direction is given by

Algorithm 1 Proposed three-term CG algorithm (MPRP-IFR Hybrid)

Require: Initial point $x_0 \in \mathbb{R}^n$, tolerance $\varepsilon > 0$, parameter $0 < \varphi < 1$

$g_0 = \nabla f(x_0)$
$d_0 = -g_0$
$k \leftarrow 0$
while $\|g_k\| > \varepsilon$ **do**
    Find α_k by (strong) Wolfe conditions
    $x_{k+1} = x_k + \alpha_k d_k$
    $g_{k+1} = \nabla f(x_{k+1})$
    if $k = 0$ **then**
        $d_{k+1} = -g_{k+1}$
    else if $k = 1$ **then**
        Compute β_1^{New}
        $d_2 = -g_2 + \beta_1^{\text{New}} d_1$
    else
        Compute $\beta_k^{\text{MPRP-IFR}}, \theta_k$ (see equations (19)–(20))
        $d_{k+1} = -g_{k+1} + \beta_k^{\text{MPRP-IFR}}\, d_k + \theta_k y_{k-1}$
    end if
    $k \leftarrow k + 1$
end while

$$d_k = \begin{cases} -g_k, & k = 0, \\ -g_k + \beta_k^{\text{MPRP-IFR}}\, d_{k-1} + \theta_k y_{k-1}, & k \geq 1, \end{cases}$$

where

$$\beta_k^{\text{MPRP-IFR}} = (1-\varphi)\,\beta_k^{\text{MPRP}} + \varphi\,\beta_k^{\text{IFR}}, \qquad 0 < \varphi < 1,$$

and $\theta_k = -\frac{g_k^T d_{k-1}}{\|d_{k-1}\|^2}$, $y_{k-1} = g_k - g_{k-1}$. The parameters β_k^{MPRP} and β_k^{IFR} are as defined in (19)–(20).

3.1 Assumptions

We make the following standard assumptions on the objective function f:

A1 The level set $\Omega = \{x \in \mathbb{R}^n \mid f(x) \leq f(x_0)\}$ is bounded, where x_0 is the initial point.

A2 In a neighborhood U of Ω, the function f is continuously differentiable and the gradient $g(x)$ is Lipschitz continuous. That is, there exists $L > 0$ such that

$$\|g(x) - g(y)\| \leq L\|x - y\|, \quad \forall x, y \in U.$$

Under these assumptions, there exists $\gamma > 0$ such that $\|g(x)\| \leq \gamma$ for all $x \in \Omega$.

3.2 Sufficient descent property

We establish that the search directions generated by our method are sufficiently descending.

Theorem 1 (Sufficient descent)**.** Let $\{x_k\}$ and $\{d_k\}$ be sequences generated by the proposed algorithm, where the step size α_k satisfies the strong Wolfe condition (5) with $0 < \delta < \sigma < \frac{1}{2}$. Then there exists a constant $c > 0$ such that

$$g_k^T d_k \leq -c\|g_k\|^2, \qquad \forall k \geq 0. \tag{21}$$

Proof. We proceed by induction.

Base case ($k = 0$): We have $d_0 = -g_0$, so $g_0^T d_0 = -\|g_0\|^2$. Thus, the claim holds with $c = 1$.

Inductive step ($k \geq 1$): By the definition of the search direction,

$$d_k = -g_k + \beta_k^{\text{MPRP-IFR}}\, d_{k-1} + \theta_k y_{k-1}.$$

Taking the inner product with g_k, we have

$$\begin{aligned} g_k^T d_k &= -\|g_k\|^2 + \beta_k^{\text{MPRP-IFR}}\ (g_k^T d_{k-1}) + \theta_k\, (g_k^T y_{k-1}) \\ &= -\|g_k\|^2 + \beta_k^{\text{MPRP-IFR}}\ (g_k^T d_{k-1}) + \theta_k\, (\|g_k\|^2 - g_k^T g_{k-1}). \end{aligned} \tag{22}$$

Recall that $\theta_k = -\frac{g_k^T d_{k-1}}{\|d_{k-1}\|^2}$, so we can substitute:

$$\theta_k (g_k^T y_{k-1}) = -\frac{g_k^T d_{k-1}}{\|d_{k-1}\|^2}\left(\|g_k\|^2 - g_k^T g_{k-1}\right)$$

This term acts to further control the projected component along y_{k-1} and, under the line search, does not increase $g_k^T d_k$.

From the strong Wolfe condition (5), we have

$$|g_k^T d_{k-1}| \leq \sigma |g_{k-1}^T d_{k-1}|, \quad \text{and since } g_{k-1}^T d_{k-1} < 0 \implies |g_k^T d_{k-1}| \leq -\sigma g_{k-1}^T d_{k-1}.$$

The definition of $\beta_k^{\text{MPRP-IFR}}$ as a convex combination of nonnegative β_k^{MPRP} and β_k^{IFR} ensures it is nonnegative and upper bounded as in [13, 8].

Combining these properties and using arguments similar to [2, Lemma 3], one can show that the cumulative effect of the last two terms in (22) does not exceed a (small) fraction of the first negative term. That is, there exists a constant $c > 0$ (dependent on σ) such that

$$g_k^T d_k \leq -c\|g_k\|^2,$$

which completes the induction. □

Lemma 1. Under the strong Wolfe line search with $\sigma < \frac{1}{2}$, the parameters β_k^{MPRP} and β_k^{IFR} are nonnegative at every iteration.

The nonnegativity of $\beta_k^{\text{MPRP-IFR}}$ guarantees that the search direction maintains sufficient descent, which is crucial for global convergence.

3.3 Global convergence

We conclude this section by establishing the global convergence of the algorithm.

Lemma 2 (Zoutendijk Condition, [19])**.** Suppose assumption **A1** is satisfied. If d_k is a descent direction at each iteration, and α_k satisfies the Wolfe conditions, then

$$\sum_{k=0}^{\infty} \frac{(g_k^T d_k)^2}{\|d_k\|^2} < \infty.$$

Lemma 3 (Uniform boundedness of β_k^{MPRP}, β_k^{IFR}, and θ_k)**.** Suppose that assumptions **A1**–**A2** hold and that the step size α_k satisfies the strong Wolfe line search with $0 < \sigma < \frac{1}{2}$. Then, the parameters β_k^{MPRP}, β_k^{IFR}, and θ_k are uniformly bounded for all $k \geq 0$. That is, there exist constants $C_\beta, C_\theta > 0$ such that

$$0 \leq \beta_k^{\text{MPRP}},\ \beta_k^{\text{IFR}} \leq C_\beta, \qquad |\theta_k| \leq C_\theta.$$

Proof. For β_k^{IFR}, as defined in (20), we have

$$\beta_k^{\text{IFR}} = \frac{\|g_k\|^2}{\|g_{k-1}\|^2} \cdot \frac{|g_k^T d_{k-1}|}{-g_{k-1}^T d_{k-1}}.$$

From [8, Theorem 1], under the strong Wolfe condition with $0 < \sigma < 1/2$, there exists $0 < C_0 < 1$ such that

$$\frac{|g_k^T d_{k-1}|}{-g_{k-1}^T d_{k-1}} \leq \sigma < \frac{1}{2}.$$

Combined with the Cauchy–Schwarz inequality and the boundedness of gradients on the level set Ω, both $\|g_k\|$ and $\|g_{k-1}\|$ are bounded above and below (away from zero unless convergence occurs). Therefore, there exists $C_\beta > 0$ such that $0 \leq \beta_k^{\mathrm{IFR}} \leq C_\beta$ for all k.

Similarly, for β_k^{MPRP}, following [13, Theorem 2.1] and the above inequalities,

$$\beta_k^{\mathrm{MPRP}} \leq 2\sigma\beta_k^{\mathrm{FR}} \leq 2\sigma\frac{\|g_k\|^2}{\|g_{k-1}\|^2} \leq C_{\beta'},$$

where $C_{\beta'} > 0$ denotes a constant that bounds the ratio $|\beta'_{k+1}|/|\beta_k|$ in the convergence analysis.

For $\theta_k = -\frac{g_k^T d_{k-1}}{\|d_{k-1}\|^2}$, using the strong Wolfe inequality, the Cauchy–Schwarz inequality, and the boundedness of search directions (from the descent lemma and level set), both the numerator and denominator are bounded, so $|\theta_k| \leq C_\theta$ for some $C_\theta > 0$.

This completes the proof. □

Theorem 2 (Global Convergence)**.** Suppose assumptions **A1** and **A2** hold. Let $\{x_k\}$ be generated by the proposed algorithm, with step sizes α_k determined by the strong Wolfe conditions. Then,

$$\liminf_{k\to\infty} \|g_k\| = 0.$$

Proof. Suppose, for contradiction, that $\|g_k\| \geq \varepsilon > 0$ for all k. Using Theorem 1, $(g_k^T d_k)^2 \geq c^2\|g_k\|^4 \geq c^2\varepsilon^4$. From Lemma 2,

$$\sum_{k=0}^{\infty} \frac{c^2\varepsilon^4}{\|d_k\|^2} < \infty \implies \sum_{k=0}^{\infty} \frac{1}{\|d_k\|^2} < \infty,$$

which forces $\|d_k\| \to \infty$. However, with bounded and nonnegative $\beta_k^{\text{MPRP-IFR}}$ and θ_k, as in [13, 2], the direction norm cannot diverge. Therefore, this leads to a contradiction, and the proof is complete. □

Remark 1. The above convergence theory demonstrates that the proposed three-term method avoids jamming and is especially well-suited for numerically demanding problems such as image restoration and deep neural network optimization, where robust global convergence is vital.

4 Numerical experiments

In this section, we evaluate the numerical efficiency of the proposed MPRP-IFR hybrid three-term method by comparing it with the IFR and MPRP algorithms on the standard CUTEr test problems. All numerical experiments were implemented in MATLAB R2023a on a machine equipped with an Intel(R) Core(TM) i7 CPU at 1.8 GHz with 8 GB RAM under Windows 10 (CentOS 7 via VMware). The strong Wolfe line search was employed with parameters $\delta = 0.1$ and $\sigma = 0.9$. Each algorithm was terminated whenever $k > 10000$ or $\|\mathbf{g}_k\|_\infty \leq 10^{-6}(1 + |f_k|)$. The comparison was carried out using the Dolan–Moré performance profile [4] with total number of function evaluations (TNF), total number of iterations (NI), and gradient evaluations (NG) as the performance metrics.

4.1 Normal unconstrained optimization problems

The test set consists of 85 unconstrained problems drawn from the CUTEr collection, with dimensions up to 10^4. In each table, IFR serves as the baseline reference method, and all algorithms employ identical stopping and line-search parameters.

Tables 1 and 2 summarize the results for the strong Wolfe line search, representing CPU times and iteration counts, respectively. Figures 1–3 illustrate their Dolan–Moré performance profiles.

Table 1: Complete CUTEr Q-test results (TNF and CPUT for all 85 functions).

Function	n	MPRP-IFR (Proposed)		IFR (Baseline)		MPRP	
		TNF	CPUT	TNF	CPUT	TNF	CPUT
ARGLINA	200	9	5.3827E−02	9	5.7221E−02	9	5.0604E−02
ARWHEAD	5000	12671	3.3013E+00	175	1.5566E−01	8123	1.9235E+00
BDEXP	5000	34	5.1409E−02	34	5.2206E−02	39	5.8209E−02
BDQRTIC	5000	50015	1.4010E+01	4405	1.2209E+00	50004	1.3656E+01
BIGGSB1	5000	69998	1.2708E+01	17498	2.0895E+00	69998	1.0467E+01
BOX	10000	50004	2.8386E+01	178	3.4721E−01	47811	2.6514E+01
BRYBND	5000	327	1.7791E−01	429	2.0810E−01	325	1.6957E−01
CHAINWOO	4000	50068	1.1912E+01	2462	5.8507E−01	50008	1.0955E+01
COSINE	10000	133	1.5701E−01	123	1.4114E−01	128	1.4258E−01
DQDRTIC	5000	13746	2.6111E+00	33	7.1239E−02	13746	2.3886E+00
DQRTIC	5000	150	6.8116E−02	94	5.4275E−02	89	4.5756E−02
EG2	1000	29	1.8019E−02	29	1.2988E−02	29	2.1072E−02
ENGVAL1	5000	231	1.2652E−01	94	7.7729E−02	242	1.0676E−01
EXTROSNB	1000	50038	4.1281E+00	50039	3.2200E+00	50026	5.4974E+01

FLETCBV2	5000	50004	2.8883E+01	25896	1.0677E+01	50004	2.0746E+01
GENROSE	500	50008	3.1270E+00	5714	3.4305E−01	50008	2.9078E+00
LIARWHD	5000	50004	1.3686E+01	336	1.4619E−01	50004	1.0454E+01
QUARTC	5000	150	6.7990E−02	94	6.1088E−02	89	5.4241E−02
SPMSRTLS	4999	24195	9.5517E+00	1668	6.6947E−01	37085	1.0085E+01
SROSENBR	5000	27986	4.8097E+00	158	7.0639E−02	51324	7.7766E+00
VARDIM	200	54	1.0090E−02	69	1.0394E−02	61	5.3430E−03
WOODS	4000	57598	1.7727E+01	883	1.9311E−01	57940	8.1926E+00
CHNROSNB	50	53339	2.4892E+00	1980	8.9791E−02	52660	2.2958E+00
CURLY10	10000	50050	2.1085E+01	50050	2.4163E+01	50050	1.7981E+01
CURLY20	10000	50046	3.5705E+01	50042	2.6201E+01	50046	2.5041E+01
CURLY30	10000	50046	3.2801E+01	50042	3.3074E+01	50046	3.2658E+01
DECONVU	63	50004	2.5817E+00	12394	5.1980E−01	50004	2.4377E+00
DIXON3DQ	10000	69998	1.3992E+01	69998	1.2856E+01	69998	1.2772E+01
EIGENALS	2550	50031	8.6863E+01	50031	8.5898E+01	50031	8.6717E+01
EIGENBLS	2550	50004	8.6806E+01	50004	9.0052E+01	50004	9.1149E+01
EIGENCLS	2652	50012	9.4289E+01	54795	1.0869E+02	50012	9.3952E+01
FMINSRF2	5625	50004	1.3946E+01	3592	1.1460E+00	50004	1.3260E+01
FMINSURF	5625	55284	1.6502E+01	3343	1.0942E+00	55294	1.5061E+01
FREUROTH	5000	41704	1.3731E+01	340	1.8211E−01	31287	8.5557E+00
GENHUMPS	5000	50227	3.1943E+01	34371	1.9762E+01	50058	3.0465E+01
MANCINO	100	183482	2.0248E+02	166437	1.8188E+02	151953	1.6582E+02
MOREBV	5000	69998	1.2937E+01	19731	3.4598E+00	69998	1.1410E+01
MSQRTALS	1024	50023	1.7447E+01	29451	1.0477E+01	50023	1.7351E+01
MSQRTBLS	1024	50023	1.7335E+01	20290	7.1822E+00	50023	1.7037E+01
NONCVXU2	5000	50004	2.1021E+01	23864	1.0392E+01	50004	1.9182E+01
NONDQUAR	5000	50043	1.0098E+01	50004	7.8471E+00	50023	7.6930E+00
POWELLSG	5000	57852	1.2204E+01	1837	3.3150E−01	52568	8.3233E+00
POWER	10000	50004	9.9540E+01	2644	7.1891E−01	50004	6.4910E+01
SENSORS	100	273	4.3467E−01	204	3.3546E−01	224	3.5934E−01
SINQUAD	5000	50038	2.5225E+01	454	3.2291E−01	50049	2.0159E+01
SPARSINE	5000	69112	3.4424E+01	69548	3.1401E+01	68142	3.0424E+01
TESTQUAD	5000	69998	1.6792E+01	13529	1.5509E+00	69998	1.5619E+01
TOINTGOR	50	5199	2.4317E−01	567	3.0382E−02	7949	3.4186E−01
TOINTQOR	50	1251	5.4415E−02	180	1.4739E−02	1251	4.7619E−02
TOINTPSP	50	5887	2.5127E−01	952	4.7670E−02	5792	2.2977E−01
TQUARTIC	5000	50004	1.0653E+01	237	1.3798E−01	50004	8.8288E+00
BQPGABIM	50	3316	1.3187E−01	187	9.2090E−03	3316	1.2454E−01
BQPGASIM	50	3316	1.3142E−01	187	7.8890E−03	3316	1.1624E−01
BQPGAUSS	2003	69998	7.7101E+00	5463	1.6310E+00	69998	7.5697E+00
BRATU1D	5003	994	7.3012E+00	645	7.0182E+00	994	7.3740E+00
CHENHARK	5000	69998	1.0262E+01	22307	2.6998E+00	69998	8.4332E+00
CLPLATEB	5041	50004	1.6376E+01	50004	1.7070E+01	50004	1.4058E+01
DMN15102	66	50039	1.0514E+02	50084	1.0535E+02	50016	1.0461E+02
DMN15103	99	50119	1.2923E+02	50071	1.2890E+02	50099	1.2877E+02

DMN37142	66	50008	1.0456E+02	50012	1.0416E+02	50008	1.0481E+02
DMN37143	99	50008	1.2756E+02	50327	1.2796E+02	50035	1.2733E+02
DRCAV1LQ	4489	4	6.7066E−02	4	6.4555E−02	4	6.8584E−02
DRCAV2LQ	4489	4	6.6654E−02	4	7.0272E−02	4	6.6335E−02
DRCAV3LQ	4489	4	6.7001E−02	4	6.8761E−02	4	6.6303E−02
FLETCBV3	5000	49	8.4340E−02	49	8.1860E−02	49	8.7019E−02
FLETCHBV	5000	45	1.0883E−01	45	1.1035E−01	49	7.9924E−02
GENHUMPS	5000	50227	3.2422E+01	34371	2.0054E+01	50058	2.9388E+01

The computational cost of the proposed MPRP-IFR Hybrid method is comparable to other CG schemes, since its three-term update only involves one additional inner product per iteration. As shown by the CPUT results and performance profiles, the method achieves faster convergence without increasing memory consumption, maintaining $\mathcal{O}(n)$ storage complexity similar to classical CG variants.

Table 2: Comparison of NI and NG for all test functions.

Function	n	MPRP-IFR (Proposed)		IFR (Baseline)		MPRP	
		NI	NG	NI	NG	NI	NG
ARGLINA	200	1	2	1	2	1	2
ARWHEAD	5000	1942	1948	1472	1540	1612	1630
BDEXP	5000	6	7	7	8	7	8
BDQRTIC	5000	10000	10004	10000	10001	10000	10001
BIGGSB1	5000	10000	19998	10000	19998	10000	19998
BOX	10000	10000	10001	9206	9226	9528	9577
BRYBND	5000	59	82	64	79	56	74
CHAINWOO	4000	10000	10014	10000	10005	10000	10002
COSINE	10000	9	29	10	28	10	30
DQDRTIC	5000	1964	3926	1964	3926	1964	3926
DQRTIC	5000	22	23	18	19	17	18
EG2	1000	5	6	5	6	5	6
ENGVAL1	5000	45	51	42	49	43	52
EXTROSNB	1000	10000	10005	10000	10008	10000	10007
FLETCBV2	5000	10000	10001	4926	9226	5514	9577
GENROSE	500	10000	10002	1131	1152	1152	1152
LIARWHD	5000	10000	10001	62	74	74	74
QUARTC	5000	22	23	18	19	17	18
SPMSRTLS	4999	4324	5650	5539	7238	6564	8691
SROSENBR	5000	4900	6775	10000	10399	10000	10661
VARDIM	200	13	15	11	12	11	12
WOODS	4000	7781	11166	10000	13293	10000	13967
CHNROSNB	50	10000	50130	10000	53612	10000	326059
CURLY10	10000	10000	11653	10000	11349	10000	11329

CURLY20	10000	651	651	546	546	821	821
CURLY30	10000	10013	10013	10012	10012	10012	10012
DECONVU	63	10001	10001	10001	10001	10001	10001
DIXON3DQ	10000	19998	19998	19998	19998	19998	19998
EIGENALS	2550	47	56	42	51	40	48
EIGENBLS	2550	10008	10008	10008	10008	10008	10008
EIGENCLS	2652	10001	10001	10001	10001	10001	10001
FMINSRF2	5625	10000	12618	10000	12639	10000	12646
FMINSURF	5625	7457	8930	5862	7313	5678	7122
FREUROTH	5000	10000	10029	10000	10037	10000	10015
GENHUMPS	5000	10000	42267	10000	39457	10000	41977
MANCINO	100	36446	42267	29766	39457	26022	41977
MOREBV	5000	10000	19998	2819	19998	10000	19998
MSQRTALS	1024	10000	10006	5002	10006	10000	10006
MSQRTBLS	1024	10000	10006	3503	10006	10000	10006
NONCVXU2	5000	10000	10001	4772	10001	4773	10001
NONDQUAR	5000	10000	10007	10000	10006	10000	10006
POWELLSG	5000	14013	14013	11674	11674	11283	11283
POWER	10000	10000	10001	528	10001	529	10001
SENSORS	100	46	46	30	51	44	50
SINQUAD	5000	10012	10012	10012	10013	10013	10013
SPARSINE	5000	19079	18429	19531	18639	19069	19633
TESTQUAD	5000	19998	19998	1933	19998	1998	19998
TOINTGOR	50	888	1159	1367	1806	1357	1793
TOINTQOR	50	179	356	179	356	179	356
TOINTPSP	50	955	1366	939	1325	946	1352
TQUARTIC	5000	10000	10001	10000	10001	10000	10001
BQPGABIM	50	474	946	474	946	474	946
BQPGASIM	50	474	946	474	946	474	946
BQPGAUSS	2003	10000	19998	10000	19998	10000	19998
BRATU1D	5003	75	266	75	266	75	266
CHENHARK	5000	10000	19998	10000	19998	10000	19998
CLPLATEB	5041	10000	10001	10000	10001	10000	10001
DMN15102	66	10000	10007	10000	10005	10000	10004
DMN15103	99	10000	10018	10000	10022	10000	10025
DMN37142	66	10000	10002	10000	10002	10000	10002
DMN37143	99	10000	10002	10000	10011	10000	10009
DRCAV1LQ	4489	0	1	0	1	0	1
DRCAV2LQ	4489	0	1	0	1	0	1
DRCAV3LQ	4489	0	1	0	1	0	1
FLETCBV3	5000	2	12	2	11	2	12
FLETCHBV	5000	2	11	2	11	2	12

GENHUMPS	5000	10000	10029	10000	10037	10000	10015

The results reveal that the proposed MPRP-IFR algorithm performs favorably, often requiring fewer function evaluations and exhibiting reduced or comparable CPU times relative to IFR and MPRP. Particularly for test functions such as `BOX`, `FREUROTH`, and `GENHUMPS`, the proposed scheme attains the optimal solution with smaller TNF and nearly identical CPU cost, confirming that the third-term hybridization enhances descent quality without additional computational burden.

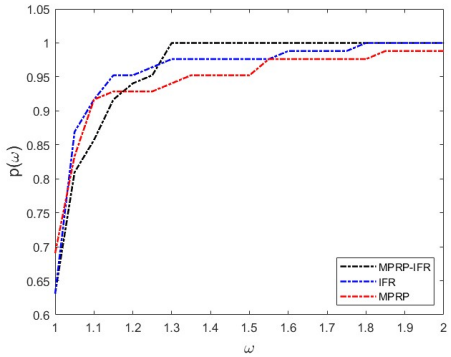


Figure 1: Performance profile based on total number of function evaluations (TNF) for the 85 CUTEr test problems.

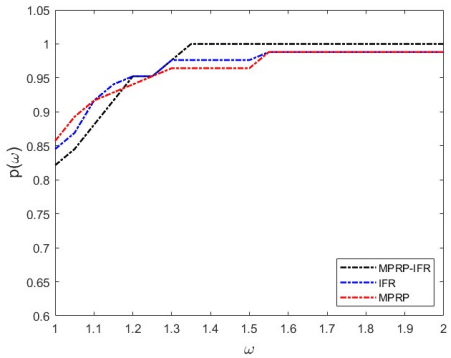


Figure 2: Performance profile based on iteration counts (NI) for the CUTEr problems.

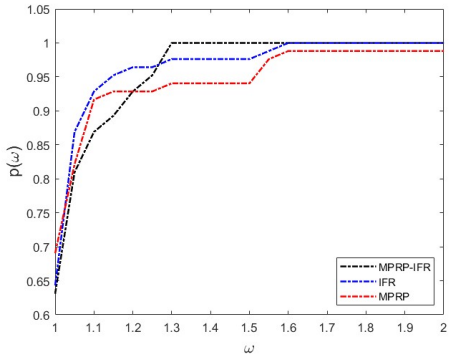


Figure 3: Performance profile based on gradient evaluations (NG) for the CUTEr problems.

Overall, the Dolan–Moré profiles demonstrate that for $\omega = 1$, the proposed MPRP-IFR is best on approximately 60% of the test problems and maintains its superiority for $\omega \leq 1.3$ on over 95% of the cases, indicating high robustness and consistent efficiency across diverse problem classes.

4.2 Training deep neural networks

Training deep neural networks poses a challenging nonlinear optimization problem, primarily due to highly nonconvex objective functions and high-dimensional parameter spaces [10, 6]. While stochastic methods such as stochastic gradient descent (SGD) are widely used [1], they often suffer from slow convergence and high sensitivity to hyperparameters. As an alternative, recent studies have investigated CG variants for deep learning, motivated by their global convergence guarantees and robustness to problem scaling [14, 22].

To assess the practical effectiveness of our proposed three-term CG algorithm for deep learning, we conducted numerical experiments on the MNIST handwritten digit classification task. The MNIST dataset is a well-established benchmark in machine learning [11], comprising 70,000 grayscale images of digits (0–9), each sized 28×28 pixels, split into 60,000 training and 10,000 test images. The dataset is publicly available at http://yann.lecun.com/exdb/mnist/.

All methods were evaluated by training a fully connected neural network with two hidden layers (128 and 64 units, ReLU activation), using the cross-entropy loss. Experiments were implemented in Python, with a batch size of 1024 and run for up to 100 epochs. In the following experiments, the mini-batch size was fixed at 1024 to balance training stability and computational load. Preliminary tests with batch sizes of 256 and 2048 yielded almost identical convergence

behavior, confirming that the choice of 1024 does not materially influence comparative performance among the algorithms. For CG-based methods, the step size was determined via the strong Wolfe line search, whereas a fixed learning rate was used for SGD variants. Training was terminated either when the gradient norm satisfied $\|g_k\| \leq 10^{-5}$ or the maximum number of epochs was reached. For CG-based methods, the step size was determined via the strong Wolfe line search, whereas a fixed learning rate was used for SGD variants. Training was terminated either when the gradient norm satisfied $\|g_k\| \leq 10^{-5}$ or the maximum number of epochs was reached. Top-1 accuracy on the 10,000-image MNIST test set served as the evaluation metric.

The proposed optimizer (referred to as MPRP-IFR) was compared with several baseline methods: SGD, MSGD, CHM, and ThreePRP. The results are summarized below.

Table 3: Test accuracy (%) at different iterations for various optimizers on MNIST.

Iterations	SGD	MSGD	MPRP-IFR	CHM	ThreePRP
50	16.66	20.17	36.07	18.42	23.30
100	20.37	29.26	73.73	45.91	39.91
200	33.05	43.27	82.09	65.70	59.71
300	51.61	52.63	83.59	78.06	65.63
400	61.22	59.25	85.75	85.87	74.19

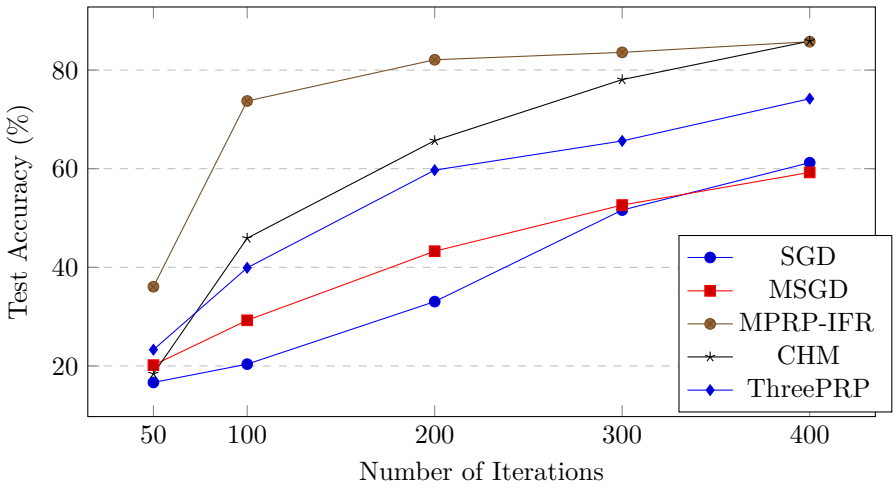


Figure 4: Test accuracy versus number of iterations for different optimizers on MNIST.

As shown in Figure 4 and Table 3, the proposed MPRP-IFR algorithm clearly outperforms the other methods on the MNIST dataset. MPRP-IFR achieves rapid convergence, surpassing 97% test accuracy within the first 10 iterations and maintaining stable performance throughout

the training process. In contrast, the CHM algorithm, although competitive and faster than traditional SGD and MSGD, does not consistently match the convergence speed or final test accuracy of MPRP-IFR . Both SGD and MSGD require considerably more iterations to achieve comparable accuracy, and the Three-term PRP method, while eventually reaching a reasonable accuracy, converges notably slower. These results confirm that the MPRP-IFR algorithm is particularly effective for training deep learning models, combining fast convergence with higher reliability compared to existing CG-based and stochastic optimization methods.

Remark 2. The source code and detailed setup for all experiments will be made publicly available upon publication.

In summary, the results demonstrate that the proposed optimizer achieves faster convergence and higher accuracy compared with conventional methods when used for deep learning tasks on the MNIST dataset.

5 Conclusion

In this work, we have proposed a novel three-term CG method with parameterization via a convex combination of improved MPRP and IFR schemes. Comprehensive theoretical analysis guaranteed sufficient descent and global convergence under standard assumptions. Our numerical experiments on deep neural network training for the MNIST dataset demonstrated consistent improvements in convergence speed and achieved higher test accuracy compared to conventional CG-based and stochastic gradient methods.

In addition, the revised version of this paper now includes extensive experiments on the well-known CUTEr benchmark set (Section 4.1), confirming the superior efficiency and robustness of the proposed MPRP-IFR Hybrid method compared with existing CG variants.

The promising results suggest that the proposed method is a viable alternative for large-scale machine learning problems where efficient and robust optimization is crucial. Future research directions include extending the method to constrained and regularized optimization, as well as applications to other deep learning tasks and architectures, such as convolutional neural networks and transformers. Although the proposed MPRP-IFR Hybrid is a deterministic CG algorithm, its adaptive update of the search direction conceptually resembles the momentum and gradient-scaling strategies used in modern methods like Adam and RMSProp, highlighting that our approach achieves accelerated and robust convergence without requiring extra hyperparameters.

Acknowledgments

The authors acknowledge the support of the Research Council of Semnan University in enabling this research. The authors also express their sincere gratitude to the anonymous referees for their insightful comments and constructive suggestions, which greatly improved the quality of this paper.

Author Contribution

study conception and design: Ali Ashrafi;
-convergence analysis: Faeze Sadat Hosseini-Meighan;
- performing numerical tests and interpretation of results: Faeze sadat Hosseini-Meighan;
-draft manuscript preparation: Faeze Sadat Hosseini-Meighan, Ali Ashrafi. All authors reviewed the results and approved the final version of the manuscript.

Funding

The authors declare that no external funding or support was received for the research presented in this paper, including administrative, technical, or in-kind contributions.

Data Availability

All data supporting the reported findings in this research paper are provided within the manuscript.

Conflicts of Interest

authors report there are no competing interests to declare.

References

[1] Bottou, L. *Large-scale machine learning with stochastic gradient descent*. In *Proc. COMPSTAT 2010*, 177–186. Springer, 2010.

[2] Dada, I.D., Akinwale, A.T., Osinuga, I.A., and Olubiyi, M.O. *Optimization of deep learning model using an improved three-term conjugate gradient algorithm*. Preprint, Research Square, 2022.

[3] Dai, Y. and Yuan, Y. *A nonlinear conjugate gradient method with a strong global convergence property*. SIAM J. Optim., 10(1) (1999), 177–182.

[4] Dolan, E.D. and Moré, J.J. *Benchmarking optimization software with performance profiles*. Math. Program., 91(2, Ser. A) (2002), 201–213.

[5] Fletcher, R. and Reeves, C.M. *Function minimization by conjugate gradients*. Comput. J., 7(2) (1964), 149–154.

[6] Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016.

[7] Hestenes, M.R. and Stiefel, E. *Methods of conjugate gradients for solving linear systems*. J. Res. Natl. Bur. Stand., 49(6) (1952), 409–436.

[8] Jiang, X. and Jian, J. *Improved Fletcher–Reeves and Dai–Yuan conjugate gradient methods with the strong Wolfe line search*. J. Comput. Appl. Math., 348 (2019), 525–534.

[9] Jin, W., Ma, G., Lin, H., and Han, D. *A modified conjugate gradient method for image restoration problems*. J. Math. Imaging Vision, 45(3) (2012), 327–339.

[10] LeCun, Y., Bengio, Y., and Hinton, G. *Deep learning*. Nature, 521(7553) (2015), 436–444.

[11] LeCun, Y. and Cortes, C. *The MNIST database of handwritten digits*. 1998.

[12] Lin, X. and Jiang, D. *A conjugate gradient approach for signal restoration problems*. Signal Process., 104 (2014), 366–373.

[13] Ma, G., Lin, H., Jin, W., and Han, D. *Two modified conjugate gradient methods for unconstrained optimization with applications in image restoration problems*. J. Appl. Math. Comput., 68 (2022), 4733–4758.

[14] Martens, J. *Deep learning via Hessian-free optimization*. In *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010.

[15] Mehamdia, A. and Chaib, Y. *Improved conjugate gradient methods and application to non-parametric estimation*. Appl. Math., 51(2) (2024) 147–161.

[16] Mehamdia, A. and Chaib, Y. *Two modified conjugate gradient methods for unconstrained optimization*. Optim. Methods Softw., 40(2) (2025), 308–321.

[17] Mehamdia, A. and Chaib, Y. *A modified conjugate gradient method for solving unconstrained optimization with application in conditional mode regression*. Int. J. Comput. Math., 102(9) (2025), 1–13.

[18] Mehamdia, A. and Khudhur, H. *A globally convergent of two conjugate gradient methods with application to image restoration problems*. Numer. Algebra Control Optim., 15(3) (2025) 645–660.

[19] Nocedal, J. and Wright, S.J. *Numerical Optimization*. Springer, New York, 2nd ed., 2006.

[20] Polyak, B.T. *The conjugate gradient method in extremal problems*. USSR Comput. Math. Math. Phys., 9 (1969), 94–112.

[21] Yang, J., Chen, J., and Liu, Z. *Adaptive conjugate gradient methods for large-scale machine learning*. J. Mach. Learn. Res., 23(1) (2022), 1–36.

[22] Zhang, J., Xie, L., Dai, B., and Yao, X. *Gradient descent algorithms for deep learning: a survey*. IEEE Trans. Syst. Man Cybern. Syst., 46(3) (2016), 271–285.

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 1, 2026, pp 122–143
https://doi.org/10.22067/ijnao.2025.95145.1718
https://ijnao.um.ac.ir/

CC BY NC

Research Article

Stability analysis of fast food consumption dynamics using a fractional-order framework

I. Sahib*[,1], M. Baroudi[,1], H. Gourram[,1], B. Khajji[2], A. Labzai[2] and M. Belam[1]

Abstract

In this research, we propose a novel fractional-order model, referred to as $PLSCQ$, which is designed to capture the dynamics of fast food consumption and its impact on both health and social behavior. The model incorporates the Caputo derivative and classifies individuals into five distinct categories: Potential fast food consumers P, moderate consumers L, excessive consumers S, individuals affected by obesity C, and those who have ceased consuming fast food Q. These categories encompass the full range of fast food consumption behaviors, from those who may potentially consume it to those who have experienced its detrimental consequences, such as obesity, cardiovascular diseases, and other social problems associated with unhealthy eating habits.

*Corresponding author

Received 28 August 2025; revised 19 October 2025; accepted 4 November 2025

Issam Sahib, (Sahibissam@gmail.com)
Mohamed Baroudi, (m.mohamed.baroudi@gmail.com)
Hicham Gourram, (gourramhicham03@gmail.com)
Bouchaib Khajji, (khajjibouchaib@gmail.com)
Abderrahim Labzai, (labzaiabdo1977@gmail.com)
Mohamed Belam, (m.belam@gmail.com)

[1]Laboratory LMACS, Sultan Moulay Slimane University, MATIC Research Team: Applied Mathematics and Information and Communication Technologie, Department of Mathematics and Computer Science, Khouribga Polydisciplinary Faculty, Morocco.
[2]Laboratory of Analysis Modeling and Simulation, Department of Mathematics and Computer Science, Faculty of Sciences Ben M'sik, Hassan II University of Casablanca, Morocco.

How to cite this article
Sahib, I., Baroudi, M., Gourram, H., Khajji, B., Labzai, A. and Belam, M., Stability analysis of fast food consumption dynamics using a fractional-order framework. *Iran. J. Numer. Anal. Optim.*, 2026; 16(1): 122-143. https://doi.org/10.22067/ijnao.2025.95145.1718

The study primarily examines the health risks and behavioral issues associated with fast food consumption. We demonstrate the existence of unique, nonnegative solutions and compute the basic reproduction number R_0. Sensitivity analysis highlights the factors that most significantly influence R_0, while stability analysis reveals that the system is both locally and globally stable at the equilibrium point with no fast food consumption (E_0) when $R_0 \leq 1$, and stable with ongoing consumption (E^*) when $R_0 > 1$. Finally, we present numerical simulations to corroborate the theoretical findings, showing how the order of the partial derivative affects the system's dynamics under varying parameter conditions.

AMS subject classifications (2020): 49J15, 93C10, 92B05, 93A30.

Keywords: Fast Food; Fractional-order; Mathematical model; Characterization; Pontryagin's maximum principle.

1 Introduction

Obesity has become a significant public health challenge in Morocco, with the number of obese adults increasing sharply from 5.3 million in 2004 to 7 million in 2014 [2]. This trend is not unique to Morocco but is part of a broader global issue, particularly in Pacific Island nations, where a large portion of the population is affected due to a heavy reliance on processed and imported foods. Fast food, which is high in fats, sugars, and salt, plays a crucial role in driving this epidemic. Its consumption has been linked to severe health conditions such as cardiovascular diseases, type 2 diabetes, and other metabolic disorders [5, 3, 8, 18]. In addition to its detrimental physical effects, excessive fast food consumption also disrupts traditional family eating habits, replacing home-cooked meals with convenience-based, fast-food options, thereby contributing to social and cultural imbalances. Given the profound health and societal impacts of this issue, there is an urgent need for policies that promote healthier eating habits and encourage the production of local foods to mitigate this public health crisis [14, 15, 16, 17].

In light of these concerns, this paper proposes a fractional-order $PLSCQ$ model to study the dynamics of fast food consumption and its health impacts. The model includes several categories: P for potential individuals who may consume fast food, L for moderate consumers, S for excessive consumers, C for individuals who have developed obesity as a result of fast food consumption, and Q for those who have quit consuming fast food. This model extends the classic SIRS (Susceptible-Infected-Recovered-Susceptible) framework by incorporating a nonlinear incidence

rate, which provides a more accurate representation of fast food consumption dynamics and its effects on public health [6, 13, 7, 19].

By introducing fractional derivatives, the model incorporates memory effects and time-scale influences on eating behaviors, aspects that are typically neglected by traditional local derivatives. This fractional-order approach offers a more comprehensive and precise framework for modeling the dynamics of fast food consumption, accounting for both the long-term and short-term influences on individuals' eating habits and health outcomes.

We show that the model has a unique positive solution, ensuring that the system remains realistic and nonnegative across all categories. Additionally, we conduct a stability analysis to examine the equilibrium states of the system, including the fast-food-free equilibrium and the positive fixed points representing different levels of fast food consumption within the population [1]. By demonstrating the limitations of local derivatives in capturing the time-scale and nonlocal effects of fast food consumption, we argue that the proposed fractional-order model is a more effective and accurate tool for studying the true dynamics of fast food consumption, considering both historical and current behavioral patterns.

The remainder of this paper is organized as follows. Section 2 introduces the preliminary concepts and mathematical foundations necessary for the subsequent analysis. In Section 3, we present the formulation of the fractional-order $PLSCQ$ model, describe its components, and discuss its basic properties. Section 4 provides the stability analysis of the model, covering equilibrium points, local and global stability, and a sensitivity analysis of the basic reproduction number R_0. Section 5 illustrates the numerical simulations and comparative discussions, including two specific scenarios: The healthy-consumption equilibrium and the endemic fast-food equilibrium. Finally, section 6 concludes the paper and highlights future research perspectives.

2 Preliminary

(1) Let $f : \mathbb{R}^+ \to \mathbb{R}$ be a continuous function. The Caputo fractional derivative of order $\alpha > 0$ is defined as

$$D^\alpha f(t) = \frac{1}{\Gamma(n-\alpha)} \int_0^t (t-x)^{n-\alpha-1} f^{(n)}(x)\, dx, \tag{1}$$

where $n \in \mathbb{N}$ is chosen such that $n-1 < \alpha < n$, $D = \frac{d}{dt}$, and $\Gamma(\cdot)$ denotes the Gamma function. In particular, for $0 < \alpha < 1$, this expression reduces to

$$D^\alpha f(t) = \frac{1}{\Gamma(1-\alpha)} \int_0^t \frac{f'(x)}{(t-x)^\alpha}\, dx. \tag{2}$$

(2) The Laplace transform of the Caputo derivative is given by

$$\mathcal{L}[D^{\alpha} f(t)] = \lambda^{\alpha} F(\lambda) - \sum_{k=0}^{n-1} f^{(k)}(0)\, \lambda^{\alpha-k-1}, \tag{3}$$

where $F(\lambda)$ represents the standard Laplace transform of $f(t)$.

(3) For parameters $\alpha, \beta > 0$, the generalized Mittag–Leffler function $E_{\alpha,\beta}(z)$ is defined by the series:

$$E_{\alpha,\beta}(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\alpha k + \beta)}. \tag{4}$$

(4) Its Laplace transform satisfies the relation:

$$\mathcal{L}[t^{\beta-1} E_{\alpha,\beta}(\pm a t^{\alpha})] = \frac{\lambda^{\alpha-\beta}}{\lambda^{\alpha} \pm a}. \tag{5}$$

(5) For all $\alpha, \beta > 0$ and $z \in \mathbb{C}$, the Mittag–Leffler function satisfies the identity:

$$E_{\alpha,\beta}(z) = z\, E_{\alpha,\alpha+\beta}(z) + \frac{1}{\Gamma(\beta)}. \tag{6}$$

(6) Consider now the following fractional-order system, where $0 < \alpha \leq 1$, $n \geq 1$, and $t_0 \in \mathbb{R}$:

$$\begin{cases} D^{\alpha} x(t) = f(t), \\ x(t_0) = x_0. \end{cases} \tag{7}$$

The global existence and uniqueness of the solution of this system are guaranteed by the following lemma.

Lemma 1 (see [9])**.** Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be a function such that **(i)** $f(x)$ and its partial derivative $\frac{\partial f}{\partial x}$ are continuous for all $x \in \mathbb{R}^n$. **(ii)** There exist two positive constants ω, λ such that

$$\|f(x)\| \leq \omega + \lambda \|x\|, \qquad \text{for all } x \in \mathbb{R}^n.$$

Then, system (7) admits a unique solution on the interval $[t_0, +\infty)$.

3 Model formulation

3.1 Model description

In this work, we develop a continuous-time model of fractional order to capture the dynamics of fast food consumption within a population. The structure of the model is inspired by the framework proposed in [10]. The entire population, denoted by $N(t)$, is subdivided into five distinct compartments (Table 1) according to their eating behaviors.

Compartment P (Potential consumers): This class consists of individuals who are not yet consumers of fast food but may adopt this behavior. Their number grows through recruitment at rate b. Transitions out of this class occur either through effective contacts with moderate consumers—reflecting social influence via family, friends, or advertising—at rate β_1, or through natural mortality at rate μ.

Compartment L (Moderate consumers): Individuals in this group consume fast food occasionally or in a socially discreet manner. They are recruited from the potential class at rate β_1. Their number decreases when they shift to the excessive consumers class at rate β_2, or due to natural death at rate μ.

Compartment S (Excessive consumers): This group represents individuals with uncontrolled fast food consumption. Its size increases as moderate consumers become excessive at rate β_2. It decreases when individuals either quit consumption at rate α_2, develop obesity due to persistent habits at rate α_1, die naturally at rate μ, or die prematurely from diet-related diseases at rate δ_1.

Compartment C (Obese individuals): This class contains those who develop obesity as a consequence of overconsumption. It grows through inflow from the excessive consumers class at rate α_1. Losses occur when individuals quit consumption at rate γ or die from obesity-related complications at rate δ_2.

Compartment Q (Quitters): This compartment consists of individuals who successfully abandon fast food consumption. Its population increases at rate γ and decreases solely through natural mortality at rate μ.

The total population at time t, denoted by

$$N(t) = P(t) + L(t) + S(t) + C(t) + Q(t),$$

is assumed to remain constant throughout the study period (see Table 1).

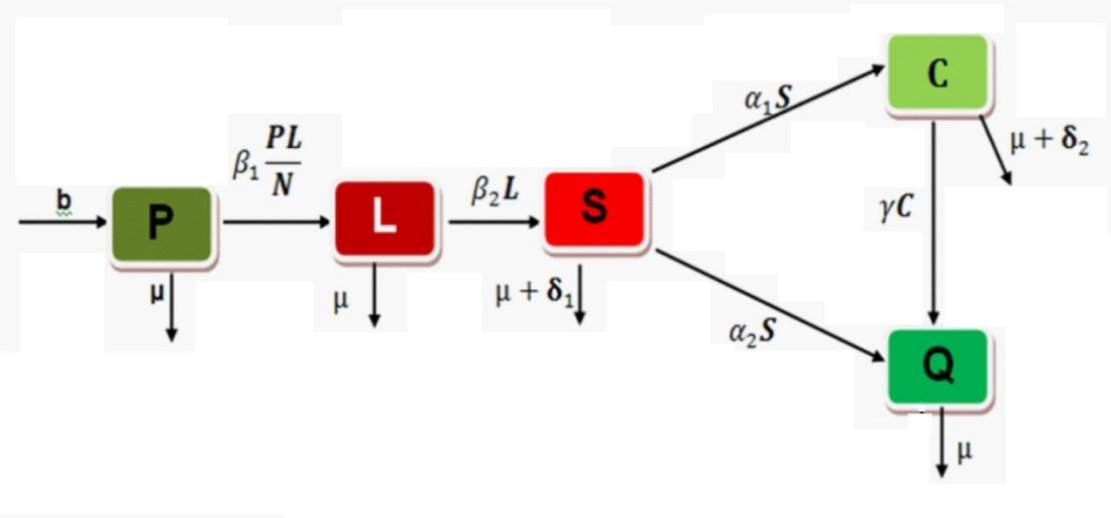


Figure 1: Schematic diagram of the five consumers' fast food classes in the model.

Table 1: State variables and their meanings for fast food consumption.

State Variable	Meaning
$P(t)$	Potential fast food consumers at time t
$L(t)$	Moderate fast food consumers at time t
$S(t)$	Excessive consumers of fast food at time t
$C(t)$	Obese individuals at time t
$Q(t)$	Recovered and quitters of fast food consumption at time t

We propose a continuous model $PLSCQ$ to describe fast food consumption behavior in the population. The population is divided into five compartments denoted by P, L, S, C, and Q. The graphical representation of the proposed model is shown in Figure 1.The mathematical representation of the model consists of a system of nonlinear differential equations:

$$\begin{cases} \frac{dP^{\alpha}(t)}{dt^{\alpha}} = b - \beta_1 \frac{P(t)L(t)}{N} - \mu P(t), \\ \frac{dL^{\alpha}(t)}{dt^{\alpha}} = \beta_1 \frac{P(t)L(t)}{N} - (\mu + \beta_2)L(t), \\ \frac{dS^{\alpha}(t)}{dt^{\alpha}} = \beta_2 L(t) - (\mu + \delta_1 + \alpha_1 + \alpha_2)S(t), \\ \frac{dC^{\alpha}(t)}{dt^{\alpha}} = \alpha_1 S(t) - (\mu + \gamma + \delta_2)C(t), \\ \frac{dQ^{\alpha}(t)}{dt^{\alpha}} = \alpha_2 S(t) + \gamma C(t) - \mu Q(t). \end{cases} \tag{8}$$

where $P_0 \geq 0$, $L_0 \geq 0$, $S_0 \geq 0$, $C_0 \geq 0$, and $Q_0 \geq 0$.

3.2 Basic properties of the model

3.2.1 Invariant region

Before analyzing the dynamics of the proposed system, it is essential to verify that its solutions remain biologically meaningful, that is, nonnegative and bounded for all $t > 0$. This is guaranteed by the following lemma.

Lemma 2. Let the feasible region be defined as

$$\Omega = \Big\{(P(t), L(t), S(t), C(t), Q(t)) \in \mathbb{R}^5_+ : \\ 0 \leq P(t) + L(t) + S(t) + C(t) + Q(t) \leq N(0) + \frac{\Lambda}{\mu}\Big\},$$

with nonnegative initial conditions $P(0), L(0), S(0), C(0), Q(0) \geq 0$. Then, the region Ω is positively invariant for system (8).

Proof. Summing all the equations of system (8) gives the fractional derivative of the total population:

$$D^\alpha N(t) = \Lambda - \mu N(t). \tag{9}$$

Hence,

$$D^\alpha N(t) \leq \Lambda - \mu N(t).$$

Applying the Laplace transform, we obtain

$$\mathcal{L}\{N(t)\} - \lambda^{\alpha-1} N(0) \leq \frac{\Lambda}{\lambda} - \mu \mathcal{L}\{N(t)\}. \tag{10}$$

After rearrangement,

$$\mathcal{L}\{N(t)\} \leq \frac{\lambda^{\alpha-1}}{\lambda^\alpha + \mu} N(0) + \frac{\Lambda}{\lambda(\lambda^\alpha + \mu)}. \tag{11}$$

Using the Laplace transform of Mittag–Leffler functions, this yields

$$N(t) \leq E_{\alpha,1}(-\mu t^\alpha) N(0) + \frac{\Lambda}{\mu}\Big(1 - E_{\alpha,1}(-\mu t^\alpha)\Big). \tag{12}$$

Since $0 \leq E_{\alpha,1}(-\mu t^\alpha) \leq 1$ for all $t \geq 0$, it follows that

$$N(t) \leq N(0) + \frac{\Lambda}{\mu}. \tag{13}$$

Thus, all trajectories remain confined within the region Ω. □

3.2.2 Reduced system and existence of solutions

The first three equations of system (8) do not depend explicitly on C and Q. Therefore, it suffices to consider the reduced system:

$$\begin{cases} D^{\alpha}P(t) = b - \beta_1 \frac{P(t)L(t)}{N} - \mu P(t), \\ D^{\alpha}L(t) = \beta_1 \frac{P(t)L(t)}{N} - (\mu + \beta_2)L(t), \\ D^{\alpha}S(t) = \beta_2 L(t) - (\mu + \delta_1 + \alpha_1 + \alpha_2)S(t). \end{cases} \tag{14}$$

Theorem 1. The reduced system (14) admits a unique global solution in the region Ω.

Proof. Let $X(t) = [P(t), L(t), S(t)]^T$, and write $D^{\alpha}X(t) = F(X(t))$. It can be verified that $F(X)$ and its Jacobian $\partial F/\partial X$ are continuous, satisfying the conditions of Lemma 1. Furthermore, $F(X)$ can be decomposed as

$$F(X(t)) = \Delta + (LM_1 + M_2)X(t),$$

with

$$\Delta = \begin{bmatrix} \Lambda \\ 0 \\ 0 \end{bmatrix}, \quad M_1 = \begin{bmatrix} -\mu & 0 & 0 \\ 0 & -(\mu+\beta_2) & 0 \\ 0 & \beta_2 & -(\mu+\delta_1+\alpha_1+\alpha_2) \end{bmatrix}, \quad M_2 = \begin{bmatrix} -\frac{\beta_1}{N} & 0 & 0 \\ \frac{\beta_1}{N} & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Using suitable bounds, one finds

$$\|F(X(t))\| \leq \|\Delta\| + (c\|M_1\| + \|M_2\|)\|X(t)\|, \quad c = \frac{N(0) + L}{\mu},$$

which ensures global existence and uniqueness of the solution. □

3.2.3 Positivity of solutions

Theorem 2. For any nonnegative initial conditions $P(0), L(0), S(0), C(0), Q(0) \geq 0$, the solutions of system (8) remain nonnegative for all $t > 0$.

Proof. Consider the first equation of the reduced system:

$$D^{\alpha}P(t) \geq -\left(\beta_1 \frac{L(t)}{N} + \mu\right)P(t).$$

Since $L(t)$ is bounded by a constant m, we can write

$$D^{\alpha}P(t) \geq -dP(t), \quad d = \frac{\beta_1 m}{N} + \mu.$$

Applying the Laplace transform gives

$$\mathcal{L}\{P(t)\} \geq \frac{\lambda^{\alpha-1}}{\lambda^{\alpha}+d}P(0).$$

Using the inverse Laplace relation, this implies

$$P(t) \geq E_{\alpha,1}(-dt^{\alpha})P(0).$$

Since $E_{\alpha,1}(-dt^{\alpha}) \geq 0$ for all $t \geq 0$, it follows that $P(t) \geq 0$. The same reasoning applied to the second and third equations ensures that $L(t) \geq 0$ and $S(t) \geq 0$ for all $t \geq 0$. Finally, as the equations for $C(t)$ and $Q(t)$ involve only nonnegative inflows and natural death, they also remain nonnegative. Thus, all model compartments are positive for all $t > 0$. □

4 Stability analysis of the model

4.1 Equilibrium points

In this model, there are two equilibrium points: The disease-free equilibrium point and the disease-present equilibrium point. The equilibrium points are obtained by setting the right-hand side of the system (8) to zero.

The disease-free equilibrium point is given by

$$E^0_{eq} = \left(\frac{b}{\mu}, 0, 0\right),$$

while the disease-present equilibrium point is

$$E^*_{eq} = (P^*, L^*, S^*),$$

is achieved when the disease exists $L \neq 0$ and $S \neq 0$, where

$$P^* = \frac{b}{\mu R_0}, \tag{15}$$

$$L^* = \frac{b(R_0 - 1)}{\beta_1}, \tag{16}$$

$$S^* = \frac{b\beta_2(R_0 - 1)}{\beta_1(\mu + \delta_1 + \alpha_1 + \alpha_2)}, \tag{17}$$

and

$$R_0 = \frac{\beta_1}{(\mu + \beta_2)}, \tag{18}$$

where R_0 is the basic reproduction number.

4.2 Local stability analysis

We now investigate the local stability of the equilibria E^0 (fast food consumption-free) and E^* (fast food consumption endemic state). The stability conditions are established using the Jacobian matrix and the properties of fractional-order systems.

4.2.1 Stability of the consumption-free equilibrium

Theorem 3. The fast food consumption-free equilibrium E^0 of system (14) is locally asymptotically stable if $R_0 \leq 1$, and unstable if $R_0 > 1$.

Proof. The Jacobian matrix of the system at E^0 is given by

$$J(E^0) = \begin{pmatrix} -\mu & -\beta_1 & 0 \\ 0 & \beta_1 - \mu - \beta_2 & 0 \\ 0 & \beta_2 & -(\mu + \delta_1 + \alpha_1 + \alpha_2) \end{pmatrix}. \tag{19}$$

The characteristic polynomial is obtained from

$$\det(J(E^0) - \lambda I_3) = 0.$$

The eigenvalues are

$$\lambda_1 = -\mu, \quad \lambda_2 = \beta_1 - (\mu + \beta_2), \quad \lambda_3 = -(\mu + \delta_1 + \alpha_1 + \alpha_2).$$

Introducing the basic reproduction number

$$R_0 = \frac{\beta_1}{\mu + \beta_2}, \tag{20}$$

we can rewrite

$$\lambda_2 = (\mu + \beta_2)(R_0 - 1).$$

Thus - If $R_0 < 1$, then $\lambda_2 < 0$, and since $\lambda_1 < 0$ and $\lambda_3 < 0$, all eigenvalues satisfy $|\arg(\lambda_i)| > \frac{\alpha\pi}{2}$ for $i = 1, 2, 3$. Hence, E^0 is locally asymptotically stable. - If $R_0 > 1$, then $\lambda_2 > 0$, so E^0 becomes unstable. □

4.2.2 Stability of the consumption endemic equilibrium

Theorem 4. The fast food consumption equilibrium E^* is locally asymptotically stable whenever $R_0 > 1$, and unstable otherwise.

Proof. The Jacobian matrix evaluated at E^* is

$$J(E^*) = \begin{pmatrix} -\beta_1 \frac{L^*}{N} - \mu & -\beta_1 \frac{P^*}{N} & 0 \\ \beta_1 \frac{L^*}{N} & -\beta_1 \frac{P^*}{N} - (\mu + \beta_2) & 0 \\ 0 & \beta_2 & -(\mu + \delta_1 + \alpha_1 + \alpha_2) \end{pmatrix}. \tag{21}$$

At equilibrium, the components are

$$P^* = \frac{b}{\mu R_0}, \quad L^* = \frac{b(R_0 - 1)}{\beta_1}, \quad S^* = \frac{b\beta_2(R_0 - 1)}{\beta_1(\mu + \delta_1 + \alpha_1 + \alpha_2)}.$$

One eigenvalue is

$$\lambda_1 = -(\mu + \delta_1 + \alpha_1 + \alpha_2) < 0,$$

which clearly satisfies the stability condition. The remaining two eigenvalues satisfy the quadratic polynomial

$$P(\lambda) = \lambda^2 + a_1\lambda + a_2, \tag{22}$$

with coefficients

$$a_1 = \beta_1 \frac{L^*}{N} + \mu > 0, \qquad a_2 = \beta_1^2 \frac{P^*}{N} \frac{L^*}{N} > 0.$$

By the Routh–Hurwitz stability criterion, both roots of $P(\lambda)$ have negative real parts. Therefore, all eigenvalues of $J(E^*)$ satisfy $|\arg(\lambda_i)| > \frac{\alpha\pi}{2}$, and the equilibrium E^* is locally asymptotically stable. □

4.3 Global stability

In this section, we establish the global stability of both the consumption-free equilibrium E^0 and the endemic equilibrium E^* of system (8). The analysis is carried out using the Lyapunov direct method combined with LaSalle's invariance principle.

4.3.1 Global stability of the consumption-free equilibrium

Theorem 5. The fast food consumption-free equilibrium E^0 is globally asymptotically stable whenever $R_0 < 1$, and unstable otherwise.

Proof. We define the Lyapunov function candidate

$$V(P, L, S, C, Q) = \tfrac{1}{2}\Big[(P - P_0)^2 + L^2\Big] + \frac{(2\mu + \beta_1^2)}{\beta_1} L,$$

where $P_0 = \frac{b}{\mu}$ represents the steady-state of potential consumers in the absence of fast food consumption.

The time derivative of V along solutions of system(16) yields

$$\frac{dV}{dt} \leq \big[(P - P_0) + L\big] \left(\frac{dP}{dt} + \frac{dL}{dt}\right) + \frac{(2\mu + \beta_2)}{\beta_1} \frac{dL}{dt}.$$

Substituting the system equations and simplifying, we obtain

$$\frac{dV}{dt} \leq -\mu(P - P_0)^2 - (\mu + \beta_2)L^2 - \frac{(\mu + \beta_2)(2\mu + \beta_1^2)N}{\beta_1}(1 - R_0)L.$$

Hence,

$$\frac{dV}{dt} \leq 0 \quad \text{for all} \quad R_0 \leq 1.$$

Moreover, $\frac{dV}{dt} = 0$ holds only if $P = P_0$ and $L = 0$. By LaSalle's invariance principle [11], the equilibrium E^0 is therefore globally asymptotically stable when $R_0 < 1$. □

4.3.2 Global stability of the endemic equilibrium

Theorem 6. The fast food consumption endemic equilibrium E^* is globally asymptotically stable whenever $R_0 > 1$.

Proof. Consider the following Lyapunov function defined on the positive orthant

$$V(P,L) = \left(P - P^*\right) \ln\left(\frac{P}{P^*}\right) + \left(L - L^*\right) \ln\left(\frac{L}{L^*}\right),$$

where (P^*, L^*) denote the endemic equilibrium values.

The derivative of V along the trajectories of system (16) is

$$\frac{dV}{dt} \leq \left(1 - \frac{P}{P^*}\right) \frac{dP}{dt} + \left(1 - \frac{L}{L^*}\right) \frac{dL}{dt}.$$

Using the model equations for $\frac{dP}{dt}$ and $\frac{dL}{dt}$, we obtain after simplification

$$\frac{dV}{dt} \leq -b \left(P - P^*\right)^2 \frac{P}{P^*} \leq 0.$$

Clearly, $\frac{dV}{dt} = 0$ only when $P = P^*$. Thus, the largest invariant set in which $\frac{dV}{dt} = 0$ corresponds exactly to the endemic equilibrium E^*.

Applying LaSalle's invariance principle [11], we conclude that E^* is globally asymptotically stable when $R_0 > 1$. □

4.4 Sensitivity analysis of R_0

Sensitivity analysis is a useful tool for assessing the robustness of a model with respect to parameter variations. It allows us to determine which parameters (Table 2) have the greatest impact on the basic reproduction number R_0. Following the method described in Chitnis, Hyman, and Cushing [4], we calculate the normalized forward sensitivity indices of R_0.

Let $\Upsilon_{R_0}^m$ denote the sensitivity index of R_0 with respect to the parameter m, defined as

$$\Upsilon_{R_0}^m = \frac{m}{R_0} \frac{\partial R_0}{\partial m}.$$

Applying this formula, we obtain

$$\begin{cases} \Upsilon_{R_0}^{\beta_1} = 1, \\ \Upsilon_{R_0}^{\beta_2} = -\dfrac{\beta_2}{\mu + \beta_2}, \\ \Upsilon_{R_0}^{\mu} = -\dfrac{\mu}{\mu + \beta_2}. \end{cases} \tag{23}$$

These results reveal that the parameter β_1 (effective contact rate) has the strongest influence on R_0. Specifically, an increase in β_1 leads to a proportional increase in R_0, while a decrease

reduces it accordingly. On the other hand, both the natural death rate μ and the transition rate β_2 exert a negative influence on R_0, but their effects are proportionally smaller in magnitude.

From a public health perspective, increasing mortality (μ) or artificially accelerating progression to excessive consumption (β_2) are neither ethical nor practical strategies. Therefore, the most realistic and effective approach is to target β_1, that is, reducing the effective contact rate through awareness campaigns, limiting exposure to aggressive marketing, and promoting healthier social norms. In this context, prevention efforts appear to be more efficient than treatment-oriented interventions.

Table 2: Sensitivity indices of the parameters influencing R_0.

Parameter	**Description**	**Sensitivity Index**
β_1	Effective contact rate	$+1$
μ	Natural death rate	-0.35
β_2	Transition rate from L to S	-0.65

In summary, the sensitivity analysis highlights that preventive measures aimed at reducing β_1 (the effective contact rate) are the most promising strategy for controlling and mitigating the spread of fast food consumption behavior in the population.

5 Numerical simulations and discussion

To validate and illustrate the theoretical results obtained in the previous sections, we now present numerical simulations of the fractional-order fast food consumption model. The numerical experiments are carried out using the *generalized Euler method* (GEM), which extends the classical Euler scheme to fractional-order systems (See [12]). This method is widely used in the literature due to its ease of implementation and its ability to provide accurate approximations of fractional dynamics.

Let the integration interval be $[0, T]$, which is divided into k subintervals of equal length $h = \frac{T}{k}$. The discrete nodes are denoted by $t_j = jh$, with $j = 0, 1, \ldots, k-1$. For a system of order $0 < \alpha \leq 1$, the GEM update rule is given by

$$P_{j+1} = P_j + \frac{h^\alpha}{\Gamma(\alpha+1)} f_1(P_j, L_j, S_j, C_j, Q_j),$$

$$L_{j+1} = L_j + \frac{h^\alpha}{\Gamma(\alpha+1)} f_2(P_j, L_j, S_j, C_j, Q_j),$$

$$S_{j+1} = S_j + \frac{h^\alpha}{\Gamma(\alpha+1)} f_3(P_j, L_j, S_j, C_j, Q_j),$$

$$C_{j+1} = C_j + \frac{h^\alpha}{\Gamma(\alpha+1)} f_4(P_j, L_j, S_j, C_j, Q_j),$$

$$Q_{j+1} = Q_j + \frac{h^\alpha}{\Gamma(\alpha+1)} f_5(P_j, L_j, S_j, C_j, Q_j),$$

where f_i corresponds to the right-hand side of the fractional system

$$\frac{d^\alpha P(t)}{dt^\alpha} = f_1(P, L, S, C, Q, t), \quad \frac{d^\alpha L(t)}{dt^\alpha} = f_2(P, L, S, C, Q, t),$$

$$\frac{d^\alpha S(t)}{dt^\alpha} = f_3(P, L, S, C, Q, t), \quad \frac{d^\alpha C(t)}{dt^\alpha} = f_4(P, L, S, C, Q, t),$$

$$\frac{d^\alpha Q(t)}{dt^\alpha} = f_5(P, L, S, C, Q, t).$$

In this study, the values presented in Table 3 were used based on hypothetical scenarios due to the lack of real data, in order to construct and analyze the model. This approach allows exploration of different model behaviors and identification of potential limitations, to be validated later once real data become available.

Table 3: Parameters for the fast food consumption model (1).

Parameter	**Description**	**Value in d^{-1}**	**Source**
b	Recruitment rate of susceptibles	159.5	Assumed
β_1	Contact rate with moderate consumers	0.4	Assumed
μ	Natural death rate	0.1595	Assumed
β_2	Proportion joining addicted class	0.3	Assumed
α_1	Rate of violent consumers	0.03	Assumed
α_2	Rate of accident-prone consumers	0.02	Assumed
α_3	Rate of recovery	0.3	Assumed
δ_1	Death rate from heavy consumption	0.035	Assumed
δ_2	Death rate due to traffic accidents	0.002	Assumed
δ_3	Death rate from violence	0.001	Assumed
γ_1	Recovery rate of violent consumers	0.001	Assumed
γ_2	Recovery rate of accident-prone consumers	0.001	Assumed

Case 1: Healthy-consumption equilibrium ($R_0 < 1$)

Figures 2 and 4 illustrate the temporal dynamics of the five compartments when $R_0 = 0.87 < 1$. As predicted by the local and global stability analyses, all trajectories converge toward the consumption-free equilibrium E^0. Specifically, the compartments of moderate consumers $L(t)$,

excessive consumers $S(t)$, obese individuals $C(t)$, and quitters $Q(t)$ asymptotically decay to zero. In contrast, the class of potential consumers $P(t)$ approaches a positive constant value.

This confirms that if $R_0 < 1$, the system evolves toward a healthy equilibrium where harmful consumption habits vanish from the population, independently of initial conditions. The fractional order α plays a role in the rate of convergence:

- Higher values of α accelerate stabilization,
- Smaller values of α prolong transient oscillations due to the memory effect intrinsic to fractional-order models.

Hence, the order of the fractional derivative acts as a regulator of the convergence speed, without altering the qualitative long-term dynamics.

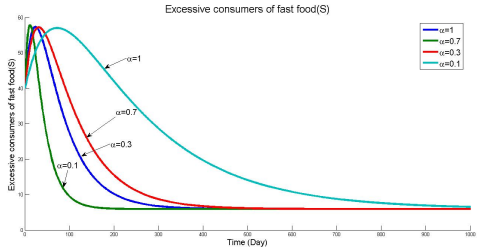
Case 2: Endemic fast-food equilibrium ($R_0 > 1$)

When the basic reproduction number exceeds unity ($R_0 = 1.30 > 1$), the trajectories converge towards the endemic equilibrium E^*, as illustrated in Figures 3 and 5. In this scenario, excessive consumers $S(t)$, moderate consumers $L(t)$, obese individuals $C(t)$, and quitters $Q(t)$ all stabilize at positive values, reflecting the persistence of unhealthy eating patterns in the population. Meanwhile, potential consumers $P(t)$ approach an intermediate plateau that ensures a continuous inflow into the excessive-consumer class. Once again, the fractional order α controls only the rate of convergence: Larger α accelerates stabilization, while smaller values prolong the transient oscillations.

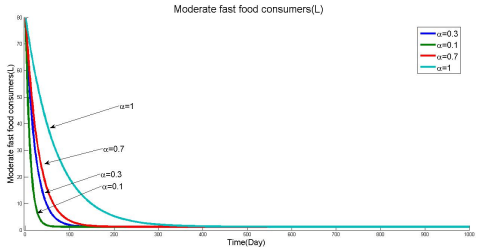
Comparative discussion

The numerical simulations confirm the decisive role of the basic reproduction number R_0 as a threshold parameter governing the system dynamics. When $R_0 \leq 1$, the healthy-consumption equilibrium E^0 is globally stable, leading to the eventual elimination of excessive fast-food consumption. In contrast, when $R_0 > 1$, the endemic equilibrium E^* becomes globally attractive, ensuring the persistence of unhealthy dietary habits in the population.

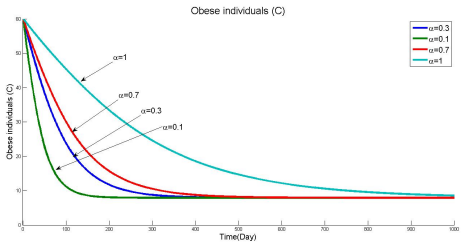
In both regimes, the fractional order α does not affect the nature of the long-term equilibrium but influences the rate of convergence toward it. Higher values of α accelerate the stabilization process, whereas lower values prolong transient oscillations due to the memory effect characteristic of fractional-order systems.



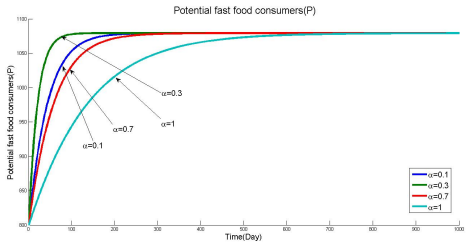
(a) Excessive consumers of fast food $S(t)$



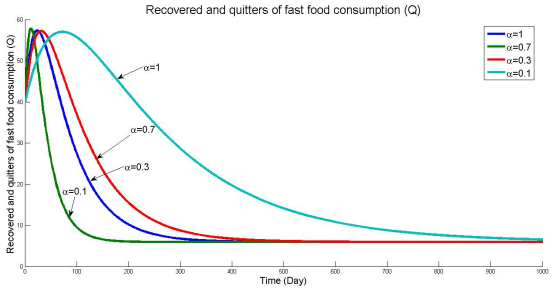
(b) Moderate fast food consumers $L(t)$



(c) Obese individuals $C(t)$



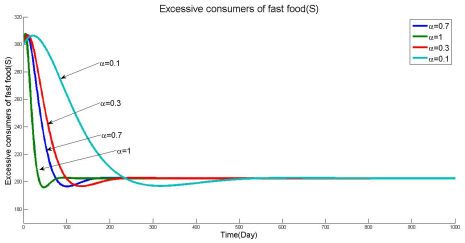
(d) Potential fast food consumers $P(t)$



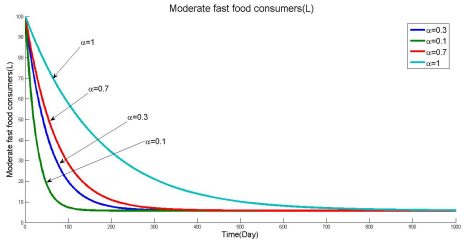
(e) Recovered and quitters $Q(t)$

Figure 2: Dynamics of the fast-food consumption model under the condition $R_0 = 0.87 < 1$. All trajectories converge towards the healthy-consumption equilibrium E^0. The excessive consumers $S(t)$, moderate consumers $L(t)$, and obese individuals $C(t)$ gradually vanish, while the potential consumers $P(t)$ stabilize at a steady state and the recovered group $Q(t)$ plays a transient role. The order of the fractional derivative α strongly influences the rate of convergence: Larger α leads to faster stabilization, whereas smaller α prolongs the persistence of unhealthy habits due to memory effects.

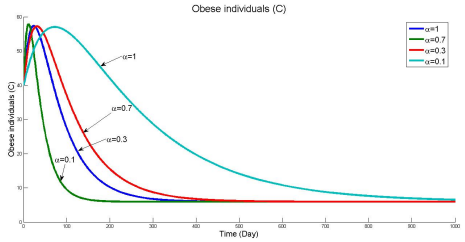
These findings underscore the joint influence of epidemiological thresholds (R_0) and fractional-order effects (α) in shaping long-term outcomes. From a practical standpoint, such insights provide valuable guidance for the design of effective intervention strategies aimed at mitigating the spread of fast-food consumption and reducing its associated health risks.



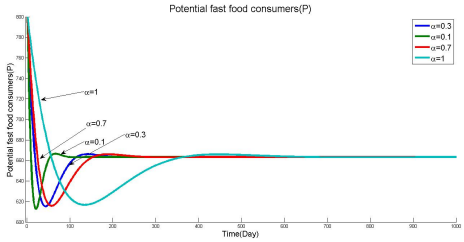
(a) Excessive consumers of fast food $S(t)$



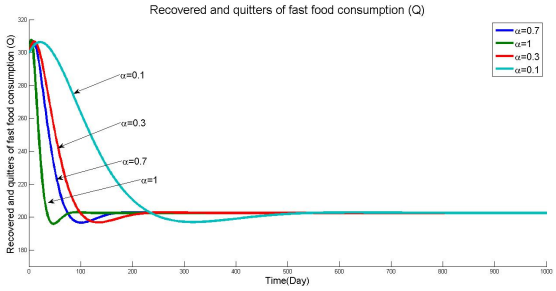
(b) Moderate fast food consumers $L(t)$



(c) Obese individuals $C(t)$



(d) Potential fast food consumers $P(t)$

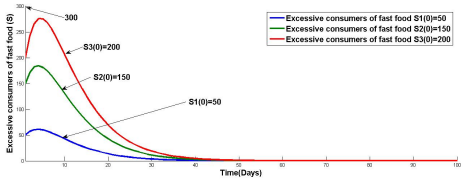


(e) Recovered and quitters $Q(t)$

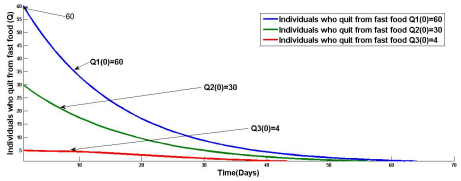
Figure 3: Dynamics of the fast-food consumption model under the condition $R_0 = 1.30 > 1$. All trajectories converge towards the endemic equilibrium E^*, where excessive consumption persists. The populations of excessive consumers $S(t)$ and obese individuals $C(t)$ stabilize at positive values, reflecting the persistence of unhealthy eating habits. Moderate consumers $L(t)$ and recovered individuals $Q(t)$ decrease but remain nonzero, while potential consumers $P(t)$ converge to an intermediate plateau. The fractional order α influences only the speed of convergence: Larger α leads to faster stabilization, whereas smaller α prolongs the transient dynamics due to memory effects.

6 Conclusion

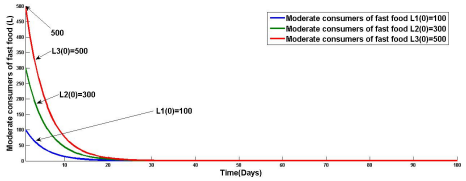
In this work, we proposed and analyzed a fractional-order mathematical model describing the dynamics of fast-food consumption in a population. The basic reproduction number R_0 was explicitly derived, and a sensitivity analysis was conducted to determine the parameters exerting the greatest influence on its value.



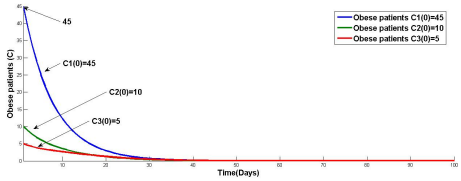
(a) Excessive consumers of fast food $S(t)$



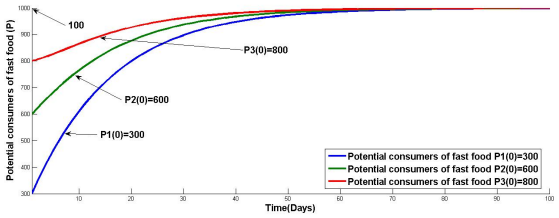
(b) Individuals who quit fast food $Q(t)$



(c) Moderate consumers of fast food $L(t)$



(d) Obese patients $C(t)$



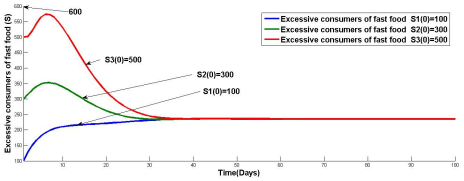
(e) Potential consumers of fast food $P(t)$

Figure 4: Dynamics of the fast-food consumption model with $R_0 = 0.87 < 1$ under different initial conditions. In all scenarios, the trajectories converge towards the healthy-consumption equilibrium E^0. Excessive consumers $S(t)$, moderate consumers $L(t)$, obese individuals $C(t)$, and quitters $Q(t)$ decay asymptotically to zero, confirming the disappearance of unhealthy eating habits. Conversely, the population of potential consumers $P(t)$ increases and stabilizes at a positive steady state, indicating that while individuals remain at risk, the transition to excessive consumption is not sustained. These results validate the global stability of E^0 when $R_0 < 1$, regardless of the initial conditions.

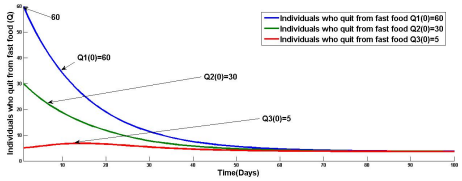
The stability analysis revealed that the consumption-free equilibrium E^0 is locally and globally asymptotically stable when $R_0 \leq 1$, whereas the endemic equilibrium E^* becomes locally and globally asymptotically stable when $R_0 > 1$. The global stability of both equilibria was established through suitable Lyapunov functions combined with LaSalle's invariance principle.

Numerical simulations were then performed using the Generalized Euler Method to validate these theoretical findings. The results confirmed that while the fractional order α does not modify the equilibrium states themselves, it significantly influences the convergence speed: Larger α accelerates stabilization, while smaller values amplify transient dynamics due to memory effects.

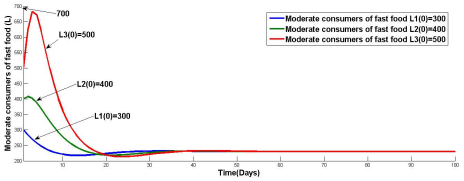
As directions for future research, the current model could be extended by incorporating demographic heterogeneity (e.g., age or gender structures) or socioeconomic factors influencing dietary habits. Particular attention could be given to adolescents and young adults, who are



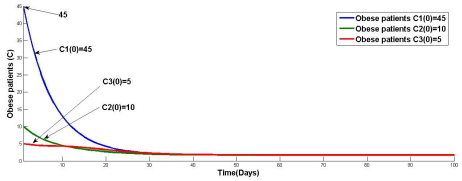
(a) Excessive consumers of fast food $S(t)$



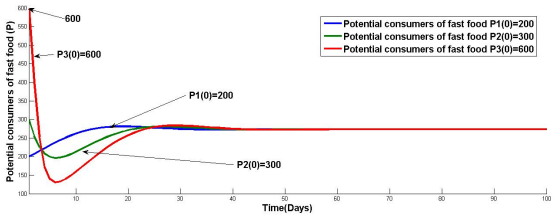
(b) Individuals who quit fast food $Q(t)$



(c) Moderate consumers of fast food $L(t)$



(d) Obese patients $C(t)$



(e) Potential consumers of fast food $P(t)$

Figure 5: Dynamics of the fast-food consumption model with $R_0 = 1.30 > 1$ under different initial conditions. All trajectories converge towards the endemic equilibrium E^*. Excessive consumers $S(t)$, moderate consumers $L(t)$, obese individuals $C(t)$, and quitters $Q(t)$ stabilize at positive values, confirming the persistence of unhealthy eating habits in the population. The potential consumers $P(t)$ decrease initially before reaching a steady state, indicating a continuous flow towards excessive consumption. These results highlight the stability of E^* when $R_0 > 1$, independent of initial conditions.

generally more vulnerable to aggressive fast-food marketing strategies and at higher risk of developing obesity-related health complications.

Funding

This research received no external funding.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] Atangana, A. *Application of fractional calculus to epidemiology,* C. Cattani, H.M. Srivastava, X.J. Yang (eds.), *Fractional Dynamics*, pp. 174–190, Walter de Gruyter, Warsaw, 2015.

[2] Cassels, S. *Overweight in the Pacific: Links between foreign dependence, global food trade, and obesity in the Federated States of Micronesia,* Global. Health, 2(1) (2006), 10.

[3] Cercato, C., and Fonseca, F.A. *Cardiovascular risk and obesity,* Diabetol. Metab. Syndr., 11 (2019), p. 74.

[4] Chitnis, N., Hyman, J.M., and Cushing, J.M. *Determining important parameters in the spread of malaria through the sensitivity analysis of a mathematical model,* Bull. Math. Biol., 70(5) (2008), 1272.

[5] Cambet, Th., Bochaton, T., Bel, F., Derimay, F., Varillon, Y., Amaz, C., Bonnefoy-Cudraz, E., Angoulvant, D., Charriere, S., Moulin, P., Mewton, N., and Bergerot, C. *Lipid goal achievements after acute myocardial infarction: the gap between real life and ESC 2019 guidelines,* Eur. J. Prev. Cardiol., 29(2) (2021), e65–e67.

[6] Fan, L., Jian-Gao, and Cao, H.-X. *Role of diet and nutritional management in non-alcoholic fatty liver disease,* J. Gastroenterol. Hepatol., 28 (2013), 81–87.

[7] Jodar, L., Santonja, F.J., and Gonzalez-Parra, G. *Modeling dynamics of infant obesity in the Region of Valencia, Spain,* Comput. Math. Appl., 56(3) (2008), 679.

[8] Gallo, G., Desideri, G., and Savoia, C. *Update on obesity and cardiovascular risk: From pathophysiology to clinical management,* Nutrients, 16(16) (2024) 2781.

[9] Hilfer, R. (ed.). *Applications of Fractional Calculus in Physics,* World Scientific, Singapore, 2000.

[10] Khajji, B., Moumine, E.M., Ferjouchia, H., Balatif, O., and Rachik, M. *Optimal control and discrete-time modelling of alcohol model with physical and psychological complications,* J. Math. Comput. Sci., 10(5) (2020), 1969–1986.

[11] La Salle, J.P. *The Stability of Dynamical Systems,* SIAM, 1976.

[12] Odibat, Z., and Momani, S. *An algorithm for the numerical solution of differential equations of fractional order,* J. Appl. Math. Inform., 26(1–2) (2008), 15–27.

[13] Safan, M., Murillo, A.L., Wadhera, D., and Castillo-Chavez, C. *Modeling the diet dynamics of children: the roles of socialization and the school environment,* Lett. Biomathematics, 5(1) (2018), 275–306.

[14] Sahib, I., Belam, M., and Labzai, A. *Mathematical modeling and optimal control strategy for a discrete-time cholera model,* Commun. Math. Biol. Neurosci., 2023 (2023), 1–20.

[15] Sahib, I., Belam, M., and Labzai, A. *The mathematical modeling and analysis of the cholera disease model,* Commun. Math. Biol. Neurosci., 2024 (2024), 1–22.

[16] Sahib, I., Belam, M., and Labzai, A. *Mathematical modeling and strategy for optimal control of diphtheria,* Results Control Optim., 7 (2024), 100156.

[17] Sahib, I., Belam, M., and Labzai, A. *Mathematical modeling and analysis of fast food consumption dynamics,* IAENG Int. J. Appl. Math., 55(7) (2025), 1–15.

[18] Signos. *The impact of processed foods on brain and metabolic health,* Signos, 2023. *Chronic disease development from poor diets,* J. Chron. Dis. Prev., 2024.

[19] van den Driessche, P., and Watmough, J. *Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission,* Math. Biosci., 180 (2002), 29–48.

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 1, 2026, pp 144–169
https://doi.org/10.22067/ijnao.2025.93541.1648
https://ijnao.um.ac.ir/

CC BY NC

Research Article

Measuring the efficiency of bi-level structured systems based on directional distance function and network DEA model

Mona Habibpoor* and Mohammadreza Alirezaee

Abstract

In practical contexts, decision-making units (DMUs) are frequently organized hierarchically, encompassing various levels of functionality, multiple subunits across different tiers, or multiple stages of production. Traditional data envelopment analysis (DEA) models often neglect these hierarchical arrangements, potentially leading to assessments that do not accurately reflect actual unit performance. This study proposes a novel DEA model specifically designed to assess the efficiency of bi-level DMUs, consisting of a leader at the upper level and a follower at the lower level. By using a directional distance function, the proposed model simultaneously minimizes inputs and maximizes outputs, offering a more flexible and accurate evaluation compared to conventional methods. The model's effectiveness is demonstrated through its application in two case studies within the banking sector, showcasing its ability to provide deeper managerial insights and improve decision-making.

*Corresponding author

Received 15 May 2025; revised 7 August 2025; accepted 4 September 2025

Mona Habibpoor
School of Mathematics and Computer Science, Iran University of Science and Technology, Narmak, Tehran, Iran. e-mail: mona_habibpoor@mathdep.iust.ac.ir

Mohammadreza Alirezaee
School of Mathematics and Computer Science, Iran University of Science and Technology, Narmak, Tehran, Iran. e-mail: mralirez@iust.ac.ir

How to cite this article
Habibpoor, M. and Alirezaee, M., Measuring the efficiency of bi-level structured systems based on directional distance function and network DEA model. *Iran. J. Numer. Anal. Optim.*, 2026; 16(1): 144-169. https://doi.org/10.22067/ijnao.2025.93541.1648

AMS subject classifications (2020): 90C08; 90B50

Keywords: Network data envelopment analysis; Bi-level structures; Directional distance function; Efficiency; Banking industry

1 Introduction

Traditional data envelopment analysis (DEA) models are often viewed as "black boxes," converting inputs to outputs without considering the internal configuration of evaluated units. These models rely on aggregating primary inputs and final outputs, potentially overlooking the complexities of unit structures. To address these shortcomings, network DEA models have emerged to integrate the internal structures of assessed units, offering a more comprehensive evaluation. Unlike conventional approaches, DEA network models are adaptable and can be tailored to different processes and various configurations of units.

Hierarchical structures within networks categorize organizations based on process intricacy and organizational scope. In large institutions like universities and banks, where hierarchical structures are prevalent, evaluating performance becomes crucial for managerial decision-making. Our study focuses on assessing bi-level units, which consist of a leader at the upper level and a follower at the lower level. The proposed model can be further expanded to evaluate various hierarchical unit typologies, thereby enhancing the applicability of DEA in organizational performance evaluations.

Some studies in the literature have investigated unit performance by examining their internal and hierarchical structures. A summary of the key literature related to hierarchical DEA is presented in Table 1. For example, Cook et al. [3] were pioneers in applying hierarchical DEA to evaluate the efficiency of power plants at various levels, a method that was later refined by Cook and Green [4] specifically for Canadian power plants. Kao [13] developed a DEA model for evaluating hierarchical units, where each level's efficiency is derived from the weighted average efficiency of its subunits at lower levels. However, these models overlook the independent performance of upper levels and intermediate outputs. Subsequent studies, such as those by Kashim, Kasim, and Abd Rahman [14], Gan et al. [8], Hendrawan, Purwanggono, and Suliantoro [12] have extended Kao's approach to various hierarchical structures. Ahmad et al. [1] expanded Kao's and Kashim's frameworks to evaluate faculty efficiency within the hierarchical structures of university faculties in Malaysia. Xiao et al. [30] extended Kao's method and utilized the cross-efficiency approach to assess the hierarchical network structures of 20 universities. They incorporated a secondary objective of maximizing-minimizing and employed the CRITIC method

for determining criterion weights. Serkani et al. [25] combined Kao's method with intuitionistic fuzzy analytic network process to evaluate the efficiency of the faculty of basic sciences of Islamic Azad University in a hierarchical structure.

Ghasemi et al. [10] developed the approaches presented for evaluating two-stage structures proposed by Liang, Cook, and Zhu [18] and Du et al. [7] and applied them to evaluate the effectiveness of hierarchical structures in schools. Luban [19] introduced a fuzzy DEA model for hierarchical systems designed to determine the activity levels of decision-making units (DMUs) in order to meet both the global targets set by the coordinating unit and the individual targets of the DMUs. Deville, Ferrier, and Leleu [6] proposed an approach to measuring the performance of hierarchical organizations, such as bank branch networks, by decomposing overall performance into technical efficiency (TE) at the lower level (branch) and allocative efficiency at the upper level (top management). Ghasemi et al. [9] developed an incremental model for multi-stage structure proposed by Cook et al. [5] and combined it with fuzzy models to evaluate the hierarchical structure in Farhangian University campuses with imprecise data.

Wu [29] suggested a bi-level DEA model to assess cost efficiency in DMUs with one leader and one follower, while Zhou et al. [33] developed a similar model for systems with one leader and multiple followers. Yao, Shuai, and Chen [31] further extended Zhou's work to assess vulnerability in regional water systems. Omrani, Mohammadi, and Emrouznejad [22] developed a bi-level multi-objective DEA model to concurrently assess the performance of DMUs across two hierarchical dimensions, utilizing fuzzy goal programming to solve their model. Walheer and Hudik [27] proposed a nonparametric method to minimize costs within multidivisional firms, with applications in the Chinese manufacturing sector. Zhang and Chen [32] extended the idea of Kao [13] to a hierarchical system with two-stage processes and suggested additive and multiplicative network DEA models for evaluating their efficiency, employing semidefinite programming to solve these models and assess performance in China's high-technology sector. However, the nonlinear nature of their model limits its applicability.

Li et al. [17] proposed a hierarchical framework to assess power plant efficiency, using a DEA-like approach for sub-unit efficiencies and aggregating them into a two-stage process. They also introduced goal programming to minimize discrepancies between projected values of plant-level and corresponding total values of sub-unit projections. However, their model may not directly apply in cases where level 2 DMUs have distinct inputs and outputs. Pachar, Gupta, and Jha [23] introduced a DEA method to compute cost efficiency in multiple retail stores, within a Stackelberg dynamic network. Zhu et al. [34] evaluated the operational performance of hotels and their departments, such as catering and occupancy, within a hierarchical framework. Habibpoor, Alirezaee, and Rashidinia [11] proposed a DEA model for evaluating the performance of bi-level units, where lower level efficiency serves as an upper limit for higher-level efficiency. Muvingi et al.

[21] discussed a multi-function parallel system with hierarchical network structures, introducing additive and nonlinear multiplicative DEA models for operational performance assessment.

Muvingi, Peer, and Lotfi [20] proposed hierarchical non-Archimedean DEA models to assess the efficiency of DMUs with a hierarchical structure. Their model was applied to analyze the locations of mobile money agents, taking into account the variations in group sizes across different regions. Shafiee and Saleh [26] extended the DEA approach to accommodate tri-level structures and proposed a goal programming method for its resolution. Kweh et al. [16] investigated the impact of environmental, social, and governance factors on the efficiency of 29 suppliers, partners, and customers of Japan Airlines from 2013 to 2022, utilizing a two-stage hierarchical network DEA technique. Additionally, Muvingi, et al. [21] analyzed the performance of hierarchical units while considering fuzzy characteristics in input and output data based on TOPSIS rankings.

Much of the existing research in network DEA has concentrated on evaluating performance in two-stage or multi-stage structures, whereas hierarchical structures have received comparatively less attention [24]. In multi-stage models, it is typically assumed that the output of each stage serves as the input for the subsequent stage. However, in real-world settings, many organizations operate within hierarchical frameworks that involve multiple functional levels or consist of several subunits across different layers. A review of the literature on hierarchical structures reveals that while some studies have examined cost performance, others have focused on operational efficiency—primarily using the methodologies proposed by Cook and Kao. However, the approaches introduced by Cook et al [3, 4] and Kao [13] tend to disregard the independent performance of higher-level units and overlook outputs and intermediate activities, thereby limiting both the accuracy and applicability of these models. Addressing these challenges requires developing robust evaluation methods tailored to the hierarchical structure of organizations.

In this study, we seek to overcome these challenges by explicitly incorporating the structural characteristics commonly observed in real-world hierarchical systems. This research offers a substantial contribution to the field of efficiency analysis within complex organizational environments—particularly in hierarchical institutions such as banks—by introducing a novel DEA approach. The proposed approach is designed to evaluate the performance of bi-level DMUs, explicitly accounting for inter-level interactions and the role of intermediate products transmitted between hierarchical layers. Our method leverages the distance function model, which offers greater flexibility compared to traditional models for measuring efficiency improvements. Specifically, this approach enables the simultaneous analysis of input contractions and output expansions, thereby supporting a more comprehensive evaluation of potential efficiency improvements across multiple dimensions. This model enables decision-makers to simultaneously evaluate input reductions and output increases, facilitating improvements in both dimensions. The distance function approach is especially effective in handling undesirable outputs, allowing

Table 1: Selected Literature on hierarchical network DEA Evaluation

Authors	**Methods**	**Application**	**Contributions**
Cook et al. [3]	Hierarchical DEA model	Power plants	Introduced hierarchical DEA to evaluate power plant efficiency across multiple levels
Cook & Green [4]	Hierarchical DEA model	Canadian power plants	Refined hierarchical DEA model for assessing power plant
Kao [13]	Hierarchical DEA model	Hierarchical units	Developed a DEA model for performance assessment at different hierarchical levels
Wu [29]	Bi-level DEA model	Cost efficiency in DMUs	Proposed a bi-level DEA model for assessing cost efficiency in systems with one leader and one follower
Zhou et al. [33]	Bi-level DEA model	Cost efficiency in Supply Chain	Developed a bi-level DEA model for systems with one leader and multiple followers
Omrani et al. [22]	Bi-level multi-objective DEA	profit and operational efficiency in Banking (Multi-dimensional performance)	Developed a multi-objective DEA model with fuzzy goal programming for hierarchical DMU performance
Walheer & Hudik [27]	Nonparametric method	Cost performance in Multidivisional firms in high-technology manufacturing industry	Proposed a method to minimize costs and reallocate resources across firms' divisions
Zhang & Chen [32]	Additive & multiplicative network DEA	TE of Hierarchical systems in high-technology industry	Suggested DEA model using semidefinite programming for hierarchical systems with two-stage processes
Li et al. [17]	DEA-like model with goal programming	Power plants	Combined DEA-like approach and goal programming to evaluate sub-unit performance and reduce discrepancies
Pachar et al. [23]	DEA with Stackelberg dynamic network	Cost efficiency in retail stores	Developed a Stackelberg-based DEA model for cost efficiency in retail stores
Zhu et al. [34]	hierarchical DEA	Operational performance in hotels and tourism	Evaluated operational performance of hotels and their department
Habibpoor et al. [11]	Bi-level DEA	Banking	Proposed a DEA model for evaluating two-level unit performance
Kremantzis et al. [15]	Additive & nonlinear multiplicative DEA	Operational performance in Higher education	Designed DEA models for operational performance in multi-function parallel systems

for their reduction alongside undesirable inputs, while potentially boosting desirable outputs. Furthermore, the direction of improvement can be customized according to specific contexts such as policy objectives or technological constraints. The results derived from this model are straightforward to interpret, representing the amount of adjustment required to reach the efficient frontier [28]. To demonstrate the practical applicability of our bi-level model, we apply it to two real-world case studies within the banking industry, providing empirical evidence of its effectiveness. Therefore, the subsequent sections of this paper are structured as follows: Section 2 introduces our proposed TE and pure efficiency (PE) models for evaluating bi-level unit performance. Section 3 presents two case studies illustrating the application and benefits of our proposed model. Finally, Section 4 provides concluding remarks.

2 Proposed models for bi-level units

The directional distance function proposed by Chambers, Chung, and Färe [2], inspired by Luenberger's benefit function, to assess TE through optimizing outputs and inputs concurrently. This method involves comparing input-output vectors (X_0, Y_0) with a reference bundle (g_X, g_Y) against a production possibility set T. The directional distance function can be defined as

$$\overrightarrow{D}(X_0, Y_0, g_X, g_Y) = \text{Max } \varphi: \quad (X_0+\varphi g_X, Y_0+\varphi g_Y) \in T. \tag{1}$$

The directional distance function's evaluation at a specific input-output bundle depends on (g_X, g_Y) and the chosen reference technology. The selected bundle (g_X, g_Y) determines the direction in which the observed bundle, if it is an interior point and is projected onto the efficient frontier of the production possibility set.

The choice of (g_X, g_Y) is arbitrary. We could opt select $(-X_0,\ Y_0)$ for (g_X, g_Y), in which case the directional distance function transforms into the following expression:

$$\overrightarrow{D}(X_0, Y_0) = \text{Max } \varphi: \{(1-\varphi)X_0, (1+\varphi)Y_0)\} \in T. \tag{2}$$

In essence, this approach aims to increase output and decrease input in proportion to φ. For instance, if φ is 10%, then we aim to expand all outputs by 10% while simultaneously reducing all inputs by 10%.

Under the assumption of constant returns to scale (CRS), the production possibility set derived from a set of j DMUs is defined as

$$T_C = \left\{ (\mathrm{X}, \mathrm{Y}) : \quad \sum_{\mathrm{j}=1}^{\mathrm{J}} \lambda_{\mathrm{j}} \mathrm{X}_{\mathrm{j}} \leq \mathrm{X}, \quad \sum_{\mathrm{j}=1}^{\mathrm{J}} \lambda_{\mathrm{j}} \mathrm{Y}_{\mathrm{j}} \geq \mathrm{Y}, \quad \lambda_{\mathrm{j}} \geq 0, \quad \mathrm{j} = 1, 2, \ldots, \mathrm{J} \right\}. \tag{3}$$

The efficiency model formulated using the directional distance function for CRS production possibility sets, $\mathrm{T_C}$, can be expressed as

$$\begin{aligned} TE_0 = Max \;\; & \varphi \\ s.t. \quad & \textstyle\sum_{\mathrm{j}=1}^{\mathrm{J}} \lambda_{\mathrm{j}} \mathrm{X}_{\mathrm{j}} \leq (1-\varphi)\mathrm{X}_0, \\ & \textstyle\sum_{\mathrm{j}=1}^{\mathrm{J}} \lambda_{\mathrm{j}} \mathrm{Y}_{\mathrm{j}} \geq (1+\varphi)\mathrm{Y}_0, \\ & \lambda_{\mathrm{j}} \geq 0, \quad j = 1, 2, \ldots, J, \end{aligned} \tag{4}$$

where φ represents the measure of technical inefficiency of the DMU_0. By implication, its efficiency equals $(1-\varphi)$. Hence, DMU_0 is efficient if and only if $\varphi^* = 0$. In other words, we aim to increase the output and reduce the input by the proportion φ.

Let $\mathrm{t} = 1 - \varphi$, the above model can be transformed into the following form:

$$\begin{aligned} TE_0 = Min \;\; & t \\ s.t. \quad & \textstyle\sum_{\mathrm{j}=1}^{\mathrm{J}} \lambda_{\mathrm{j}} \mathrm{X}_{\mathrm{j}} \leq \mathrm{t}\mathrm{X}_0 \\ & \textstyle\sum_{\mathrm{j}=1}^{\mathrm{J}} \lambda_{\mathrm{j}} \mathrm{Y}_{\mathrm{j}} \geq (2-\mathrm{t})\mathrm{Y}_0, \\ & \lambda_{\mathrm{j}} \geq 0, \quad j = 1, 2, \ldots, J, \end{aligned} \tag{5}$$

where t is the efficiency score of the DMU_0. The dual problem of (5) is the following model:

$$\begin{aligned} TE_0 = Max \;\; & 2\ \mathrm{u}\mathrm{Y}_0 \\ s.t. \quad & \mathrm{u}\mathrm{Y}_{\mathrm{j}} - \mathrm{v}\mathrm{X}_{\mathrm{j}} \leq 0, \quad \mathrm{j} = 1, 2, \ldots, \mathrm{J}, \\ & \mathrm{v}\mathrm{X}_0 + \mathrm{u}\mathrm{Y}_0 = 1, \\ & u,\ v \geq 0. \end{aligned} \tag{6}$$

It is evident that model (6) is derived from the conversion of the following fractional model:

$$\begin{aligned} TE_0 = Max \;\; & \tfrac{2\ \mathrm{u}\mathrm{Y}_0}{\mathrm{v}\mathrm{X}_0 + \mathrm{u}\mathrm{Y}_0} \\ s.t. \quad & \tfrac{2\mathrm{u}\mathrm{Y}_{\mathrm{j}}}{\mathrm{v}\mathrm{X}_{\mathrm{j}} + \mathrm{u}\mathrm{Y}_{\mathrm{j}}} \leq 1, \quad \mathrm{j} = 1, 2, \ldots, \mathrm{J}, \\ & u,\ v \geq 0. \end{aligned} \tag{7}$$

The above model measures the traditional TE of DMU_0 by considering the aggregate inputs X_0 and final outputs Y_0, utilizing the directional distance function. In conventional model, DMUs are treated as black boxes, where their performance is evaluated by considering the primary inputs and final outputs, overlooking their internal structures. However, many DMUs

exhibit hierarchical arrangements, with each unit containing lower level subunits or different functional levels.

Consider the bi-level structure depicted in Figure 1. In this framework, the leader at the upper level uses a portion of the inputs to produce independent final outputs and intermediate outputs for their subordinates. The remaining inputs and intermediate outputs are then allocated to the follower at the lower level. The follower uses both sets of inputs to generate their final outputs. Essentially, after receiving inputs from the leader, the followers act independently and demonstrate their performance based on these inputs. Thus, the leader's performance depends on the optimal performance of the follower at the lower level. As a result, a unit's overall performance is achieved based on the optimal performance of both the leader and the follower.

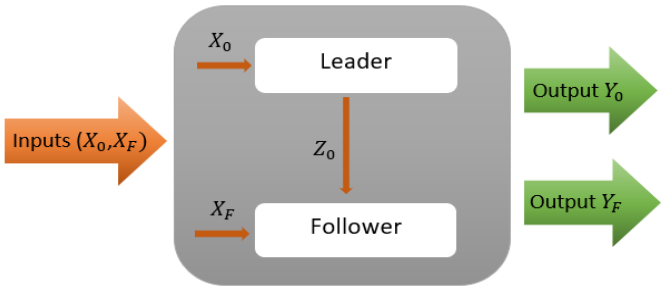


Figure 1: *The internal structure of* DMU_0.

Prior to formulating the model, we define the following symbols:

J: The total number of DMUs

j: The jth DMU possessing a bi-level structure

X_{Lj}: The input of the leader j

Y_{Lj}: The final output of the leader j

Z_{Lj}: The intermediate measure from leader j to follower j

X_{Fj}: The input of follower j

Y_{Fj}: The final output of follower j

v_1: The input weight vector of the leader

u_1: The output weight vectors of the leader

v_2: The input weight vector of the follower

u_2: The output weight vectors of the follower

w: The weight vector corresponding the intermediate measure from the leader to follower

2.1 Proposed TE model for bi-level units

The TE of the DMU_0's follower at the lower level based on directional distance function can be formulated as follows:

$$\begin{aligned} TE_{F0} = \quad & Max \quad \frac{2u_2Y_{F0}}{v_2X_{F0}+ wZ_0+u_2Y_{F0}} \\ & s.t. \quad \frac{2u_2Y_{Fj}}{v_2X_{Fj}+wZ_j+u_2Y_{Fj}} \leq 1, \qquad j = 1, 2, \ldots, J, \\ & \qquad u_2 \geq 0, \quad v_2 \geq 0, \quad w \geq 0. \end{aligned} \tag{8}$$

This model features a bounded objective function with a value of one and also has a feasible solution, ensuring the existence of a finite optimal solution. Model (8) can be transformed into a linear model for easier solution. The equivalent linear model of model (8) is

$$\begin{aligned} & TE_{F0} = Max \quad 2u_2Y_{F0} \\ & s.t. \quad u_2Y_{Fj} - v_2X_{Fj} - wZ_j = 0, \qquad j = 1, 2, \ldots, J, \\ & \qquad v_2X_{F0}+ wZ_0+u_2Y_{F0} = 1, \\ & \qquad u_2 \geq 0, \quad v_2 \geq 0, \quad w \geq 0. \end{aligned} \tag{9}$$

In the hierarchical structure, the follower at the lower level operates independently based on assigned input resources, while the leader's efficiency is assessed relative to the optimal performance of the follower under their supervision. Therefore, in terms of mathematical expression, the optimal solution of the upper level model should be feasible for both the upper and lower levels, and the lower level efficiency score should remain at TE^*_{F0}, which indicates the optimal solution of the lower level problem (9). As a result, the proposed DEA model for evaluating the TE of the leader at the upper level can be stated as follows:

$$\begin{aligned} & TE_{L0} = Max \quad \frac{2u_1Y_{L0} +2wZ_0}{v_1X_{L0} +u_1Y_{L0}+ wZ_0} \\ & s.t. \qquad \frac{2u_1Y_{Lj} + 2wZ_j}{v_1X_{Lj}+u_1Y_{Lj} + wZ_j} \leq 1, , \qquad j = 1, 2, \ldots, J, \\ & \qquad \frac{2u_2Y_{Fj}}{v_2X_{Fj}+wZ_j+u_2Y_{Fj}} \leq 1, \qquad j = 1, 2, \ldots, J, \\ & \qquad \frac{2u_2Y_{F0}}{v_2X_{F0}+ wZ_0+u_2Y_{F0}} = TE^*_{F0}, \\ & \qquad u_1 \geq 0, \quad v_1 \geq 0, \quad u_2 \geq 0, \quad v_2 \geq 0, \quad w \geq 0. \end{aligned} \tag{10}$$

According to its first constraint ($\frac{2u_1Y_{Lj}\ +\ 2wZ_j}{v_1X_{Lj}+u_1Y_{Lj}\ +\ wZ_j} \leq 1,\ \ j=1,2,\ldots,J$), the objective function of the above model is constrained between zero and one. Additionally, if ($u_2,\ v_2, w$) is the optimal solution of the lower level problem, then ($\frac{Y_{F0}}{Y_{L0}}\ u_2, \frac{X_{Fj}}{X_{L0}}\ v_2, w$) serves as a feasible solution for the higher-level model. Therefore, the upper level model is feasible and has a finite optimal solution. Similarly, the equivalent linear form of model (10) can be expressed as follows:

$$\begin{aligned} TE_{L0} = Max\ \ & 2u_1Y_{L0} + 2wZ_0 \\ s.t.\qquad & u_1Y_{Lj} + wZ_j - v_1X_{Lj} \leq 0, \qquad j=1,2,\ldots,J, \\ & u_2Y_{Fj} - v_2X_{Fj} - wZ_j \ \leq 0, \qquad j=1,2,\ldots,J, \\ & v_1X_{L0} + u_1Y_{L0} + wZ_0 = 1, \\ & 2u_2Y_{F0} - TE^*_{F0}\,(\,v_2X_{F0} + wZ_0 + u_2Y_{F0}) = 0, \\ & u_1 \geq 0, \quad v_1 \geq 0, \quad u_2 \geq 0, \quad v_2 \geq 0, \quad w \geq 0. \end{aligned} \tag{11}$$

After solving the model (11) and obtaining the upper level TE (TE_{L0}), we can determine the overall TE of bi-level unit DMU_0.

The overall TE of the DMU_0, $0 \epsilon \{1,2,\ldots,J\}$, can be expressed as follows:

$$\begin{aligned} TE_0 &= \frac{2u_1Y_{L0}+2u_2Y_{F0}+2wZ_0}{u_1Y_{L0}\ +\ u_2Y_{F0}+v_1X_{L0}\ +\ v_2X_{F0}+2\ wZ_0} \\ &= \frac{v_1X_{L0}\ +u_1Y_{L0}+\ wZ_0}{u_1Y_{L0}\ +\ u_2Y_{F0}+v_1X_{L0}\ +\ v_2X_{F0}+2\ wZ_0} \times \frac{2u_1Y_{L0}+2wZ_0}{v_1X_{L0}\ +u_1Y_{L0}+\ wZ_0} \\ &\quad + \frac{v_2X_{F0}+\ wZ_0+u_2Y_{F0}}{u_1Y_{L0}\ +\ u_2Y_{F0}+v_1X_{L0}\ +\ v_2X_{F0}+2\ wZ_0} \times \frac{2u_2Y_{F0}}{v_2X_{F0}+\ wZ_0+u_2Y_{F0}} \\ &= \frac{1}{1+t}TE_{L0} + \frac{t}{1+t}TE_{F0}, \end{aligned} \tag{12}$$

where $t = u_2^*Y_{F0}+v_2^*X_{F0}+\ w^*Z_0$ is obtained from solving model (11). Thus, the TE of the whole unit is computed as the weighted average of the upper and lower levels' technical efficiencies. A unit is deemed technically efficient if $TE_0 = 1$.

A notable advantage of our proposed approach is that if both the upper and lower levels are efficient, the entire unit will also be efficient. This feature distinguishes our method from previous research suggestions. Another significant advantage of our model is that we calculate the efficiencies of the upper and lower levels, considering their hierarchical relationship and internal interactions. Subsequently, we evaluate the overall efficiency of the bi-level system by integrating the impacts of both levels' performances. Furthermore, our model is adaptable to multi-level structures and can be customized for scenarios where units exhibit variable returns to scale (VRS).

2.2 Proposed PE model for bi-level units

The lower level PE model based on the directional distance function for $DMU_0, 0 \in \{1, 2, \ldots, J\}$ can be represented as follows:

$$\begin{aligned} PE_{F0} = Max \quad & \varphi_F \\ s.t. \quad & \sum_{j=1}^{J} \lambda_j X_{Fj} \leq (1-\varphi_F) X_{F0}, \\ & \sum_{j=1}^{J} \lambda_j Z_j \leq (1-\varphi_F) Z_0, \\ & \sum_{j=1}^{J} \lambda_j Y_{Fj} \geq (1+\varphi_F) Y_{F0}, \\ & \sum_{j=1}^{J} \lambda_j = 1, \\ & \lambda_j \geq 0, \quad j = 1, 2, \ldots, J, \end{aligned} \tag{13}$$

where φ_F is the measure of pure inefficiency of the DMU_0's follower. In fact, the follower of DMU_0 is purely efficient if and only if ${\varphi_F}^* = 0$.

Now, let $t = 1-\varphi_F$, the above model can be converted into the following model:

$$\begin{aligned} PE_{F0} = Min \quad & t \\ s.t. \quad & \sum_{j=1}^{J} \lambda_j X_{Fj} \leq t X_{F0}, \\ & \sum_{j=1}^{J} \lambda_j Z_j \leq t Z_0, \\ & \sum_{j=1}^{J} \lambda_j Y_{Fj} \geq (2-t) Y_{F0}, \\ & \sum_{j=1}^{J} \lambda_j = 1, \\ & \lambda_j \geq 0, \quad j = 1, 2, \ldots, J, \end{aligned} \tag{14}$$

where t is the PE score of the follower in DMU_0. The dual model of (14) is the following model:

$$\begin{aligned} PE_{F0} = Max \quad & 2u_2 Y_{F0} - u_{02} \\ s.t. \quad & u_2 Y_{Fj} - v_2 X_{Fj} - w Z_j - u_{02} \leq 0, \quad j = 1, 2, \ldots, J, \\ & v_2 X_{F0} + w Z_0 + u_2 Y_{F0} = 1, \\ & u_2 \geq 0, \quad v_2 \geq 0, \quad w \geq 0, \end{aligned} \tag{15}$$

where u_{02} is free in sign variable corresponding to the lower level problem. It is evident that model (15) is derived from the conversion of the following fractional model:

$$\begin{aligned} PE_{F0} = Max \quad & \frac{2u_2 Y_{F0} - u_{02}}{v_2 X_{F0} + w Z_0 + u_2 Y_{F0}} \\ s.t. \quad & \frac{2u_2 Y_{Fj} - u_{02}}{v_2 X_{Fj} + w Z_j + u_2 Y_{Fj}} \leq 1, \quad j = 1, 2, \ldots, J, \\ & u_2 \geq 0, \quad v_2 \geq 0, \quad w \geq 0, \\ & u_{02} \text{ free in sign.} \end{aligned} \tag{16}$$

In a similar manner, the optimal solution of the upper level model should be feasible for both the upper and lower levels, and the lower level efficiency score should remain at $\mathrm{PE}^{*}_{\mathrm{F0}}$, which indicates the optimal solution of the lower level problem (15). So, the upper level PE of the DMU_0, $0 \in \{1,\ 2,\ \ldots,\ \mathrm{J}\}$ can be represented as follows:

$$
\begin{aligned}
PE_{L0} = Max\ \ & \tfrac{2u_1Y_{L0}\ +2wZ_0-u_{01}}{v_1X_{L0}\ +u_1Y_{L0}+\ wZ_0} \\
s.t.\ \ & \tfrac{2u_1Y_{Lj}\ +\ 2wZ_j-u_{01}}{v_1X_{Lj}+u_1Y_{Lj}\ +\ wZ_j} \le 1, \qquad j = 1,2,\ldots,J, \\
& \tfrac{2u_2Y_{Fj}-u_{02}}{v_2X_{Fj}+wZ_j+u_2Y_{Fj}} \le 1, \qquad j = 1,2,\ldots,J, \\
& \tfrac{2u_2Y_{F0}-u_{02}}{v_2X_{F0}+\ wZ_0+u_2Y_{F0}} = \mathrm{PE}^{*}_{\mathrm{F0}}, \\
& u_1 \ge 0, \quad v_1 \ge 0, \quad u_2 \ge 0, \quad v_2 \ge 0, \quad w \ge 0,
\end{aligned}
\tag{17}
$$

where u_{01} is free in sign variable corresponding to the upper level problem. The equivalent linear reformulation of the model (17) can be written as follows:

$$
\begin{aligned}
PE_{L0} = Max\ \ & 2u_1Y_{L0}\ +\ 2wZ_0-u_{01} \\
s.t.\ \ & u_1Y_{Lj}\ +wZ_j\ -v_1X_{Lj}-u_{01} \le 0, \qquad j = 1,2,\ldots,J, \\
& u_2Y_{Fj}-v_2X_{Fj}-wZ_j\ \ -u_{02} \le 0, \qquad j = 1,2,\ldots,J, \\
& v_1X_{L0}\ +u_1Y_{L0}+\ wZ_0 = 1, \\
& 2u_2Y_{F0}-u_{02}-\mathrm{PE}^{*}_{\mathrm{F0}}\,(\ v_2X_{F0}+\ wZ_0+u_2Y_{F0}) = 0, \\
& u_1 \ge 0, \quad v_1 \ge 0, \quad u_2 \ge 0, \quad v_2 \ge 0, \quad w \ge 0.
\end{aligned}
\tag{18}
$$

After solving the model (18) and obtaining the upper level PE ($\mathrm{PE}_{\mathrm{L0}}$), we can determine the overall PE of bi-level unit DMU_0. The overall PE of the DMU_0, $0\epsilon\{1,2,\ldots,\mathrm{J}\}$, can be expressed as follows:

$$
\begin{aligned}
PE_0 &= \tfrac{2u_1Y_{L0}\ +\ 2u_2Y_{F0}+2wZ_0\ -\ u_{01}-u_{02}}{u_1Y_{L0}\ +\ u_2Y_{F0}+v_1X_{L0}\ +\ v_2X_{F0}+2\ wZ_0} \\
&= \tfrac{v_1X_{L0}\ +u_1Y_{L0}+\ wZ_0}{u_1Y_{L0}\ +\ u_2Y_{F0}+v_1X_{L0}\ +\ v_2X_{F0}+2\ wZ_0} \times \tfrac{2u_1Y_{L0}+2\ wZ_0\ -\ u_{01}}{v_1X_{L0}\ +u_1Y_{L0}+\ wZ_0} \\
&\quad + \tfrac{v_2X_{F0}+\ wZ_0+u_2Y_{F0}}{u_1Y_{L0}\ +\ u_2Y_{F0}+v_1X_{L0}\ +\ v_2X_{F0}+2\ wZ_0} \times \tfrac{2u_2Y_{F0}\ -\ u_{02}}{v_2X_{F0}+\ wZ_0+u_2Y_{F0}} \\
&= \tfrac{1}{1+t}\mathrm{PE}_{\mathrm{L0}} + \tfrac{t}{1+t}\mathrm{PE}_{\mathrm{F0}},
\end{aligned}
\tag{19}
$$

where $t = u_2^* Y_{F0} + v_2^* X_{F0} + w^* Z_0$ is obtained from solving model (18). So, the overall PE is computed as the weighted average of the upper and lower levels' PEs. A unit is deemed PE if $PE_0 = 1$. If both the upper and lower levels are pure efficient, then the whole unit will also be pure efficient. Our suggested model can be further developed to accommodate different interactions that may occur between upper and lower levels.

3 Empirical examples and discussion

Case study 1. A case study of an Iranian commercial bank

The banking industry holds a critical position within any economy, fulfilling essential functions such as deposit acceptance, transaction intermediation, payment facilitation, and credit allocation. Moreover, banks play a pivotal role in maintaining economic stability through the implementation of monetary policies. Hence, ensuring the vitality and efficiency of the banking system is paramount, as poor performance can potentially trigger financial and economic crises. To avert such scenarios, bank managers must consistently assess the performance of their branches and devise strategies for improvement. Given the broad spectrum of banking operations across diverse economic sectors, it becomes crucial to comprehensively evaluate all dimensions of their performance.

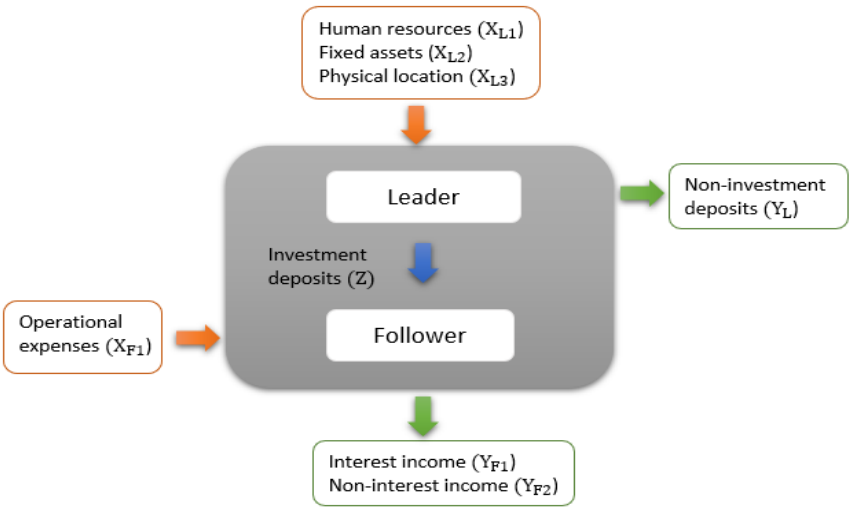


Figure 2: Bi-level structure in the case study of Iranian commercial bank.

In the realm of banking industry performance analysis, two primary perspectives define inputs: The production perspective and the intermediation perspective. From the production

perspective, banks are viewed as producers utilizing human and physical resources to generate various deposit accounts. Consequently, deposits are considered outputs within this framework. Conversely, the intermediation perspective portrays banks as facilitators of financial services. Here, banks gather deposits from customers and utilize operational expenses (e.g., deposit interest costs and other operational expenditures) to transform these funds into loans and services, thereby earning interest and non-interest income.

Recognizing the multifaceted activities of banks in both realms, managers often seek to evaluate branch performance by considering multiple facets simultaneously. This article adopts a bi-level framework that integrates both the production and intermediation approaches. In this study, we conceptualize the bank as a bi-level unit, functioning as a producer at the upper level and an intermediary at the lower level, as depicted in Figure 2. We apply this framework to assess the performance of 24 branches of a commercial bank in Iran within this bi-level structure.

Table 2: Data sets of the leaders and followers.

Branch	x_{L1} (Person)	x_{L2} (1 billion Rials)	x_{L3} (m∧2)	z_1 (10 billion Rials)	y_{L1} (100 million Rials)	x_{F1} (100 million Rials)	y_{F1} (10 billion Rials)	y_{F1} (10 billion Rials)
1	2.50	1.87	814.00	49.20	30.94	3.61	15.25	13.67
2	3.75	2.12	244.20	11.81	31.25	3.33	13.27	3.28
3	4.25	2.91	350.02	12.79	30.31	2.37	13.42	3.55
4	2.75	1.82	179.08	8.36	30.94	3.67	12.96	2.32
5	6.50	1.48	284.90	10.33	30.63	4.06	13.12	2.87
6	6.75	1.43	268.62	11.81	30.00	4.00	12.81	3.42
7	10.50	2.47	195.36	19.19	16.56	2.88	10.52	5.06
8	19.00	4.68	431.42	27.55	10.31	0.28	9.61	7.38
9	24.25	4.58	179.08	19.68	23.75	0.45	14.18	5.19
10	16.00	4.93	195.36	15.74	4.06	0.17	10.07	3.96
11	4.25	1.38	211.64	10.82	30.94	4.23	13.12	3.01
12	21.00	4.24	219.78	25.58	23.44	0.79	14.95	6.84
13	25.00	4.73	203.50	23.62	0.94	0.28	9.30	6.01
14	3.50	1.58	179.08	11.32	30.63	3.95	13.12	3.14
15	8.00	1.68	187.22	11.32	25.63	3.84	12.20	3.14
16	1.50	0.69	81.40	4.43	30.94	5.02	12.81	1.23
17	9.50	2.56	284.90	22.14	30.31	2.82	14.03	6.15
18	2.75	1.82	179.08	7.87	30.63	3.72	12.81	2.32
19	3.00	1.23	170.94	13.28	28.44	5.64	13.12	3.69
20	6.00	1.73	252.34	9.84	29.69	3.95	12.51	3.42
21	7.25	1.73	260.48	12.30	29.69	4.12	12.96	3.55
22	10.50	1.97	195.36	16.24	15.94	3.10	9.76	4.51
23	21.50	4.54	488.40	24.60	11.25	0.45	9.30	6.97
24	23.00	4.09	170.94	20.66	21.88	0.34	13.73	4.51

The leader's inputs at the production level include human resources (x_{L1}), fixed assets (x_{L2}), and physical locations (x_{L3}). The leader's outputs consist of investment deposits (z_1) as an intermediary output from leader to follower and non-investment deposits (y_{L1}) as an independent output. Investment deposits are the output for the production level and serve as input for the intermediation level. At the intermediary level, the inputs include invested deposits (z_1) and operational expenses (x_{F1}). The outputs include interest income (y_{F1}) and non-interest income (y_{F2}). Table 2 presents the datasets of the input and output indicators for the leaders and followers.

If the black box approach is utilized, then the bank's bi-level structure is disregarded. Figure 3 displays the black box structure for our scenario. Essentially, we will have a DMU with four inputs and three outputs. To assess the bank's performance in this manner, four inputs (human resources, fixed assets, physical location, and operational expenses) are employed to generate three outputs (non-investment deposits, interest income, and non-interest income).

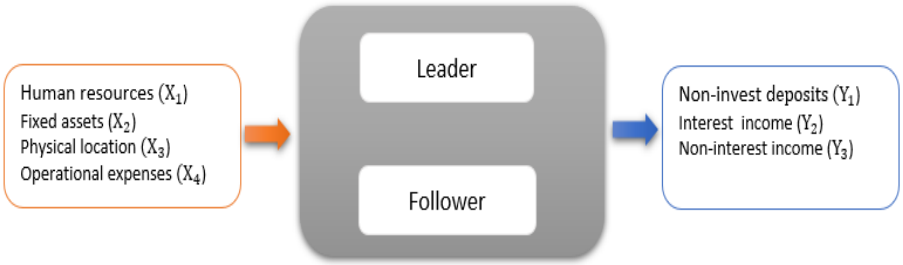


Figure 3: Black-box structure in the case study of an Iranian commercial bank.

The results of TE scores for the selected branches using both black-box and proposed models are presented in Table 3. Additionally, the upper level TE (TE_L) and lower level TE (TE_F) scores for the selected branches using the suggested model are also displayed in Table 3.

Table 4 displays the PE calculation results for all chosen branches using both black-box and suggested models. In Table 4, PEC_L and PEC_F represent the upper level and lower level PE scores using our model, respectively.

The following results were derived from the examination of the tables and the conduct of this study:

Increasing discrimination power in TE assessment:

The findings demonstrate that while the conventional model identifies 15 units as technically efficient, our model designates only 1 unit as efficient. Specifically, the black-box model categorizes regions into 10 groups, whereas the proposed model divides them into 24 categories.

Table 3: TE results of the black-box and proposed models.

DMU	Black-Box TE	Ranking	Proposed TE	TE_L	TE_F	Ranking
1	1	1	0.987	1	0.974	4
2	1	1	0.884	0.814	0.968	18
3	1	1	0.843	0.709	1	22
4	1	1	0.931	0.841	0.997	12
5	0.995	3	0.835	0.750	0.950	23
6	0.994	4	0.871	0.814	0.943	19
7	1	1	0.964	1	0.927	7
8	1	1	0.905	0.821	1	15
9	1	1	0.982	0.967	1	5
10	1	1	0.920	0.844	1	14
11	0.996	2	0.895	0.857	0.938	16
12	1	1	0.994	1	0.988	3
13	1	1	0.980	0.985	0.975	6
14	1	1	0.931	0.917	0.945	11
15	0.945	10	0.893	0.859	0.931	17
16	1	1	1	1	1	1
17	1	1	0.961	0.957	0.964	8
18	0.993	5	0.920	0.819	1	13
19	1	1	0.947	1	0.889	10
20	0.963	7	0.835	0.725	1	24
21	0.954	9	0.856	0.791	0.937	20
22	0.986	6	0.954	0.966	0.942	9
23	0.959	8	0.851	0.739	1	21
24	1	1	0.995	1	0.991	2

This results in a %41 improvement in the efficiency separation of units, highlighting the superior discrimination power of our model compared to the traditional approach.

Increasing discrimination power in PE assessment:

Our proposed model shows a significant increase in the discrimination of PEs when compared to the black-box model. As shown in Table 3, while the black-box model, assuming VRS, groups regions into 8 categories, our model classifies them into 20 categories. This enhanced discrimination offers a more nuanced perspective on unit rankings, providing greater differentiation between units. Thus, our model delivers a more precise evaluation of unit performance than conventional methods.

Increasing discrimination compared to traditional approaches:

For units deemed inefficient by both methods, the differences in their efficiency scores and rankings highlight the fundamental differences between the black-box and bi-level approaches. This comparison illustrates how evaluating units by considering their various functional aspects simultaneously produces different results than conventional methods, which fail to account for interactions and internal structures. A key factor contributing to this enhanced discrimination is the integration of the leader-follower relationship inherent in bi-level structures. Addition-

Table 4: PE results of the black-box and proposed models.

DMU	Black-Box PE	Ranking	Proposed PE	PE_L	PE_F	Ranking
1	1	1	1	1	1	1
2	1	1	1	1	0.993	1
3	1	1	0.970	0.969	1	11
4	1	1	0.994	0.994	1.000	4
5	0.996	2	0.985	0.985	0.990	8
6	1	1	0.965	0.965	0.977	13
7	1	1	0.965	1	0.931	12
8	1	1	0.925	0.857	1	17
9	1	1	0.991	0.985	1	6
10	1	1	0.924	0.866	1	18
11	0.996	3	0.996	0.996	0.988	3
12	1	1	1	1	1	1
13	1	1	0.984	0.993	0.976	10
14	1	1	0.993	0.994	0.986	5
15	0.970	6	0.893	0.870	0.942	19
16	1	1	1	1	1	1
17	1	1	0.997	1	0.993	2
18	0.995	4	0.984	0.984	1	9
19	1	1	0.990	1	0.980	7
20	0.966	7	0.951	0.951	1	16
21	0.960	8	0.952	0.952	0.983	15
22	1	1	0.956	0.967	0.944	14
23	0.980	5	0.861	0.751	1	20
24	1	1	1	1	1	1

ally, the inclusion of intermediate outputs, often overlooked in traditional models, acts as logical constraints that further improve the model's discriminative ability. Another strength of our approach is its ability to simultaneously reduce input consumption and increase output production, reinforcing its effectiveness in assessing unit efficiency.

Identify sources of inefficiency:

The results from our model offer a deeper understanding of the sources of inefficiency in branches compared to traditional models. For example, Table 3 shows that branches 20 and 23 are deemed technically inefficient by both models. However, while the black-box model does not pinpoint the source of inefficiency, our model reveals that the inefficiency stems from the first level (production). This insight suggests that managers at these branches should focus on improving performance at the upper level, reducing costs, and attracting more deposits. Furthermore, applying our model enables a detailed evaluation of bank branches' performance across both production and intermediation levels. The results indicate which branches, at the production level, could attract more deposits by reducing costs like human resource expenses, and which branches, at the intermediation level, have the potential to increase their income. Analyzing both functional levels provide actionable insights into how branches can improve by optimizing both

costs and income. A key innovation of our model is that a unit is only considered efficient if both its upper and lower levels are efficient. This criterion emphasizes the need to address deficiencies at all levels to achieve optimal performance, a perspective often overlooked in previous studies.

Managerial guidelines:

Based on the findings of this case study, decision-makers can consider the following recommendations to improve branch efficiency:

Targeted allocation of input resources at upper and lower levels: Using insights from the bi-level approach, managers can allocate resources more effectively. Addressing inefficiencies at the resource absorption level can lead to optimized input usage and improved cost management. For instance, reallocating resources from less efficient branches to those that are more effective at marketing and intermediation can enhance overall operational efficiency.

Modifying strategic plans to attract more revenue: This could include revising operational processes, improving customer engagement, using new technologies to streamline financial processes, or creating programs to attract more customers and increase non-interest income. Strategic changes like these can help improve both non-interest income and overall branch performance.

In summary, these findings highlight the superiority of our approach in accurately classifying and evaluating units, uncovering inefficiencies often overlooked by traditional black-box models. The comprehensive evaluation offered by our model underscores the importance of adopting advanced DEA techniques for more accurate performance assessment in the banking sector. By enhancing discrimination and considering critical structural aspects, our approach provides a robust framework for evaluating and improving the efficiency of organizations, particularly those with hierarchical structures.

Case study 2. A case study of US Commercial Banks

To demonstrate the capabilities and practical utility of our proposed model, we apply it to analyze US commercial banks, a topic previously investigated by Liang, Cook, and Zhu [18]. Their study assesses the performance of these banks using a two-stage production process that focuses on profitability and marketability (see Figure 4). They assumed that all outputs from the first stage serve as inputs to the second stage, implying that the upper level does not generate independent outputs (i.e., Y_L=0). Additionally, the inputs at the lower level are solely intermediate outputs of the leader, without utilizing independent inputs (i.e., X_F=0). They introduced a non-cooperative model for two scenarios: One where the first stage acts as the leader and another where the second stage assumes leadership.

We focus on the first scenario to compare our model with theirs. Here, the first stage is the leader, with inputs including the number of employees, assets ($millions), and equity ($million), while the second stage involves three outputs: Market value ($millions), earnings per share ($),

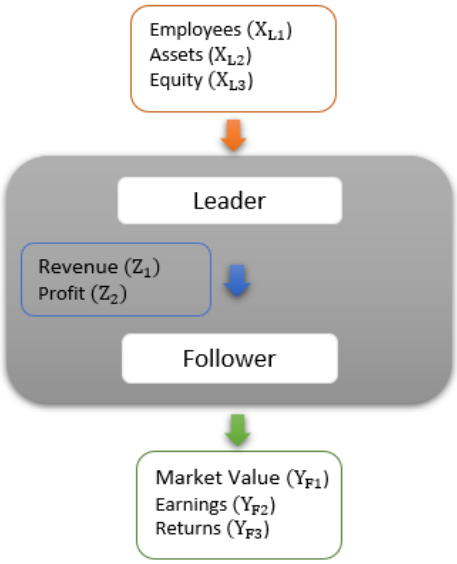


Figure 4: Bi-level structure in the case study of US commercial banks.

and returns to investors (%). Additionally, there are two intermediate measures between the stages: Profit ($millions) and revenue ($millions). The data statistics are presented in Table 5.

Table 5: Data statistics of US commercial bank.

Data	Min	Max	Ave	STD
Employees (x_{L1})	4900.00	95288.00	28945.70	20872.11
Assets (x_{L2})	13228.90	256853.00	85019.90	64686.99
Equity (x_{L3})	1265.10	20222.00	6348.16	4717.13
Revenue (z_1)	2565.40	31690.00	7865.10	6418.33
Profit (z_2)	215.00	3464.00	942.30	732.58
Market Value (y_{F1})	3218.00	33221.70	10888.15	6820.65
Earnings (y_{F2})	1.19	20.37	4.88	3.60
Returns (y_{F3})	28.30	108.50	58.37	17.35

Table 6 compares the TE results obtained from the black-box model, Liang's model, and our proposed model. It's important to note that in the conventional model, DEA scores for the overall process are determined by considering employees, assets, and equity as inputs, and market value, earnings, and returns as outputs.

Table 6 illustrates the rankings of TE scores based on our new model, the black-box model, and Liang's model, revealing non-identical rankings among them. The black-box model recognizes five units (5, 16, 24, 29, and 30) as efficient units. However, our proposed model and Liang's model identify units 16 and 30 as efficient and rank them first. Following these units, our

Table 6: TE results comparison between our model, Liang's model, and the black-box model.

Bank	Black-box TE	Ranking	Liang 's TE	Ranking	Proposed TE	TE_L	TE_F	Ranking
1	0.749	14	0.449	18	0.808	0.912	0.654	16
2	0.622	25	0.363	28	0.730	0.745	0.705	27
3	0.694	18	0.422	22	0.779	0.814	0.724	22
4	0.762	12	0.427	21	0.813	0.900	0.686	14
5	1	1	0.606	10	0.891	1	0.756	7
6	0.725	17	0.418	23	0.787	0.816	0.742	21
7	0.688	20	0.394	27	0.765	0.796	0.714	25
8	0.738	15	0.475	16	0.793	0.798	0.785	20
9	0.681	21	0.449	19	0.794	0.800	0.786	19
10	0.668	22	0.339	29	0.639	0.470	1	29
11	0.691	19	0.416	25	0.722	0.642	0.865	28
12	0.629	24	0.454	17	0.804	0.802	0.806	18
13	0.733	16	0.446	20	0.732	0.578	1	26
14	0.652	23	0.417	24	0.774	0.789	0.750	23
15	0.891	6	0.642	8	0.808	0.730	0.925	15
16	1	1	1	1	1	1	1	1
17	0.985	2	0.714	6	0.892	0.922	0.855	6
18	0.938	4	0.797	4	0.942	0.890	1	3
19	0.797	9	0.584	12	0.854	0.827	0.891	10
20	0.854	7	0.592	11	0.834	0.797	0.886	11
21	0.796	10	0.567	13	0.830	0.754	0.938	12
22	0.759	13	0.520	15	0.823	0.775	0.892	13
23	0.575	26	0.398	26	0.771	0.760	0.788	24
24	1	1	0.826	2	0.918	0.855	0.992	5
25	0.954	3	0.740	5	0.919	0.850	1	4
26	0.770	11	0.531	14	0.804	0.733	0.910	17
27	0.895	5	0.655	7	0.882	0.789	1	8
28	0.833	8	0.620	9	0.868	0.852	0.889	9
29	1	1	0.809	3	0.945	0.895	1	2
30	1	1	1	1	1	1	1	1

proposed model ranks DMUs 29, 18, and 25 as the top units, respectively, whereas Liang's model ranks units 24, 29, and 18 after units 16 and 30. Notably, significant disparities in rankings are observed for certain DMUs, including 4, 5, 8, 11, 13, 15 and 24, between our proposed model and Liang's model.

This comparison underlines the applicability of our proposed model for evaluating two-stage structures. The findings demonstrate improved separability relative to the traditional model and reveal differences in the ranking of inefficient units when compared to the Liang model. It is important to note that the Liang model is input-oriented and focuses on reducing input indicators, while our approach employs a directed distance function that simultaneously considers input reduction and output enhancement.

This feature makes our efficiency scores higher than Liang's efficiency scores because our directed function approach seeks a more accessible and appropriate set of benchmarks for inefficient units. Another usefulness of our model is that it can be generalized to evaluate the performance

Table 7: The PE results of our proposed model and black-box model.

Bank	Black-box PE	Ranking	Proposed PE	PE_L	PE_F	Ranking
1	1	1	1	1	1	1
2	0.933	6	0.956	0.905	1	4
3	0.909	9	0.933	0.886	0.973	9
4	0.929	8	0.929	0.962	0.901	10
5	1	1	0.976	1	0.956	2
6	0.870	10	0.861	0.815	0.898	18
7	0.810	15	0.851	0.843	0.859	19
8	0.937	5	0.935	0.869	0.992	8
9	0.969	2	0.941	0.910	0.966	5
10	0.695	20	0.706	0.620	1	27
11	0.784	18	0.720	0.563	0.933	26
12	0.931	7	0.925	0.901	0.943	11
13	0.818	13	0.800	0.664	1	24
14	0.730	19	0.820	0.842	0.794	23
15	1	1	0.900	0.803	1	14
16	1	1	1	1	1	1
17	1	1	0.940	1	0.874	6
18	1	1	0.962	0.917	1	3
19	0.957	3	0.913	0.899	0.924	13
20	0.858	11	0.880	0.874	0.888	17
21	0.814	14	0.840	0.771	0.942	20
22	0.788	17	0.828	0.778	0.907	22
23	0.588	21	0.797	0.787	0.816	25
24	1	1	0.936	0.888	0.992	7
25	1	1	0.921	0.862	1	12
26	0.804	16	0.837	0.774	0.953	21
27	0.951	4	0.899	0.816	1	15
28	0.848	12	0.884	0.843	0.967	16
29	1	1	1	1	1	1
30	1	1	1	1	1	1

of organizations that have undesirable outputs. By selecting an appropriate direction and identifying effective criteria, it becomes possible to assess the performance of these organizations in a more accurate and realistic manner.

Liang's model assumes CRS, whereas our proposed model can be designed for VRS. The efficiency results under the VRS assumption using our PE model and the black-box PE model are reported in Table 7.

The results reveal that the black-box PE identifies 10 units as purely efficient, while the bi-level model designates only 4 units as such. This demonstrates the enhanced resolution of our proposed model compared to the traditional model.

4 Conclusions

This paper introduced a new DEA model specifically designed to evaluate bi-level units, characterized by a leader at the upper level and followers at the lower level. By incorporating the interactions between different hierarchical levels and intermediate outputs, our model provided a detailed and comprehensive evaluation of DMU performance in hierarchical settings. It established a logical evaluation framework based on the leader-follower dynamic; an aspect previously underexplored in research. The proposed model can be easily extended for evaluating various hierarchical setups.

Another distinguishing feature of our model is its evaluation criterion based on simultaneously reducing input consumption and increasing output production. This criterion enhances the precision and reliability of performance assessments compared to traditional methods, marking a substantial advancement in achieving precise assessments within hierarchical organizations.

We applied our model to assess the efficiency of banking units from two different performance aspects and examined the effect of each aspect in evaluating the overall branch performance. The results showed that our approach provides a more accurate and meaningful assessment of branches compared to traditional models by creating more discrimination power in branch ranking. We found that conventional DEA models often fail to evaluate all functional aspects. The conventional DEA models integrate all functional aspects of a system as the sum of its initial inputs and final outputs, which leads to ignoring some intermediate activities. This oversight leads to ambiguous performance results and an inability to identify inefficiencies. Furthermore, our model demonstrated its effectiveness and comprehensiveness in evaluating units with two-stage processes. Consequently, our approach offers managers a deeper insight for decision-making within hierarchical structures. Overall, our findings represent a meaningful advancement in the field of performance evaluation, for complex hierarchical organizations.

Future research could further develop this approach by applying it across diverse industries and organizational structures. Additionally, addressing scenarios involving undesirable outputs or data uncertainty would be valuable directions for extending the model's applicability.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Ahmad, A.M., Khurizan, N.S., Awang, N., Taib, M.F., Yahya, M.F., and Shuhut, R. *Efficiency assessment through hierarchical network data envelopment analysis approach: perspectives of university faculties*, Malays. J. Math. Sci., 19(2) (2025) 707–726.

[2] Chambers, R.G., Chung, Y., and Färe, R. *Benefit and distance functions*, J. Econ. Theory, 70(2) (1996) 407–419.

[3] Cook, W.D., Chai, D., Doyle, J. and Green, R.H. *Hierarchies and groups in DEA*, J. Prod. Anal., 10(2) (1998) 177–198.

[4] Cook, W.D., and Green, R.H. *Evaluating power plant efficiency: A hierarchical model*, Comput. Oper. Res., 32(4) (2005) 813–823.

[5] Cook, W.D., Zhu, J., Bi, G. and Yang, F. *Network DEA: additive efficiency decomposition*, Eur. J. Oper. Res., 207(1) (2010) 1122–1129.

[6] Deville, A., Ferrier, G.D., and Leleu, H. *Measuring the performance of hierarchical organizations: An application to bank efficiency at the regional and branch levels*, Manage. Account. Res., 25(1) (2014) 30–44.

[7] Du, J., Zhu, J., Cook, W.D. and Huo, J. *DEA models for parallel systems: Game-theoretic approaches*, Asia-Pac. J. Oper. Res., 32(2) (2015) 1–22.

[8] Gan, G., Lee, H.S., Lee, L., Wang, X. and Wang, Q. *Network hierarchical DEA with an application to international shipping industry in Taiwan*, J. Oper. Res. Soc., 71(6) (2020) 991–1002.

[9] Ghasemi, N., Najafi, E., Lotfi, F.H. and Sobhani, F.M.*Assessing the performance of organizations with the hierarchical structure using data envelopment analysis: An efficiency analysis of Farhangian University*, Measurement, 156(1) (2020) 107609.

[10] Ghasemi, N., Najafi, E., Lotfi, F.H. and Sobhani, F.M. *Measuring efficiency of organizations with hierarchical structures using DEA and game theory*, J. Ind. Eng. Manage. Stud., 8(2) (2021) 1–25.

[11] Habibpoor, M., Alirezaee, M.R., and Rashidinia, J. *Developing a bi-level structure for evaluation of regional bank branch managers focusing on their consumption*, Croat. Oper. Res. Rev., 13(1) (2022) 137–147.

[12] Hendrawan, B.M., Purwanggono, B., and Suliantoro, H. *Efficiency measurements of Diponegoro university faculty using the data envelopment analysis hierarchical network model*, IOP Conf. Ser. Mater. Sci. Eng., 1072(1) (2021) 012021.

[13] Kao, C. *Efficiency measurement for hierarchical network systems*, Omega, 51(1) (2015) 121–127.

[14] Kashim, R., Kasim, M.M., and Abd Rahman, R. *Measuring efficiency of a university faculty using a hierarchical network data envelopment analysis model*, J. Inf. Commun. Technol., 17(4) (2018) 569–585.

[15] Kremantzis, M.D., Beullens, P., Kyrgiakos, L.S. and Klein, J. *Measurement and evaluation of multi-function parallel network hierarchical DEA systems*, Socio-Econ. Plann. Sci., 84 (2022) 101428.

[16] Kweh, Q.L., Ting, I.W.K., Asif, J. and Lu, W.M.*Environmental, social, and governance and hierarchical network data envelopment analysis firm efficiencies of Japan Airlines' supply chain*, Bus. Strateg. Environ., 33(7) (2024) 7060–7076.

[17] Li, W., Cook, W.D., Li, Z. and Zhu, J. *Efficiency measurement for hierarchical situations*, J. Oper. Res. Soc., 72(3) (2020) 654–662.

[18] Liang, L., Cook, W.D., and Zhu, J. *DEA models for two-stage processes: Game approach and efficiency decomposition*, Naval Res. Logist., 55(7) (2008) 643–652.

[19] Luban, F. *Measuring efficiency of a hierarchical organization with fuzzy DEA method*, Econ. Ser. Manag., 12(1) (2009) 87–97.

[20] Muvingi, J., Peer, A.I., and Lotfi, F.H. *Hierarchical non-Archimedean DEA models: Application on mobile money agents locations in the city of Harare*, Int. J. Data Sci., 5(3) (2020) 181–202.

[21] Muvingi, J., Peer, A.I., Jablonský, J., Azizi, H. and Lotfi, F.H. *Hierarchical fuzzy DEA model with double frontiers combined with TOPSIS technique: application on mobile money agents locations*, OPSEARCH, 61(3) (2024) 1154–1191.

[22] Omrani, H., Mohammadi, S., and Emrouznejad, A. *A bi-level multi-objective data envelopment analysis model for estimating profit and operational efficiency of bank branches*, RAIRO Oper. Res., 53(5) (2019) 1633–1648.

[23] Pachar, N., Gupta, A., and Jha, P.C. *Bilevel programming DEA approach for efficiency evaluation: A case study of Indian electronics retail stores*, Yugosl. J. Oper. Res., 30(1) (2020) 461–481.

[24] Ratner, S.V., Shaposhnikov, A.M., and Lychev, A.V. *Network DEA and its applications (2017–2022): A systematic literature review*, Mathematics, 11(9) (2023) 2141.

[25] Serkania, E.S., Lotfi, F.H. , Najafi, E., and Namin, M.A. *Efficiency measurement for hierarchical network systems using network DEA and intuitionistic fuzzy ANP*, Scientia Iranica, 29(4) (2022) 2252–2269.

[26] Shafiee, M., and Saleh, H. *Goal programming method for solving tri-level data envelopment analysis*, in Analytical Decision Making and Data Envelopment Analysis (Edalatpanah, S., Hosseinzadeh Lotfi, F., Kerstens, K., and Wanke, P., eds.), Springer, Infosys Sci. Found. Ser., (2024).

[27] Walheer, B., and Hudik, M. *Reallocation of resources in multidivisional firms: A nonparametric approach*, Int. J. Prod. Econ., 214(1) (2019) 196–205.

[28] Wang, K., Xian, Y., Lee, C.H., Wei, Y., and Huang, Z. *On selecting directions for directional distance functions in a non-parametric framework: a review*, Ann. Oper. Res., 278(1) (2019) 43–76.

[29] Wu, D.D. *Bi-level programming data envelopment analysis with constrained resource*, Eur. J. Oper. Res., 207(1) (2010) 856–864.

[30] Xiao, S., Kremantzis, M.D., Kyrgiakos, L.S., Vlontzos, G. and Pardalos, P.M. *Embracing fairness within a cross-efficiency hierarchical network DEA system*, Oper. Res. J., 24(1) (2024) 4.

[31] Yao, L., Shuai, Y., and Chen, X. *Regional water system vulnerability evaluation: a bi-level DEA approach*, J. Hydrol., 589 (2020) 125160.

[32] Zhang, L., and Chen, K. *Hierarchical network systems: An application to high-technology industry in China*, Omega, (2017) 1–14.

[33] Zhou, X., Luo, R., Tu, Y., Lev, B. and Pedrycz, W. *Data envelopment analysis for bi-level systems with multiple followers*, Omega, 77 (2018) 180–188.

[34] Zhu, W., Lin, H., Yu, Y. and Ghiyasi, M. *Performance measurement of Taiwan hotels with hierarchical network DEA with shared inputs*, Int. J. Appl. Decis. Sci., 16(4) (2023) 474–507.

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 1, 2026, pp 170–187
https://doi.org/10.22067/ijnao.2025.94352.1679
https://ijnao.um.ac.ir/

CC BY NC

Research Article

Qualitative and Numerical Analysis of a Time-Fractional Ebola Model in Population Dynamics

G. Saini, B. Ghosh*, and S. Chand

Abstract

Mathematical modeling of infectious diseases plays a vital role in understanding population dynamics and guiding control strategies. In this study, we formulate and analyze time-fractional Ebola virus model incorporating the Caputo derivative of order $\sigma \in (0, 1)$ to capture memory and nonlocal effects in disease transmission. This formulation is important as it extends classical integer-order models to more realistically describe the persistence and delayed response of infection dynamics. A comprehensive qualitative analysis is presented, establishing the existence, uniqueness, and Hyers–Ulam stability of the solution of the considered model. To obtain the numerical approximation, a robust finite difference scheme based on the classical

*Corresponding author

Received 10 July 2025; revised 18 October 2025; accepted 15 November 2025

Gaurav Saini
Center for Data Science, Department of Computer Science and Engineering, Siksha 'O' Anusandhan (Deemed to be University), India. e-mail: gauravsaini@soa.ac.in

Bappa Ghosh
Center for Artificial Intelligence and Machine Learning, Department of Computer Science and Engineering, Siksha 'O' Anusandhan (Deemed to be University), India. e-mail: bg12.ghosh@gmail.com

Sunita Chand
Department of Mathematics, Siksha 'O' Anusandhan (Deemed to be University), India. e-mail: sunitachand@soa.ac.in

How to cite this article
Saini, G., Ghosh, B. and Chand, S., Qualitative and Numerical Analysis of a Time-Fractional Ebola Model in Population Dynamics . *Iran. J. Numer. Anal. Optim.*, 2026; 16(1): 170-187. https://doi.org/10.22067/ijnao.2025.94352.1679

L1 discretization is developed. The resulting nonlinear system is efficiently solved using the Newton–Raphson iterative technique. The numerical results demonstrate that fractional-order models provide enhanced flexibility and accuracy compared to traditional integer-order formulations. Specifically, variation in the fractional order significantly influences the rate of infection and recovery, highlighting the crucial role of memory effects in epidemic modeling. These findings confirm that incorporating fractional derivatives offers improved insight into Ebola virus dynamics and can support more effective intervention and treatment strategies.

AMS subject classifications (2020): Primary 26A33, 37M05 ; Secondary 92D25, 65L05.

Keywords: Ebola virus, fractional order model, Caputo derivative, L1 technique, Qualitative analysis.

1 Introduction

Fractional calculus generalizes classical calculus by extending the concept of derivatives and integrals to noninteger (arbitrary) orders. This framework is particularly well-suited for describing systems with memory, hereditary effects, and long-range temporal dependencies. As a result, fractional differential equations (FDEs) have emerged as powerful tools in modeling a wide range of complex phenomena across various disciplines, including viscoelastic materials [23], anomalous diffusion processes [17], quantitative finance, and population dynamics [10, 9]. For comprehensive discussions on the theoretical foundations and applications of fractional calculus, one may refer to [18, 5] and references therein. Although FDEs often involve increased computational complexity due to their nonlocal nature, they offer a more accurate representation of systems where classical integer-order models fall short. The growing interest in fractional modeling reflects its ability to bridge the gap between empirical observations and theoretical predictions, making it a vital component of modern applied mathematics and scientific computing.

Population dynamics models provide mathematical frameworks to analyze how population sizes evolve under the influence of biological interactions, environmental conditions, and disease transmission [11]. These models are instrumental in fields such as ecology, epidemiology, and resource management, where they aid in understanding the spread of infectious diseases and formulating effective control and intervention strategies [12, 22]. Ebola virus, a highly virulent pathogen belonging to the Filoviridae family, causes Ebola virus disease (EVD), a severe and often fatal illness in humans and nonhuman primates. First identified in 1976 during simultaneous outbreaks in Sudan and the Democratic Republic of Congo near the Ebola River, the virus has since led to several deadly epidemics across sub-Saharan Africa. Ebola is primarily transmitted

through direct contact with the blood, secretions, organs, or other bodily fluids of infected individuals or contaminated materials. The incubation period ranges from 2 to 21 days, and symptoms include sudden onset of fever, fatigue, muscle pain, vomiting, diarrhea, and, often, internal and external bleeding. Currently, treatment for EVD is primarily supportive and includes rehydration, oxygen supplementation, and the management of clinical complications. In recent years, experimental therapies and vaccines have been introduced, among which the rVSV-ZEBOV vaccine has demonstrated notable efficacy in clinical trials. Mathematical models of Ebola virus transmission, often based on compartmental structures like the SEIR framework, are crucial tools for understanding outbreak dynamics and informing strategies for intervention, quarantine, vaccination, and public health preparedness.

The classical Ebola virus model for $t > 0$ is given as follows:

$$\begin{cases} \partial_t S(t) = -\alpha S(t)I(t) + \beta R(t) - \gamma N, \\ \partial_t I(t) = \alpha S(t)I(t) - \epsilon I(t) - \delta I(t), \\ \partial_t R(t) = \delta I(t) - \beta R(t), \\ \partial_t D(t) = \epsilon I(t) + \gamma N, \end{cases} \tag{1}$$

subject to the initial conditions

$$S(0) = S_0,\ I(0) = I_0,\ R(0) = R_0,\ D(0) = D_0\ (see\ Table\ 1). \tag{2}$$

Table 1: Description of symbols used in the model

Symbol	**Description**
$\frac{d}{dt} \equiv \partial_t$	Derivative with respect to time t
$S(t)$	Susceptible population at time t
$I(t)$	Infected population at time t
$R(t)$	Recovered population at time t
$D(t)$	Cumulative death population at time t
N	Total population (assumed constant)
α	Infection rate
β	Rate at which recovered become susceptible
γ	Natural death rate
ϵ	Ebola-induced death rate
δ	Recovery rate

The classical model effectively represents the basic transmission mechanisms. However, it overlooks the memory and hereditary effects that influence disease evolution. To address this limitation, we extend the system to a fractional-order framework using the Caputo derivative. In this work, we study the following fractional-order Ebola virus model (FEVM) for $t > 0$:

$$\begin{cases} \mathcal{D}_t^\sigma S(t) = -\alpha S(t)I(t) + \beta R(t) - \gamma N, \\ \mathcal{D}_t^\sigma I(t) = \alpha S(t)I(t) - \epsilon I(t) - \delta I(t), \\ \mathcal{D}_t^\sigma R(t) = \delta I(t) - \beta R(t), \\ \mathcal{D}_t^\sigma D(t) = \epsilon I(t) + \gamma N, \end{cases} \tag{3}$$

subject to the initial conditions mentioned in (2). $\mathcal{D}_t^\sigma$ denotes the Caputo derivative of order $\sigma \in (0, 1)$, defined by [20]

$$\mathcal{D}_t^\sigma \phi(t) = \int_0^t \frac{(t-z)^{-\sigma}\, \partial_z \phi(z)}{\Gamma(1-\sigma)} dz.$$

Mathematical modeling has significantly contributed to understanding the transmission dynamics of the Ebola virus, evaluating intervention strategies, and predicting outbreak scenarios. Due to the nonlinear and nonlocal characteristics of fractional-order systems, finding exact solutions is typically challenging [13]. Therefore, numerous numerical and semi-analytical techniques have been developed to approximate solutions effectively [15]. Selected contributions include: Pontryagin's maximum principle is used by Bonyah, Badu, and Asiedu-Addo [7] to derive the necessary conditions for optimality in a compartmental Ebola model. These conditions were implemented numerically via a fourth-order Runge–Kutta method within an iterative forward-backward framework. A dual modeling framework, deterministic and stochastic, was explored by Berge et al. [6]. The deterministic system is handled using a standard Runge–Kutta method, while the stochastic variant employs a continuous-time Markov chain to capture the random nature of disease transmission. To model fractional dynamics, Singh [21] developed an iterative numerical scheme based on the Grünwald–Letnikov fractional derivative. In another study, Ahmad, Rafiq, and Abbas [4] investigated the SVEIR model using the classical fourth-order Runge–Kutta method and a nonstandard finite difference scheme. Srivastava et al. developed three Ebola virus models based on the fractal–fractional derivative with power-law, exponential, and Mittag–Leffler kernels. They applied Lagrange polynomial interpolation to construct numerical schemes for each model and compared their behaviors under varying parameters. The conformable fractional differential transform method, a power-series-based technique, was applied by Nazir et al. [19] to obtain approximate solutions of a nonlinear Ebola model. To solve a Caputo-type fractional model, Derakhshan [8] proposed the Sinc–Legendre collocation method, demonstrating high accuracy and numerical stability in fractional-order equations. Lagrange

polynomial interpolation was also used by Adu et al. [3] to approximate the trajectory of a fractional Ebola system, emphasizing the method's capability to capture nonlocal dynamics. In a separate approach, Abbas et al. [2] incorporated the Grünwald–Letnikov derivative into the classical fourth-order Runge–Kutta framework to solve a fractional-order Ebola model. Abah et al. [1] utilized the homotopy perturbation method to analyze a nonlinear Ebola model incorporating mitigation strategies such as quarantine and public awareness.

Motivated by the existing literature, we have conducted a qualitative study on an Ebola virus model involving an arbitrary-order derivative, focusing on its analytical solution existence, uniqueness, and stability. Additionally, we propose an efficient iterative difference scheme based on the standard L1 technique to approximate the solution of the nonlinear system (3). Theoretical analysis and illustrative test examples validate the accuracy and efficiency of the proposed method. The structure of the paper is as follows: The existence of the analytical solution of (3) is discussed in Section 2, while Section 3 establishes its uniqueness. Stability results for (3) are derived in Section 3.1. The construction of the numerical scheme for solving (3) is provided in Section 4. Section 5 focuses on numerical simulations, and concluding remarks are presented in Section 6.

2 Existence of the solution

In this section, we analyze the existence of a solution for the FEVM using fixed-point techniques. Applying the Riemann–Liouville fractional integral operator to both sides of system (3), we obtain the equivalent integral formulation:

$$\begin{cases} S(t) = S_0 + \displaystyle\int_0^t \frac{G_1(z, S(z))}{\Gamma(\sigma)(t-z)^{\sigma-1}}\, dz, \quad I(t) = I_0 + \int_0^t \frac{G_2(z, I(z))}{\Gamma(\sigma)(t-z)^{\sigma-1}}\, dz, \\ R(t) = R_0 + \displaystyle\int_0^t \frac{G_3(z, R(z))}{\Gamma(\sigma)(t-z)^{\sigma-1}}\, dz, \quad D(t) = D_0 + \int_0^t \frac{G_4(z, D(z))}{\Gamma(\sigma)(t-z)^{\sigma-1}}\, dz, \end{cases}$$

where S_0, I_0, R_0, D_0 are the given initial conditions, and the nonlinear functions $G_i(t, \cdot)$ are defined as

$$\begin{cases} G_1(t, S(t)) = -\alpha S(t) I(t) + \beta R(t) - \gamma N, \\ G_2(t, I(t)) = \alpha S(t) I(t) - \epsilon I(t) - \delta I(t), \\ G_3(t, R(t)) = \delta I(t) - \beta R(t), \quad G_4(t, D(t)) = \epsilon I(t) + \gamma N. \end{cases}$$

Assumption (C_1) : We assume that the functions $S(z), I(z), R(z), D(z), S^1(z), I^1(z), R^1(z), D^1(z) \in L^1[0,1]$, and that the following bounds hold for some positive real constants q_1, q_2, q_3:

$$\|I\| \leq q_1, \quad \|S\| \leq q_2, \quad \|R\| \leq q_3.$$

Theorem 1. The kernels G_1, G_2, G_3, G_4 defined in the FEVM satisfy the Lipschitz conditions under the assumption (C_1), provided that $0 \leq \eta_i < 1$ where $i = 1, 2, 3, 4$.

Proof. First, we prove that the kernel $G_1(t, S)$ satisfies the Lipschitz condition. We compute

$$\begin{aligned}\|G_1(t, S^1) - G_1(t, S^2)\| &= \| - \alpha S^1 I + \beta R - \gamma N - (-\alpha S^2 I + \beta R - \gamma N)\| \\ &= \|\alpha I(S^2 - S^1)\| \leq \alpha q_1 \|S^2 - S^1\| \leq \eta_1 \|S^2 - S^1\|,\end{aligned}$$

where $\eta_1 = \alpha q_1$. Thus, G_1 satisfies the Lipschitz condition provided $\eta_1 < 1$. Next, we consider the kernel $G_2(t, I)$

$$\begin{aligned}\|G_2(t, I^1) - G_2(t, I^2)\| &= \|\alpha S I^1 - \epsilon I^1 - \delta I^1 - (\alpha S I^2 - \epsilon I^2 - \delta I^2)\| \\ &= \|(\alpha S - \delta - \epsilon)(I^1 - I^2)\| \leq (\alpha q_2 + \delta + \epsilon)\|I^1 - I^2\| \\ &\leq \eta_2 \|I^1 - I^2\|,\end{aligned}$$

where $\eta_2 = \alpha q_2 + \epsilon + \delta$. Now, for the kernel $G_3(t, R)$

$$\begin{aligned}\|G_3(t, R^1) - G_3(t, R^2)\| &= \|\delta I - \beta R^1 - (\delta I - \beta R^2)\| \leq \beta \|R^1 - R^2\| \\ &\leq \eta_3 \|R^1 - R^2\|,\end{aligned}$$

where $\eta_3 = \beta$. Finally, for $G_4(t, D)$, we have

$$\|G_4(t, D^1) - G_4(t, D^2)\| = \|\epsilon I + \gamma N - (\epsilon I + \gamma N)\| = 0.$$

Thus G_4 is constant with respect to D, and hence Lipschitz condition with $\eta_4 = 0$. Therefore, all kernels G_i satisfy the Lipschitz condition with constants $\eta_i < 1$ for $i = 1, 2, 3, 4$, which completes the proof. □

Now, we reformulate the system using the kernels G_i, assuming zero initial conditions $S(0) = I(0) = R(0) = D(0) = 0$. Then, the system becomes

$$\begin{cases} S(t) = \displaystyle\int_0^t \frac{G_1(z, S(z))}{\Gamma(\sigma)(t-z)^{1-\sigma}}\, dz, & I(t) = \displaystyle\int_0^t \frac{G_2(z, I(z))}{\Gamma(\sigma)(t-z)^{1-\sigma}}\, dz, \\ R(t) = \displaystyle\int_0^t \frac{G_3(z, R(z))}{\Gamma(\sigma)(t-z)^{1-\sigma}}\, dz, & D(t) = \displaystyle\int_0^t \frac{G_4(z, D(z))}{\Gamma(\sigma)(t-z)^{1-\sigma}}\, dz. \end{cases}$$

Let us define the following recursive sequences to approximate the solution:

$$\begin{cases} S_k(t) = \displaystyle\int_0^t \frac{G_1(z, S_{k-1}(z))}{\Gamma(\sigma)(t-z)^{1-\sigma}}\, dz, & I_k(t) = \displaystyle\int_0^t \frac{G_2(z, I_{k-1}(z))}{\Gamma(\sigma)(t-z)^{1-\sigma}}\, dz, \\ R_k(t) = \displaystyle\int_0^t \frac{G_3(z, R_{k-1}(z))}{\Gamma(\sigma)(t-z)^{1-\sigma}}\, dz, & D_k(t) = \displaystyle\int_0^t \frac{G_4(z, D_{k-1}(z))}{\Gamma(\sigma)(t-z)^{1-\sigma}}\, dz, \end{cases} \tag{4}$$

where $S_0(t), I_0(t), R_0(t), D_0(t) \equiv 0$, initialized by the given initial conditions.

Theorem 2. The FEVM (3) admits a solution if the following condition holds:

$$\eta = \max\{\eta_1, \eta_2, \eta_3, \eta_4\} < 1.$$

Proof. We begin by estimating the difference between two successive approximations of the susceptible population as follows:

$$\begin{aligned} \|DS_{k+1}(t)\| &= \|S_{k+1}(t) - S_k(t)\| \\ &= \left\| \int_0^t \frac{[G_1(z, S_k(z)) - G_1(z, S_{k-1}(z))]\, dz}{\Gamma(\sigma)(t-z)^{1-\sigma}} \right\| \\ &\leq \int_0^t \frac{\|G_1(z, S_k(z)) - G_1(z, S_{k-1}(z))\|\, dz}{\Gamma(\sigma)(t-z)^{1-\sigma}}. \end{aligned}$$

Similarly, for the infected population,

$$\begin{aligned} \|DI_{k+1}(t)\| &= \|I_{k+1}(t) - I_k(t)\| \\ &= \left\| \int_0^t \frac{[G_2(z, I_k(z)) - G_2(z, I_{k-1}(z))]\, dz}{\Gamma(\sigma)(t-z)^{1-\sigma}} \right\| \end{aligned}$$

$$\leq \int_0^t \frac{\|G_2(z, I_k(z)) - G_2(z, I_{k-1}(z))\|\, dz}{\Gamma(\sigma)(t-z)^{1-\sigma}}.$$

For the recovered population,

$$\begin{aligned}\|DR_{k+1}(t)\| &= \|R_{k+1}(t) - R_k(t)\| \\ &= \left\| \int_0^t \frac{[G_3(z, R_k(z)) - G_3(z, R_{k-1}(z))]\, dz}{\Gamma(\sigma)(t-z)^{1-\sigma}} \right\| \\ &\leq \int_0^t \frac{\|G_3(z, R_k(z)) - G_3(z, R_{k-1}(z))\|\, dz}{\Gamma(\sigma)(t-z)^{1-\sigma}}.\end{aligned}$$

For the death compartment,

$$\|DD_{k+1}(t)\| = \|D_{k+1}(t) - D_k(t)\| = \left\| \int_0^t \frac{[G_4(z, D_k(z)) - G_4(z, D_{k-1}(z))]}{\Gamma(\sigma)(t-z)^{1-\sigma}} dz \right\| = 0.$$

Now, we show the convergence of the sequences using the recursive bound. For instance,

$$\|S_{n+1}(t) - S(t)\| = \left\| \int_0^t \frac{[G_1(z, S_n(z)) - G_1(z, S(z))]}{\Gamma(\sigma)(t-z)^{1-\sigma}} dz \right\| \leq \frac{\eta_1 t^\sigma}{\Gamma(1+\sigma)} \|S_n - S\|.$$

Iterating the above inequality leads to

$$\|S_{n+1}(t) - S(t)\| \leq \left(\frac{\eta_1 t^\sigma}{\Gamma(1+\sigma)} \right)^n \|S_1 - S\|.$$

Similarly, we get

$$\begin{aligned}\|I_{n+1}(t) - I(t)\| &\leq \left(\frac{\eta_2 t^\sigma}{\Gamma(1+\sigma)} \right)^n \|I_1 - I\|, \\ \|R_{n+1}(t) - R(t)\| &\leq \left(\frac{\eta_3 t^\sigma}{\Gamma(1+\sigma)} \right)^n \|R_1 - R\|, \\ \|D_{n+1}(t) - D(t)\| &= 0.\end{aligned}$$

Since $\eta = \max\{\eta_1, \eta_2, \eta_3, \eta_4\} < 1$, it follows that all sequences converge uniformly. Hence the system admits a unique continuous solution. □

3 Uniqueness of the solution

Next, we demonstrate the uniqueness of solutions for the FEVM.

Theorem 3 (Uniqueness of solution)**.** The FEVM (3) possesses a unique solution if the following inequality is satisfied:

$$\frac{\eta_i}{\Gamma(\sigma)} < 1 \quad \text{for} \quad i = 1, 2, 3, 4.$$

Proof. Assume, to the contrary, that there exists a distinct solution to the system (3), denoted by $S^\star(t), I^\star(t), R^\star(t), D^\star(t)$. Then, both this solution and the original one must satisfy the corresponding integral formulations of the model

$$S(t) = \int_0^t \frac{G_1(z, S(z))dz}{\Gamma(\sigma)(t-z)^{1-\sigma}}, \qquad S^\star(t) = \int_0^t \frac{G_1(z, S^\star(z))dz}{\Gamma(\sigma)(t-z)^{1-\sigma}}.$$

Subtracting and taking norms, we get

$$\begin{aligned} \|S(t) - S^\star(t)\| &= \left\| \int_0^t \frac{[G_1(z, S(z)) - G_1(z, S^\star(z))]\, dz}{\Gamma(\sigma)(t-z)^{1-\sigma}} \right\| \\ &\leq \int_0^t \frac{\|G_1(z, S(z)) - G_1(z, S^\star(z))\|\, dz}{\Gamma(\sigma)(t-z)^{1-\sigma}} \leq \frac{\eta_1 t^\sigma}{\Gamma(1+\sigma)} \|S - S^\star\|. \end{aligned}$$

If $\frac{\eta_1 t^\sigma}{\Gamma(1+\sigma)} < 1$, then it follows that $\|S - S^\star\| = 0$. Hence $S(t) = S^\star(t)$ for $t \in [0, T]$. Similarly, for the infected compartment,

$$\begin{aligned} \|I(t) - I^\star(t)\| &\leq \int_0^t \frac{\|G_2(z, I(z)) - G_2(z, I^\star(z))\| dz}{\Gamma(\sigma)(t-z)^{1-\sigma}} \leq \frac{\eta_2 t^\sigma}{\Gamma(1+\sigma)} \|I - I^\star\| \\ &\Rightarrow I(t) = I^\star(t). \end{aligned}$$

For the recovered compartment,

$$\begin{aligned} \|R(t) - R^\star(t)\| &\leq \int_0^t \frac{\|G_3(z, R(z)) - G_3(z, R^\star(z))\| dz}{\Gamma(\sigma)(t-z)^{1-\sigma}} \leq \frac{\eta_3 t^\sigma}{\Gamma(1+\sigma)} \|R - R^\star\| \\ &\Rightarrow R(t) = R^\star(t). \end{aligned}$$

Finally, since G_4 is constant with respect to D, we have $\|D(t) - D^\star(t)\| = 0 \Rightarrow D(t) = D^\star(t)$.

This however means, all components of the solution coincide. This proves that the solution is unique. □

3.1 Stability of the solution

Definition 1 (Hyers–Ulam stability [14])**.** The fractional integral system described in (4) is said to be stable if there exist constants $\delta_i > 0$ for $i = 1, 2, 3, 4$, such that for every $\zeta_i > 0$, the following inequalities hold:

$$\left| S(z) - \int_0^t \frac{G_1(z, S(z))dz}{\Gamma(\sigma)(t-z)^{1-\sigma}} \right| \leq \zeta_1, \left| I(z) - \int_0^t \frac{G_2(z, I(z))dz}{\Gamma(\sigma)(t-z)^{1-\sigma}} \right| \leq \zeta_2,$$
$$\left| R(z) - \int_0^t \frac{G_3(z, R(z))dz}{\Gamma(\sigma)(t-z)^{1-\sigma}} \right| \leq \zeta_3, \left| D(z) - \int_0^t \frac{G_4(z, D(z))dz}{\Gamma(\sigma)(t-z)^{1-\sigma}} \right| \leq \zeta_4.$$

Then, there exist functions $\bar{S}(z), \bar{I}(z), \bar{R}(z), \bar{D}(z)$ satisfying the inequalities

$$\left| \bar{S}(z) - \int_0^t \frac{G_1(z, \bar{S}(z))dz}{\Gamma(\sigma)(t-z)^{1-\sigma}} \right| \leq \zeta_1, \left| \bar{I}(z) - \int_0^t \frac{G_2(z, \bar{I}(z))dz}{\Gamma(\sigma)(t-z)^{1-\sigma}} \right| \leq \zeta_2,$$
$$\left| \bar{R}(z) - \int_0^t \frac{G_3(z, \bar{R}(z))dz}{\Gamma(\sigma)(t-z)^{1-\sigma}} \right| \leq \zeta_3, \left| \bar{D}(z) - \int_0^t \frac{G_4(z, \bar{D}(z))dz}{\Gamma(\sigma)(t-z)^{1-\sigma}} \right| \leq \zeta_4,$$

such that

$$\left| S(z) - \bar{S}(z) \right| = \int_0^t \frac{\left| G_1(z, S(z)) - G_1(z, \bar{S}(z)) \right| dz}{\Gamma(\sigma)(t-z)^{1-\sigma}} \leq \frac{\eta_1 t^\sigma}{\Gamma(1+\sigma)} \left\| S - \bar{S} \right\|.$$

Let $\zeta_1^\star = \frac{t^\sigma}{\Gamma(1+\sigma)} \left\| S - \bar{S} \right\|$ and $\delta_1 = \eta_1$, then we have $\left| S(z) - \bar{S}(z) \right| \leq \zeta_1^\star \delta_1$.

Similarly, we obtain

$$\left| I(z) - \bar{I}(z) \right| \leq \zeta_2^\star \delta_2, \quad \left| R(z) - \bar{R}(z) \right| \leq \zeta_3^\star \delta_3, \quad \left| D(z) - \bar{D}(z) \right| \leq \zeta_4^\star \delta_4.$$

Theorem 4 (Hyers-Ulam stability)**.** Under the given assumptions, the FEVM (3) is stable.

Proof. According to the uniqueness Theorem 3, the model (3) has a unique solution $S(z), I(z), R(z), D(z)$. Let the approximate solution of the model be denoted by $S^\star(z), I^\star(z), R^\star(z)$ and $D^\star(z)$. Then,

we have

$$|S(z) - S^\star(z)| = \int_0^t \frac{|G_1(z, S(z)) - G_1(z, S^\star(z))|\, dz}{\Gamma(\sigma)(t-z)^{1-\sigma}} \leq \frac{\eta_1 t^\sigma}{\Gamma(1+\sigma)} \|S - S^\star\|.$$

Let $\eta_1^\star = \frac{t^\sigma}{\Gamma(1+\sigma)} \|S - S^\star\|$ and $\delta_1^\star = \eta_1$. Then $|S(z) - S^\star(z)| \leq \eta_1^\star \delta_1^\star$. Similarly, for $I(z)$

$$|I(z) - I^\star(z)| = \int_0^t \frac{|G_2(z, I(z)) - G_2(z, I^\star(z))|\, dz}{\Gamma(\sigma)(t-z)^{1-\sigma}} \leq \frac{\eta_2 t^\sigma}{\Gamma(1+\sigma)} \|I - I^\star\|.$$

Let $\eta_2^\star = \frac{t^\sigma}{\Gamma(1+\sigma)} \|I - I^\star\|$, and $\delta_2^\star = \eta_2$. Then $|I(z) - I^\star(z)| \leq \eta_2^\star \delta_2^\star$. For $R(z)$

$$|R(z) - R^\star(z)| = \int_0^t \frac{|G_3(z, R(z)) - G_3(z, R^\star(z))|}{\Gamma(\sigma)(t-z)^{1-\sigma}} dz \leq \frac{\eta_3 t^\sigma}{\Gamma(1+\sigma)} \|R - R^\star\|.$$

Let $\eta_3^\star = \frac{t^\sigma}{\Gamma(1+\sigma)} \|R - R^\star\|$, and $\delta_3^\star = \eta_3$. Then $|R(z) - R^\star(z)| \leq \eta_3^\star \delta_3^\star$. Finally, for $D(z)$, since G_4 is independent of D, we trivially have $|D(z) - D^\star(z)| = 0$, which satisfies Hyers–Ulam stability with $\eta_4^\star = 0$, $\delta_4^\star = 0$. Thus, all state variables satisfy the Hyers–Ulam condition, and the model is Hyers–Ulam stable. □

4 The discretized problem

Let M be a positive integer and define a uniform partition of the interval $[0, T]$, $T > 0$ as $\{t_j = j\tau \mid j = 0, 1, \ldots, M\}$, where the step size is $\tau = T/M$. The Caputo fractional derivative of order σ applied to $S(t)$ at the grid point t_j for $j = 1, 2, \ldots, M$ is expressed as

$$\mathcal{D}_t^\sigma S(t_j) = \frac{1}{\Gamma(1-\sigma)} \sum_{k=0}^{j-1} \int_{t_k}^{t_{k+1}} (t_j - s)^{-\sigma} \partial_s S(s)\, ds,$$

which can be approximated using the L1 scheme [16] as

$$\begin{aligned} \mathcal{D}_M^\sigma S(t_j) &= \frac{1}{\Gamma(1-\sigma)} \sum_{k=0}^{j-1} \frac{1}{\tau} [S(t_{k+1}) - S(t_k)] \int_{t_k}^{t_{k+1}} (t_j - s)^{-\sigma} ds \\ &= \sum_{k=0}^{j-1} [S(t_{k+1}) - S(t_k)]\, d_{j-k} \end{aligned}$$

$$= S(t_j)d_1 + \sum_{k=1}^{j-1}(d_{j-k+1} - d_{j-k})S(t_k) - S(t_0)d_j, \tag{5}$$

where $d_p = \frac{\tau^{-\sigma}}{\Gamma(2-\sigma)}[p^{1-\sigma} - (p-1)^{1-\sigma}]$ for $p = 1, 2, \ldots, M$. Similarly, the L1 discretization of $\mathcal{D}_t^\sigma I(t)$, $\mathcal{D}_t^\sigma R(t)$ and $\mathcal{D}_t^\sigma D(t)$ at t_j are given by

$$\begin{aligned}
\mathcal{D}_M^\sigma I(t_j) &= \sum_{k=0}^{j-1} [I(t_{k+1}) - I(t_k)]\, d_{j-k}, \\
\mathcal{D}_M^\sigma R(t_j) &= \sum_{k=0}^{j-1} [R(t_{k+1}) - R(t_k)]\, d_{j-k}, \\
\mathcal{D}_M^\sigma D(t_j) &= \sum_{k=0}^{j-1} [D(t_{k+1}) - D(t_k)]\, d_{j-k}.
\end{aligned}$$

Substituting the above into the model (3), we obtain the discrete system

$$\begin{cases}
\mathcal{D}_M^\sigma S(t_j) = -\alpha S(t_j)I(t_j) + \beta R(t_j) - \gamma N, \\
\mathcal{D}_M^\sigma I(t_j) = \alpha S(t_j)I(t_j) - \epsilon I(t_j) - \delta I(t_j), \\
\mathcal{D}_M^\sigma R(t_j) = \delta I(t_j) - \beta R(t_j), \\
\mathcal{D}_M^\sigma D(t_j) = \epsilon I(t_j) + \gamma N.
\end{cases} \tag{6}$$

Rewriting in simplified form, we get

$$\begin{cases}
(d_1 + \alpha I(t_j))S(t_j) - \beta R(t_j) + \gamma N + \sum_{k=1}^{j-1}(d_{j-k+1} - d_{j-k})S(t_k) - S(t_0)d_j = 0, \\
(d_1 + \epsilon + \delta - \alpha S(t_j))I(t_j) + \sum_{k=1}^{j-1}(d_{j-k+1} - d_{j-k})I(t_k) - I(t_0)d_j = 0, \\
(d_1 + \beta)R(t_j) - \delta I(t_j) + \sum_{k=1}^{j-1}(d_{j-k+1} - d_{j-k})R(t_k) - R(t_0)d_j = 0, \\
(d_1)D(t_j) - \epsilon I(t_j) - \gamma N + \sum_{k=1}^{j-1}(d_{j-k+1} - d_{j-k})D(t_k) - D(t_0)d_j = 0.
\end{cases} \tag{7}$$

Define the system compactly for $j = 1, 2, \ldots, M$ as

$$\begin{cases} F_1(S_j, I_j, R_j, D_j) = 0, & F_2(S_j, I_j, R_j, D_j) = 0, \\ F_3(S_j, I_j, R_j, D_j) = 0, & F_4(S_j, I_j, R_j, D_j) = 0, \\ S_j, I_j, R_j, D_j \geq 0, \end{cases} \tag{8}$$

where S_j denotes the approximate value of $S(t)$ at time t_j. Similar notation is used for I_j, R_j, and D_j. The nonlinear algebraic system (8) is solved using the Newton–Raphson method, with the initial guess

$$\begin{bmatrix} S_j^{(0)} \\ I_j^{(0)} \\ R_j^{(0)} \\ D_j^{(0)} \end{bmatrix} = \begin{bmatrix} S_{j-1} \\ I_{j-1} \\ R_{j-1} \\ D_{j-1} \end{bmatrix}.$$

Then, the nth iteration is computed by

$$\begin{bmatrix} S_j^{(n)} \\ I_j^{(n)} \\ R_j^{(n)} \\ D_j^{(n)} \end{bmatrix} = \begin{bmatrix} S_j^{(n-1)} \\ I_j^{(n-1)} \\ R_j^{(n-1)} \\ D_j^{(n-1)} \end{bmatrix} - J_{(n-1)}^{-1} \begin{bmatrix} F_1^{(n-1)} \\ F_2^{(n-1)} \\ F_3^{(n-1)} \\ F_4^{(n-1)} \end{bmatrix},$$

where $J_{(n-1)}$ is the Jacobian matrix computed using partial derivatives of F_i with respect to the unknowns. Finally, the solution at step j is approximated as

$$\begin{bmatrix} S_j \\ I_j \\ R_j \\ D_j \end{bmatrix} \approx \begin{bmatrix} S_j^{(n)} \\ I_j^{(n)} \\ R_j^{(n)} \\ D_j^{(n)} \end{bmatrix}.$$

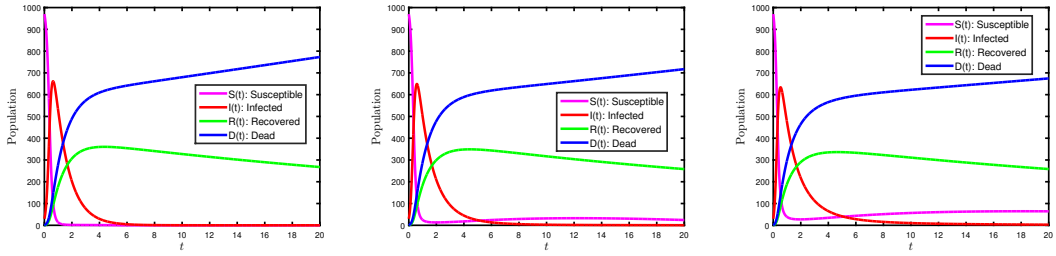
5 Numerical illustration and discussions

The following test example is presented to validate the accuracy and efficiency of the proposed numerical scheme. All simulations are carried out using MATLAB R2016a.

Example 1. Consider the model problem (3) with initial conditions as given in [21]: $\alpha = 0.01$, $\beta = 0.02$, $\gamma = 0.01$, $N = 1000$, $\epsilon = 0.6$, $\delta = 0.4$, $S(0) = 970$, $I(0) = 30$, $R(0) = 0$, $D(0) = 0$.

To evaluate the model (3), a comparative numerical experiment was conducted using three different fractional orders: 0.99, 0.95, and 0.9. The simulation results for each case are depicted in

Figures 1. Specifically, Figure 1a corresponds to $\sigma = 0.99$, Figure 1b to $\sigma = 0.95$, and Figure 1c to $\sigma = 0.9$. Across all three cases, the system dynamics are qualitatively similar, confirming the consistency and robustness of the proposed method. These observations align with previous results reported in [21]. Further, comparative analysis is presented in Figures 2, which show the influence of varying fractional orders ($\sigma = 0.90$, 0.95, 0.98, and 0.99) on the dynamics of the individual compartments. Figure 2a illustrates the dynamics of the susceptible population $S(t)$, which exhibits a decreasing trend as σ increases. This implies that higher fractional orders slightly accelerate the depletion of the susceptible population. The infected population $I(t)$ in Figure 2b grows sharply at first but then falls to zero more quickly when σ gets closer to 0.99. This indicates that infection is recovered and controlled more quickly at higher fractional orders. The recovered population $R(t)$, shown in Figure 2c, increases monotonically with time and stabilizes more quickly as σ increases, indicating a faster transition from infection to recovery. Finally, the evolution of the cumulative death population $D(t)$ is illustrated in Figure 2d, where a gradual stabilization follows an initially rapid growth. The figure indicates a pronounced increase at higher fractional orders, highlighting the significant influence of σ on mortality accumulation. Although the system's qualitative behavior remains consistent for $0.90 \leq \sigma \leq 0.99$, small changes in σ produce notable quantitative shifts, highlighting the model sensitivity to the fractional order.

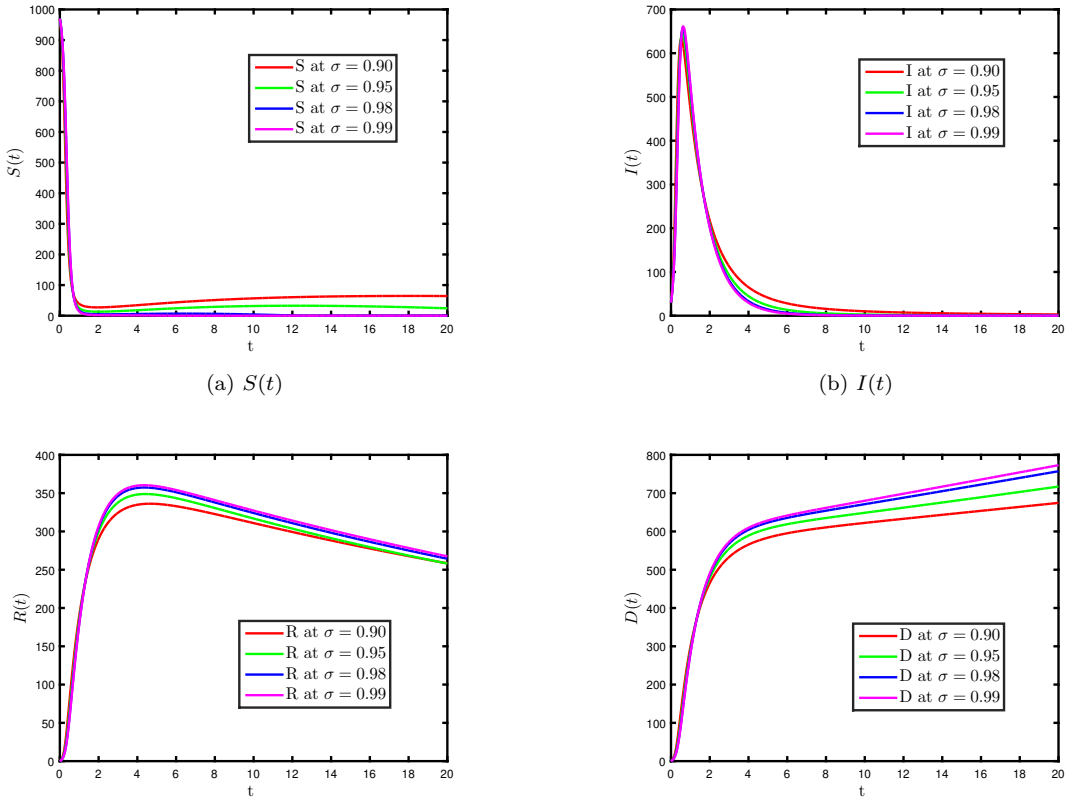


(a) $\sigma = 0.99$ (b) $\sigma = 0.95$ (c) $\sigma = 0.9$

Figure 1: Solutions of $S(t), I(t), R(t)$, and $D(t)$.

6 Conclusion

In this study, we developed and analyzed a FEVM transmission, focusing on the assessment of its long-term dynamics. We examined the existence, uniqueness, and stability of the model solutions using fractional calculus, specifically employing the Caputo derivative. Numerical simulations were carried out using an iterative scheme with initial conditions. To explore the transmission behavior of the Ebola virus, we analyzed the system under varying fractional orders and param-



(a) $S(t)$ (b) $I(t)$

(c) $R(t)$ (d) $D(t)$

Figure 2: Comparison of $S(t)$, $I(t)$, $R(t)$, and $D(t)$ for different σ.

eter sensitivities. The results of the numerical experiments illustrate that FEVM yields solutions that are not only consistent with biological expectations but also more flexible and realistic than their integer-order counterparts. These findings underscore the significance of incorporating fractional derivatives into epidemiological modeling, as they offer a more accurate representation of memory and hereditary properties inherent in infectious disease dynamics.

Funding

This research received no external funding.

Conflicts of interest

The authors declare that there are no conflicts of interest.

References

[1] Abah, R.T., Zhiri, A.B., Oshinubi, K. and Adeniji, A. *Mathematical analysis and simulation of Ebola virus disease spread incorporating mitigation measures*. Frankl. Open, 6 (2024), 100066.

[2] Abbas, N., Zanib, S.A., Ramzan, S., Nazir, A. and Shatanawi, W. *A conformable mathematical model of Ebola virus disease and its stability analysis*. Heliyon, 10(16) (2024) e35818.

[3] Adu, I.K., Wireko, F.A., Sebil, C. and Asamoah, J.K.K. *A fractal–fractional model of Ebola with reinfection*. Results Phys., 52 (2023), 106893.

[4] Ahmad, W., Rafiq, M. and Abbas, M. *Mathematical analysis to control the spread of Ebola virus epidemic through voluntary vaccination*. Eur. Phys. J. Plus, 135(10) (2020), 775.

[5] Saini, G., Ghosh, B., Chand, S. and Mohapatra, J. *An iterative-based difference scheme for nonlinear fractional integro-differential equations of Volterra type*. Partial Differ. Equ. Appl. Math., 13 (2025), 101138.

[6] Berge, T., Chapwanya, M., Lubuma, J.S. and Terefe, Y.A. *A mathematical model for Ebola epidemic with self-protection measures*. J. Biol. Syst., 26(1) (2018), 107–131.

[7] Bonyah, E., Badu, K. and Asiedu-Addo, S.K. *Optimal control application to an Ebola model*. Asian Pac. J. Trop. Biomed., 6(4) (2016), 283–289.

[8] Derakhshan, M.H. *The stability analysis and numerical simulation based on Sinc Legendre collocation method for solving a fractional epidemiological model of the Ebola virus*. Partial Differ. Equ. Appl. Math., 3 (2021), 100037.

[9] Ghosh, B. and Mohapatra, J. *Numerical simulation for two species time fractional weakly singular model arising in population dynamics*. Int. J. Model. Simul., 45(4) (2025), 1383–1396.

[10] Gunasekar, T., Manikandan, S., Govindan, V., Ahmad, J., Emam, W. and Al-Shbeil, I. *Symmetry analyses of epidemiological model for Monkeypox virus with Atangana–Baleanu fractional derivative*. Symmetry, 15(8) (2023), 1605.

[11] Gunasekar, T., Manikandan, S., Suba, M. and Akgül, A. *A fractal–fractional mathematical model for COVID-19 and tuberculosis using Atangana–Baleanu derivative*. Math. Comput. Model. Dyn. Syst., 30(1) (2024), 857–881.

[12] Gunasekar, T., Manikandan, S., Haque, S., Suba, M. and Mlaiki, N. *Fractal–fractional mathematical modeling of monkeypox disease and analysis of its Ulam–Hyers stability*. Bound. Value Probl., 2025(1) (2025), 20.

[13] Hussain, S., Arora, G. and Kumar, R. *Analytical and numerical treatments of a coagulation population balance model*. Chaos Solit. Fractals, 192 (2025), 115974.

[14] Jung, S.M. *Hyers–Ulam stability of linear differential equations of first order*. Appl. Math. Lett., 17(10) (2004), 1135–1140.

[15] Kaushik, S., Kumar, R. and da Costa, F.P. *Theoretical analysis of a discrete population balance model with sum kernel*. Port. Math., 80(3) (2023), 343–367.

[16] Lin, Y. and Xu, C. *Finite difference/spectral approximations for the time-fractional diffusion equation*. J. Comput. Phys., 225(2) (2007), 1533–1552.

[17] Mainardi, F. *Fractional calculus and waves in linear viscoelasticity: An introduction to mathematical models*. World Scientific, 2010.

[18] Manikandan, S., Gunasekar, T., Kouidere, A., Venkatesan, K.A., Shah, K. and Abdeljawad, T. *Mathematical modelling of HIV/AIDS treatment using Caputo–Fabrizio fractional differential systems*. Qual. Theory Dyn. Syst., 23(4) (2024), 149.

[19] Nazir, A., Ahmed, N., Khan, U., Mohyud-Din, S.T., Nisar, K.S. and Khan, I. *An advanced version of a conformable mathematical model of Ebola virus disease in Africa*. Alex. Eng. J., 59(5) (2020), 3261–3268.

[20] Podlubny, I. *Fractional Differential Equations*. Academic Press, 1999.

[21] Singh, H. *Analysis for fractional dynamics of Ebola virus model*. Chaos Solit. Fractals, 138 (2020), 109992.

[22] Suba, M., Venkatesan, A., Gunasekar, T., Manikandan, S., Raghavendran, P., Santra, S.S. and Karmakar, S. *Advancing Zika virus control with fractional order optimal control with Wolbachia-infected Aedes aegypti mosquitoes*. In *International Conference on Mathematical Modelling, Applied Analysis and Computation*, Springer, (2024), 160–183.

[23] Yang, X.J., Gao, F. and Ju, Y. *General fractional derivatives with applications in viscoelasticity*. Academic Press, 2020.

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 1, 2026, pp 188–219
https://doi.org/10.22067/ijnao.2025.94641.1689
https://ijnao.um.ac.ir/

CC BY NC

Research Article

Transdermal drug delivery via a hollow microneedle considering the skin's mechanical properties: Analytical and numerical study

E. Azhdari* and A. Emami

Abstract

Hollow microneedles have a significant impact on transdermal drug delivery by overcoming the stratum corneum layer of the skin and delivering drugs directly to living skin tissues. In this paper, using a mathematical model, the diffusion of nanocarriers, free drugs, and bound drugs in the microneedle channel, different skin layers, and the bloodstream is described. The mechanical properties of the skin, which affect drug delivery due to the elasticity and elastin present in the dermis of the skin, are considered. The qualitative behavior of the model is studied using energy estimation. The well-posedness of the fully coupled system is examined. Numerical simulations are used to study the model's numerical behavior and sensitivity analysis.

AMS subject classifications (2020): Primary 92B99; Secondary 65M12.

*Corresponding author

Received 26 July 2025; revised 9 October 2025; accepted 18 October 2025

Ebrahim Azhdari
Department of Mathematics and Computer Science, Faculty of Basic Sciences, Salman Farsi University of Kazerun, Kazerun, Iran. e-mail: e.azhdari@kazerunsfu.ac.ir

Aram Emami
Department of Mathematics and Computer Science, Faculty of Basic Sciences, Salman Farsi University of Kazerun, Kazerun, Iran. e-mail: emami@kazerunsfu.ac.ir

How to cite this article
Azhdari, E. and Emami, A., Transdermal drug delivery via a hollow microneedle considering the skin's mechanical properties: Analytical and numerical study. *Iran. J. Numer. Anal. Optim.*, 2026; 16(1): 188-219. https://doi.org/10.22067/ijnao.2025.94641.1689

Keywords: Mathematical model; Numerical simulation; Viscoelastic properties; Energy estimate; Hollow microneedles.

1 Introduction

The transdermal drug delivery system has emerged as a powerful, acceptable, and general approach for new drug delivery systems [27]. It has gained great interest in practice due to many advantages such as less invasiveness, ease of application, and so on [19, 32]. A drug applied to the skin's surface must penetrate the stratum corneum (SC) and viable epidermis (VE) in order to reach the blood (BL) vessel-containing papillary dermis (PD), which is determined by the anatomical makeup of the skin. The rest of the drug may enter the lymphatic system or penetrate into the reticular dermis (RD). The topmost layer of the skin, known as the SC, serves as a barrier to protect the skin and stops additional drug penetration because its cells are made up of densely packed, flat, dead keratinocytes without nuclei [28]. This problem doubles the importance of dermal drug delivery by new methods.

To overcome this issue, hollow microneedles (MNs) are used. In general, MNs are divided into five categories, namely solid, hollow, coated, dissolving, and swellable MNs [7, 20, 34]. Hollow MNs contain a channel (CH). After passing through the SC, the MN delivers the drug directly to the underlying layers with the help of the CH, as shown in [7]. For the treatment to be effective, instead of freely transporting the drug to the desired sites, the drug is placed in nanocarriers and the nanocarriers are transported to the desired site through the CH [17]. This prevents the drug molecules from being quickly destroyed by the body.

The dynamic properties of skin tissue, which vary among patients, can affect transdermal drug delivery [18, 29]. It is challenging to isolate these factors experimentally alone. Mathematical modeling can be of great help. Many mathematical models have been studied in the field of transdermal drug delivery [2, 3, 5, 6, 22]. Azhdari and Emami [3] have simulated transdermal drug delivery using polymeric MNs. The effect of MN swelling in the skin on drug delivery has been investigated and studied numerically and analytically. A model explored for drug delivery through hollow MNs was examined by Benbrook and Zhan [5], focusing on the influence of various characteristics of hollow MNs, nanocarriers, clinical environments, and environmental variables . The model presented by Zhan et al. does not address the mechanical properties of the skin, which are influenced by a protein called elastin, which is found in the dermis layer of the skin and has elasticity and stretch properties. In this paper, we study these important properties and show their effect on drug delivery.

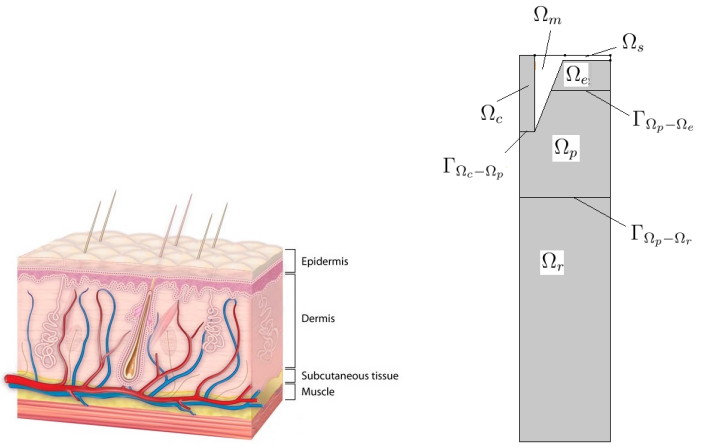
The summary of this paper is as follows: In section 2, the anatomy of the skin is discussed. The mathematical model of the problem and the initial and boundary conditions, as well as the viscoelastic effect are presented in section 3. In section 4, the qualitative behavior of the model, using energy estimation, and the well-posedness of the fully coupled system are discussed. Numerical results, numerical simulations, numerical error analysis, and convergence study are presented in section 5. Finally, in section 6, conclusions are presented.

2 The anatomy of the skin

The skin is the covering and protective part of the body, which in addition to having a significant impact on our appearance, also indicates our internal health [21]. The skin is composed of three main layers (Figure 1a), each of which includes a sublayer and appendages that have their own functions and play a role in maintaining the overall structure of the skin [1]. These layers of the skin, from surface to depth, include: epidermis, dermis, and hypodermis. In the case of the epidermis, it is mostly composed of dead cells, and the hypodermis plays a major role in warming the body. The dermis is the second layer of skin located below the epidermis. The dermis of the skin is divided into two areas: one is the PD, and the other is the RD [30]. This layer is composed of connective tissue and contains a large amount of nervous tissue and blood vessels. When it comes to thickness, the dermis is significantly thicker than the epidermis and is responsible for the skin's strength and elasticity. The geometry of the model is shown in Figure 1b. According to this figure, to make sense, we indicate the domains as follows: SC (Ω_s), MN (Ω_m), CH (Ω_c), VE (Ω_e), PD and BL (Ω_p), and RD (Ω_r). Also, $15\mu m$ for SC, $100\mu m$ for VE, $350\mu m$ for PD, and $800\mu m$ for RD are the actual skin layer thickness values that the model is based on. Also, the length of the MN is $250\mu m$ and the width of the CH is $30\mu m$. The cylindrical CH is located in the center of the MN.

3 Mathematical model

The model includes equations that describe the transport of infusate, interstitial fluid, and drug inside the MN CH, skin tissues, and circulatory system. The notations have been included in Table 1 to facilitate the reading of the parameters.



(a) The layers of the skin include: epidermis, dermis, and hypodermis.

(b) The geometry of the model: SC (Ω_s), MN (Ω_m), CH (Ω_c), VE (Ω_e), PD and BL (Ω_p) and RD (Ω_r).

Figure 1: The anatomy of the skin and geometry of the model.

3.1 Fluid transfer model

The fluid inside the MN CH and skin tissues such as VE, PD, and RD, which act as a porous medium, are described with the help of the continuity equation and the moment conservation equation as follows:

$$\nabla \cdot \mathbf{v} = 0, \tag{1}$$

$$\rho \frac{\partial \mathbf{v}}{\partial t} + \rho(\mathbf{v} \cdot \nabla \mathbf{v}) = -\nabla p + \mu \nabla^2 \mathbf{v}, \tag{2}$$

where $\mathbf{v}$ is the velocity, p is the pressure, ρ is the density, and μ is the viscosity.

3.2 Drug delivery model

In this section, a mathematical model for different types of drugs, which are delivered in the form of nanocarriers, released drugs (free drugs), and encapsulated drugs (bound drugs), is presented. The drug is first introduced into the skin tissue in the form of nanocarrier capsules through the

Table 1: Parameter notation and description.

Parameter Notation	Parameter Description
n	Nanocarrier
f	Free drugs
b	Bound drugs
MN (Ω_m)	Microneedle
$e\ VE$ (Ω_e)	Viable epidermis
$c\ SC$ (Ω_s)	Stratum corneum
$p\ PD$ (Ω_p)	Papillary dermis
$r\ RD$ (Ω_r)	Reticular dermis
$c\ CH$ (Ω_c)	Channel
$C_{i,j}$	Concentration of $i = n, f, b$ in Ω_j $j = c, e, p, r, bl$
$D_{i,j}$	Diffusion coefficient of $i = n, f$ in Ω_j $j = e, c, p, r, bl$
$k_{d,i}$	Physical degradation rate of free drug in Ω_j $j = c, e, p, r$
$k_{rel,i}$	Drug release rate in Ω_j $j = e, p, r, bl$
k_{pb}	Drug association rate with protein
k_{pu}	Drug dissociation rate from protein
$k_{clr,i}$	Rate of plasma clearance of $i = n, f$
R_{ly}	Rate of fluid exchange between the PD and lymph
k_{Ex}	Constant dependent
R_{bl}	Rate of fluid exchange
σ_f	Osmotic reflection coefficient of drug
E_1	Stiffness coefficient of the single spring
E_2	Stiffness coefficient of the spring
η	Stiffness coefficient of the dashpot
D_v	Viscoelastic diffusion coefficient
σ	Stress generated by the skin
$\mathbf{v}$	Velocity
p	Pressure
ρ	Density
μ	Viscosity

MN CH. After contact with the skin tissues, the nanocarriers begin to release the drug, and the remaining nanocarriers and released drugs are transported to the adjacent tissues, which depends on how far the MN is inserted into the skin. In this paper, we assume that the tip of the MN passes through the SC and VE and is located in the PD. Nanocarriers and free drugs are subject to elimination in all layers of the skin corresponding to the skin tissues due to biological and physical reactions. In particular, some of the nanocarriers and free drugs in the PD layer may enter the circulatory and lymphatic systems. The transfer of nanocarriers and free drugs within the MN CH, skin layers and bloodstream, determined through the convection, diffusion, degradation and release equations, is expressed as follows:

$$\frac{\partial C_{n,c}}{\partial t} = \nabla \cdot (D_{n,c} \nabla C_{n,c}) - \nabla \cdot (\mathbf{v} C_{n,c}),\ in\ \Omega_c \times (0, T], \tag{3}$$

$$\frac{\partial C_{f,c}}{\partial t} = \nabla \cdot (D_{f,c} \nabla C_{f,c}) - \nabla \cdot (\mathbf{v} C_{f,c}) - k_{d,c} C_{f,c}, \ in\ \Omega_c \times (0,T], \tag{4}$$

$$\frac{\partial C_{n,e}}{\partial t} = \nabla \cdot (D_{n,e} \nabla C_{n,e}) - \nabla \cdot (\mathbf{v} C_{n,e}) - k_{rel,e} C_{n,e}, \ in\ \Omega_e \times (0,T], \tag{5}$$

$$\frac{\partial C_{f,e}}{\partial t} = \nabla \cdot (D_{f,e} \nabla C_{f,e}) - \nabla \cdot (\mathbf{v} C_{f,e}) + k_{rel,e} C_{n,e} - k_{d,e} C_{f,e}, \quad in\ \Omega_e \times (0,T], \tag{6}$$

$$\frac{\partial C_{n,p}}{\partial t} = \nabla \cdot (D_{n,p} \nabla C_{n,p}) - \nabla \cdot (\mathbf{v} C_{n,p}) - R_{ly} C_{n,p} - k_{rel,p} C_{n,p}, \quad in\ \Omega_p \times (0,T], \tag{7}$$

$$\frac{\partial C_{f,p}}{\partial t} = \nabla \cdot (D_{f,p} \nabla C_{f,p}) - \nabla \cdot (\mathbf{v} C_{f,p}) + k_{rel,p} C_{n,p} - k_{d,p} C_{f,p} - R_{ly} C_{f,p} - k_{pb} C_{f,p} + k_{pu} C_{b,p} - Ex[C_{f,p}, C_{f,bl}], \ in\ \Omega_p \times (0,T], \tag{8}$$

$$\frac{dC_{b,p}}{dt} = k_{pb} C_{f,p} - k_{pu} C_{b,p}, \ in\ \Omega_p \times (0,T], \tag{9}$$

$$\frac{\partial C_{n,r}}{\partial t} = \nabla \cdot (D_{n,r} \nabla C_{n,r}) - \nabla \cdot (\mathbf{v} C_{n,r}) - k_{rel,r} C_{n,r}, \ in\ \Omega_r \times (0,T], \tag{10}$$

$$\frac{\partial C_{f,r}}{\partial t} = \nabla \cdot (D_{f,r} \nabla C_{f,r}) - \nabla \cdot (\mathbf{v} C_{f,r}) + k_{rel,r} C_{n,r} - k_{d,r} C_{f,r} \quad in\ \Omega_r \times (0,T], \tag{11}$$

$$\frac{dC_{n,bl}}{dt} = -k_{rel,bl} C_{n,bl} - k_{clr,n} C_{n,bl}, \ in\ \Omega_p \times (0,T], \tag{12}$$

$$\frac{dC_{f,bl}}{dt} = k_{rel,bl} C_{n,bl} - (k_{pb} C_{f,bl} - k_{pu} C_{b,bl}) - k_{clr,f} C_{f,bl} + Ex[C_{f,p}, C_{f,bl}] \quad in\ \Omega_p \times (0,T], \tag{13}$$

$$\frac{dC_{b,bl}}{dt} = k_{pb}C_{f,bl} - k_{pu}C_{b,bl},\ in\ \Omega_p \times (0,T], \tag{14}$$

where $C_{i,j}$ are nanocarrier concentration, free drugs concentration and bound drugs concentration in CH, VE, PD, RD, BL, for $i = n, f, b$ and $j = c, e, p, r, bl$, respectively, $D_{i,j}$ are the diffusion coefficient of nanocarrier and free drugs in CH, VE, PD, RD, for $i = n, f$ and $j = c, e, p, r$, respectively, $k_{d,i}$ are the physical degradation rate of free drug in CH, VE, PD, RD, for $i = c, e, p, r$, respectively, $k_{rel,i}$ are the drug release rate in VE, PD, RD, BL for $i = e, p, r, bl$, respectively, R_{ly} is the rate of fluid exchange between the PD and lymph, k_{pb} is the drug association rate with protein, k_{pu} is the drug dissociation rate from protein, and $k_{clr,n}$ and $k_{clr,f}$ are the plasma clearance rate.

In (8) and (13), the free drugs exchange rate between the PD and BL is defined as

$$Ex[C_{f,p}, C_{f,bl}] = k_{Ex}(C_{f,p} - C_{f,bl}) - R_{bl}(1 - \sigma_f)C_{f,bl}, \tag{15}$$

where k_{Ex} is a constant dependent on the vascular permeability of nanocarriers and capillary surface area per unit of tissue volume, R_{bl} is the rate of fluid exchange between the PD and the blood, and σ_f is the osmotic reflection coefficient of drug.

The following can be mentioned as a biological implication of the equations. Diffusion and convection with flow of interstitial fluid are required in VE, while in MN, convection, infusion, and diffusion are required to transport drug nanocarriers. In PD, transport of nanocarriers is dependent not only on diffusion, convection and release of the drug but also determined by exchange between the tissue and circulatory systems, whereas in RD, transport is mediated by convection, diffusion, and local release of the drug. Exchange between BL and PD, local release of the drug and plasma clearance all affect the blood concentration of the drug nanocarriers.

3.3 Viscoelastic effects

Elastin is a protein found naturally in the dermis of the skin and has elasticity and stretch properties. This protein allows the skin to return to its original shape after being stretched. This phenomenon causes the skin to become resistant to the drug's entry into the PD through the MN CH. Therefore, to describe the viscoelastic behavior of the skin, it is necessary to add the term $\nabla.(D_v \nabla\sigma)$ to the right-hand side of (7) and (8). Keeping the sign used by Ferreira et al. [13], σ denotes the stress generated by the skin and D_v is called the viscoelastic diffusion coefficient. Therefore, (7) and (8) change as follows after this explanation:

$$\frac{\partial C_{n,p}}{\partial t} = \nabla \cdot (D_{n,p}\nabla C_{n,p}) + \nabla.(D_v \nabla \sigma) - \nabla \cdot (\mathbf{v} C_{n,p}) - R_{ly} C_{n,p} \\ -k_{rel,p} C_{n,p}, \tag{16}$$

$$\frac{\partial C_{f,p}}{\partial t} = \nabla \cdot (D_{f,p}\nabla C_{f,p}) + \nabla.(D_v \nabla \sigma) - \nabla \cdot (\mathbf{v} C_{f,p}) + k_{rel,p} C_{n,p} - k_{d,p} C_{f,p} \\ -R_{ly} C_{f,p} - k_{pb} C_{f,p} + k_{pu} C_{b,p} - Ex[C_{f,p}, C_{f,bl}]. \tag{17}$$

Many viscoelastic models obtained under different experimental conditions are presented here [24, 31]. In what follows, we present one of the simplest models, also known as the Zener model, which is linear [4, 26]. The model presented is used to describe the viscoelastic behavior of skin due to the presence of elastin in the dermis layer. The delay times used in this model are consistent with those predicted for viscoelastic behavior in living tissues [25]. The system is made up of a spring in parallel with a spring-dashpot in series, and the connection between stress and strain is depicted as follows:

$$\frac{d\sigma}{E_2 \times dt} + \frac{\sigma}{\eta} = (1 + \frac{E_1}{E_2}) \times \frac{d\epsilon}{dt} + \frac{E_1}{\eta} \times \epsilon, \tag{18}$$

where ϵ is the strain created by the nanocarrier and drug molecules. E_1 represents the sole spring's stiffness and E_2 and η represent the spring-dashpot pair's stiffness and damping factor, respectively.

To demonstrate that the skin prevents the drug from entering the PD, the terms on the right side of (18) must be multiplied by a negative sign, meaning that σ and ϵ have opposite signs.

As stated in [4, 12], in (18), we assume the relationship between strain and the concentration of drug and nanocarriers as the following linear relationship:

$$\epsilon = f(C_i), \qquad i = n, p \ and \ f, p. \tag{19}$$

The diffusion of solute through viscoelastic materials does not obey the Fick Law. This is because, in biological terms, for example, the concentration of the solvent may change the physical structure of the tissue (e.g., cartilage, hydrogel, or cell-filled matrix). The viscoelastic matrix is in opposition to the Brownian motion of the molecules, which can be measured by the stress reaction to the strain. Several authors have proposed a general model in which the flux is due to two phenomena: the concentration gradient of a solute and the stress gradient created by the matrix [13, 15]. Therefore, to measure the stress created in (16)–(17), we represent it with

the simplest Zener model that includes both instantaneous elasticity and delayed viscoelasticity. To solve (18), we provide the strain ϵ based on the concentration C_i for $i = n, p$ and $i = f, p$.

Many biological materials are porous, hydrated structures. The concentration of fixed charges on proteoglycans attracts counterions, creating an osmotic pressure difference. Increasing the concentration of C_i for $i = n, p$ and $i = f, p$ ions within the tissue directly leads to volumetric strain and swelling. This relationship is often linear over a range of physiologically relevant concentrations. Mathematically, in the Taylor expansion, if the deviation is small, then the terms with higher orders can be eliminated. So we get a linear relationship between strain and concentration.

3.4 Initial and boundary conditions

The mathematical model describing the MNs with holes inserted into the skin and solved in two dimensions for drug delivery. Throughout the entire computational domain, the initial concentrations of the free and bound drugs are taken to be zero, whereas the nanocarrier concentration in CH is taken to be the starting concentration at the start.

No liquid or drug passes through the wall of the MN cavity. Also, the CH surface is a hard wall and the drug flux is zero. The drug is taken to be zero in the vicinity of the computational domain (Ω_e, Ω_p and Ω_r). Therefore, we have the following boundary condition:

$$D\nabla C \cdot \eta = 0, \tag{20}$$

where η is the exterior unit normal, D represents the diffusion coefficient of free drugs or nanocarriers, and C represents the concentration.

Boundary conditions are applied at the interface boundaries between different layers of the skin, $\Gamma_{\Omega_p-\Omega_e}$ and $\Gamma_{\Omega_p-\Omega_r}$, as well as at the interface boundary between the CH and the skin tissue, $\Gamma_{\Omega_c-\Omega_p}$, for drug and fluid transport are as follows:

$$\begin{cases} D_1\nabla C_1 \cdot \eta_1 = \iota_{1,2}(C_1 - C_2), \\ D_2\nabla C_2 \cdot \eta_2 = -D_1\nabla C_1 \cdot \eta_1, \end{cases} \tag{21}$$

where η_1 and η_2 are the exterior unit normal, D_1 and D_2 are the diffusion coefficient of nanocarriers or free drugs in the domains with an interface boundary, C_1 and C_2 are the concentration in the domains with an interface boundary, and $\iota_{1,2}$ is the mass transfer coefficient of the interfaces.

The MN surface is a hard wall, so the flow velocity at the MN boundary is zero. For the various layer boundaries of the skin, we have assumed that the flow velocity at these boundaries is zero. This indicates that the skin's boundary walls are regarded as rigid, establishing a stationary no-slip boundary condition on the walls. Therefore boundary conditions for the flow velocity in all domains are assumed to be zero, $\mathbf{v} = 0$.

4 Well-posedness

In this section, we use the energy estimate to arrive at the stability of the solution for a finite time interval, from which the existence of the solution also follows. We also show the uniqueness of the solution.

4.1 Energy estimates

In this section, we will discuss an energy estimate that gives us an overview of the qualitative behavior of the problem. First, we solve (18) to get

$$\begin{aligned}\sigma(t) = \sigma(0)e^{-\frac{E_2}{\eta}t} - (E_1+E_2)\epsilon(t) + (E_1+E_2)\epsilon(0)e^{-\frac{E_2}{\eta}t} \\ +\frac{E_2^2}{\eta}\int_0^t \epsilon(s)e^{-\frac{E_2}{\eta}(t-s)}ds.\end{aligned} \tag{22}$$

By Applying (22) in (16) and (17) and considering (19), we get

$$\begin{aligned}\frac{\partial C_{n,p}}{\partial t} = \nabla\cdot(D^*_{n,p}\nabla C_{n,p}) + \frac{D_v E_2^2}{\eta}\int_0^t e^{-\frac{E_2}{\eta}(t-s)}\nabla^2 C_{n,p}ds \\ -\nabla\cdot(\mathbf{v}C_{n,p}) - R_{ly}C_{n,p} - k_{rel,p}C_{n,p},\end{aligned} \tag{23}$$

$$\begin{aligned}\frac{\partial C_{f,p}}{\partial t} = \nabla\cdot(D^*_{f,p}\nabla C_{f,p}) + \frac{D_v E_2^2}{\eta}\int_0^t e^{-\frac{E_2}{\eta}(t-s)}\nabla^2 C_{f,p}ds - \nabla\cdot(\mathbf{v}C_{f,p}) \\ +k_{rel,p}C_{n,p} - k_{d,p}C_{f,p} - R_{ly}C_{f,p} - k_{pb}C_{f,p} \\ +k_{pu}C_{b,p} - Ex[C_{f,p}, C_{f,bl}],\end{aligned} \tag{24}$$

where $D^*_{i,p} = D_{i,p} - D_v(E_1+E_2)$ for $i = n, f$.

Note that the new interface boundary conditions for (23)–(24) are following

$$\begin{cases} (D_1\nabla C_1 + \dfrac{D_v E_2^2}{\eta}\displaystyle\int_0^t e^{-\frac{E_2}{\eta}(t-s)}\nabla C_1 ds)\cdot\eta_1 = \iota_{1,2}(C_1 - C_2), \\ D_2\nabla C_2\cdot\eta_2 = -(D_1\nabla C_1 + \dfrac{D_v E_2^2}{\eta}\displaystyle\int_0^t e^{-\frac{E_2}{\eta}(t-s)}\nabla C_1 ds)\cdot\eta_1. \end{cases} \tag{25}$$

By multiplying (3)–(6), (9)–(14) and (23)–(24) by $\phi_{i,j}$ for $i = n, f, b$ and $j = c, e, p, r, bl$, we obtain

$$\begin{aligned}(\frac{\partial C_{n,c}}{\partial t}, \phi_{n,c}) = &-(D_{n,c}\nabla C_{n,c}, \nabla\phi_{n,c}) + (\mathbf{v}C_{n,c}, \nabla\phi_{n,c}) \\ &+ \int_{\partial\Omega_c} D_{n,c}\nabla C_{n,c}\cdot\phi_{n,c}ds - \int_{\partial\Omega_c}\mathbf{v}C_{n,c}\cdot\phi_{n,c}ds,\end{aligned} \tag{26}$$

$$\begin{aligned}(\frac{\partial C_{f,c}}{\partial t}, \phi_{f,c}) = &-(D_{f,c}\nabla C_{f,c}, \nabla\phi_{f,c}) + (\mathbf{v}C_{f,c}, \nabla\phi_{f,c}) - k_{d,c}(C_{f,c}, \phi_{f,c}) \\ &+ \int_{\partial\Omega_c} D_{f,c}\nabla C_{f,c}\cdot\phi_{f,c}ds - \int_{\partial\Omega_c}\mathbf{v}C_{f,c}\cdot\phi_{f,c}ds,\end{aligned} \tag{27}$$

$$\begin{aligned}(\frac{\partial C_{n,e}}{\partial t}, \phi_{n,e}) = &-(D_{n,e}\nabla C_{n,e}, \nabla\phi_{n,e}) + (\mathbf{v}C_{n,e}, \nabla\phi_{n,e}) - k_{rel,e}(C_{n,e}, \phi_{n,e}) \\ &+ \int_{\partial\Omega_e} D_{n,e}\nabla C_{n,e}\cdot\phi_{n,e}ds - \int_{\partial\Omega_e}\mathbf{v}C_{n,e}\cdot\phi_{n,e}ds,\end{aligned} \tag{28}$$

$$\begin{aligned}(\frac{\partial C_{f,e}}{\partial t}, \phi_{f,e}) = &-(D_{f,e}\nabla C_{f,e}, \nabla\phi_{f,e}) + (\mathbf{v}C_{f,e}, \nabla\phi_{f,e}) + k_{rel,e}(C_{n,e}, \phi_{f,e}) \\ &+ \int_{\partial\Omega_e} D_{f,e}\nabla C_{f,e}\cdot\phi_{f,e}ds \\ &- \int_{\partial\Omega_e}\mathbf{v}C_{f,e}\cdot\phi_{f,e}ds - k_{d,e}(C_{f,e}, \phi_{f,e}),\end{aligned} \tag{29}$$

$$(\frac{dC_{b,p}}{dt}, \phi_{b,p}) = k_{pb}(C_{f,p}, \phi_{b,p}) - k_{pu}(C_{b,p}, \phi_{b,p}), \tag{30}$$

$$(\frac{\partial C_{n,r}}{\partial t},\phi_{n,r}) = -(D_{n,r}\nabla C_{n,r},\nabla\phi_{n,r}) + (\mathbf{v}C_{n,r},\nabla\phi_{n,r}) - k_{rel,r}(C_{n,r},\phi_{n,r}) + \int_{\partial\Omega_r} D_{n,r}\nabla C_{n,r}\cdot\phi_{n,r}ds - \int_{\partial\Omega_r}\mathbf{v}C_{n,r}\cdot\phi_{n,r}ds, \tag{31}$$

$$(\frac{\partial C_{f,r}}{\partial t},\phi_{f,r}) = -(D_{f,r}\nabla C_{f,r},\nabla\phi_{f,r}) + (\mathbf{v}C_{f,r},\nabla\phi_{f,r}) + \int_{\partial\Omega_r} D_{f,r}\nabla C_{f,r}\cdot\phi_{f,r}ds - \int_{\partial\Omega_r}\mathbf{v}C_{f,r}\cdot\phi_{f,r}ds - k_{d,r}(C_{f,r},\phi_{f,r}) + k_{rel,r}(C_{n,r},\phi_{f,r}), \tag{32}$$

$$(\frac{dC_{n,bl}}{dt},\phi_{n,bl}) = -k_{rel,bl}(C_{n,bl},\phi_{n,bl}) - k_{clr,n}(C_{n,bl},\phi_{n,bl}), \tag{33}$$

$$(\frac{dC_{f,bl}}{dt},\phi_{f,bl}) = k_{rel,bl}(C_{n,bl},\phi_{f,bl}) - k_{pb}(C_{f,bl},\phi_{f,bl}) + k_{pu}(C_{b,bl},\phi_{f,bl}) - k_{clr,f}(C_{f,bl},\phi_{f,bl}) + k_{Ex}(C_{f,p},\phi_{f,bl}) - k_{Ex}(C_{f,bl},\phi_{f,bl}) - R_{bl}(1-\sigma_f)(C_{f,bl},\phi_{f,bl}), \tag{34}$$

$$(\frac{dC_{b,bl}}{dt},\phi_{b,bl}) = k_{pb}(C_{f,bl},\phi_{b,bl}) - k_{pu}(C_{b,bl},\phi_{b,bl}), \tag{35}$$

$$(\frac{\partial C_{n,p}}{\partial t},\phi_{n,p}) = -(D^*_{n,p}\nabla C_{n,p},\nabla\phi_{n,p}) + \int_{\partial\Omega_p} D^*_{n,p}\nabla C_{n,p}\cdot\phi_{n,p}ds - \int_{\partial\Omega_p}\mathbf{v}C_{n,p}\cdot\phi_{n,p}ds - k_{rel,p}(C_{n,p},\phi_{n,p}) - R_{ly}(C_{n,p},\phi_{n,p}) - \frac{D_vE_2^2}{\eta}\int_0^t e^{-\frac{E_2}{\eta}(t-s)}(\nabla C_{n,p},\nabla\phi_{n,p})ds + (\mathbf{v}C_{n,p},\nabla\phi_{n,p}) + \frac{D_vE_2^2}{\eta}\int_{\partial\Omega_p}\int_0^t e^{-\frac{E_2}{\eta}(t-s)}(\nabla C_{n,p},\phi_{n,p})d, \tag{36}$$

$$(\frac{\partial C_{f,p}}{\partial t},\phi_{f,p}) = -(D^*_{f,p}\nabla C_{f,p},\nabla\phi_{f,p}) + (\mathbf{v}C_{f,p},\nabla\phi_{f,p}) + k_{pu}(C_{b,p},\phi_{f,p}) - \int_{\partial\Omega_p}\mathbf{v}C_{f,p}\cdot\phi_{f,p}ds + k_{rel,p}(C_{n,p},\phi_{f,p}) - R_{ly}(C_{f,p},\phi_{f,p})$$

$$
\begin{aligned}
&-\frac{D_v E_2^2}{\eta}\int_0^t e^{-\frac{E_2}{\eta}(t-s)}(\nabla C_{f,p},\nabla\phi_{f,p})ds - k_{Ex}(C_{f,p},\phi_{f,p})\\
&+\frac{D_v E_2^2}{\eta}\int_{\partial\Omega_p}\int_0^t e^{-\frac{E_2}{\eta}(t-s)}(\nabla C_{f,p},\phi_{f,p})ds + k_{Ex}(C_{f,bl},\phi_{f,p})\\
&+R_{bl}(1-\sigma_f)(C_{f,bl},\phi_{f,p}) - k_{d,p}(C_{f,p},\phi_{f,p}) - k_{pb}(C_{f,p},\phi_{f,p})\\
&+\int_{\partial\Omega_p} D^*_{f,p}\nabla C_{f,p}\cdot\phi_{f,p}ds.
\end{aligned}
\tag{37}
$$

By taking $\phi_{i,j}=C_{i,j}$ for $i=n,f,b$ and $j=c,e,r,p,bl$, and summing up the equations, and considering the boundary and interface boundary conditions, we obtain

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\sum_{\substack{i=n,f\\ j=c,p,e,r,bl}}\|C_{i,j}\|^2_{L^2(\Omega_j)} + \frac{1}{2}\frac{d}{dt}\sum_{i=p,bl}\|C_{b,i}\|^2_{L^2(\Omega_p)}\\
&+\sum_{\substack{i=n,f\\ j=c,e,r}} D_{i,j}\|\nabla C_{i,j}\|^2_{[L^2(\Omega_j)]^2} + \sum_{i=n,f} D^*_{i,p}\|\nabla C_{i,p}\|^2_{[L^2(\Omega_p)]^2}\\
&+\sum_{i=c,e,p,r} k_{d,i}\|C_{f,i}\|^2_{L^2(\Omega_i)} + \sum_{i=e,p,r,bl} k_{rel,i}\|C_{n,i}\|^2_{L^2(\Omega_i)}\\
&+\sum_{i=n,f} R_{ly}\|C_{i,p}\|^2_{L^2(\Omega_p)} + \sum_{i=p,bl} k_{pb}\|C_{f,i}\|^2_{L^2(\Omega_p)} + R_{bl}(1-\sigma_f)\|C_{f,bl}\|^2_{L^2(\Omega_p)}\\
&+\sum_{i=n,f} k_{clr,i}\|C_{i,bl}\|^2_{L^2(\Omega_p)} + \sum_{i=p,bl} k_{Ex}\|C_{f,i}\|^2_{L^2(\Omega_p)} + \sum_{i=p,bl} k_{pu}\|C_{b,i}\|^2_{L^2(\Omega_p)}\\
&+\sum_{\substack{i=c,p\\ j=p,e,r\\ k=n,f}} \iota_{i,j}\|[C_{k,i,j}]\|^2_{L^2(\Gamma_{i,j})}\\
&\quad=\sum_{\substack{i=n,f\\ j=c,p,e,r}}\big(\mathbf{v}C_{i,j},\nabla C_{i,j}\big) + 2k_{Ex}\big(C_{f,p},C_{f,bl}\big)\\
&\qquad+R_{bl}(1-\sigma_f)\big(C_{f,bl},C_{f,p}\big) + \sum_{i=p,bl} k_{pb}\big(C_{b,i},C_{f,i}\big)\\
&\qquad+\sum_{i=p,e,r,bl} k_{rel,i}\big(C_{n,i},C_{f,i}\big) + \sum_{i=p,bl} k_{pu}\big(C_{b,i},C_{f,i}\big)\\
&\qquad-\sum_{i=n,f}\frac{D_v E_2^2}{\eta}\int_0^t e^{-\frac{E_2}{\eta}(t-s)}(\nabla C_{i,p}(s),\nabla C_{i,p}(t))ds.
\end{aligned}
\tag{38}
$$

Note that we consider $\sigma(0)$, $C_{n,p}(0)$ and $C_{f,p}(0)$ as constants. In (38), the symbol $[C_{k,i,j}]=(C_{k,i}-C_{k,j})$ represents the exchange of concentrations of nanocarrier and drug at the interface boundary between the two domains. It is shown with this symbol for convenience. Only ordered

pairs of $(c,p),(p,e)$, and (p,r) are permitted in the final term on the left side of (38). It is important to note that in (38), wherever $L^2(\Omega_{bl})$ appears, it means $L^2(\Omega_p)$.

Use Cauchy's inequality with epsilon to get

$$\sum_{\substack{i=n,f\\ j=c,p,e,r}} (\mathbf{v}C_{i,j}, \nabla C_{i,j}) \leq \sum_{\substack{i=n,f\\ j=c,p,e,r}} \big(\frac{1}{4\epsilon_{i,j}^2}\|\mathbf{v}\|_\infty^2 \|C_{i,j}\|^2_{L^2(\Omega_j)} + \epsilon_{i,j}^2 \|\nabla C_{i,j}\|^2_{[L^2(\Omega_j)]^2}\big),$$

$$\sum_{i=p,e,r,bl} k_{rel,i}\big(C_{n,i}, C_{f,i}\big) \leq \sum_{i=p,e,r,bl} \big(\frac{k_{rel,i}}{4\epsilon_{rel,i}^2}\|C_{n,i}\|^2_{L^2(\Omega_i)} + \epsilon_{rel,i}^2 \|C_{f,i}\|^2_{L^2(\Omega_i)}\big),$$

$$\sum_{i=p,bl} k_{pu}\big(C_{b,i}, C_{f,i}\big) \leq \sum_{i=p,bl} \big(\frac{k_{pu}}{4\epsilon_{pu,i}^2}\|C_{b,i}\|^2_{L^2(\Omega_p)} + \epsilon_{pu,i}^2 \|C_{f,i}\|^2_{L^2(\Omega_p)}\big),$$

$$\sum_{i=p,bl} k_{pb}\big(C_{b,i}, C_{f,i}\big) \leq \sum_{i=p,bl} \big(\frac{k_{pb}}{4\epsilon_{pb,i}^2}\|C_{b,i}\|^2_{L^2(\Omega_p)} + \epsilon_{pb,i}^2 \|C_{f,i}\|^2_{L^2(\Omega_p)}\big),$$

$$2k_{Ex}\big(C_{f,p}, C_{f,bl}\big) \leq \frac{k_{Ex}}{2\epsilon_{Ex}^2}\|C_{f,p}\|^2_{L^2(\Omega_p)} + \epsilon_{Ex}^2 \|C_{f,bl}\|^2_{L^2(\Omega_p)},$$

$$R_{bl}(1-\sigma_f)\big(C_{f,bl}, C_{f,p}\big) \leq \frac{R_{bl}(1-\sigma_f)}{4\epsilon_{bl}^2}\|C_{f,p}\|^2_{L^2(\Omega_p)} + \epsilon_{bl}^2 \|C_{f,bl}\|^2_{L^2(\Omega_p)},$$

$$\begin{aligned} &-\sum_{i=n,f} \frac{D_v E_2^2}{\eta}\int_0^t e^{-\frac{E_2}{\eta}(t-s)}(\nabla C_{i,p}(s), \nabla C_{i,p}(t))ds \\ &\quad\leq \sum_{i=n,f}\big(\epsilon_{i,\sigma}^2\|\nabla C_{i,p}(t)\|^2 + \frac{D_v^2 E_2^3}{8\epsilon_{i,\sigma}^2\eta}\int_0^t \|\nabla C_{i,p}(s)\|^2 ds\big). \end{aligned}$$

Multiplying (38) by two and using above inequalities, we obtain

$$\frac{d}{dt}\sum_{\substack{i=n,f\\ j=c,p,e,r,bl}} \|C_{i,j}\|^2_{L^2(\Omega_j)} + \frac{d}{dt}\sum_{i=p,bl}\|C_{b,i}\|^2_{L^2(\Omega_p)} + \sum_{i=n,f} 2R_{ly}\|C_{i,p}\|^2_{L^2(\Omega_p)}$$

$$
\begin{aligned}
&+\sum_{\substack{i=n,f\\ j=c,e,r}} 2\big(D_{i,j}-\epsilon_{i,j}^2-\epsilon_{i,\sigma}^2\big)\|\nabla C_{i,j}\|^2_{[L^2(\Omega_j)]^2}+\sum_{i=e,p,r,bl} 2k_{rel,i}\|C_{n,i}\|^2_{L^2(\Omega_i)}\\
&+\sum_{i=n,f} 2\big(D^*_{i,p}-\epsilon_{i,p}^2-\epsilon_{i,\sigma}^2\big)\|\nabla C_{i,p}\|^2_{[L^2(\Omega_p)]^2}+\sum_{i=c,e,p,r} 2k_{d,i}\|C_{f,i}\|^2_{L^2(\Omega_i)}\\
&+\sum_{i=p,bl} 2k_{pb}\|C_{f,i}\|^2_{L^2(\Omega_p)}+\sum_{i=p,bl} 2k_{pu}\|C_{b,i}\|^2_{L^2(\Omega_p)}\\
&+\sum_{i=n,f} 2k_{clr,i}\|C_{i,bl}\|^2_{L^2(\Omega_p)}+\sum_{i=p,bl} 2k_{Ex}\|C_{f,i}\|^2_{L^2(\Omega_p)}\\
&+2R_{bl}(1-\sigma_f)\|C_{f,bl}\|^2_{L^2(\Omega_p)}+\sum_{\substack{i=c,p\\ j=p,e,r\\ k=n,f}} 2\iota_{i,j}\|[C_{k,i,j}]\|^2_{L^2(\Gamma_{i,j})}\\
&\le \sum_{\substack{i=n,f\\ j=c,p,e,r}} \frac{1}{2\epsilon_{i,j}^2}\|\mathbf{v}\|^2_\infty\|C_{i,j}\|^2_{L^2(\Omega_j)}+\sum_{i=p,e,r,bl}\frac{k_{rel,i}}{2\epsilon^2_{rel,i}}\|C_{n,i}\|^2_{L^2(\Omega_i)}\\
&\quad+\sum_{i=p,e,r,bl} 2\epsilon^2_{rel,i}\|C_{f,i}\|^2_{L^2(\Omega_i)}+\sum_{i=p,bl}\Big(\frac{k_{pu}}{2\epsilon^2_{pu,i}}+\frac{k_{pb}}{2\epsilon^2_{pb,i}}\Big)\|C_{b,i}\|^2_{L^2(\Omega_p)}\\
&\quad+\sum_{i=p,bl} 2(\epsilon^2_{pu,i}+\epsilon^2_{pb,i})\|C_{f,i}\|^2_{L^2(\Omega_p)}+\Big(\frac{k_{Ex}}{\epsilon^2_{Ex}}+\frac{R_{bl}(1-\sigma_f)}{2\epsilon^2_{bl}}\Big)\|C_{f,p}\|^2_{L^2(\Omega_p)}\\
&\quad+2(\epsilon^2_{Ex}+\epsilon^2_{bl})\|C_{f,bl}\|^2_{L^2(\Omega_p)}+\sum_{i=n,f}\frac{D_v^2E_2^3}{4\epsilon^2_{i,\sigma}\eta}\int_0^t\|\nabla C_{i,p}(s)\|^2_{L^2(\Omega_p)}ds. \qquad (39)
\end{aligned}
$$

Fixing in (39) $\epsilon_{i,j}$ such that $(D_{i,j}-\epsilon^2_{i,j}-\epsilon^2_{i,\sigma})>0$ and $(D^*_{i,p}-\epsilon^2_{i,p}-\epsilon^2_{i,\sigma})>0$ for $i=n,f$ and $j=c,e,r$.

By integrating from (39) and using the Gronwall lemma, we get Theorem 1.

Theorem 1. If $\mathcal{C}=(C_{i,j})$ for $i=n,f,b$ and $j=c,p,e,r,bl$ is a solution of the variational formulation of (3)–(14), respectively, then a positive constant Φ that is independent of time exists such that

$$
\varepsilon(\mathcal{C}(t))\le\frac{1}{\Phi_{\min}}\varepsilon(\mathcal{C}(0))e^{\Phi t}, \qquad (40)
$$

where

$$
\begin{aligned}
\mathcal{C}=&\sum_{\substack{i=n,f\\ j=c,p,e,r,bl}}\|C_{i,j}\|^2_{L^2(\Omega_j)}+\sum_{i=p,bl}\|C_{b,i}\|^2_{L^2(\Omega_p)} \qquad (41)\\
&+\int_0^t\sum_{\substack{i=n,f\\ j=c,p,e,r}}\|\nabla C_{i,j}\|^2_{[L^2(\Omega_j)]^2}ds+\int_0^t\sum_{\substack{i=n,f\\ j=p,e,r}}\|C_{i,j}\|^2_{L^2(\Omega_j)}ds
\end{aligned}
$$

$$+\int_0^t \sum_{i=p,bl} \|C_{b,i}\|^2_{L^2(\Omega_p)} ds + \int_0^t \|C_{f,bl}\|^2_{L^2(\Omega_p)} ds$$
$$+\int_0^t \|C_{f,c}\|^2_{L^2(\Omega_c)} ds + \int_0^t \sum_{\substack{i=c,p \\ j=p,e,r \\ k=n,f}} 2\iota_{i,j} \|[C_{k,i,j}]\|^2_{L^2(\Gamma_{i,j})} ds,$$

$$\Phi = \frac{\Phi_{Max}}{\Phi_{\min}}, \tag{42}$$

$$\begin{aligned}\Phi_{\min} = \min \Big\{ & 1, 2\big(D_{i,j} - \epsilon^2_{i,j} - \epsilon^2_{i,\sigma}\big)_{\substack{i=n,f \\ j=c,e,r}}, 2\big(D^*_{i,p} - \epsilon^2_{i,p} - \epsilon^2_{i,\sigma}\big)_{i=n,f}, \\ & (2\iota_{i,j})_{\substack{i=c,p \\ j=p,e,r}}, 2(k_{rel,p} + R_{ly}), 2k_{rel,e}, 2k_{rel,r}, \\ & 2(k_{rel,bl} + k_{clr,bl}), 2(k_{d,p} + R_{ly} + k_{pb} + k_{Ex}), \\ & 2k_{d,e}, 2k_{d,r}, 2k_{d,c}, 2k_{pu}, \\ & 2(k_{pb} + k_{clr,bl} + k_{Ex} + R_{bl}(1 - \sigma_f))\Big\},\end{aligned}$$

and

$$\begin{aligned}\Phi_{Max} = \max \Big\{ & \Big(\frac{1}{2\epsilon^2_{i,j}} \|\mathbf{v}\|^2_\infty\Big)_{\substack{i=n,f \\ j=c,p,e,r}}, \Big(\frac{k_{rel,i}}{2\epsilon^2_{rel,i}}\Big)_{i=p,e,r}, \Big(\frac{D_v^2 E_2^3}{4\epsilon^2_{i,\sigma}\eta}\Big)_{i=n,f}, \\ & (2\epsilon^2_{rel,i})_{i=p,e,r}, \Big(\frac{k_{pu}}{2\epsilon^2_{pu,i}} + \frac{k_{pb}}{2\epsilon^2_{pb,i}}\Big)_{i=p,bl}, \\ & 2(\epsilon^2_{pu,i} + \epsilon^2_{pb,i})_{i=p,bl}, \Big(\frac{k_{Ex}}{\epsilon^2_{Ex}} + \frac{R_{bl}(1-\sigma_f)}{2\epsilon^2_{bl}}\Big), 2(\epsilon^2_{Ex} + \epsilon^2_{bl})\Big\}.\end{aligned}$$

From Theorem 1, we can conclude that the mathematical model of the problem is stable in a finite time interval. Also, the existence of a solution to weak problem (26)–(37) is a direct consequence of the energy estimate (40).

4.2 Uniqueness of the solution

In this section, we examine the uniqueness of the fully coupled system (fluid flow and drug transport). Two distinct solutions, $(\hat{\mathbf{v}}, \hat{C}_{i,j}, \hat{\sigma})$ and $(\tilde{\mathbf{v}}, \tilde{C}_{i,j}, \tilde{\sigma})$ for $i = n, f, b$ and $j = c, e, p, r, bl$, are examined for (1)–(2), (3)–(6), (9)–(14), (16)–(17) and (18). After subtracting them, use the test functions $\phi_{\mathbf{v}}$, $\phi_{i,j}$ and ϕ_σ to create a weak form. In the equations, enter $\phi_{\mathbf{v}} = w_{\mathbf{v}}$, $\phi_{i,j} = w_{i,j}$

and $\phi_\sigma = w_\sigma$, where $w_{\mathbf{v}} = \hat{\mathbf{v}} - \tilde{\mathbf{v}}$, $w_{i,j} = \hat{C}_{i,j} - \tilde{C}_{i,j}$ and $w_\sigma = \hat{\sigma} - \tilde{\sigma}$, then integrate by part to obtain

$$\begin{aligned}
&\sum_{\substack{i=n,f\\ j=c,p,e,r,bl}} (\frac{\partial w_{i,j}}{\partial t}, w_{i,j}) + 2(\frac{\partial w_\sigma}{\partial t}, w_\sigma) + (\frac{\partial w_{\mathbf{v}}}{\partial t}, w_{\mathbf{v}}) \\
&= - \sum_{\substack{i=n,f\\ j=c,p,e,r}} (D_{i,j}\nabla w_{i,j}, \nabla w_{i,j}) + \sum_{\substack{i=n,f\\ j=c,p,e,r}} (\mathbf{v} w_{i,j}, \nabla w_{i,j}) \\
&\quad + \sum_{\substack{i=n,f\\ j=c,p,e,r}} (w_{\mathbf{v}}\tilde{c}_{i,j}, \nabla w_{i,j}) + \sum_{i=n,f} \big((E_1+E_2)(\frac{\partial w_{i,p}}{\partial t}, w_\sigma) + \frac{E_1E_2}{\eta}(w_{i,p}, w_\sigma)\big) \\
&\quad - \sum_{i=c,e,p,r} k_{d,i}(w_{f,i}, w_{f,i}) - \sum_{i=e,p,r,bl} k_{rel,i}(w_{n,i}, w_{n,i}) - \sum_{i=n,f} R_{ly}(w_{i,p}, w_{i,p}) \\
&\quad - \sum_{i=p,bl} k_{pb}(w_{f,i}, w_{f,i}) - \sum_{i=p,bl} k_{pu}(w_{b,i}, w_{b,i}) - \sum_{i=n,f} k_{clr,i}(w_{i,bl}, w_{i,bl}) \\
&\quad - \sum_{i=p,bl} k_{Ex}(w_{f,i}, w_{f,i}) - R_{bl}(1-\sigma_f)(w_{f,bl}, w_{f,bl}) - \sum_{i=p,e,r,bl} k_{rel,i}\big(w_{n,i}, w_{f,i}\big) \\
&\quad - \sum_{i=p,bl} (k_{pu}+k_{pb})\big(w_{b,i}, w_{f,i}\big) - (2k_{Ex} + R_{bl}(1-\sigma_f))\big(w_{f,p}, w_{f,bl}\big) \\
&\quad -(w_{\mathbf{v}}.\nabla\tilde{\mathbf{v}}, w_{\mathbf{v}}) - \frac{\mu}{\rho}(\nabla w_{\mathbf{v}}, \nabla w_{\mathbf{v}}) - \sum_{i=n,f} D_v(\nabla w_\sigma, \nabla w_{i,p}) - \frac{2E_2}{\eta}(w_\sigma, w_\sigma).
\end{aligned}$$

Using

$$(f,g) \le \frac{1}{2\epsilon}\|f\|^2 + \frac{\epsilon}{2}\|g\|^2,$$

and

$$|\int_\Omega (w_{\mathbf{v}}.\nabla\tilde{\mathbf{v}}) w_{\mathbf{v}} ds| \le \|w_{\mathbf{v}}\|^2_{L^4}\|\nabla\tilde{\mathbf{v}}\|^2_{L^2} \le A\|w_{\mathbf{v}}\|_{L^2}\|\nabla w_{\mathbf{v}}\|_{L^2}\|\nabla\tilde{\mathbf{v}}\|^2_{L^2},$$

we get

$$\begin{aligned}
&\sum_{\substack{i=n,f\\ j=c,p,e,r,bl}} \frac{\partial}{\partial t}\|w_{i,j}\|^2_{L^2(\Omega_j)} + \frac{\partial}{\partial t}\|w_{\mathbf{v}}\|^2_{L^2(\Omega_p)} + 2\frac{\partial}{\partial t}\|w_\sigma\|^2_{L^2(\Omega_p)} \\
&\le \sum_{\substack{i=n,f\\ j=c,e,r}} \big(\epsilon_{i,j} + \epsilon_{v_{i,j}} - 2D_{i,j}\big)\|\nabla w_{i,j}\|^2_{[L^2(\Omega_j)]^2} + \sum_{i=n,f} \epsilon_{1_{i,p}}\|\nabla w_\sigma\|^2 \\
&\quad + \sum_{i=n,f} \big(\epsilon_{i,p} + \epsilon_{v_{i,p}} - 2D_{i,p} + \frac{D_v}{\epsilon_{1_{i,p}}}\big)\|\nabla w_{i,p}\|^2_{[L^2(\Omega_p)]^2} \\
&\quad +(\epsilon_v - \frac{2\mu}{\rho})\|\nabla w_{\mathbf{v}}\|^2_{[L^2(\Omega_j)]^2} - \sum_{i=c,e,p,r} 2k_{d,i}\|w_{f,i}\|^2_{L^2(\Omega_i)}
\end{aligned}$$

$$
\begin{aligned}
&- \sum_{i=e,p,r,bl} 2k_{rel,i}\|w_{n,i}\|^2_{L^2(\Omega_i)} - \sum_{i=n,f} 2R_{ly}\|w_{i,p}\|^2_{L^2(\Omega_p)} \\
&- \sum_{i=p,bl} 2k_{pb}\|w_{f,i}\|^2_{L^2(\Omega_p)} - \sum_{i=p,bl} 2k_{pu}\|w_{b,i}\|^2_{L^2(\Omega_p)} - \sum_{i=n,f} 2k_{clr,i}\|w_{i,bl}\|^2_{L^2(\Omega_p)} \\
&- \sum_{i=p,bl} 2k_{Ex}\|w_{f,i}\|^2_{L^2(\Omega_p)} - 2R_{bl}(1-\sigma_f)\|w_{f,bl}\|^2_{L^2(\Omega_p)} \\
&+ \sum_{\substack{i=n,f \\ j=c,p,e,r}} \big(\frac{1}{\epsilon_{i,j}}\|\mathbf{v}\|^2_{\infty}\|w_{i,j}\|^2_{L^2(\Omega_j)} + \frac{1}{\epsilon_{v_{i,j}}}\|\tilde{C}_{i,j}\|^2\|w_{\mathbf{v}}\|^2\big) \qquad (43) \\
&+ \sum_{i=p,e,r,bl} \frac{k_{rel,i}}{\epsilon_{rel,i}}\|w_{n,i}\|^2_{L^2(\Omega_i)} + \sum_{i=p,e,r,bl} 2\epsilon_{rel,i}\|w_{f,i}\|^2_{L^2(\Omega_i)} \\
&+ \sum_{i=p,bl} \big(\frac{k_{pu}}{\epsilon_{pu,i}} + \frac{k_{pb}}{\epsilon_{pb,i}}\big)\|w_{b,i}\|^2_{L^2(\Omega_p)} + \sum_{i=p,bl} 2(\epsilon_{pu,i}+\epsilon_{pb,i})\|w_{f,i}\|^2_{L^2(\Omega_p)} \\
&+ \frac{2k_{Ex}+R_{bl}(1-\sigma_f)}{\epsilon_{bl}}\|w_{f,p}\|^2_{L^2(\Omega_p)} + 2\epsilon_{bl}\|w_{f,bl}\|^2_{L^2(\Omega_p)} + \frac{A^2}{\epsilon_v}\|w_{\mathbf{v}}\|^2\|\nabla\tilde{\mathbf{v}}\|^2 \\
&+ \sum_{i=n,f} \big(\epsilon_{2_{i,p}}\|\frac{\partial w_{i,p}}{\partial t}\|^2 + \frac{(E_1+E_2)}{\epsilon_{2_{i,p}}}\|w_\sigma\|^2 + \epsilon_{3_{i,p}}\|w_{i,p}\|^2 + \frac{E_1E_2}{\eta\epsilon_{3_{i,p}}}\|w_\sigma\|^2\big)
\end{aligned}
$$

The $\epsilon_i s$ constants can be chosen to be negative $\big(\epsilon_{i,j}+\epsilon_{v_{i,j}}-2D_{i,j}\big)$, $\big(\epsilon_{i,p}+\epsilon_{v_{i,p}}-2D_{i,p}+\frac{D_v}{\epsilon_{1_{i,p}}}\big)$ and $(\epsilon_v - \frac{2\mu}{\rho})$, and the $\epsilon_{1_{i,p}}$ and $\epsilon_{2_{i,p}}$ coefficients can be ignored in (43). Therefor there exists $\alpha \in L^2(0,T)$ such that

$$
\begin{aligned}
&\frac{\partial}{\partial t}\big(\sum_{\substack{i=n,f,b \\ j=c,p,e,r,bl}} \|w_{i,j}\|^2 + \|w_{\mathbf{v}}\|^2 + \|w_\sigma\|^2\big) \\
&\leq \alpha(t)\big(\sum_{\substack{i=n,f,b \\ j=c,p,e,r,bl}} \|w_{i,j}\|^2 + \|w_{\mathbf{v}}\|^2 r + \|w_\sigma\|^2\big).
\end{aligned}
$$

In other words, $\|w_{\mathbf{v}}\|$, $\|w_\sigma\|$ and $\|w_{i,j}\|$ $i = n, f, b$ $j = c, p, e, r, bl$ are obviously zero at $= 0$. We can thus prove the uniqueness of the solution by applying the Gronwall lemma, which yields $\|w_{\mathbf{v}}\|^2 = \|w_\sigma\|^2 = \|w_{i,j}\|^2 = 0$ $i = n, f, b$ $j = c, p, e, r, bl$.

5 Numerical simulations

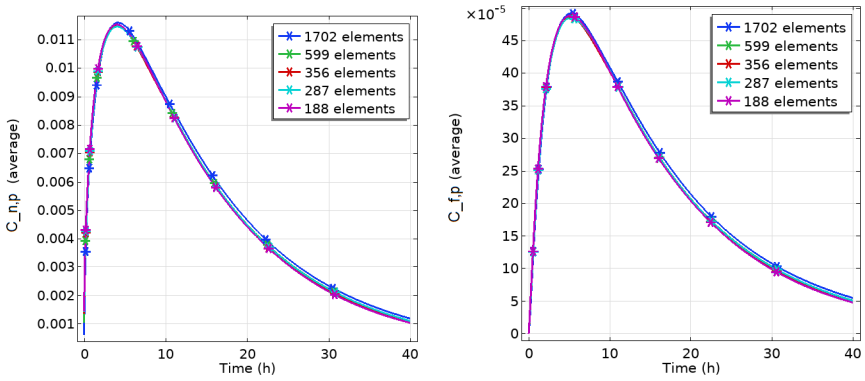
5.1 Numerical methods and independence of the solution on the mesh

Equations (1)–(21) were solved in the geometry shown in Figure 1b. The numerical results presented in what follows are obtained using COMSOL Multiphysics 5.6. The piecewise linear finite element space P_1 is used for the concentration and P_2 for flow equations, and the backward differential formula with order between 1 and 2 with time step length 0.01 is used to take the integral in time. A triangular mesh was created automatically. Table 2 provides the values used for the parameters in the numerical simulation.

Table 2: Values of the parameters for drug properties.

Parameter and Symbol	Values	Units	Reference
Diffusivity in CH($D_{f,c}, D_{n,c}$)	3.8×10^{-10}	m^2/s	[5, 35]
Diffusivity in skin tissues ($D_{i,j}$ $i = n, f$ $j = e, p, r, bl$)	1×10^{-10}	m^2/s	[5, 23]
Density(ρ)	1000	kg/m^3	[5]
Drug degradation rate(k_d)	5.6×10^{-6}	$1/s$	[5, 33]
Drug release rate(k_{rel})	1×10^{-4}	$1/s$	[5]
Drug association rate with protein(k_{pb})	0.833	$1/s$	[5, 10]
Drug dissociation rate from protein(k_{pu})	0.278	$1/s$	[5, 10]
Constant dependent(k_{Ex})	2.3×10^{-3}	$1/s$	[5]
Rate of fluid exchange(R_{ly})	4.2×10^{-7}	$1/s$	[5]
Rate of fluid exchange(R_{bl})	1.19×10^{-5}	$1/s$	[5]
Rate of plasma clearance(k_{clr})	5×10^{-5}	$1/s$	[5, 23]
Osmotic reflection coefficient of drug(σ_f)	0.15		[5]
Viscosity(μ)	0.78	$Pa.s$	[5]
Stiffness coefficient of the single spring(E_1)	1.2294×10^{-1}	Pa	[4, 15]
Stiffness coefficient of the spring(E_2)	1.7239×10^{-1}	Pa	[4, 15]
Stiffness coefficient of the dashpot(η)	1.0532×10^{-1}	$N/m.s$	[4, 15]
Viscoelastic diffusion coefficient(D_v)	5×10^{-12}	$mol/m^3.s.Pa$	[4, 15]

The various meshes are taken into consideration to ensure that the numerical results are independent of the mesh size. Using convergence tests, the mesh density's adequacy has been examined. For two different meshes, the convergence of the solutions and the mesh independence have been investigated. To better demonstrate this assertion, Figure 2 and Tables 3 and 4 compare various meshes for the average concentrations of nanocarriers and free drug in the PD, respectively. Notably, the reference solution is defined by a mesh consisting of 1702 elements. Figure 2 indicates that the numerical results are independent of mesh size, as the changes between the meshes are nearly indistinguishable.



(a) (b)

Figure 2: Average nanocarrier concentrations (a) and free drug concentrations (b) in the PD to demonstrate that numerical results are independent of meshing.

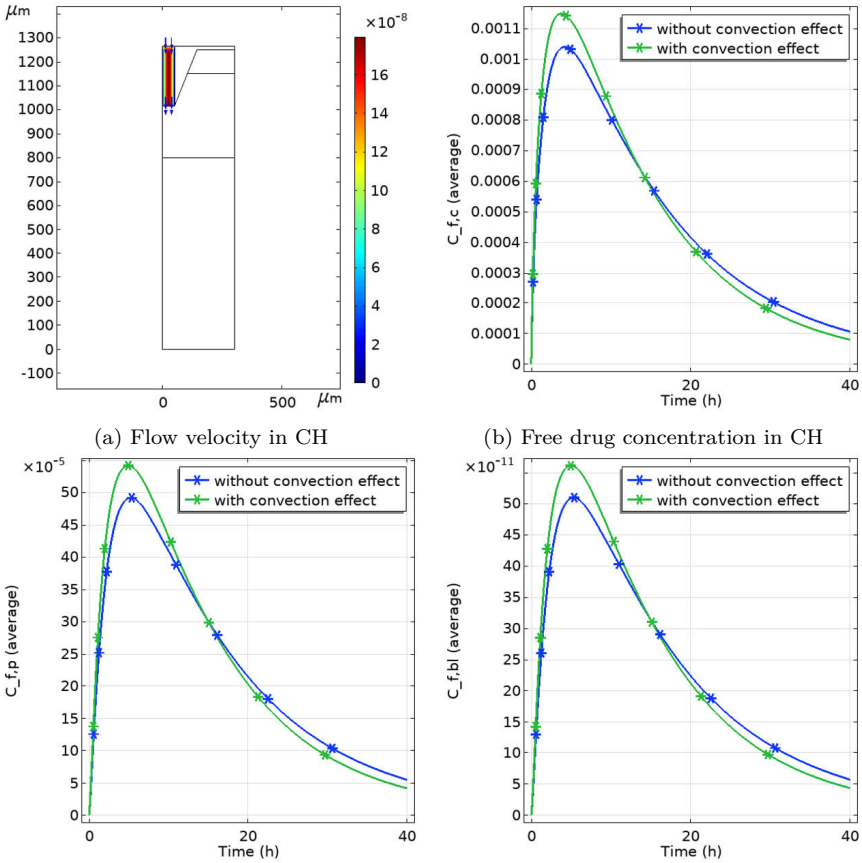
Table 3: Different errors for nanocarriers concentration in PD.

Number of elements	L^∞-norm	L^2-norm
599	4.26×10^{-4}	1.10×10^{-2}
356	9.34×10^{-4}	1.83×10^{-2}
287	9.73×10^{-4}	1.95×10^{-2}
188	1.03×10^{-3}	1.96×10^{-2}

Table 4: Different errors for free drug concentration in PD.

Number of elements	L^∞-norm	L^2-norm
599	1.09×10^{-5}	4.65×10^{-4}
356	1.65×10^{-5}	7.82×10^{-4}
287	1.73×10^{-5}	8.40×10^{-4}
188	1.76×10^{-5}	8.45×10^{-4}

Figure 3a shows a flow velocity diagram in CH, with arrows indicating the flow direction. The average free drug concentrations in CH, PD and BL, computed with and without convective transport, are plotted in Figure 3b,3c, and 3d to show how convection affects drug transport in various domains. Convection has an increasing impact on drug transport, according to these plots.



(a) Flow velocity in CH

(b) Free drug concentration in CH

(c) Free drug concentration in PD

(d) Free drug concentration in BL

Figure 3: Effect of flow velocity on the drug transport

5.2 Convergence study

In what follows, we assume that the domains of $\Omega_i, i = c, e, p, r$ are polygonal lines. Suppose that $\Im_{i,h}, i = c, e, p, r, h \to 0$ is a sequence of admissible triangles from $\Omega_i, i = c, e, p, r$, respectively, with $\max_{\triangle \in \Im_{i,h}} diam(\triangle) \to 0$. We assume that the triangles on the common boundary are

congruent. By $V_{i,h}$ we denote the subspace $H^1(\Omega_i)$ defined by the piecewise linear functions of u_h induced by the triangulation of $\Im_{i,h}$ for $i = c, e, p, r$. Let $C_{i,j,h} \in \prod V_{k,h}$ for $i = n, f, b$, $j = c, e, p, r, bl$ and $k = c, e, p, r$ be the approximate piecewise linear solutions of (3)–(6), (9)–(14) and (16)–(17). Then, considering $e_{i,j} = C_{i,j} - C_{i,j,h}$ for $i = n, f, b$, $j = c, e, p, r, bl$ and following the procedure for proving in section 4.2, we get the following results:

$$\begin{aligned}
&\frac{d}{dt}\Big(\sum_{\substack{i=n,f\\ j=c,p,e,r,bl}} \|e_{i,j}\|^2_{L^2(\Omega_j)} + \sum_{j=p,bl} \|e_{b,j}\|^2_{L^2(\Omega_p)}\Big)\\
&+ \sum_{\substack{i=n,f\\ j=c,e,r,p}} 2\big(D_{i,j} - \epsilon^2_{i,j,1} - \epsilon^2_{i,j,2}\big)\|\nabla e_{i,j}\|^2_{[L^2(\Omega_j)]^2}\\
&\quad\le \sum_{\substack{i=n,f\\ j=c,e,r,p}} \big(\frac{1}{2\epsilon^2_{i,j,1}}\|\mathbf{v}_h\|^2_\infty + A_{i,j}\big)\|e_{i,j}\|^2\\
&\qquad + \sum_{\substack{i=n,f\\ j=c,e,r,p}} \frac{1}{2\epsilon^2_{i,j,2}}\|e_{\mathbf{v}}\|^2_\infty\|c_{i,j}\|^2_\infty + \big(\frac{D_v}{2\epsilon^2_{n,p,3}} + \frac{D_v}{2\epsilon^2_{f,p,6}}\big)\|\nabla e_\sigma\|_{[L^2(\Omega_p)]^2}
\end{aligned} \tag{44}$$

where $A_{f,c} = -2k_{d,c}$, $A_{n,e} = \frac{k_{rel,e}}{2\epsilon^2_{f,e,3}} - 2k_{rel,e}$, $A_{f,e} = 2\epsilon^2_{f,e,3} - 2k_{d,e}$,
$A_{b,p} = \frac{k_{pu}}{2\epsilon^2_{f,p,3}} - 2k_{pu}$, $A_{n,r} = \frac{k_{rel,r}}{2\epsilon^2_{f,r,3}} - 2k_{rel,r}$, $A_{f,r} = 2\epsilon^2_{f,r,3} - 2k_{d,r}$,
$A_{n,bl} = \frac{k_{rel,bl}}{2\epsilon^2_{f,bl}} - 2k_{rel,bl} - 2k_{clr,n}$, $A_{b,bl} = \frac{k_{pu}}{2\epsilon^2_{f,bl,2}} + 2\epsilon^2_{b,bl} - 2k_{pu}$,
$A_{n,p} = \frac{k_{rel,p}}{2\epsilon^2_{f,p,4}} + 2\epsilon^2_{n,p,3} - 2R_{ly} - 2k_{rel,p}$,
$A_{f,p} = \frac{k_{Ex}}{2\epsilon^2_{f,bl,3}} + 2\epsilon^2_{f,p,3} + 2\epsilon^2_{f,p,4} + 2\epsilon^2_{f,p,5} + 2\epsilon^2_{f,p,6} - 2R_{ly} - 2k_{Ex} - 2k_{d,p} - 2k_{pb}$,
$A_{f,bl} = \frac{k_{pb}}{2\epsilon^2_{f,bl}} + \frac{(k_{Ex}+R_{bl}(1-\sigma_f))}{2\epsilon^2_{f,p,5}} + 2\epsilon^2_{f,bl,2} + 2\epsilon^2_{f,bl} + 2\epsilon^2_{b,bl,3} - 2k_{Ex} - 2k_{pb} - 2k_{clr,f} - 2R_{bl}(1-\sigma_f)$.

Put $g(\mathbf{v}_h(t)) = \big(\frac{1}{2\epsilon^2_{i,j,1}}\|\mathbf{v}_h\|^2_\infty + A_{i,j}\big)$, then inequality (44) leads to

$$\begin{aligned}
&\sum_{\substack{i=n,f\\ j=c,p,e,r,bl}} \|e_{i,j}\|^2_{L^2(\Omega_j)} + \sum_{j=p,bl} \|e_{b,j}\|^2_{L^2(\Omega_p)}\\
&+ \int_0^t e^{\int_s^t g(\mathbf{v}_h(\mu))d\mu} \sum_{\substack{i=n,f\\ j=c,e,r,p}} 2\big(D_{i,j} - \epsilon^2_{i,j,1} - \epsilon^2_{i,j,2}\big)\|\nabla e_{i,j}\|^2_{[L^2(\Omega_j)]^2} ds\\
&\quad\le e^{\int_s^t g(\mathbf{v}_h(\mu))d\mu}\Big(\sum_{\substack{i=n,f\\ j=c,p,e,r,bl}} \|e_{i,j}(0)\|^2_{L^2(\Omega_j)} + \sum_{j=p,bl} \|e_{b,j}(0)\|^2_{L^2(\Omega_p)}\Big)\\
&\qquad + \int_0^t e^{\int_s^t g(\mathbf{v}_h(\mu))d\mu}\Big(\sum_{\substack{i=n,f\\ j=c,e,r,p}} \frac{1}{2\epsilon^2_{i,j,2}}\|e_{\mathbf{v}}\|^2_\infty\|c_{i,j}\|^2_\infty
\end{aligned}$$

$$+\Big(\frac{D_v}{2\epsilon_{n,p,3}^2}+\frac{D_v}{2\epsilon_{f,p,6}^2}\Big)\|\nabla e_\sigma\|^2_{[L^2(\Omega_p)]}\Big)ds.$$

The following conclusion can be reached by using the approximation that was employed in [14, 16].

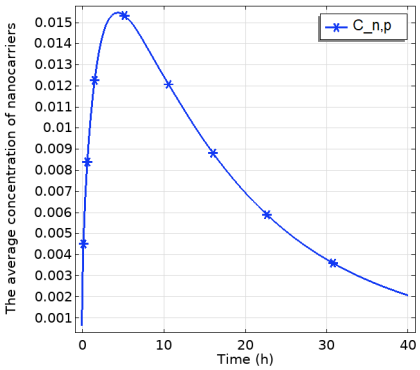
$$\sum_{\substack{i=n,f\\ j=c,p,e,r,bl}}\|e_{i,j}\|^2_{L^2(\Omega_j)}+\sum_{j=p,bl}\|e_{b,j}\|^2_{L^2(\Omega_p)}+\int_0^t\sum_{\substack{i=n,f\\ j=c,p,e,r,bl}}\|\nabla e_{i,j}\|^2ds\leq\mathbb{C}h^4,$$

where $\mathbb{C}$ is the positive constant that is independent of h and t.

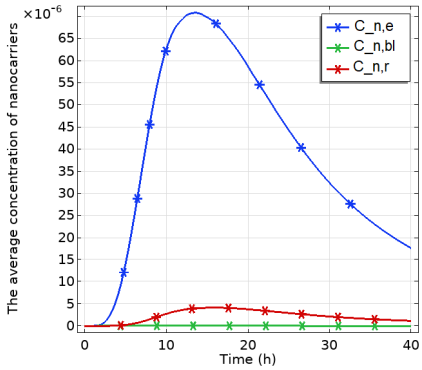
5.3 Numerical simulations and sensitive analysis

Figures 4a and 4b show the concentration of nanocarrier in the circulatory system and skin tissues and how that concentration changes over time. The findings indicate a gradual decrease in this concentration after it first rises in all skin components. The PD has a higher concentration than the VE and RD, respectively, because the needle is located there. The last location of this drug delivery process is the circulatory system, which has the lowest concentration. It is clear from the figure that at the beginning of diffusion, an immediate jump in concentration occurs inside the PD compared to the VE and RD. As shown in Figure 4c, the concentration of free drug is displayed along the skin layers. It is evident, nevertheless, that most of the drug is present in the skin layers as bound drug. Figure 4d illustrates the percentage of the drug that enters the bloodstream as both bound and free.

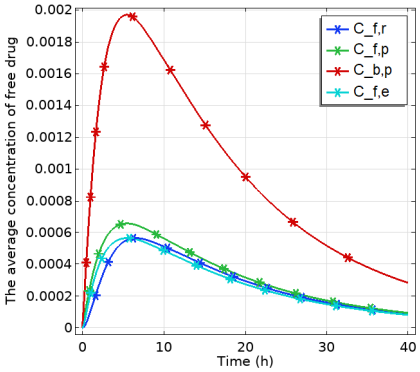
Figures 5 and 6 demonstrate the impact of drug release rate on drug delivery outcomes in skin and blood. It is evident that quick drug release can successfully lower the amount of nanocarriers in BL and all skin layers (Figure 5). In contrast, the concentration of free drug varies. Because nanocarriers penetrate the PD, an increased rate of release enables the drug to diffuse more quickly, which raises the concentration of free drug quickly. However, more free drug can be moved from the PD to the VE and RD due to faster release.



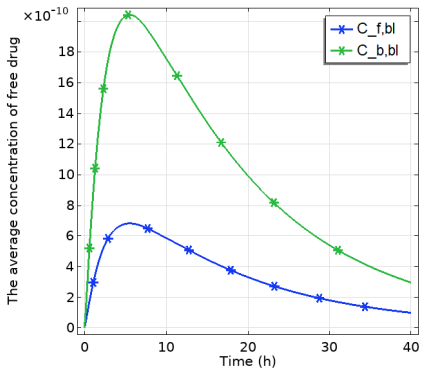
(a) The concentration of nanocarriers in PD



(b) The concentration of nanocarriers in VE, RD, BL



(c) The concentration of released drugs in skin

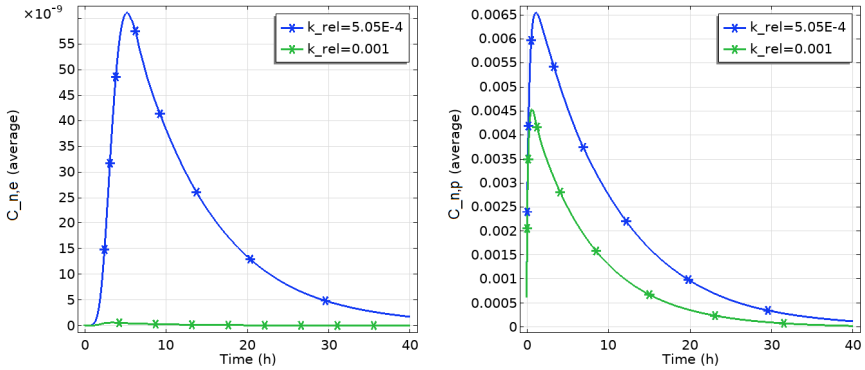


(d) The concentration of released drugs in blood

Figure 4: Drug concentrations in blood and skin tissues over time.

5.4 Dependence on viscoelastic properties and comparing the model

In the previous sections, we assumed a constant value for the viscoelastic diffusion coefficient. In this section, we want to analyze the dependence of the concentrations of the nanocarrier and free drug on D_v. Figure 7 illustrates how the viscoelastic diffusion coefficient D_v affects the release of the nanocarrier and, consequently, the free drug in PD and BL. As the figures demonstrate, we can anticipate less nanocarrier accumulation and, as a result, less free drug accumulation in PD and BL as D_v rises. This is due to the fact that the elastin present in the dermis increases the resistance of the dermis to the entry of the drug through the MN CH. Ultimately, the amount of drug that enters the bloodstream decreases as the viscoelastic diffusion coefficient increases. This important property was not mentioned in [5].



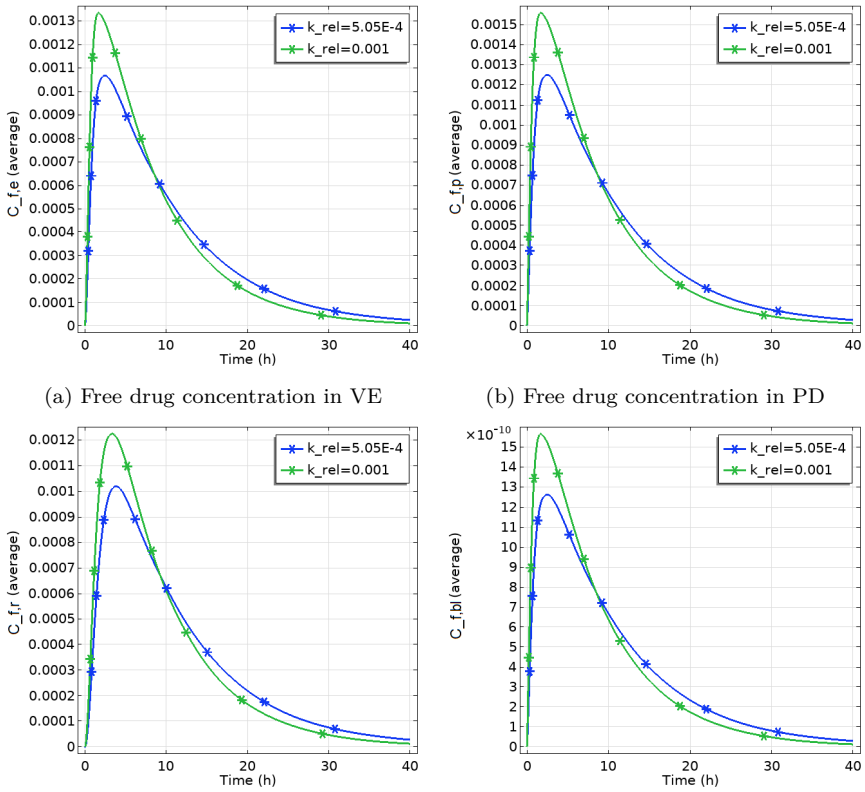
(a) Nanocarriers concentration in VE (b) Nnocarriers concentration in PD

(c) Nanocarriers concentration in RD (d) Nanocarriers concentration in BL

Figure 5: The effect of drug release rate (k_{rel}) on the nanocarriers concentration in skin and BL.

The comparison between the model presented in this paper and Benbrook and Zhan's model presented in [5] is illustrated in Figure 8 and Table 5. Figure 8 shows the average concentration of the drug in the bloodstream. This figure illustrates how the viscoelastic term, which was added to our model to reflect the skin's viscoelastic property, which acts as a barrier, causes the concentration of nanocarriers and free drug to be lower than those in Zhan's model. Table 5 shows a quantitative comparison of nanocarrier and free drug concentrations in PD and BL between our model and Zhan's model. These values were recorded at a set time of 20 hours. From this table, it is clear that the viscoelastic property affects the transport of drugs and acts as a barrier to drug delivery.

To confirm our work and the effectiveness of the model, the results of this model were compared in Figure 9a with the experimental results in [9]. As can be seen, our results are in line with experimental results demonstrating that the model is valid. The viscoelasticity of the skin degrades as a result of diseases and hydration loss brought on by aging [11]. According to



(a) Free drug concentration in VE

(b) Free drug concentration in PD

(c) Free drug concentration in RD

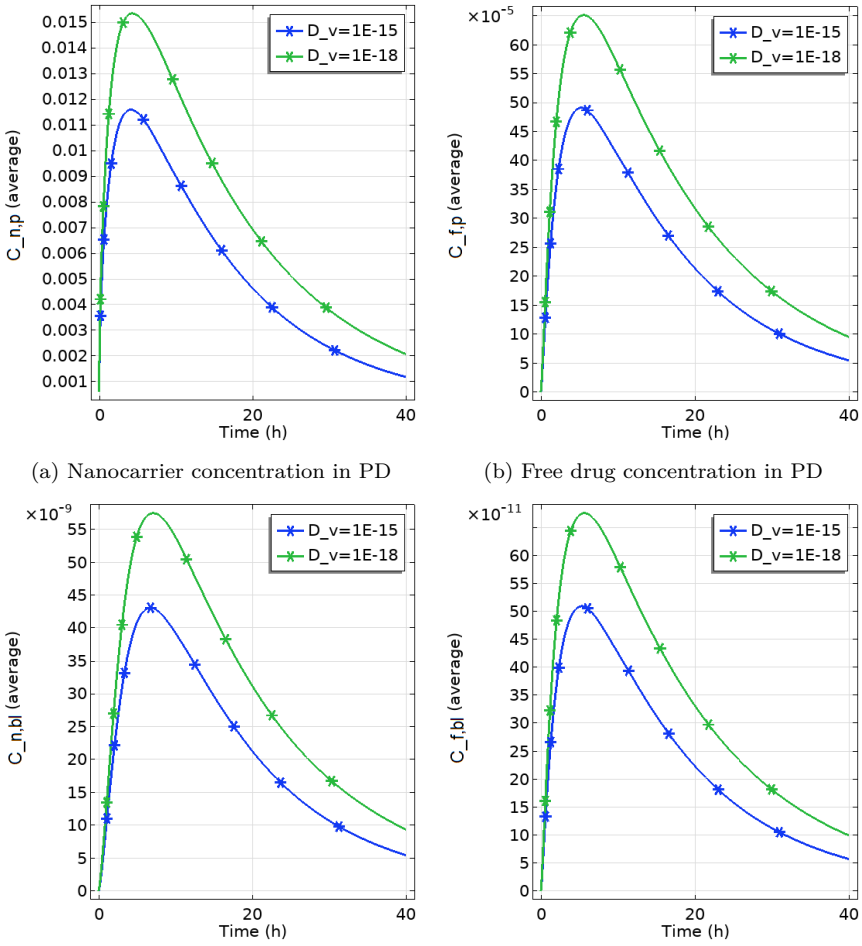
(d) Free drug concentration in BL

Figure 6: The effect of drug release rate (k_{rel}) on the free drug concentration in skin and BL.

[8], which explores how age affects the viscoelastic properties of the skin, three important and influential factors are reduced with age: the immediate contraction of the skin after release of negative pressure, the total contraction after the first cycle, and the ratio of the instantaneous contraction to the maximum deformation, which measures the skin's biological elasticity. This results in a reduction of stiffness in the Zener model. This paper shows that, on average, the parameters decline by 0.6 units between 20 and 60 years of age. To illustrate how drug delivery changes with age, we consider in Figure 9b two different values of stiffness and observe that viscoelastic skin properties decrease with age, leading to increased drug delivery into the skin.

Table 5: Quantitative comparison of our model with Zhan's model at a set time of 20 hours.

	$C_{n,p}$	$C_{f,p}$	$C_{n,bl}$	$C_{f,bl}$
Benbrook and Zhan's model [5]:	0.0069526	3.1846×10^{-4}	3.1249×10^{-8}	3.3212×10^{-10}
Our model:	0.0046546	2.1459×10^{-4}	2.1340×10^{-8}	2.2393×10^{-10}



(a) Nanocarrier concentration in PD

(b) Free drug concentration in PD

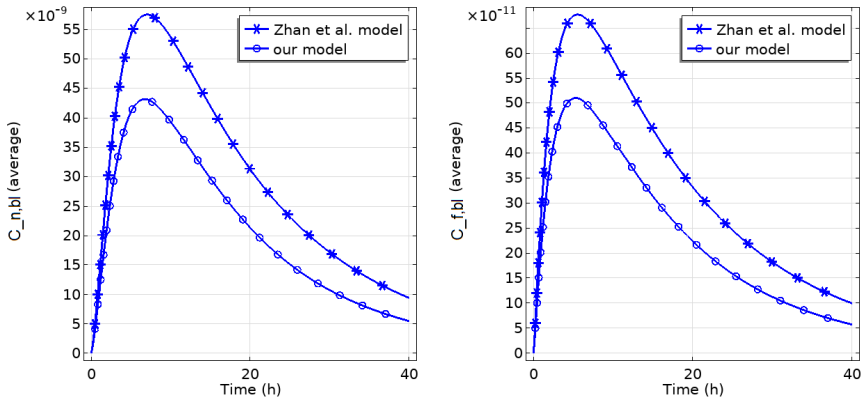
(c) Nanocarrier concentration in BL

(d) Free drug concentration in BL

Figure 7: The impact of the viscoelastic diffusion coefficient (D_v) on the various PD concentrations.

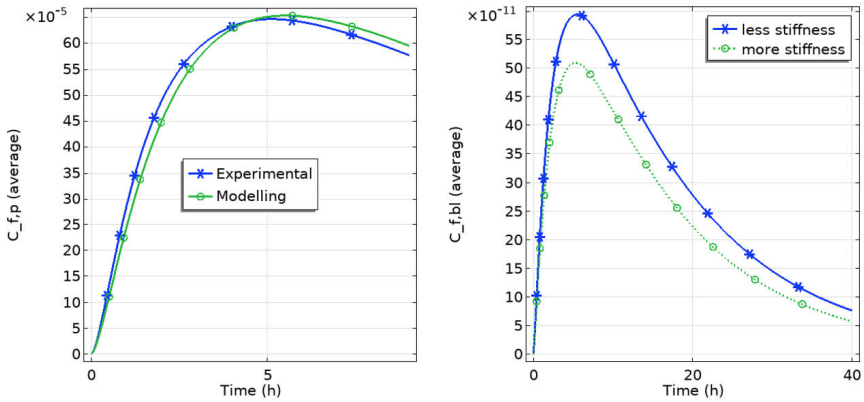
6 Conclusion

A mathematical model incorporating ordinary and partial equations, along with initial and boundary conditions, has been developed to analyze the transport of nanocarriers and drugs via a hollow MN embedded in the skin. It describes drug diffusion into the nanocarrier, free drug, and bound drug states. The objective is to facilitate drug entry into the bloodstream using a nanocarrier positioned in the MN CH. This model accounts for the skin's mechanical characteristics influenced by elastin in the dermis and demonstrates significant effects on drug delivery through numerical results. Some conclusions from this analytical and numerical study are presented.



(a) (b)

Figure 8: Comparison of the nanocarriers concentration (a) and the free drugs concentration (b) in BL in our model with the model presented in [5].



(a) (b)

Figure 9: Model validation. (a) Comparison of model predicted free drug concentration in PD with laboratory measurements. (b) Effect of age, disease state and skin hydration on viscoelastic properties.

- Using energy estimation, the stability of the model solution and its uniqueness can be proven (Theorem 1);
- By taking into account various meshes and comparing the convergence of solutions using the L^2 and L^∞ norms, it is possible to determine the independence of numerical results from mesh size (Figure 2 and Tables 3 and 4);
- Considering the viscoelasticity of the skin due to the presence of elastin in the dermis, we see that it has a significant effect on drug transport. As shown in Figure 7, less drug enters the PD and, eventually, less drug enters the bloodstream as D_v rises (Figure 7);

- The negative effect of the skin properties as a result of diseases and hydration loss brought on by aging on viscoelasticity and thus on drug delivery to skin was investigated (Figure 9b);

- Comparing our model with the model presented by [5] (Figure 8), this difference value is related to considering the viscoelastic effect on the skin (Equations (16) and (17)).

As a future work, we plan to consider the three-dimensional model in real terms so that we can get closer to reality.

Declarations

Conflict of interest There is no conflicting of interest.

Funding

This research received no external funding.

Conflicts of Interest

All authors jointly worked on the results and they read and approved the final manuscript, and also declare that they have no competing interests.

References

[1] Abdo, J.M., Sopko, N.A. and Milner, S.M. *The applied anatomy of human skin: A model for regeneration*. Wound Med., 28 (2020), 100179.

[2] Anissimov, Y.G. and Roberts, M.S. *Modelling dermal drug distribution after topical application in human*. Pharm Res., 28 (2011), 2119–2129.

[3] Azhdari, E. and Emami, A. *Swelling effect of microneedle in the skin drug delivery using Stephan condition: Analytical and numerical study*. Ric. Mat., (2024), 1–28.

[4] Azhdari, E. and Naghipoor, J. *The effect of viscoelasticity of the tissue on the magneto-responsive drug delivery system*. J. Math. Biol., 84(13) (2022), 1–26.

[5] Benbrook, N. and Zhan, W. *Mathematical modelling of hollow microneedle-mediated transdermal drug delivery*. Drug Deliv. Transl. Res., 15 (2025). 3226–3251.

[6] Calcutt, J.J., Roberts, M.S. and Anissimov, Y.G. *Predicting viable skin concentration: Modelling the subpapillary plexus*. Pharm Res., 39 (2022), 783–793.

[7] Cárcamo-Martínez, Á., Mallon, B., Domínguez-Robles, J., Vora, L.K., Anjani, Q.K. and Donnelly, R.F. *Hollow microneedles: A perspective in biomedical applications*. Int. J. Pharm., 599 (2021), 120455.

[8] Chen, D., Yin, S., Lu, X., Fu, H., Gao, H. and Zhang, S. *Research on the correlation between skin elasticity evaluation parameters and age*. Cosmetics, 11(6) (2024), 205.

[9] Davis, S.P., Martanto, W., Allen, M.G. and Prausnitz, M.R. *Hollow metal microneedles for insulin delivery to diabetic rats*. IEEE Trans. Biomed. Eng., 52 (2005), 909–915.

[10] Eikenberry, S. *A tumor cord model for Doxorubicin delivery and dose optimization in solid tumors*. Theor. Biol. Med. Model., 6 (2009), 16.

[11] Everett, J.S. and Sommers, M.S. *Skin viscoelasticity: Physiologic mechanisms, measurement issues, and application to nursing science*. Biol. Res. Nurs., 15(3) (2013), 338–346.

[12] Ferreira, J.A., Grassi, M., Gudino, E. and de Oliveira, P. *A 3d model for mechanistic control drug release*. SIAM J. Appl. Math., 74 (2014), 620–633.

[13] Ferreira, J.A., Grassi, M., Gudiuo, E. and de Oliveira, P. *A new look to non-fickian diffusion*. Appl. Math. Model., 39 (1) (2015), 194–204.

[14] Ferreira, J.A., Jordão, D. and Pinto, L. *Drug delivery enhanced by ultrasound: Mathematical modeling and simulation*. Comput. Math. Appl., 107 (2022), 57–69.

[15] Ferreira, J.A., Naghipoor, J. and de Oliveira, P. *A coupled non-fickian model of a cardiovascular drug delivery system*. Math. Med. Biol., 33(3) (2016), 329–357.

[16] Ferreira, J.A., Pinto, L. and Santos, R. *Numerical analysis of a porous-elastic model for convection enhanced drug delivery*. J. Comput. Appl. Math., 339 (2022), 113719.

[17] Goyal, R., Macri, L.K., Kaplan, H.M. and Kohn, J. *Nanoparticles and nanofibers for topical drug delivery*. J. Control Release, 240 (2016), 77–92.

[18] Jacquet, E., Chambert, J., Pauchot, J. and Sandoz, P. *Intra-and inter-individual variability in the mechanical properties of the human skin from in vivo measurements on 20 volunteers*. Skin Res. Technol., 23 (2017), 491–499.

[19] Lee, H., Song, C., Baik, S., Kim, D., Hyeon, T. and Kim, D-H. *Deviceassisted transdermal drug delivery*. Adv. Drug Deliv. Rev., 45 (2018), 127–135.

[20] Li, M., Vora, L.K., Peng, K. and Donnelly, R.F. *Trilayer microneedle array assisted transdermal and intradermal delivery of dexamethasone*. Int. J. Pharm., 612 (2022), 121295.

[21] Lotfollahi, Z. *The anatomy, physiology and function of all skin layers and the impact of ageing on the skin*. Wound Pract. Res., 32(1) (2024), 6–10.

[22] Machekposhti, S.A., Soltani, M., Najafizadeh, P., Ebrahimi, S.A. and Chen, P. *Biocompatible polymer microneedle for topical/dermal delivery of tranexamic acid*. J. Control Release, 261 (2017), 87–92.

[23] McLean, K. and Zhan, W. *Mathematical modelling of nanoparticle-mediated topical drug delivery to skin tissue*. Int. J. Pharm., 611 (2022), 121322.

[24] Muliana, A. and Rajagopal, K.R. *Modeling the response of nonlinear viscoelastic biodegradable polymeric stents*. Int. J. Solids Struct., 49 (2012), 989–1000.

[25] Nekouzadeh, A., Pryse, K.M., Elson, E.L. and Genin, G.M. *A simplified approach to quasilinear viscoelastic modeling*. J. Biomech., 40 (2007), 3070–3078.

[26] Palacio-Torralba, J., Hammer, S., Good, D.W., McNeill, S.A., Stewart, G.D., Reuben, R.L. and Chen, Y. *Quantitative diagnostics of soft tissue through viscoelastic characterization using time-based instrumented palpation*. J. Mech. Behav. Biomed. Mater., 41 (2015), 149–160.

[27] Phatale, V., Vaiphei, K.K., Jha, S., Patil, D., Agrawal, M. and Alexander, A. *Overcoming skin barriers through advanced transdermal drug delivery approaches*. J. Control Release, 80 (2022), 351–361.

[28] Prausnitz, M.R., Mitragotri, S. and Langer, R. *Current status and future potential of transdermal drug delivery*. Nat. Rev. Drug Discov., 3 (2004), 115–124.

[29] Prausnitz, M.R. and Langer, R. *Transdermal drug delivery*. Nat. Biotechnol., 26 (2008), 1261–1268.

[30] Rippa, A.L., Kalabusheva, E.P. and Vorotelyak, E.A. *Regeneration of dermis: Scarring and cells involved*. Cells, 8(6) (2019), 607.

[31] Soares, J.S. and Zunino, P. *A mixture model for water uptake, degradation, erosion and drug release from polydisperse polymeric networks*. Biomaterials, 31 (2010), 3032–3042.

[32] Wu, Y., Hutton, A.R.J., Pandya, A.K., Patravale, V.B. and Donnelly, R.F. *Microneedle and polymeric films: Delivery of proteins, peptides and nucleic acids*. Drug Deliv. Target., (2023), 93–111.

[33] Wu, D.C. and Ofner, C.M. *Adsorption and degradation of doxorubicin from aqueous solution in polypropylene containers*. AAPS PharmSciTech, 14 (2013), 74–77.

[34] Yadav, P.R., Munni, M.N., Campbell, L., Mostofa, G., Dobson, L., Shittu, M., Pattanayek, S.K., Uddin, v. and Das, v. *Translation of polymeric microneedles for treatment of human diseases: Recent trends, progress, and challenges*. J. Math. Biol., 84(8) (2021), 1132.

[35] Zhan, W. *Mathematical modelling of drug delivery to solid tumour*. Imperial College London, 2014.

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 1, 2026, pp 220–242
https://doi.org/10.22067/ijnao.2025.94069.1670
https://ijnao.um.ac.ir/

CC BY NC

Research Article

An improved search direction based on algebraic equivalent transformation technique for convex quadratic optimization

A. Kraria*, B. Merikhi and D. Benterki

Abstract

This work presents an improved interior point algorithm with full Newton step for convex quadratic optimization. Based on the technique of algebraic equivalent transformation, we first propose a new search direction for convex quadratic optimization with the aim of improving the algorithmic complexity of the proposed algorithm. We then perform a complete theoretical study of convergence and complexity, proving that our algorithm is well-defined, converge quadratically and achieves the best known polynomial complexity bounds established for primal-dual interior point methods. Following this, we conduct comparative numerical

*Corresponding author

Received 15 June 2025; revised 21 October 2025; accepted 22 November 2025

Aicha Kraria
Laboratory of Fundamental and Numerical Mathematics, Department of Mathematics, Faculty of Sciences, Setif 1 University-Ferhat Abbas, Setif, 19000, Algeria. e-mail: aicha.kraria@univ-setif.dz

Bachir Merikhi
Laboratory of Fundamental and Numerical Mathematics, Department of Mathematics, Faculty of Sciences, Setif 1 University-Ferhat Abbas, Setif, 19000, Algeria. e-mail: bmerikhi@univ-setif.dz

Djamel Benterki
Laboratory of Fundamental and Numerical Mathematics, Department of Mathematics, Faculty of Sciences, Setif 1 University-Ferhat Abbas, Setif, 19000, Algeria. e-mail: djbenterki@univ-setif.dz

How to cite this article
Kraria, A. Merikhi, B. and Benterki, D., An improved search direction based on algebraic equivalent transformation technique for convex quadratic optimization. *Iran. J. Numer. Anal. Optim.*, 2026; 16(1): 220-242. https://doi.org/10.22067/ijnao.2025.94069.1670

tests to evaluate the efficiency of the algorithm. The theoretical and numerical results are encouraging and clearly confirm our purpose.

AMS subject classifications (2020): 90C51; 90C20.

Keywords: Interior point algorithm; Convex quadratic programming; Algebraic transformation; Search direction.

1 Introduction

Since the 1980s, interior point methods have seen great development and diffusion in the solution of optimization problems, and primarily for the solution of convex problems, such as linear, convex quadratic, and convex nonlinear programming, due to their theoretical and practical effectiveness. However, for nonconvex problems, several approaches have been developed, such as interior point algorithm for nonconvex optimization (see, e.g., [9, 22, 28]) and difference of convex functions algorithm (DCA) (see, e.g., [1, 20, 25]).

In this work, we are interested in solving a convex quadratic optimization (CQO) problem, using interior point methods.

One of the crucial steps in these methods is calculating the search direction. Several techniques have been developed for this purpose. One is based on the so-called self-regular and kernel barrier functions, which requires two types of iteration: inner iterations and outer iterations, and for each inner iteration an appropriate step size is calculated (see, e.g., [5, 6, 8, 21]). Another technique is based on the so-called algebraic equivalent transformation (AET). Recently, Darvay [10, 11] introduced a new interior point algorithm for linear optimization (LO) based on the AET technique, for the first time. The main idea of this technique is to transform the nonlinear equation in the system that defines the central path by $\psi\left(\frac{xs}{\mu}\right) = \psi(e)$, where ψ is a continuously differentiable and invertible function, then apply Newton's method to find new search directions. The advantage of this technique is that the algorithm uses a full Newton step size at each iteration. Later on, Achache [2] generalized this technique to CQO, while Zhang, Bai, and Wang [33] generalized it to linearly constrained convex optimization. Due to the effectiveness of this technique, Several other works have applied this idea to other optimization problems (e.g., [3, 4, 12, 13, 15, 17, 18, 23, 26, 29].

Another transformation was introduced by Darvay and Takács [14] based on applying the function $\psi(t) = t^2$ to both sides of the equation $\frac{xs}{\mu} = \sqrt{\frac{xs}{\mu}}$ instead of $\frac{xs}{\mu} = e$, which resulted in a new search direction for LO. Later, Zhang et al. [34] extended this work to $P_*(\kappa)$-linear

complementarity problem. In [19], Kheirfam and Nasrollahi introduced a new interior point algorithm for LO based on the above AET technique using the parametric function $\psi(t) = t^q$, $q \geq 2$. Subsequently, Zaoui et al. [30, 32] proposed the new functions $\psi(t) = t^{\frac{3}{2}}$ and $\psi(t) = t^{\frac{7}{4}}$, respectively, for LO, and it was proved that the algorithm using this first function is better than the other one. Recently, Zaoui, Benterki, and Khelladi [31] extended the algorithm given by Darvay and Takács [14] to CQO.

Inspired by the previous works, we first present an interior point algorithm for CQO based on a new function $\psi(t) = t^{\frac{5}{4}}$, to enhance the performance of primal-dual algorithms. Next, we prove that this algorithm is well-defined and has a polynomial complexity. Finally, we perform comparative numerical tests between our approach and that of Zaoui, Benterki, and Khelladi [31], to confirm our theoretical results and evaluate the effectiveness of our algorithm. Our aim is to propose, develop, and analyze an interior point algorithm based on a new search direction for CQO.

This paper is divided into six sections. In Section 2, we introduce some notions related to interior point methods for CQO problems. In Section 3, we focus on calculating the search direction using the AET technique, and present the corresponding algorithm. Section 4 is dedicated to analyzing the algorithm's convergence and complexity. In Section 5, we perform comparative numerical tests to show the efficiency of our algorithm. Finally, a general conclusion is given in Section 6.

2 Preliminaries

We consider the standard CQO problem (P) and its Lagrangian dual (D)

$$\min\Big\{\frac{1}{2}x^TQx+c^Tx : Ax = b, x \geq 0\Big\}, \tag{P}$$

$$\max\Big\{-\frac{1}{2}x^TQx + b^Ty : A^Ty + s = Qx + c, s \geq 0\Big\}, \tag{D}$$

where $Q \in \mathbb{R}^{n\times n}$ is a positive semidefinite symmetric matrix, $A \in \mathbb{R}^{m\times n}$ is a full rank matrix, $b \in \mathbb{R}^m$ and $c \in \mathbb{R}^n$.

We call both problems (P) and (D) the primal-dual problem. Throughout this paper, we assume that the set of strictly feasible solutions $\mathcal{F}^0$ of the primal-dual problem is nonempty, where

$$\mathcal{F}^0 = \{(x, y, s) \in \mathbb{R}^{n+m+n} : Ax = b, A^Ty + s = Qx + c, x > 0, s > 0\}.$$

Solving the primal-dual problem is equivalent to solve the following nonlinear system given by Karush–Kuhn–Tucker conditions

$$\begin{aligned} Ax &= b, \quad x \geq 0, \\ -Qx + A^T y + s &= c, \quad s \geq 0, \\ xs &= 0. \end{aligned} \tag{1}$$

In primal-dual interior point method, the equation $xs = 0$ is replaced by the parametrized equation $xs = \mu e$, where $\mu > 0$ and $e = (1, 1, \ldots, 1)^T \in \mathbb{R}^n$. This gives the following system

$$\begin{aligned} Ax &= b, \quad x > 0, \\ -Qx + A^T y + s &= c, \quad s > 0, \\ xs &= \mu e, \end{aligned} \tag{2}$$

which is equivalent to the system given in (3);

$$\begin{aligned} Ax &= b, \quad x > 0, \\ -Qx + A^T y + s &= c, \quad s > 0, \\ \frac{xs}{\mu} &= \sqrt{\frac{xs}{\mu}}, \end{aligned} \tag{3}$$

with $\frac{xs}{\mu} = \left(\frac{x_1 s_1}{\mu}, \frac{x_2 s_2}{\mu}, \ldots, \frac{x_n s_n}{\mu}\right)^T$ and $\sqrt{\frac{xs}{\mu}} = \left(\sqrt{\frac{x_1 s_1}{\mu}}, \sqrt{\frac{x_2 s_2}{\mu}}, \ldots, \sqrt{\frac{x_n s_n}{\mu}}\right)^T$.

It is proved that system (2) has a unique solution $(x(\mu), y(\mu), s(\mu))$, for each $\mu > 0$. The set of all these solutions forms the so-called central path of (P) and (D). If $\mu \to 0$, then the central path converges to the optimal solutions of (P) and (D).

3 New search direction for CQO

Based on the work introduced by Darvay and Takács [14], we write system (3) as follows:

$$\begin{aligned} Ax &= b, \quad x > 0, \\ -Qx + A^T y + s &= c, \quad s > 0, \\ \psi\Big(\frac{xs}{\mu}\Big) &= \psi\Big(\sqrt{\frac{xs}{\mu}}\Big), \end{aligned} \tag{4}$$

where ψ is an invertible continuously differentiable function defined on $(L, +\infty)$, and L is determined by the condition $2t\psi'(t^2) - \psi'(t) > 0$.

Let $(x, y, s) \in \mathcal{F}^0$ be the current iterate. Applying Newton's method to system (4), we find the new iteration $x_+ = x + \Delta x$, $y_+ = y + \Delta y$, $s_+ = s + \Delta s$, where $(\Delta x, \Delta y, \Delta s)$ is the solution of the following system:

$$\begin{aligned} A\Delta x &= 0, \\ -Q\Delta x + A^T\Delta y + \Delta s &= 0, \\ \frac{1}{\mu}(s\Delta x + x\Delta s) &= \frac{\psi\left(\sqrt{\frac{xs}{\mu}}\right) - \psi\left(\frac{xs}{\mu}\right)}{\psi'\left(\frac{xs}{\mu}\right) - \frac{\sqrt{\mu}e}{2\sqrt{xs}}\psi'\left(\sqrt{\frac{xs}{\mu}}\right)}. \end{aligned} \tag{5}$$

We introduce the following notations:

$$v = \sqrt{\frac{xs}{\mu}}, \quad d_x = \frac{v\Delta x}{x}, \quad d_s = \frac{v\Delta s}{s}. \tag{6}$$

Hence, system (5) can be written as follows:

$$\begin{aligned} \bar{A}d_x &= 0, \\ -\bar{Q}d_x + \bar{A}^T\Delta_y + d_s &= 0, \\ d_x + d_s &= p_v, \end{aligned} \tag{7}$$

with

$$p_v = \frac{2\psi(v) - 2\psi(v^2)}{2v\psi'(v^2) - \psi'(v)}, \quad \bar{A} = \frac{1}{\mu}Adiag\Big(\frac{x}{v}\Big), \quad \bar{Q} = \frac{1}{\mu}diag\Big(\frac{x}{v}\Big)Qdiag\Big(\frac{x}{v}\Big),$$

and $diag\big(\frac{x}{v}\big)$ is the $n \times n$ diagonal matrix with diagonal entries $\frac{x_i}{v_i}$.

In this paper, we consider a new function $\psi : \left(\frac{1}{\sqrt[5]{16}}, +\infty\right) \to \mathbb{R}$, $\psi(t) = t^{\frac{5}{4}}$. Then we have

$$p_v = \frac{8v - 8v^{\frac{9}{4}}}{10v^{\frac{5}{4}} - 5e}. \tag{8}$$

At each iteration, we define a proximity measure

$$\delta(x, s, \mu) = \frac{\|p_v\|}{2} = \left\| \frac{4v - 4v^{\frac{9}{4}}}{10v^{\frac{5}{4}} - 5e} \right\|, \tag{9}$$

where $\|\cdot\|$ represents the Euclidean norm.

Remark 1. Let $(d_x, \Delta y, d_s)$ be a solution of system (7). Then

$$d_x^T d_s = d_x^T \bar{Q} d_x = \frac{1}{\mu} d_x^T diag\Big(\frac{x}{v}\Big) Q diag\Big(\frac{x}{v}\Big) d_x \geq 0, \tag{10}$$

because Q is a positive semidefinite matrix.

We detail below the algorithm steps for this approach.

Algorithm 1 Primal-dual algorithm for CQO

Input
An accuracy parameter $\epsilon > 0$;
An update parameter $0 < \theta < 1$ (default $\theta = \frac{1}{7\sqrt{n}}$);
A barrier parameter $\mu^0 = \frac{(x^0)^T s^0}{n}$;
An initial point $(x^0, y^0, s^0) \in \mathcal{F}^0$, such that $\delta(x^0, s^0; \mu^0) < \beta = \frac{1}{4}$;
begin
$x := x^0$; $y := y^0$; $s := s^0$; $\mu := \mu^0$; $k := 0$;
While $x^T s > \epsilon$ **do**
$\mu := (1 - \theta)\mu$;
Compute $(\Delta x, \Delta y, \Delta s)$ from system (5), then put
$x := x + \Delta x$; $y := y + \Delta y$; $s := s + \Delta s$; $k := k + 1$;
end While
end.

In order to facilitate the analysis of Algorithm 1, we introduce the following notation: $q_v = d_x - d_s$. From (7), we have

$$d_x = \frac{p_v + q_v}{2}, \qquad d_s = \frac{p_v - q_v}{2}, \tag{11}$$

and

$$\|p_v\|^2 = \|d_x\|^2 + \|d_s\|^2 + 2 d_x^T d_s = \|q_v\|^2 + 4 d_x^T d_s. \tag{12}$$

Using (10) and (12), we deduce that

$$\|p_v\|^2 \geq \|q_v\|^2. \tag{13}$$

Remark 2. In contrast to the case of CQO, in the LO case, the previous relation (13) becomes $\|p_v\|^2 = \|q_v\|^2$.

4 Analysis of convergence and complexity

In this section, we present a comprehensive theoretical study of the convergence and polynomial complexity of Algorithm 1 that we have conducted.

In the first lemma, we prove the strict feasibility of the iterates.

Lemma 1. Let $(x, y, s) \in \mathcal{F}^0$. The new iteration

$$(x_+, y_+, s_+) = (x, y, s) + (\Delta x, \Delta y, \Delta s)$$

is strictly feasible if $\delta(x, s, \mu) < 1$.

Proof. Let $0 \leq \alpha \leq 1$. We define $x_+(\alpha), s_+(\alpha)$ as follows:

$$x_+(\alpha) = x + \alpha\Delta x,\ s_+(\alpha) = s + \alpha\Delta s.$$

Hence

$$\frac{1}{\mu}x_+(\alpha)s_+(\alpha) = \frac{xs}{\mu} + \frac{\alpha}{\mu}(x\Delta s + s\Delta x) + \frac{\alpha^2}{\mu}\Delta x\Delta s.$$

Using (6), (7) and (11), we obtain

$$\begin{aligned}\frac{1}{\mu}x_+(\alpha)s_+(\alpha) &= v^2 + \alpha v(d_x + d_s) + \alpha^2 d_x d_s \\ &= v^2 + \alpha v p_v + \alpha^2\Big(\frac{p_v^2 - q_v^2}{4}\Big) \\ &= (1-\alpha)v^2 + \alpha(v^2 + v p_v) + \alpha^2\Big(\frac{p_v^2 - q_v^2}{4}\Big). \end{aligned} \tag{14}$$

We calculate the expression of $v^2 + vp_v$ by using (8) as follows:

$$v^2 + vp_v = \frac{2v^{\frac{13}{4}} + 3v^2}{10v^{\frac{5}{4}} - 5e} = e + \frac{2v^{\frac{13}{4}} + 3v^2 - 10v^{\frac{5}{4}} + 5e}{10v^{\frac{5}{4}} - 5e}. \tag{15}$$

After studying the function $\dfrac{2t^{\frac{13}{4}} + 3t^2}{10t^{\frac{5}{4}} - 5}$ for all $t > \frac{1}{\sqrt[5]{16}}$, we find that $\dfrac{2t^{\frac{13}{4}} + 3t^2}{10t^{\frac{5}{4}} - 5} \geq 1$. This is equivalent to

$$v^2 + vp_v \geq e. \tag{16}$$

From (14) and (16), we get

$$\frac{1}{\mu}x_+(\alpha)s_+(\alpha) \geq (1-\alpha)v^2 + \alpha e + \alpha^2\Big(\frac{p_v^2 - q_v^2}{4}\Big)$$

$$\geq (1-\alpha)v^2 + \alpha\left(e - \left((1-\alpha)\frac{p_v^2}{4} + \alpha\frac{q_v^2}{4}\right)\right).$$

We have

$$\left\|(1-\alpha)\frac{p_v^2}{4} + \alpha\frac{q_v^2}{4}\right\|_\infty \leq (1-\alpha)\frac{\|p_v^2\|_\infty}{4} + \alpha\frac{\|q_v^2\|_\infty}{4}$$
$$\leq (1-\alpha)\frac{\|p_v\|^2}{4} + \alpha\frac{\|q_v\|^2}{4},$$

by using (13), we obtain

$$\left\|(1-\alpha)\frac{p_v^2}{4} + \alpha\frac{q_v^2}{4}\right\|_\infty \leq (1-\alpha)\frac{\|p_v\|^2}{4} + \alpha\frac{\|p_v\|^2}{4}$$
$$= \frac{\|p_v\|^2}{4} = \delta^2.$$

If $\delta < 1$, then $e - \left((1-\alpha)\frac{p_v^2}{4} + \alpha\frac{q_v^2}{4}\right) > 0$, which implies that

$$x_+(\alpha)s_+(\alpha) > 0.$$

As $x_+(\alpha), s_+(\alpha)$ are continuous functions, they do not change sign on the interval $[0,1]$, and since $x_+(0) = x > 0$, $s_+(0) = s > 0$, this implies that $x_+(1) = x + \Delta x = x_+ > 0$ and $s_+(1) = s + \Delta s = s_+ > 0$, which proves the lemma. □

Lemma 2. Let $v > \frac{1}{\sqrt[5]{16}}e$ and $v_+ = \sqrt{\frac{x_+s_+}{\mu}}$. If $\delta = \delta(x,s,\mu) < \frac{1}{4}$, then

$$v_+ > \frac{1}{\sqrt[5]{16}}e.$$

Proof. By replacing α with 1 in (14), we obtain

$$\frac{x_+s_+}{\mu} = v_+^2 = v^2 + vp_v + \frac{p_v^2}{4} - \frac{q_v^2}{4}, \tag{17}$$

from (16), we find

$$v_+^2 \geq e + \frac{p_v^2}{4} - \frac{q_v^2}{4} \geq e - \frac{q_v^2}{4},$$

which implies that

$$(v_+)_i^2 \geq 1 - \frac{\|q_v\|_\infty^2}{4} \geq 1 - \frac{\|q_v\|^2}{4}.$$

From (13), we deduce that

$$(v_+)_i^2 \geq 1 - \frac{\|p_v\|^2}{4} = 1 - \delta^2, \quad \text{for all } i = 1:n,$$

which implies that

$$\min(v_+) \geq \sqrt{1-\delta^2}. \tag{18}$$

Since $\delta = \delta(x,s,\mu) < \frac{1}{4}$, then

$$\min(v_+) > \frac{1}{\sqrt[5]{16}}. \tag{19}$$

□

Lemma 3. [13]
Let us consider a decreasing function $f :]d, +\infty[\to \mathbb{R}_+^*$, with $d > 0$. If we take the vector $v \in \mathbb{R}_+^n$, where $\min(v) = \min_i(v_i) \geq d$, then

$$\|f(v)(e - v^2)\| \leq f(\min(v))\|e - v^2\| \leq f(d)\|e - v^2\|.$$

In the following lemma, we prove the quadratic convergence of the iterate.

Lemma 4. If $\delta = \delta(x,s,\mu) < \frac{1}{4}$, then

$$\delta(x_+, s_+, \mu) \leq \frac{160(2\sqrt{2} + \sqrt[8]{759375})(-15\sqrt[8]{15} + 4\sqrt{30})}{-200 + 375\sqrt[4]{15}}\delta^2.$$

Proof. Let $v > \frac{1}{\sqrt[5]{16}}e$. From (9), we have

$$\delta(x_+, s_+, \mu) = \frac{\|p_{v_+}\|}{2} = \left\| \frac{4v_+ - 4v_+^{\frac{9}{4}}}{\left(10v_+^{\frac{5}{4}} - 5e\right)(e - v_+^2)} \right\|, v_+ \neq e.$$

We divide the expression of $\delta(x_+, s_+, \mu)$ into two terms:

1. $e - v_+^2$.
2. $f_1(v_+) = \dfrac{4v_+ - 4v_+^{\frac{9}{4}}}{\left(10v_+^{\frac{5}{4}} - 5e\right)(e - v_+^2)}$.

First, we compute the expression $e - v_+^2$ using (14) as follows:

$$\begin{aligned} e - v_+^2 &= e - \frac{x_+ s_+}{\mu} \\ &= e - \left((v^2 + v p_v) + \frac{p_v^2}{4} - \frac{q_v^2}{4} \right), \end{aligned}$$

from (15), we have

$$\begin{aligned} e - v_+^2 &= \frac{q_v^2}{4} - \left(\frac{2v^{\frac{13}{4}} + 3v^2 - 10v^{\frac{5}{4}} + 5e}{10v^{\frac{5}{4}} - 5e} + \frac{p_v^2}{4} \right) \\ &= \frac{q_v^2}{4} - \frac{p_v^2}{4} \left(\frac{36v^{\frac{9}{2}} - 12v^{\frac{13}{4}} + v^2 - (10v^{\frac{5}{4}} - 5e)^2}{16(v - v^{\frac{9}{4}})^2} \right). \end{aligned} \tag{20}$$

It can be shown that $0 \leq \left(\frac{36v^{\frac{9}{2}} - 12v^{\frac{13}{4}} + v^2 - (10v^{\frac{5}{4}} - 5e)^2}{16(v - v^{\frac{9}{4}})^2} \right) \leq 3e$, for all $v > \frac{1}{\sqrt[5]{16}} e, v \neq e$. Therefore

$$\begin{aligned} \|e - v_+^2\| &\leq \frac{\|q_v^2\|}{4} + \left\| \frac{p_v^2}{4} \left(\frac{36v^{\frac{9}{2}} - 12v^{\frac{13}{4}} + v^2 - (10v^{\frac{5}{4}} - 5e)^2}{16(v - v^{\frac{9}{4}})^2} \right) \right\| \\ &\leq \frac{\|q_v\|^2}{4} + 3\frac{\|p_v\|^2}{4}, \end{aligned}$$

by using (9) and (13), we get

$$\|e - v_+^2\| \leq \frac{\|p_v\|^2}{4} + 3\frac{\|p_v\|^2}{4} = 4\delta^2. \tag{21}$$

Second, we consider the function $f_1(t) = \dfrac{4t - 4t^{\frac{9}{4}}}{\left(10t^{\frac{5}{4}} - 5\right)(1 - t^2)}$. Since $f_1'(t) \leq 0$ for all $t > \frac{1}{\sqrt[5]{16}}$; $t \neq 1$, the function f_1 is decreasing on $\left] \frac{1}{\sqrt[5]{16}}, +\infty \right[\setminus \{1\}$. By applying Lemma 3, (18), and (21), we obtain

$$\delta(x_+, s_+, \mu) \leq f_1\left(\sqrt{1 - \delta^2}\right) \|e - v_+^2\| \leq 4 f_1\left(\sqrt{1 - \delta^2}\right) \delta^2.$$

If $\delta < \frac{1}{4}$, then $f_1\left(\sqrt{1 - \delta^2}\right) \leq f_1\left(\frac{\sqrt{15}}{4}\right)$. Thus

$$\begin{aligned} \delta(x_+, s_+, \mu) &\leq 4 f_1\left(\frac{\sqrt{15}}{4}\right) \delta^2 \\ &= \frac{160(2\sqrt{2} + \sqrt[8]{759375})(-15\sqrt[8]{15} + 4\sqrt{30})}{-200 + 375\sqrt[4]{15}} \delta^2 \\ &\approx 2.128\delta^2. \end{aligned}$$

□

In the next lemma, we give an upper bound of the duality gap.

Lemma 5. If $\delta = \delta(x, s, \mu) < \frac{1}{4}$, then the duality gap satisfies the following inequality:

$$x_+^T s_+ \leq \mu\Big(n + \frac{3}{16}\Big).$$

Proof. Let $v > \frac{1}{\sqrt[5]{16}}e$. We have

$$v_+^2 = \frac{x_+ s_+}{\mu}.$$

Then

$$x_+^T s_+ = \sum_{i=1}^{n}(x_+)_i(s_+)_i = \mu\sum_{i=1}^{n}(v_+^2)_i.$$

According to (17), we have

$$x_+^T s_+ = \mu\sum_{i=1}^{n}\left((v^2 + vp_v)_i + \frac{(p_v^2)_i}{4} - \frac{(q_v^2)_i}{4}\right). \tag{22}$$

Applying (15) yields

$$\begin{aligned} v^2 + vp_v &= e + \frac{2v^{\frac{13}{4}} + 3v^2 - 10v^{\frac{5}{4}} + 5e}{10v^{\frac{5}{4}} - 5e} \\ &= e + \left(\frac{20v^{\frac{9}{2}} + 20v^{\frac{13}{4}} - 15v^2 - (10v^{\frac{5}{4}} - 5e)^2}{16(v - v^{\frac{9}{4}})^2}\right)\frac{p_v^2}{4} \\ &= e + \left(\frac{36v^{\frac{9}{2}} - 12v^{\frac{13}{4}} + v^2 - (10v^{\frac{5}{4}} - 5e)^2}{16(v - v^{\frac{9}{4}})^2} - e\right)\frac{p_v^2}{4}. \end{aligned}$$

As shown in the proof of Lemma 4, $\dfrac{36v^{\frac{9}{2}} - 12v^{\frac{13}{4}} + v^2 - (10v^{\frac{5}{4}} - 5e)^2}{16(v - v^{\frac{9}{4}})^2} \leq 3e$, so

$$v^2 + vp_v \leq e + 2\frac{p_v^2}{4}. \tag{23}$$

By using (22) and (23), we deduce that

$$\begin{aligned} x_+^T s_+ &\leq \mu\sum_{i=1}^{n}\left(1 + 2\frac{(p_v^2)_i}{4} + \frac{(p_v^2)_i}{4} - \frac{(q_v^2)_i}{4}\right) \\ &= \mu\left(n + 3\frac{\|p_v\|^2}{4} - \frac{\|q_v\|^2}{4}\right) \\ &\leq \mu(n + 3\delta^2). \end{aligned}$$

If $\delta < \frac{1}{4}$, then

$$x_+^T s_+ \leq \mu\Big(n + \frac{3}{16}\Big).$$

□

The following lemma shows that the algorithm is well defined.

Lemma 6. If $\delta < \frac{1}{4}$ and $\theta = \frac{1}{7\sqrt{n}}$, then

$$\delta(x_+, s_+, \mu_+) < \frac{1}{4}.$$

Proof. We introduce the notation $\bar{v} = \sqrt{\frac{x_+ s_+}{\mu_+}}$. Then, from (19) we have

$$\bar{v} = \frac{1}{\sqrt{1-\theta}} v_+ > \frac{1}{\sqrt[5]{16}} e, \ 0 < \theta < 1.$$

From the definition of δ, we have

$$\begin{aligned}
\delta(x_+, s_+, \mu_+) &= \frac{\|p_{\bar{v}}\|}{2} \\
&= \left\| \frac{4\bar{v} - 4\bar{v}^{\frac{9}{4}}}{(10\bar{v}^{\frac{5}{4}} - 5e)(e - \bar{v}^2)} (e - \bar{v}^2) \right\| \\
&= \left\| f_1(\bar{v})(e - \bar{v}^2) \right\|.
\end{aligned}$$

In addition, from (18) we have

$$\min(\bar{v}) = \frac{\min(v_+)}{\sqrt{1-\theta}} \geq \frac{\sqrt{1-\delta^2}}{\sqrt{1-\theta}}.$$

If $\delta < \frac{1}{4}$, then

$$\min(\bar{v}) > \frac{\sqrt{15}}{4}.$$

Since f_1 is decreasing (see Lemma 4), we get

$$\begin{aligned}
\delta(x_+, s_+, \mu_+) &= \left\| f_1(\bar{v})(e - \bar{v}^2) \right\| \\
&\leq f_1\Big(\frac{\sqrt{15}}{4}\Big) \|e - \bar{v}^2\| \\
&= f_1\Big(\frac{\sqrt{15}}{4}\Big) \left\| e - \frac{v_+^2}{1-\theta} \right\|
\end{aligned}$$

$$\leq \frac{1}{1-\theta} f_1\Big(\frac{\sqrt{15}}{4}\Big)\Big(\|\theta e\| + \|e - v_+^2\|\Big).$$

By using (21), we obtain

$$\delta(x_+, s_+, \mu_+) \leq \frac{1}{1-\theta} f_1\Big(\frac{\sqrt{15}}{4}\Big)\Big(\theta\sqrt{n} + 4\delta^2\Big).$$

Since $\delta < \frac{1}{4}$, then

$$\delta(x_+, s_+, \mu_+) \leq \frac{1}{1-\theta} f_1\Big(\frac{\sqrt{15}}{4}\Big)\Big(\theta\sqrt{n} + \frac{1}{4}\Big). \tag{24}$$

Suppose that $\theta = \frac{1}{7\sqrt{n}}$. Then

$$1 - \theta \geq \frac{6}{7}, \tag{25}$$

for all $n \geq 1$.

Finally, by using (24) and (25), we deduce that

$$\begin{aligned}\delta(x_+, s_+, \mu_+) &\leq \frac{7}{6} f_1\Big(\frac{\sqrt{15}}{4}\Big)\Big(\frac{1}{7} + \frac{1}{4}\Big)\\ &< \frac{1}{4}.\end{aligned}$$

□

In the last lemma, we give a lower bound on the number of iterations required for optimality.

Lemma 7. Let $(x^0, y^0, s^0) \in \mathcal{F}^0$, let $\mu^{(0)} = \frac{(x^0)^T s^0}{n}$, let $\delta < \frac{1}{4}$, and let $x^{(k+1)}, s^{(k+1)}$ be the vectors obtained after $k+1$ iterations. The inequality $(x^{(k+1)})^T s^{(k+1)} \leq \epsilon$ is satisfied if

$$k \geq \frac{1}{\theta} \ln \frac{\mu^{(0)}(n + \frac{3}{16})}{\epsilon}.$$

Proof. After k iterations, we find $\mu^{(k)} = (1-\theta)^k \mu^{(0)}$. Using this and Lemma 5, we obtain

$$(x^{(k+1)})^T s^{(k+1)} \leq \mu^{(k)}\left(n + \frac{3}{16}\right) = \mu^{(0)}(1-\theta)^k\left(n + \frac{3}{16}\right).$$

Then, the inequality $(x^{(k+1)})^T s^{(k+1)} \leq \epsilon$ is satisfied if

$$\mu^{(0)}(1-\theta)^k\left(n + \frac{3}{16}\right) \leq \epsilon,$$

this is equivalent to

$$\ln(1-\theta)^k + \ln \mu^{(0)}\left(n + \frac{3}{16}\right) \leq \ln \epsilon,$$

that is,

$$-k\ln(1-\theta) \geq \ln \frac{\mu^{(0)}\left(n+\frac{3}{16}\right)}{\epsilon}.$$

As $-\ln(1-\theta) \geq \theta$, for all $0 < \theta < 1$, so $-k\ln(1-\theta) \geq k\theta$. Thus, the inequality

$$-k\ln(1-\theta) \geq \ln \frac{\mu^{(0)}\left(n+\frac{3}{16}\right)}{\epsilon}$$

is satisfied if

$$k\theta \geq \ln \frac{\mu^{(0)}\left(n+\frac{3}{16}\right)}{\epsilon},$$

hence

$$k \geq \frac{1}{\theta} \ln \frac{\mu^{(0)}\left(n+\frac{3}{16}\right)}{\epsilon}.$$

□

Theorem 1. If $\theta = \frac{1}{7\sqrt{n}}$, then the algorithm gives an approximate solution of the primal-dual problem after at most

$$7\sqrt{n} \ln \frac{\mu^{(0)}\left(n+\frac{3}{16}\right)}{\epsilon}.$$

Proof. Replacing θ with $\frac{1}{7\sqrt{n}}$ in Lemma 7 gives the result. □

5 Comparative numerical experiments

To evaluate the effectiveness of Algorithm 1 using the new function $\psi(t) = t^{\frac{5}{4}}$, we conduct comparative numerical tests between our function and that of Zaoui, Benterki, and Khelladi [31]. We solve five fixed-size examples and two variable-size examples. Moreover, we test some real-examples from Maros-Mészáros test set [24]. The algorithm was implemented in MATLAB on a computer with an Intel i3 processor and 4 GB of RAM, with the following parameters: $\epsilon = 10^{-4}$ and $\theta = 0.5$. In the table of results, we denote by

- Nbr: the number of iterations required for optimality.

- Time: the computation time in seconds.
- $x^T s$: the duality gap at final iteration.

Example 1. [16] Consider

$$A = \begin{pmatrix} 1 & 1 & 2 & 1 \end{pmatrix} \quad , b = \begin{pmatrix} 3 \end{pmatrix}, \quad Q = \begin{pmatrix} 4 & 2 & 2 & 0 \\ 2 & 4 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad c = \begin{pmatrix} -8 \\ -6 \\ -4 \\ 0 \end{pmatrix}.$$

The initial point is

$$x^0 = \begin{pmatrix} 0.5 & 0.5 & 0.5 & 1 \end{pmatrix}^T, \quad y^0 = \begin{pmatrix} -5 \end{pmatrix}, \quad s^0 = \begin{pmatrix} 1 & 2 & 8 & 5 \end{pmatrix}^T.$$

Example 2. [27] Consider

$$A = \begin{pmatrix} 1 & 1 & 1 & 0 \\ -1 & 2 & 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \quad Q = \begin{pmatrix} 2 & -2 & 0 & 0 \\ -2 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad c = \begin{pmatrix} -2 \\ -6 \\ 0 \\ 0 \end{pmatrix}.$$

The initial point is

$$x^0 = \begin{pmatrix} 0.5145 & 0.8764 & 0.6090 & 0.7617 \end{pmatrix}^T, \quad y^0 = \begin{pmatrix} -5.4801 & -1.1789 \end{pmatrix}^T,$$
$$s^0 = \begin{pmatrix} 1.5776 & 4.3148 & 5.4801 & 1.1789 \end{pmatrix}^T.$$

Example 3. [8] Consider

$$A = \begin{pmatrix} 1 & 1.2 & 1 & 1.8 & 0 \\ 3 & -1 & 1.5 & -2 & 1 \\ -1 & 2 & -3 & 4 & 2 \end{pmatrix}, \quad b = \begin{pmatrix} 9.31 \\ 5.45 \\ 6.60 \end{pmatrix},$$

$$Q = \begin{pmatrix} 20 & 1.2 & 0.5 & 0.5 & -1 \\ 1.2 & 32 & 1 & 1 & 1 \\ 0.5 & 1 & 14 & 1 & 1 \\ 0.5 & 1 & 1 & 15 & 1 \\ -1 & 1 & 1 & 1 & 16 \end{pmatrix}, \quad c = \begin{pmatrix} 1 \\ -1.5 \\ 2 \\ 1.5 \\ 3 \end{pmatrix}.$$

The initial point is

$$x^0 = \begin{pmatrix} 2.4539 & 0.7875 & 1.5838 & 2.4038 & 1.3074 \end{pmatrix}^T,$$
$$y^0 = \begin{pmatrix} 20.5435 & 9.4781 & 4.3927 \end{pmatrix}^T,$$
$$s^0 = \begin{pmatrix} 7.1215 & 7.9763 & 8.3150 & 6.8686 & 7.9750 \end{pmatrix}^T.$$

Example 4. [16] Consider

$$A = \begin{pmatrix} 1 & 2 & 1 & 1 & 1 & 0 & 0 \\ 3 & 1 & 2 & -1 & 0 & 1 & 0 \\ 0 & -1 & -4 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 5 \\ 4 \\ -1.5 \end{pmatrix},$$

$$Q = \begin{pmatrix} 2 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad c = \begin{pmatrix} -1 \\ -3 \\ 1 \\ -1 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

The initial point is

$$x^0 = \begin{pmatrix} 0.3599 & 0.5425 & 0.8142 & 2.0703 & 0.6706 & 2.8197 & 2.2993 \end{pmatrix}^T,$$
$$y^0 = \begin{pmatrix} -2.2576 & -1.1007 & -1.5314 \end{pmatrix}^T,$$
$$s^0 = \begin{pmatrix} 4.4655 & 1.6272 & 2.6719 & 3.0413 & 2.2577 & 1.1008 & 1.5315 \end{pmatrix}^T.$$

Example 5. [8] Consider

$$A = \begin{pmatrix} 1 & -1 & 1.9 & 1.25 & 1.2 & 0.4 & -0.7 & 1.06 & 1.5 & 1.05 \\ 1.3 & 1.2 & 0.15 & 2.15 & 1.25 & 1.5 & 0.4 & 1.52 & 1.3 & 1 \\ 1.5 & -1.1 & 3.5 & 1.25 & 1.8 & 2 & 1.95 & 1.2 & 1 & -1 \end{pmatrix}, \quad b = \begin{pmatrix} 11.651 \\ 16.672 \\ 21.295 \end{pmatrix},$$

$$Q = \begin{pmatrix} 30 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 21 & 0 & 1 & -1 & 1 & 0 & 1 & 0.5 & 1 \\ 1 & 0 & 15 & -0.5 & -2 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & -0.5 & 30 & 3 & -1 & 1 & -1 & 0.5 & 1 \\ 1 & -1 & -2 & 3 & 27 & 1 & 0.5 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & 1 & 16 & -0.5 & 0.5 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0.5 & -0.5 & 8 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & 1 & 0.5 & 1 & 24 & 1 & 1 \\ 1 & 0.5 & 1 & 0.5 & 1 & 0 & 1 & 1 & 39 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 11 \end{pmatrix}, \quad c = \begin{pmatrix} -0.5 \\ -1 \\ 0 \\ 0 \\ -0.5 \\ 0 \\ 0 \\ -1 \\ -0.5 \\ -1 \end{pmatrix}.$$

The initial point is

$$x^0 = \begin{pmatrix} 0.9491 & 0.6121 & 1.8477 & 1.8115 & 1.2511 & 2.5211 & 1.5062 & 1.5658 & 0.8207 & 1.1289 \end{pmatrix}^T,$$
$$y^0 = \begin{pmatrix} 4.3800 & 19.9367 & 4.5679 \end{pmatrix}^T,$$
$$s^0 = \begin{pmatrix} 3.8902 & 4.4625 & 3.9788 & 3.6606 & 3.9018 & 3.5561 & 3.8760 & 3.7190 & 3.9138 & 4.3396 \end{pmatrix}^T.$$

Table 1 summarizes the numerical results obtained for the problems described above.

Table 1: Numerical results obtained for fixed size examples.

	$\psi(t) = t^{\frac{5}{4}}$			$\psi(t) = t^2$		
Examples	Nbr	Time	$x^T s$	Nbr	Time	$x^T s$
Example 1	18	0.0080	5.28×10^{-5}	20	0.0289	7.12×10^{-5}
Example 2	17	0.0076	8.88×10^{-5}	20	0.0108	5.99×10^{-5}
Example 3	20	0.0082	8.03×10^{-5}	23	0.0164	5.62×10^{-5}
Example 4	18	0.0081	9.60×10^{-5}	21	0.0248	7.16×10^{-5}
Example 5	20	0.0082	6.79×10^{-5}	22	0.0105	9.15×10^{-5}

Example 6. [8] Consider

$$A(i,j) = \begin{cases} 1 & \text{if } (j = i \text{ or } j = i + m), \\ 0 & \text{otherwise.} \end{cases} \quad Q(i,j) = \begin{cases} 1 & \text{if } j = i, \\ 0 & \text{otherwise.} \end{cases}$$

$$c(i) = \begin{cases} -1 & \text{if } i = 1 : m, \\ 0 & \text{if } i = m + 1 : n,\ n = 2m, \end{cases}$$

$$b(i) = 2 \quad \text{for } i = 1 : m.$$

The initial point is

$$x^0(i) = 1 \qquad \text{for } i = 1 : n,$$
$$y^0(i) = -1 \quad \text{for } i = 1 : m,$$
$$s^0(i) = \begin{cases} 1 & \text{if } \ i = 1 : m, \\ 2 & \text{if } \ i = m+1 : n. \end{cases}$$

The results obtained for Example 6 are summarized in the following Table 2.

Table 2: Numerical results obtained for Example 6.

	$\psi(t) = t^{\frac{5}{4}}$			$\psi(t) = t^2$		
Size	Nbr	Time	$x^T s$	Nbr	Time	$x^T s$
$n = 100$	21	0.0454	9.43×10^{-5}	24	0.0673	7.13×10^{-5}
$n = 200$	22	0.1284	9.43×10^{-5}	25	0.1304	7.36×10^{-5}
$n = 600$	24	1.4369	7.07×10^{-5}	27	1.68704	5.86×10^{-5}
$n = 1000$	25	5.9973	5.89×10^{-5}	27	6.6444	9.77×10^{-5}
$n = 2000$	26	37.678	5.89×10^{-5}	29	48.034	5.17×10^{-5}
$n = 4000$	27	172.71	5.89×10^{-5}	30	249.95	5.32×10^{-5}
$n = 6000$	27	585.21	8.84×10^{-5}	30	670.68	7.98×10^{-5}
$n = 7000$	28	959.09	5.15×10^{-5}	30	1059.28	9.31×10^{-5}

Example 7. [7] Consider

$$A(1, i) = 1 \qquad \text{for all } i = 1 : n,$$
$$Q(i, j) = \begin{cases} 2 & \text{if } \ j = i, \\ 0 & \text{otherwise}, \end{cases}$$

$$c(i) = -1 \qquad \text{for all } \ i = 1 : n,$$
$$b = 1.$$

The results obtained for Example 7 are summarized in Table 3.

We then tested some examples from Maros-Meszaros test set using different values of θ: $\{0.1, 0.3, 0.5\}$ and also its theoretical value $(\frac{1}{7\sqrt{n}})$, which is denoted by θ_{theo} in the table. The results for these examples are summarized in Table 4.

Table 3: Numerical results obtained for Example 7.

	$\psi(t) = t^{\frac{5}{4}}$			$\psi(t) = t^2$		
Size	Nbr	Time	$x^T s$	Nbr	Time	$x^T s$
$n = 1000$	15	1.8535	7.00×10^{-5}	17	2.2843	7.62×10^{-5}
$n = 2000$	15	13.850	7.01×10^{-5}	17	16.015	7.63×10^{-5}
$n = 4000$	15	65.623	7.02×10^{-5}	17	71.749	7.64×10^{-5}
$n = 6000$	15	185.75	7.05×10^{-5}	17	226.80	7.69×10^{-5}
$n = 8000$	15	479.02	7.04×10^{-5}	17	562.85	7.65×10^{-5}
$n = 10000$	15	1127.8	7.08×10^{-5}	17	1211.1	7.67×10^{-5}

Table 4: Numerical results for some examples from Maros-Mészáros test set

		$\psi(t) = t^{\frac{5}{4}}$			$\psi(t) = t^2$		
Example	θ	Nbr	Time	$x^T s$	Nbr	Time	$x^T s$
Tame	$\theta = 0.1$	95	0.0100	9.04×10^{-5}	95	0.0100	9.11×10^{-5}
	$\theta = 0.3$	28	0.0088	9.79×10^{-5}	29	0.0096	7.89×10^{-5}
	$\theta = 0.5$	15	0.0079	8.04×10^{-5}	17	0.0080	8.72×10^{-5}
	θ_{theo}	94	0.0100	9.04×10^{-5}	94	0.0119	9.10×10^{-5}
Genhs28	$\theta = 0.1$	110	0.0128	9.30×10^{-5}	110	0.0129	9.37×10^{-5}
	$\theta = 0.3$	33	0.0100	8.23×10^{-5}	33	0.0103	9.47×10^{-5}
	$\theta = 0.5$	18	0.0093	5.03×10^{-5}	20	0.0099	6.40×10^{-5}
	θ_{theo}	250	0.0131	9.57×10^{-5}	250	0.0166	9.59×10^{-5}
Hs51	$\theta = 0.1$	103	0.0098	9.73×10^{-5}	103	0.0100	9.80×10^{-5}
	$\theta = 0.3$	31	0.0086	8.40×10^{-5}	29	0.0088	9.66×10^{-5}
	$\theta = 0.5$	17	0.0085	5.03×10^{-5}	19	0.0087	5.95×10^{-5}
	θ_{theo}	164	0.0100	9.94×10^{-5}	94	0.0114	9.97×10^{-5}
Zecevic2	$\theta = 0.1$	95	0.0098	9.04×10^{-5}	95	0.093	9.10×10^{-5}
	$\theta = 0.3$	28	0.0086	9.79×10^{-5}	29	0.0087	7.88×10^{-5}
	$\theta = 0.5$	15	0.0036	8.04×10^{-5}	17	0.0084	8.72×10^{-5}
	θ_{theo}	94	0.0094	9.03×10^{-5}	94	0.0098	9.10×10^{-5}
Lotschd	$\theta = 0.1$	112	0.0138	9.04×10^{-5}	112	0.0139	9.11×10^{-5}
	$\theta = 0.3$	33	0.0105	9.88×10^{-5}	34	0.0110	7.95×10^{-5}
	$\theta = 0.5$	18	0.0100	6.03×10^{-5}	20	0.0101	7.44×10^{-5}
	θ_{theo}	278	0.0202	9.88×10^{-5}	278	0.0193	9.89×10^{-5}

Notes

The numerical tests we have performed show that our new function $\psi = t^{\frac{5}{4}}$ is more efficient than the one proposed in [31]. In all examples, we record a reduction in the number of iterations and in the computation time for our function $\psi = t^{\frac{5}{4}}$ compared with $\psi = t^2$. This shows that the algorithm using our function is the fastest.

6 General conclusion

We are interested in solving a CQO problem using a primal-dual interior point method based on the AET technique. We proposed a new function ψ, and conducted a complete theoretical study on the convergence and complexity of the proposed algorithm. We then performed numerical experiments comparing our results with those of Zaoui, Benterki, and Khelladi [31]. The results obtained using our function are encouraging and reinforce the effectiveness of our algorithm.

Future research might aim to develop other functions with powers between $\frac{1}{2}$ and 1 in order to improve the algorithmic complexity and numerical behavior of primal-dual interior point methods for convex quadratic programming. In addition, the algorithm could be extended to other optimization problems, such as nonlinear programming problems, semidefinite problems, and complementarity problems.

Acknowledgements

The authors are very grateful and would like to thank the Editor-in-Chief and the anonymous referees for their comments and suggestions which helped to improve the presentation of this paper.

Funding

This research received no external funding.

Conflicts of Interest

All authors jointly worked on the results and they read and approved the final manuscript, and also declare that they have no competing interests.

References

[1] Abbaszadehpeivasti, H., de Klerk, E. and Zamani, M. *On the rate of convergence of the difference-of-convex algorithm (DCA)*, J. Optim. Theory. Appl., 202 (2024), 475–496.

[2] Achache, M. *A new primal-dual path-following method for convex quadratic programming*, Comput. Appl. Math., 25(1) (2006), 97–110.

[3] Asadi, S. and Mansouri, H. *A new full-Newton step $O(n)$ infeasible interior-point algorithm for $\mathcal{P}_*(\kappa)$-horizontal linear complementarity problems*, Comput. Sci. J. Mold., 64 (2014), 37–61.

[4] Asadi, S., Mansouri, H. and Darvay, Z. *An infeasible full-NT step IPM for horizontal linear complementarity problem over Cartesian product of symmetric cones*, Optim., 66(2) (2016), 225–250.

[5] Bai, Y.Q., El Ghami, M. and Roos, C. *A new efficient large-update primal-dual interior-point method based on a finite barrier*, SIAM J. Optim., 13(3) (2002), 766–782.

[6] Bouafia, M., Benterki, D. and Yassine, A. *Complexity analysis of interior point methods for linear programming based on a parameterized kernel function*, RAIRO Oper. Res., 50(4-5) (2016), 935–949.

[7] Boudjellal, N. and Benterki, D. *A new full-Newton step feasible interior point method for convex quadratic programming*, Optimization, 73 (2024), 1571–1588.

[8] Boudjellal, N., Roumili, H. and Benterki, D. *Complexity analysis of interior point methods for convex quadratic programming based on a parameterized kernel function*, Bol. Soc. Parana. Mat., 40 (2022), 1–16.

[9] Chinchilla, R., Yang, G. and Hespanha, J.P. *Newton and interior-point methods for (constrained) nonconvex–nonconcave minmax optimization with stability and instability guarantees*, Math. Control Signals Syst., 36 (2024), 381–421.

[10] Darvay, Z. *A new algorithm for solving self-dual linear optimization problems*, Studia Univ. Babeș-Bolyai Informatica, 47 (2002), 15–26.

[11] Darvay, Z. *New interior point algorithms in linear programming*, Adv. Model. Optim., 5(1) (2003), 51–92.

[12] Darvay, Z., Papp, I.M. and Takács, P.R. *An infeasible full-Newton step algorithm for linear optimization with one centering step in major iteration*, Studia Univ. Babes-Bolyai Informatica, 59 (1) (2014), 28–45.

[13] Darvay, Z., Papp, I. M. and Takács P. R. *Complexity analysis of a full-Newton step interior-point method for linear optimization*, Period. Math. Hungar., 73 (2016), 27–42.

[14] Darvay, Z. and Takács, P. R. *New method for determining search directions for interior-point algorithms in linear optimization*, Optim. Lett., 12 (2018), 1099–1116.

[15] Grimes, W. and Achache, M. *A path-following interior-point algorithm for monotone LCP based on a modified Newton search direction*, RAIRO Oper. Res., 57(3) (2023), 1059–1073.

[16] Hock, W. *Test examples for nonlinear programming codes*. Springer Berlin, Heidelberg, 1981.

[17] Kheirfam, B. *A new infeasible interior-point method based on Darvay's technique for symmetric optimization*, Ann. Oper. Res., 211 (2013), 209–224.

[18] Kheirfam, B. *New complexity analysis of a full Nesterov–Todd step interior-point method for semidefinite optimization*, Asian-Eur. J. Math., 10(4) (2017), 1750070.

[19] Kheirfam, B. and Nasrollahi, A. *An extension for identifying search directions for interior-point methods in linear optimization*, Asian-Eur. J. Math., 13 (2020), 2050014.

[20] LeThi, H.A., Le, H.M., Phan, D.N. and Tran, B. *Novel DCA based algorithms for a special class of nonconvex problems with application in machine learning*, Appl. Math. Comput., 409 (2021), 125904.

[21] Li, X. and Zhang, M. *Interior-point algorithm for linear optimization based on a new trigonometric kernel function*, Oper. Res. Lett., 43(5) (2015), 471–475.

[22] Lukšan, L., Matonoha. C. and Vlček, J. *Interior-point method for non-linear non-convex optimization*, Numerical linear algebra with applications, 11(5-6) (2004), 431–453.

[23] Mansouri, H. and Pirhaji, M. *A polynomial interior-point algorithm for monotone linear complementarity problems*, J. Optim. Theory Appl., 157 (2013), 451–461.

[24] Maros, I. and Mészáros, C. *A repository of convex quadratic programming problems*, Optimization methods and software, 11(1-4) (1999), 671–681.

[25] Phan, D.N. and Thi, H.A. *Difference-of-convex algorithm with extrapolation for nonconvex, nonsmooth optimization problems*, Math. Oper. Res., 49(3) (2024), 1973–1985.

[26] Roumili, H. and Kebbiche, Z. *A weighted target-following algorithm for linearly constrained convex optimization*, Int. J. Open Problems Compt. Math., 5(4) (2012), 25–33.

[27] Tahmasebzadeh, S., Navidi, H. and Malek, A. *Novel interior point algorithms for solving nonlinear convex optimization problems*, Adv. Oper. Res., 2015 (2015).

[28] Vanderbei, R.J. and Shanno, D.F. *An interior-point algorithm for nonconvex nonlinear programming*, Comput. Optim. Appl., 13 (1999) , 231–252.

[29] Wang, G.Q. *A new polynomial interior-point algorithm for the monotone linear complementarity problem over symmetric cones with full NT-steps*, Asia-Pac. J. Oper. Res., 29(2) (2012), 1250015.

[30] Zaoui, B., Benterki, D. and Khelladi, S. *New efficient descent direction of a primal-dual path-following algorithm for linear programming*, Stat. Optim. Inf. Comput., 12 (2024), 1098–1112.

[31] Zaoui, B., Benterki, D. and Khelladi, S. *Efficient descent direction of a primal-dual interior point algorithm for convex quadratic optimization*, J. Inf. Optim. Sci., (2025). Available online.

[32] Zaoui, B., Benterki, D., Kraria, A. and Raouache, H. *Interior-point algorithm for linear programming based on a new descent direction*, RAIRO Oper. Res., 57 (2023), 2473–2491.

[33] Zhang, M., Bai Y.Q. and Wang G.Q. *A new primal-dual path-following interior-point algorithm for linearly constrained convex optimization*, Journal of Shanghai University (English Edition), 12 (2008), 475–480.

[34] Zhang, M., Huang K., Li M. and Lv Y. *A new full-Newton step interior-point method for $P_*(\kappa)$-LCP based on a positive-asymptotic kernel function*, J. Appl. Math. Comput., 64 (2020), 313–330.

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 1, 2026, pp 243–263
https://doi.org/10.22067/ijnao.2025.91431.1578
https://ijnao.um.ac.ir/

CC BY NC

Research Article

Quasilinearization method to solve several classes of nonlinear Lane–Emden equations by Gegenbauer polynomials

F. Sheikhi and B. Ghazanfari*,

Abstract

This paper presents a comprehensive numerical approach for solving various types of singular nonlinear Lane–Emden equations. The proposed method begins by applying the quasilinearization method, to transform the nonlinear differential equation into a sequence of linear equations. Subsequently, these linear equations are solved using the collocation method with Gegenbauer functions. A notable advantage of Gegenbauer polynomials lies in their adjustable parameter, when appropriately selected; this parameter enhances their effectiveness in obtaining numerical solutions for specific problems. After establishing the convergence of the method, we present numerical examples to demonstrate its effectiveness. A comparison of the results with those obtained using other basis functions highlights the superior performance and applicability of these polynomials.

AMS subject classifications (2020): 34A25; 34A34; 42C10; 65L05.

*Corresponding author

Received 31 December 2024; revised 13 April 2025; accepted 7 May 2025

Fateme Sheikhi
Department of Mathematics, Lorestan University, Khorramabad, 68151-44316, Iran. e-mail: sheikhi.fa@fs.lu.ac.ir

Bahman Ghazanfari
Department of Mathematics, Lorestan University, Khorramabad, 68151-44316, Iran. e-mail:ghazanfari.ba@lu.ac.ir

How to cite this article
Sheikhi, F. and Ghazanfari, B., Quasilinearization method to solve several classes of nonlinear Lane–Emden equations by Gegenbauer polynomials. *Iran. J. Numer. Anal. Optim.*, 2026; 16(1): 243-263. https://doi.org/10.22067/ijnao.2025.91431.1578

Keywords: Nonlinear Volterra integro-differential equation; Legendre polynomial; Error analysis.

1 Introduction

We investigate the following nonlinear differential equations, which fall under the classification of Lane–Emden type equations:

$$u''(x) + \frac{\tau}{x}u'(x) + g(x,u) = z(x), \qquad x \geq 0,\ \tau \geq 1, \tag{1}$$

subject to the initial conditions

$$u(0) = \lambda, \qquad u'(0) = 0. \tag{2}$$

It should be noted that $g(x,u)$ is continuous and z belongs to $C[0,1]$. These equations incorporate various specific forms of $g(x,u)$, which model a wide range of problems in mathematics, physics, and astrophysics, including phenomena such as thermal explosions and stellar structures, among others [5, 18]. This type of equation is named after the astrophysicists Jonathan Homer Lane and Robert Emden. It describes the variation of density as a function of radial distance for a polytrope.

Nowadays, various scientific problems are formulated in infinite or semi-infinite domains [2]. To address these challenges, several spectral methods have been proposed [2, 3]. Additionally, a variety of approaches have been introduced in the literature for solving these types of equations. The emphasis on numerical methods stems from the challenges associated with obtaining accurate and efficient analytical solutions. In many scientific applications, the advantages of these methods over analytical or semi-analytical approaches have been clearly demonstrated.

Researchers have employed various numerical methods to solve these equations. For instance, Wazwaz [21] applied the Adomian decomposition method; Mandelzweig and Tabakin [13] utilized the quasilinearization method; Yousefi [22] implemented the Legendre wavelets method; Parand and Khaleghi [16] used the rational Chebyshev (second kind) collocation method; Bhrawy and Alofi [1] applied the Jacobi–Gauss collocation method, and Nasab et al. [12] employed the Chebyshev wavelet finite difference method.

The advantages of Gegenbauer polynomials include their flexibility, high accuracy, and adaptability to a wide range of problems.

• These polynomials are characterized by a tunable parameter, which allows them to be tailored

to specific applications, such as solving differential equations or approximating functions.
• Their orthogonal properties make them particularly effective for spectral methods, ensuring rapid convergence.
• Additionally, Gegenbauer polynomials can handle singularities and boundary conditions more effectively than many other basis functions.
They are a powerful tool in computational mathematics and physics [14, 19].

Since no prior work has explored the use of Gegenbauer polynomials as basis functions in the quasilinearization method (QLM), we propose a novel approach for solving Lane–Emden equations by integrating the QLM with these polynomials as basis functions.

Our approach involves a systematic pattern for solving Lane–Emden equations. Initially, the parameter of the Gegenbauer polynomials is estimated to achieve high-accuracy approximate solutions for problems with known exact solutions. Subsequently, this parameter, in conjunction with the QLM, is employed to derive approximate solutions for a broader class of equations. After establishing the convergence of the method, we present several illustrative examples to demonstrate its effectiveness. Finally, we compare our results with those obtained using alternative methods, highlighting the efficiency and applicability of the proposed scheme.

This study is structured as follows:
In Section 2, the definitions and preliminary concepts used throughout the article are briefly outlined. The proposed technique is described in Section 3. In Section 4, our scheme is applied to various numerical examples. Finally, the conclusion is presented in the last section.

2 Preliminaries

We provide a brief introduction to the definitions and formulations of polynomials used in the method outlined in this article. Furthermore, we discuss several types of Lane–Emden equations.

2.1 Some types of polynomials

2.1.1 Bernoulli polynomials

Initially, Bernoulli numbers [6] can be derived using the following expression:

$$\frac{x}{e^x-1}=\sum_{n=0}^{\infty}\delta_n\frac{x^n}{n!}. \tag{3}$$

It is also evident that

$$\delta_{2k+1}=0, \qquad k=1,2,3,\ldots.$$

The Bernoulli polynomials, named after Jacob Bernoulli, of order n are combinations of Bernoulli numbers and binomial coefficients, and are expressed as follows [17, 4]:

$$\beta_n(x)=\sum_{k=0}^{n}\binom{n}{k}\delta_{n-k}x^k, \qquad n=0,1,\ldots, \tag{4}$$

where δ_{n-k} represents the Bernoulli numbers.

2.1.2 Fractional order of rational Bernoulli functions

The rational Bernoulli functions, for $L>0$, are defined as follows: [4]:

$$RB_m(x;L)=\beta_m\Big(\frac{x}{x+L}\Big)=\sum_{k=0}^{n}\binom{n}{k}\delta_{n-k}(\frac{x}{x+L})^k, \qquad m=0,1,2,\ldots. \tag{5}$$

By applying the algebraic mapping $x\to x^\alpha$, these functions can be expressed in fractional order as follows [17]:

$$FRB_n(x,L)=R_n(x^\alpha,L)=\sum_{k=0}^{n}\binom{n}{k}\beta_k\Big(\frac{x^\alpha}{x^\alpha+L}\Big)^{n-k}, \qquad n=0,1,\ldots. \tag{6}$$

2.1.3 Rational Bessel functions

The solutions to Bessel's differential equation of order n are known as Bessel functions. The equation is expressed as follows:

$$x^2\frac{d^2u(x)}{dx^2}+x\frac{du(x)}{dx}+(x^2-n^2)u(x)=0, \qquad x\in(-\infty,\infty). \tag{7}$$

The Bessel polynomials, which form an orthogonal sequence of polynomials [15], are defined as follows:

$$B_n(x)=\sum_{r=0}^{\lfloor\frac{N-n}{2}\rfloor}\frac{(-1)^r}{r!(n+r)!}\Big(\frac{x}{2}\Big)^{2r+n}, \qquad x\in[0,1], \tag{8}$$

where $n \in \mathbb{N}$, and the number of bases of the Bessel polynomials is denoted as N.

Rational Bessel functions (RBs), derived using the algebraic mapping $x \to \frac{x}{x+L}$ in (8), are expressed as follows:

$$RB_n(x;L) = B_n(\frac{x}{x+L}) = \sum_{r=0}^{\lfloor \frac{N-n}{2} \rfloor} \frac{(-1)^r}{r!(n+r)!} \Big(\frac{x}{2(x+L)}\Big)^{2r+n}, \quad n = 0, 1, \ldots, N. \tag{9}$$

2.1.4 Exponential Bessel functions

The exponential Bessel functions (EBs), similar to the previous definition [15], are derived from (8) using the algebraic mapping $x \to 1 - e^{-\frac{x}{L}}$. They are defined as follows:

$$EB_n(x;L) = B_n(1 - e^{-\frac{x}{L}}) = \sum_{r=0}^{\lfloor \frac{N-n}{2} \rfloor} \frac{(-1)^r}{r!(n+r)!} \Big(1 - e^{-\frac{x}{L}}\Big)^{2r+n}, \quad n = 0, 1, \ldots, N. \tag{10}$$

In both definitions, the parameter L is a constant and acts as the scaling factor for the mapping.

2.1.5 Gegenbauer polynomials

Gegenbauer polynomials, also known as ultra-spherical polynomials $C_j^{\alpha}(x)$, are defined for non-negative integers $j \in \mathbb{Z}^+$ and are associated with the parameter $\alpha > -1/2$. These polynomials are named after Leopold Gegenbauer. Their analytical form is given by

$$C_j^{\alpha}(x) = \sum_{k=0}^{j} (-1)^{j-k} \frac{j!\Gamma(\alpha + 1/n)\Gamma(j+k+2\alpha)}{\Gamma(k+\alpha+1/2)\Gamma(j+2\alpha)(j-k)!k!L^k} (\frac{x+1}{2})^k, \tag{11}$$

where $j = 0, 1, 2, \ldots$.
When $\alpha = 0$, the Gegenbauer polynomial $C_j^0(x)$ reduces to the Chebyshev polynomials of the first kind, denoted as $T_j(x)$. For $\alpha = \frac{1}{2}$, the polynomial $C_j^{\frac{1}{2}}(x)$ corresponds to the Legendre polynomials $L_j(x)$. Furthermore, when $\alpha = 1$, the polynomial $C_j^1(x)$ is equal to $\frac{1}{j+1}U_j(x)$, where $U_j(x)$ represents the Chebyshev polynomials of the second type [9].

2.2 Some types of Lane–Emden equations

Lane–Emden type equations play a significant role in various fields of science and technology, leading to different forms of $g(x,u)$ and $z(x)$. For the case where $g(x,u) = u^m(x)$, $\tau = 2$, $z(x) = 0$, $\lambda = 1$ in (1) and (2), we obtain the standard Lane–Emden equation. This equation is used to model the thermal behavior of a spherical gas cloud due to the mutual attraction of its molecules, following the principles of classical thermodynamics. It is expressed as follows:

$$u''(x) + \frac{2}{x}u'(x) + u^m(x) = 0, \qquad u(0) = 1, \qquad u'(0) = 0.$$

These equations have analytical solutions for $m = 0, 1, 5$. Another variant of the Lane–Emden equation, introduced by Bonner, is characterized by $\tau = 2$, $g(x,u) = e^{u(x)}$, $z(x) = 0$, and $\lambda = 0$. This variant is used to model isothermal gas spheres and is expressed as follows:

$$u''(x) + \frac{2}{x}u'(x) + e^{u(x)} = 0, \qquad u(0) = 0, \qquad u'(0) = 0.$$

This equation models frictional gas spheres embedded in a pressurized medium, representing the maximum mass for hydrostatic equilibrium. Due to its physical similarities with the first type of Lane–Emden equation, it is often referred to as the second type of Lane–Emden equation.

3 Description of the present technique

3.1 The quasilinearization method

QLM is a powerful method for solving nonlinear differential equations. The process involves the following key steps:

• **Linearization:** At each iteration, the nonlinear components of the differential equation are linearized around the current approximate solution, often using a Taylor series expansion.

• **Iteration:** The linearized equation is solved to obtain a new approximate solution. This method is applied iteratively, generating a sequence of approximations that progressively converge to the accurate solution of the nonlinear problem.

• **Convergence:** Under certain conditions, the sequence of solutions converges to the exact solution of the original nonlinear differential equation.

• **Applications:** QLM is widely used in fields such as physics, engineering, and applied mathematics, especially for problems where direct methods are challenging to apply.

This method is particularly effective for solving initial value or boundary value problems and can be efficiently implemented using numerical techniques.

Kalaba and Bellman [11] were the pioneers of the QLM. They developed this method as a systematic strategy to address nonlinear problems by leveraging the tools of linear analysis. Over the decades, the technique has found widespread application in fields such as engineering, physics, and economics, particularly in scenarios where nonlinear behaviors are dominant [3].

In this section, we focus our discussion on nonlinear ordinary differential equations involving a single variable, defined over the semi-infinite interval $[0,\infty)$, as presented below:

$$L^n u(x) = f\Big(u(x), u^{(1)}(x), u^{(2)}(x), \ldots, u^{(n-1)}(x), x\Big), \tag{12}$$

under the initial conditions

$$p_k\Big(u(0), u^{(1)}(0), \ldots, u^{(n-1)}(0)\Big) = 0, \qquad k = 1, \ldots, l, \tag{13}$$

and

$$p_k\Big(u(b), u^{(1)}(b), \ldots, u^{(n-1)}(b)\Big) = 0, \qquad k = l+1, \ldots, n. \tag{14}$$

In this context, L^n denotes a linear nth order ordinary differential operator, while f and p_i, for $i = 1, \ldots, n$, are nonlinear functions of $u(x)$ and its derivatives. This technique generates the iterative estimate $u_{r+1}(x)$ to the solution of (12) as a solution to the subsequent linear differential equation:

$$\begin{aligned} L^n u_{r+1}(x) =& f\Big(u_r(x), u_r^{(1)}(x), \ldots, u_r^{(n-1)}(x), x\Big) \\ &+ \sum_{s=0}^{n-1} \Big(u_{r+1}^{(s)}(x) - u_r^{(s)}(x)\Big) f_{u^{(s)}}\Big(u_r(x), u_r^{(1)}(x), \ldots, u_r^{(n-1)}(x), x\Big), \end{aligned} \tag{15}$$

with the following linearized initial conditions:

$$\begin{aligned} \sum_{s=0}^{n-1} \Big(u_{r+1}^{(s)}(0) - u_r^{(s)}(0)\Big) p_{k_{u^{(s)}}}\Big(u_r(0), u_r^{(1)}(0), \ldots, u_r^{(n-1)}(0), 0\Big) = 0, \\ k = 1, \ldots, l, \end{aligned} \tag{16}$$

and

$$\begin{aligned} \sum_{s=0}^{n-1} \Big(u_{r+1}^{(s)}(b) - u_r^{(s)}(b)\Big) p_{k_{u^{(s)}}}\Big(u_r(b), u_r^{(1)}(b), \ldots, u_r^{(n-1)}(b), b\Big) = 0, \\ k = l+1, \ldots, n, \end{aligned} \tag{17}$$

where $u_r^{(0)}(x) = u_r(x)$. The initial approximation $u_0(x)$ is chosen based on mathematical or physical principles. Additionally, the following functions

$$f_{u^{(s)}} = \frac{\partial^s f}{\partial u^{(s)}},$$

and

$$p_{k\,u^{(s)}} = \frac{\partial^s p_k}{\partial u^{(s)}},$$

where $s = 1, 2, \ldots, n-1$, represent the partial derivatives of the respective functions as follows:

$$f(u_r(x), u_r^{(1)}(x), \ldots, u_r^{(n-1)}(x), x),$$

$$p_{k\,u^{(s)}}(u_r(x), u_r^{(1)}(x), \ldots, u_r^{(n-1)}(x), x).$$

3.2 Application of the methods

Consider the Lane–Emden equation (1), subject to the initial conditions specified in (2). The proposed approach is based on the QLM, which transforms the original nonlinear equation (1) into a sequence of linear equations.

The Lane–Emden type equations are formulated as follows:

$$u''(x) + \frac{\tau}{x}u'(x) + g(x,u) = z(x), \qquad \tau > 1,\ x \geq 0, \tag{18}$$

under the initial conditions

$$u(0) = \lambda, \qquad u'(0) = 0. \tag{19}$$

Applying the QLM to the Lane–Emden type equations yields an iterative estimate, denoted as $u_{r+1}(x)$, which is expressed as follows:

$$\begin{aligned} &u''_{r+1}(x) + \frac{\tau}{x}u'_r(x) + f(x, u_r(x)) \\ &- (u_{r+1}(x) - u_r(x))f_u(x, u_r(x)) + \frac{\tau}{x}(u'_{r+1}(x) - u'_r(x)) = 0, \end{aligned} \tag{20}$$

under the initial conditions:

$$u_{r+1}(0) = \lambda, \qquad u'_{r+1}(0) = 0, \tag{21}$$

We begin by selecting a reasonable initial approximation for the function $y(x)$, denoted as $y_0(x)$. Here, we set $y_0(x) = A$. Since $y_0(x)$ is a known function of x, (20) is reformulated into a linear

differential equation with variable coefficients. The proposed solution for the $(r+1)$th estimate is initially defined as follows:

$$S_{n,r}(x) = \sum_{i=0}^{n} a_{i,r} f_i(x; L), \tag{22}$$

where $f_i(x; L)$ represent the basic functions, n is a positive integer, and $a_{i,r}$ are the unknown coefficients. The basis functions, defined in Section 2, are employed in the collocation method.

Our goal is to determine the coefficients $a_{i,r}$ as Gegenbauer functions. The initial conditions in (20) are then satisfied through the following relation:

$$u_{n,r+1}(x) = \lambda + x^2 S_{n,r}(x). \tag{23}$$

By substituting (23) in (20), the residual function at each iteration of the QLM is formulated as follows:

$$\begin{aligned} Res_r(x) =& u''_{n,r+1}(x) + \frac{\tau}{x} u'_{n,r}(x) + f(x, u_{n,r}(x)) - z(x) \\ & - (u_{n,r+1}(x) - u_{n,r}(x)) f_u(x, u(x)) + \frac{\tau}{x}(u'_{n,r+1}(x) - u'_{n,r}(x)). \end{aligned} \tag{24}$$

Next, we collocate (24) at the points provided in [17]:

$$x_t = L \left(\frac{1 - \cos(\frac{(2j-1)\pi}{(N+1)2})}{2} \right)^{\frac{1}{\gamma}}, \qquad t = 1, 2, \ldots, N+1, \tag{25}$$

where these points correspond to the roots of the generalized first-kind Chebyshev functions on the range $[0, L]$, with the parameter γ for the collocation points set to $\frac{1}{2}$. By substituting (25) into (24), $N+1$ algebraic equations can be obtained and solved. We obtain $N+1$ algebraic equations, which can then be solved. The primary objective of the collocation method is to determine the unknown coefficients $a_{i,r}$, by minimizing the residual function.

3.3 Convergence and existence of the method

In this section, the issues of convergence and existence will be briefly discussed [20]. It has been demonstrated that the sequence of functions $u_r(x)$ converges to the exact solution of the original equation when the value of x_c, or the interval $[0, x_c]$, is sufficiently small.

To demonstrate that the presented technique achieves quadratic convergence to the solution of (12) with the boundary conditions specified in (13), we simplify our analysis to the second-

order differential equation specified given by

$$u''_{r+1}(x) = f(u_r(x), x) + (u_{r+1}(x) - u_r(x)) f_{u_r}(u_r(x), x). \tag{26}$$

The boundary conditions for this equation are defined as follows:

$$u_{r+1}(0) = u_{r+1}(x_c) = 0. \tag{27}$$

For further details, refer to [11].

3.4 Existence

To prove the existence of the sequence defined by (15), we use an integral representation to express the solution to this system. This can be achieved by applying Green's function. Specifically, the Green's function $G(x, \theta)$ for the homogeneous system described by

$$u''_{r+1}(x) = 0, \qquad u_{r+1}(0) = u_{r+1}(x_c) = 0, \tag{28}$$

can be easily derived as follows [7, 8]:

$$G(x, \theta) = \frac{x(\theta - x_c)}{x_c}, \qquad 0 \leq x \leq \theta, \tag{29}$$

$$G(x, \theta) = \frac{\theta(x - x_c)}{x_c}, \qquad \theta \leq x \leq x_c. \tag{30}$$

If we express (26) in the form

$$u''_{r+1}(x) = f(x), \tag{31}$$

then (26) can be represented by the following linear integral equation:

$$u_{r+1}(x) = \int_0^{x_c} G(x, \theta)[f(u_r(x), x) + (u_{r+1}(x) - u_r(x)) f_{u_r}(u_r(x), x)] d\theta. \tag{32}$$

To estimate the bound on $u_{r+1}(x)$, we consider the absolute values in the integral representation (32) as follows:

$$\begin{aligned} |u_{r+1}(x)| \leq \int_0^{x_c} |G(x, \theta)| [& |f(u_r(x), x)| + |u_r(x)||f_{u_r}(u_r(x), x)| \\ & + |u_{r+1}(x)||f_{u_r}(u_r(x), x)|] d\theta. \end{aligned} \tag{33}$$

Let us denote the maximum values of $\max|f(u_r(x),x)|$ and $\max|f_{u_r}(u_r(x),x)|$ by M. Thus, we can rewrite the inequality as follows:

$$|u_{r+1}(x)|\leq \int_0^{x_c} |G(x,\theta)|[M+M|u_r(x)|+M|u_{r+1}(x)|]d\theta. \tag{34}$$

Now, assuming that $|u_0(x)|\leq 1$ for $r=0$, we can simplify (34) to

$$|u_1(x)|\leq \int_0^{x_c} |G(x,\theta)|[2M+M|u_1(x)|]d\theta. \tag{35}$$

This inequality provides a way to estimate the bound on $|u_1(x)|$ in terms of the Green's function and the maximum values of the functions involved. To further analyze this, we can isolate $|u_1(x)|$ on one side of the inequality and derive a more explicit bound.

From (29) and (30), we can determine that

$$max|G(x,\theta)| = \frac{x_c}{4}. \tag{36}$$

Using this result, we can rewrite the bound on $|u_1(x)|$ as follows:

$$|u_1(x)|\leq \frac{x_c}{4}\int_0^{x_c} [2M+M|u_1(x)|]d\theta. \tag{37}$$

Integrating both sides of (37) and solving it for $|u_1(x)|$, give

$$|u_1(x)|\leq \frac{M\frac{x_c^2}{2}}{1-M\frac{x_c^2}{4}}. \tag{38}$$

If we choose x_c to be sufficiently small, then we can ensure that $|u_1(x)|$ remains bounded. Specifically, if

$$x_c^2\leq \frac{4}{3M},$$

then it follows that

$$|u_1(x)|\leq 1.$$

We can apply the same procedure iteratively to $u_2(x), u_3(x), \ldots$, and conclude that, under the condition $x_c^2\leq \frac{4}{3M}$,

$$u_r(x)\leq 1, \qquad 0\leq x\leq x_c.$$

Thus, we can conclude that the sequence $u_r(x)$ is bounded for sufficiently small intervals $[0, x_c]$.

3.5 Convergence

To prove the quadratic convergence to the solution of (26) and (27), we rewrite (26) in a more explicit form. Start from

$$u_r''(x) = f(u_{r-1}(x), x) + f_{u_{r-1}}(u_{r-1}(x), x)(u_r(x) - u_{r-1}(x)). \tag{39}$$

By subtracting (39) from (26), the following expression for the second derivative of $u_{r-1}(x)$ is derived:

$$\begin{aligned} u_{r+1}''(x) =& f(u_r(x), x) - f(u_{r-1}(x), x) - f_{u_{r-1}}(u_{r-1}(x), x)(u_r(x) - u_{r-1}(x)) \\ &+ f_{u_r}(u_r(x), x)(u_{r+1}(x) - u_r(x)), \end{aligned} \tag{40}$$

Here, $(u_{r+1}(x) - u_r(x))$ is the unknown quantity we want to analyze.

By applying the mean value theorem, the difference $f(u_r(x), x) - f(u_{r-1}(x), x)$ can be expressed as follows:

$$\begin{aligned} f(u_r(x), x) - f(u_{r-1}(x), x) - (u_r(x) - u_{r-1}(x)) f_{u_{r-1}}(u_{r-1}(x), x) \\ = \frac{1}{2}(u_r(x) - u_{r-1}(x))^2 f_{vv}(v), \end{aligned} \tag{41}$$

where v lies between $u_{r-1}(x)$ and $u_r(x)$. Substituting (41) into (40), we obtain

$$\begin{aligned} u_{r+1}(x) - u_r(x) = \int_0^{x_c} G(x, \theta)[\frac{1}{2}(u_r(x) - u_{r-1}(x))^2 f_{vv}(v) \\ + (u_{r+1}(x) - u_r(x)) f_{u_r}(u_r(x), x)] d\theta. \end{aligned} \tag{42}$$

Let $m_1 = max|f_{vv}(v)|$. From this, we can derive the following inequality:

$$|u_{r+1}(x) - u_r(x)| \leq \frac{x_c}{4} \int_0^{x_c} [\frac{m_1}{2}(u_r(x) - u_{r-1}(x)^2 + m|u_{r+1}(x) - u_r(x)|] d\theta. \tag{43}$$

Rearranging (43), we find

$$|u_{r+1}(x) - u_r(x)| \leq \frac{m_1 \frac{x_c^2}{8}}{1 - m\frac{x_c^2}{4}} (u_r(x) - u_{r-1}(x))^2 = k(|u_r(x) - u_{r-1}(x)|)^2. \tag{44}$$

Equation (44) indicates that if there is convergence, then it is quadratic. To further illustrate this, we can rewrite the inequality recursively:

$$|u_{r+1}(x) - u_r(x)| \leq k(|u_r(x) - u_{r-1}(x)|)^2,$$

which can be expanded as

$$|u_r(x) - u_{r-1}(x)| \leq k(|u_{r-1}(x) - u_{r-2}(x)|)^2,$$

$$|u_{r-1}(x) - u_{r-2}(x)| \leq k(|u_{r-2}(x) - u_{r-3}(x)|)^2,$$

and so on, leading to:

$$|u_2(x) - u_1(x)| \leq k(|u_1(x) - u_0(x)|)^2.$$

This recursive relationship shows that each successive error term is bounded by a constant k multiplied by the square of the previous error term.

To proceed with the analysis using simple substitution, we introduce a new variable to simplify the expressions. Let us define

$$|u_{r+1}(x) - u_r(x)| \leq k(|u_r(x) - u_{r-1}(x)|)^2 \leq k(k(|u_{r-1}(x) - u_{r-2}(x)|)^2)^2 \leq \cdots .$$

To derive the inequality, we start from the established recursive relationship:

$$|u_{r+1}(x) - u_r(x)| \leq k[k^{2^r-2}(|u_1(x) - u_0(x)|)^{2^r}] = \frac{[k(u_1(x) - u_0(x)]^{2^r}}{k}. \tag{45}$$

If the quantity $[k(u_1(x) - u_0(x))] < 1$, then as r increases. The right-hand side of (45) will approach zero. Consequently, $u_r(x)$ will approach to a function $u(x)$ and (32) reduces to

$$u(x) = \int_0^{x_c} G(x, \theta) f(u(x)) d\theta. \tag{46}$$

Therefore, the function $u(x)$ satisfies the original equation. Note that if the value of x or the interval $[0, x]$ is sufficiently small, the quantity $k(|u_1(x) - u_0(x)|)$ will be less than one. It is also important to note that this quantity depends on the maximum value of $|u_1(x) - u_0(x)|$. Therefore, if the interval $[0, x_c]$ is excessively large, then we can theoretically select a better initial approximation $u_0(x)$ to ensure that the maximum difference between the absolute values of $u_1(x)$ and $u_0(x)$ is small enough to keep $[k(|u_1(x) - u_0(x)|)]$ less than one.

4 Illustrative examples

In this section, we address various nonlinear Lane–Emden differential equations using the QLM and collocation technique with Gegenbauer polynomials (with parameter $\alpha = 0.45$). Subsequently, we compare the results obtained from the proposed method with those derived from alternative approaches and exact solutions.

Example 1. For $\tau = 2$, $g(x,u) = u^m(x)$, $z(x) = 0$, $\lambda = 1$, (1) and (2) reduce to the standard Lane–Emden equations:

$$u''(x) + \frac{2}{x}u'(x) + u^m(x) = 0, \qquad x > 0, \tag{47}$$

under the initial conditions:

$$u(0) = 1, \qquad u'(0) = 0,$$

where m is a constant. Replacing $m = 0, 1, 5$ into (47) yields the following exact solutions:
1. $m = 0$:

$$u(x) = 1 - \frac{1}{3!}x^2,$$

2. $m = 1$:

$$u(x) = \frac{\sin(x)}{x},$$

and
3. $m = 5$:

$$u(x) = \frac{1}{\sqrt{1 + \frac{x^2}{3}}}.$$

In cases where an exact analytical solution is unavailable, we use (20) and (21) to obtain

$$u''_{r+1}(x) + \frac{2}{x}u'_{r+1}(x) + mu_{r+1}(x)u_r^{(m-1)}(x) - (m-1)u_r^{(m)}(x) = 0, \tag{48}$$

under the following initial conditions:

$$u_{r+1}(0) = 1, \qquad u'_{r+1}(0) = 0, \qquad r = 0, 1, \ldots, \text{max-iteration}. \tag{49}$$

Table 1: Analysis of the first zeros of $u(x)$ obtained for $m = 2, 3$ and 4, and $N = 50$ of Example 1.

m	x	FRB [17]	RBs [15]	EBs [15]	$C^{\alpha}, \alpha = 0.45$	[10]	[16]
2	0.1	0.9983349985	0.9983349985	0.9983349985	0.9983349985	0.9983350	0.99833499
	0.5	0.9593527158	0.9593527158	0.9593527158	0.9593527158	0.9593527	0.95935270
	1.0	0.8486541114	0.8486541114	0.8486541114	0.8486541114	0.8486541	0.84865410
	3.0	0.2418240830	0.2418240830	0.2418240830	0.2418240830	0.2418241	0.24182408
	4.0	0.0488401499	0.0488401499	0.0488401499	0.0488401499	0.0488401	0.04884013
	4.3	0.0068109	0.006810943	0.0068109432	0.00681094	0.0068109	0.00681093
3	0.1	0.9983358295	0.9983358295	0.9983358295	0.9983358295	0.9983358	0.99833582
	0.5	0.9598390699	0.9598390699	0.9598390699	0.9598390699	0.9598391	0.95983906
	1.0	0.8550575685	0.8550575685	0.8550575685	0.8550575685	0.8550576	0.85505756
	3.0	0.3592265006	0.3592265006	0.35922650	0.3592265006	0.35922650	0.35922650
	4.0	0.2092816133	0.2092816133	0.20928161	0.2092816133	0.20928161	0.20928161
	4.3	0.1754138726	0.1754138726	0.17541387	0.1754138726	0.17541387	0.17541387
4	0.1	0.9983366613	0.9983366595	0.9983366595	0.9983366595	0.9.983367	1.0158508
	0.5	0.9603108972	0.9603109023	0.9603109023	0.9603109023	0.9.603109	0.9603108
	1.0	0.8608141488	0.8608138122	0.8608138122	0.8608138122	0.8608138	0.8608138
	3.0	0.4400560799	0.4400506915	0.4400506915	0.4400506923	0.44005068	0.4400555
	4.0	0.3180503821	0.3180424290	0.3180424290	0.3180424441	0.31804312	0.3180442
	4.3	0.2900786481	0.2900624969	0.2900624969	0.2900621154	0.29006281	0.2900622

We assume an initial guess $u_0(x) = 1$. Based on the current approach, we define the following residual function:

$$Res_r(x) = u''_{n,r+1}(x) + \frac{2}{x}u'_{n,r+1}(x) + mu_{n,r+1}(x)u^{(m-1)}_{n,r}(x) - (m-1)u^{(m)}_{n,r}(x). \tag{50}$$

We apply the collocation method to (50) at $N + 1$ points, as specified in (25), to construct a system of $N + 1$ algebraic equations. These equations are solved using Newton's iterative method. Numerical results approximated for $n = 50$ with nonlinear exponent values $m = 2, 3, 4$, are reported to 10 decimal places in Table 1 and compared with references [15, 17, 10, 16]. Additionally, the absolute errors and maximum absolute errors of the approximate solutions for $L = 4.35$, and exponents $m = 0, 1, 5$, are presented in Tables 2 and 3, respectively. Finally, CPU times of $U_{10}(x)$ for different values of x and m are presented in Table 4.

Example 2. For $\tau = 2$, $g(x, u) = e^{u(x)}$, $z(x) = 0$, $\lambda = 0$, (1) and (2) reduce to

$$u''(x) + \frac{2}{x}u'(x) + e^{u(x)} = 0, \qquad x \geq 0 \tag{51}$$

under the following initial conditions:

$$u(0) = 0, \qquad u'(0) = 0. \tag{52}$$

Using (20) and (21), we obtain

$$u''_{r+1}(x) + \frac{2}{x}u'_{r+1}(x) + e^{u_r(x)}\Big(u_{r+1}(x) - u_r(x) + 1\Big) = 0, \tag{53}$$

Table 2: Analysis of the absolute errors in the approximate solutions of Example 1 for $N = 50$, $L = 4.35$, and 10th iteration.

m	x	FRB [17]	RBs [15]	EBs [15]	$C^\alpha, \alpha = 0.45$
0	0.1	0	$1.65\,10^{-16}$	$4.04\,10^{-17}$	$1.69\,10^{-53}$
	0.5	$3.00\,10^{-60}$	$1.17\,10^{-17}$	$1.22\,10^{-17}$	$1.66\,10^{-46}$
	1.0	$2.00\,10^{-60}$	$7.93\,10^{-18}$	$3.30\,10^{-19}$	$6.73\,10^{-42}$
	3.0	$2.95\,10^{-58}$	$7.90\,10^{-18}$	$9.22\,10^{-19}$	$2.18\,10^{-34}$
	4.0	$6.18\,10^{-58}$	$7.89\,10^{-18}$	$9.21\,10^{-19}$	$2.33\,10^{-32}$
	4.3	$9.1\,10^{-58}$	$7.89\,10^{-18}$	$9.21\,10^{-19}$	$2.90\,10^{-34}$
1	0.1	$1.36\,10^{-23}$	$1.65\,10^{-16}$	$4.04\,10^{-17}$	$1.13\,10^{-53}$
	0.5	$3.39\,10^{-22}$	$1.13\,10^{-17}$	$1.22\,10^{-17}$	$1.11\,10^{-46}$
	1.0	$1.97\,10^{-21}$	$6.65\,10^{-18}$	$1.81\,10^{-19}$	$4.52\,10^{-42}$
	3.0	$1.07\,10^{-20}$	$3.45\,10^{-19}$	$4.05\,10^{-20}$	$1.46\,10^{-34}$
	4.0	$3.05\,10^{-20}$	$1.50\,10^{-18}$	$1.75\,10^{-19}$	$1.56\,10^{-32}$
	4.3	$8.60\,10^{-21}$	$1.68\,10^{-18}$	$1.96\,10^{-19}$	$1.31\,10^{-34}$
5	0.1	$1.56\,10^{-16}$	$1.65\,10^{-16}$	$4.04\,10^{-17}$	$4.14\,10^{-33}$
	0.5	$8.04\,10^{-15}$	$1.01\,10^{-17}$	$1.21\,10^{-17}$	$1.09\,10^{-25}$
	1.0	$8.52\,10^{-14}$	$3.35\,10^{-18}$	$2.03\,10^{-19}$	$1.69\,10^{-20}$
	3.0	$3.84\,10^{-13}$	$1.99\,10^{-18}$	$2.32\,10^{-19}$	$7.40\,10^{-13}$
	4.0	$1.71\,10^{-12}$	$2.14\,10^{-18}$	$2.50\,10^{-19}$	$7.61\,10^{-11}$
	4.3	$2.38\,10^{-12}$	$2.11\,10^{-18}$	$2.47\,10^{-19}$	$1.98\,10^{-12}$

Table 3: The maximum absolute errors of Example 1 for $N = 50$ and $m = 0, 1, 5$.

m	FRB [17]	RBs [15]	EBs [15]	$C^\alpha, \alpha = 0.45$
0	$6.18\,10^{-58}$	$1.65\,10^{-16}$	$4.04\,10^{-17}$	$2.33\,10^{-32}$
1	$3.05\,10^{-20}$	$1.65\,10^{-16}$	$4.04\,10^{-17}$	$1.56\,10^{-32}$
5	$4.773\,10^{-12}$	$1.65\,10^{-16}$	$4.04\,10^{-17}$	$7.61\,10^{-11}$

Table 4: The CPU times of Example 1 by the presented method for $u_{10}(x)$ in different values of x and m.

x	$m = 0$	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$
0.1	$2.88\,ms^*$	$3.52\,ms$	$3.16\,ms$	$3.59\,ms$	$3.00\,ms$	$2.29\,ms$
0.5	$3.48\,ms$	$4.45\,ms$	$3.55\,ms$	$3.71\,ms$	$3.53\,ms$	$2.8\,ms$
1.0	$875.00\,us^{**}$	$1.07\,ms$	$828.10\,us$	$853.10\,us$	$857.80\,us$	$760.90\,us$
3.0	$2.78\,ms$	$3.01\,ms$	$2.90\,ms$	$3.67\,ms$	$3.02\,ms$	$2.38\,ms$
4.0	$3.23\,ms$	$3.19\,ms$	$3.36\,ms$	$3.21\,ms$	$3.18\,ms$	$2.55\,ms$
4.3	$5.27\,ms$	$5.07\,ms$	$4.88\,ms$	$4.97\,ms$	$5.18\,ms$	$4.16\,ms$

* ms (millisecond),
** us (microsecond).

Table 5: The derived values of $u(x)$ for Example 2 via the presented method.

x	FRB [17]	RBs [15]	EBs [15]	$C^{\alpha}, \alpha = 0.45$	[21]	[16]	$CPUTime$
0.1	−0.001665833	−0.001665833	−0.001665833	−0.001665833	−0.001665833	−0.001666418	1.60 ms
0.2	−0.006653367	−0.006653367	−0.006653367	−0.006653367	−0.006653367	−0.006653971	2.03 ms
0.5	−0.041153957	−0.041153957	−0.041153957	−0.041153957	−0.041153956	−0.041154515	1.86 ms
1.0	−0.158827677	−0.158827677	−0.158827677	−0.158827677	−0.158827353	−0.158828173	568.70 us
1.5	−0.338019424	−0.338019424	−0.338019424	−0.338019424	−0.338013110	−0.338019830	2.20 ms
2.0	−0.559823004	−0.559823004	−0.559823004	−0.559823004	−0.559962660	−0.559823312	1.55 ms
2.5	−0.806340870	−0.806340870	−0.806340870	−0.806340870	−0.810019671	−0.806341084	2.35 ms

under the initial conditions,

$$u_{r+1}(0) = 0, \qquad u'_{r+1}(0) = 0, \qquad r = 0, 1, \ldots, \text{max-iteration}. \tag{54}$$

We set $u_0(x) = 0$, which implies that the initial guess satisfies the boundary conditions. Using the proposed method, we construct the following residual function:

$$Res_r(x) = u''_{n,r+1}(x) + \frac{2}{x}u'_{n,r+1}(x) + e^{u_{n,r+1}(x)}\Big(u_{n,r+1}(x) - u_{n,r}(x) + 1\Big). \tag{55}$$

In Table 5, the CPU times and approximate values of $u(x)$ obtained from the presented method with $n = 50$, at the 10th iteration, and $L = 2.5$, are compared with those obtained in [17, 15, 21, 16].

Example 3. For $\tau = 2$, $g(x, u) = \sinh(u(x))$, $\lambda = 1$ and $z(x) = 0$, (1) and (2) reduce to

$$u''(x) + \frac{2}{x}u'(x) + \sinh(u(x)) = 0, \qquad x \geq 0, \tag{56}$$

under the following initial conditions:

$$u(0) = 1, \qquad u'(0) = 0.$$

Using the QLM in (20) and (21), the iterative approximation $u_{r+1}(x)$ is determined as the solution to the following differential equation:

$$u''_{r+1}(x) + \frac{2}{x}u'_{r+1}(x) + \sinh(u_r(x)) + (u_{r+1}(x) - u_r(x))\cosh(u_r(x)) = 0, \tag{57}$$

under the initial conditions:

$$u_{r+1}(0) = 1, \qquad u'_{r+1}(0) = 0, \qquad r = 0, 1, \ldots, \text{max-iteration}. \tag{58}$$

We assume $u_0(x) = 1$. Using the proposed approach, we define the following residual function:

Table 6: The derived values of $u(x)$ for Example 3 via the presented method.

x	FRB [17]	RBs [15]	EBs [15]	$C^{\alpha}, \alpha = 0.45$	[21]	[16]	$CPUTime$
0.1	0.998042841	0.998042841	0.998042841	0.998042841	0.9980428	0.998042841	2.49 ms
0.2	0.992189434	0.992189434	0.992189434	0.992189434	0.9921894	0.992189434	2.80 ms
0.5	0.951961092	0.951961092	0.951961092	0.951961092	0.9519611	0.951961092	2.91 ms
1.0	0.818242928	0.818242928	0.040263157	0.818242928	0.8182516	0.818242928	592.20 us
1.5	0.625438763	0.625438763	0.625438763	0.625438763	0.62589160	0.625438763	3.03 ms
2.0	0.406622887	0.40662288	0.406622887	0.406622887	0.41366910	0.406622887	786.00 us

$$\begin{aligned} Res_r(x) =& u''_{r+1}(x) + \frac{2}{x}u'_{r+1}(x) + \sinh(u_r(x)) \\ &+ (u_{r+1}(x) - u_r(x))\cosh(u_r(x)) = 0. \end{aligned} \tag{59}$$

In Table 6, the CPU times and approximate values of $u(x)$ obtained from the presented method with $n = 50$, at the 10th iteration, and $L = 2$, are compared with those obtained in references [17, 15, 21, 16].

Example 4. If $\tau = 2$, $g(x, u) = \sin(u(x))$, $\lambda = 1$ and $z(x) = 0$, then (1) and (2) have been defined as follows:

$$u''(x) + \frac{2}{x}u'(x) + \sin(u(x)) = 0, \tag{60}$$

under the following initial conditions:

$$u(0) = 1, \qquad u'(0) = 1.$$

By using the QLM, we obtain

$$u''_{r+1}(x) + \frac{2}{x}u'_{r+1}(x) + \sin(u_r(x)) + \cos(u_r(x))\Big(u_{r+1}(x) - u_r(x)\Big) = 0, \tag{61}$$

under the initial conditions:

$$u_{r+1}(0) = 1, \qquad u'_{r+1}(0) = 0, \qquad r = 0, 1, 2, \ldots, \text{max-iteration}. \tag{62}$$

We consider $u_0(x) = 1$. Based on the present technique, we formulate the residual function as follows:

$$\begin{aligned} Res_r(x) =& u''_{n,r+1}(x) + \frac{2}{x}u'_{n,r+1}(x) + \sin(y_{n,r}(x)) \\ &+ \cos(u_{n,r}(x))\Big(u_{n,r+1}(x) - u_{n,r}(x)\Big). \end{aligned} \tag{63}$$

In Table 7, the CPU times and approximate values of $u(x)$ obtained from the presented method with $n = 40$, at the 10th iteration, and $L = 2$, are compared with those obtained in [17, 15, 21, 16].

Table 7: The derived values of $u(x)$ for Example 4 via the presented method.

x	FRB [17]	RBs [15]	EBs [15]	$C^{\alpha}, \alpha = 0.45$	[21]	[16]	$CPUTime$
0.1	0.998597927	0.998597927	0.998597927	0.998597927	0.9985979	0.998605142	$1.21\,ms$
0.2	0.994396264	0.994396264	0.994396264	0.994396264	0.9943962	0.994406270	$1.28\,ms$
0.5	0.965177780	0.965177780	0.965177780	0.965177780	0.96517778	0.965188168	$1.40\,ms$
1.0	0.863681125	0.863681125	0.863681125	0.863681125	0.86368110	0.863688130	$184.30\,us$
1.5	0.705045233	0.705045233	0.705045233	0.705045233	0.70504192	0.705052410	$1.49\,ms$
2.0	0.506463627	0.506463627	0.506463627	0.506463627	0.50637203	0.50646875	$484.40\,us$

5 Conclusion

We have proposed an efficient and accurate numerical technique to approximate solutions for singular nonlinear Lane–Emden equations with initial conditions. Specifically, we determined the parameter of Gegenbauer functions and combined the QLM with the collocation technique to solve various types of Lane–Emden equations.

The primary advantage of this approach lies in its ability to transform the nonlinear differential equation into a system of linear algebraic equations, which are simpler to solve. Furthermore, the collocation method minimizes the residual functions, enabling the determination of unknown coefficients. To compute the solutions and absolute errors, we developed a code in Maple software, which efficiently handled all computations.

Acknowledgements

I would like to thank the editor and esteemed reviewers for their valuable and helpful comments in improving our manuscript.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Bhrawy, A.H. and Alofi, A.S. *A Jacobi–Gauss collocation method for solving nonlinear Lane–Emden type equations* Commun. Nonlinear Sci. Numer. Simul., 17 (1) (2012), 62–70.

[2] Boyd, J.P. *Spectral methods using rational basis functions on an infinite interval* J. Comput. Phys., 69 (1) (1987), 112–142.

[3] Boyd, J.P. *Orthogonal rational functions on a semi-infinite interval* J. Comput. Phys., 70 (1) (1987), 63–88.

[4] Calvert, V., Mashayekhi, S. and Razzaghi, M. *Solution of Lane–Emden type equations using rational Bernoulli functions* Math. Methods Appl. Sci., 39 (5) (2016), 1268–1284.

[5] Chandrasekhar, S. *Introduction to the study of stellar structure* Dover Publ., New York (Dover reprint ed.) 1967.

[6] Cheon, G.S. *A note on the Bernoulli and Euler polynomials* Appl. Math. Lett., 16 (3) (2003), 365–368.

[7] Coddington, E. and Levinson, N. *Theory of ordinary differential Equations* McGraw-Hill, New York (1955).

[8] Courant, R. and Hilbert, D. *Methods of mathematical physics* Wiley (first English edn., rev.) 1953.

[9] Doha, E. H. *The coefficients of differentiated expansions and derivatives of ultraspherical polynomials* Comput. Math. Appl., 21 (2-3) (1991), 115–122.

[10] Horedt, G.P. *Seven-digit tables of Lane-Emden functions* Astrophys. Space Sci., 126 (1986), 357–408.

[11] Kalaba, R. *On nonlinear differential equations, the maximum operation, and monotone convergence* J. Math. Mech., 8 (1959), 519–574.

[12] Kazemi Nasab, A., Kilicman, A., Atabakan, Z. P. and Leong, W. *A numerical approach for solving singular nonlinear Lane–Emden type equations arising in astrophysics* New Astron., 34 (2015), 178–186.

[13] Mandelzweig, V.B. and Tabakin, F. *Quasilinearization approach to nonlinear problems in physics with application to nonlinear ODEs* Comput. Phys. Commun., 141 (2) (2001), 268–281.

[14] Olver, S. and Townsend, A. *A fast and well-conditioned spectral method* SIAM Rev., 55 (3) (2013), 462–489.

[15] Parand, K. and Ghaderi, A. *Two efficient computational algorithms to solve the singularly perturbed Lane-Emden problem* arXiv:1708.07384v1 (2017).

[16] Parand, K. and Khaleqi, S. *The rational Chebyshev of second kind collocation method for solving a class of astrophysics problems* Eur. Phys. J. Plus, 131 (2016), 1–24.

[17] Parand, K., Yousefi, H. and Delkhosh, M. *A numerical approach to solve Lane-Emden-type equations by the fractional order of rational Bernoulli functions* Rom. J. Phys., 62 (2017), 1–24.

[18] Richardson, O.W. *The emission of electricity from hot bodies* Longmans, Green and Co., London (2nd ed.) 1921.

[19] Sakar, F.M., Hussain, S. and Ahmad, I. *Application of Gegenbauer polynomials for biunivalent functions defined by subordination* Turk. J. Math., 46 (3) (2022), 1089–1098.

[20] Stanley, E. and Lee, E.S. *Quasilinearization and Invariant Imbedding* Math. Sci. Eng., 41 (1968), 9–39.

[21] Wazwaz, A.M. *A new algorithm for solving differential equations of Lane–Emden type* Appl. Math. Comput., 118 (2001), 287–310.

[22] Yousefi, S.A. *Legendre wavelets method for solving differential equations of Lane–Emden type* Appl. Math. Comput., 181 (2) (2006), 1417–1422.

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 1, 2026, pp 264–282
https://doi.org/10.22067/ijnao.2025.95643.1739
https://ijnao.um.ac.ir/

CC BY NC

Research Article

An efficient hybrid three-term conjugate gradient method with practical applications

Y.E. Hemici*, S. Khelladi and D. Benterki

Abstract

The conjugate gradient method is widely used for large-scale unconstrained optimization due to its efficiency in iteration number and computing time. In this paper, we present a new parameter β_k by hybridizing the Liu–Storey parameter β_k^{LS} and its modification β_k^{DLS}. We also use a hybrid three-term technique ensuring sufficient descent based on strong Wolfe conditions. We prove that the new direction possesses the descent property and that the corresponding algorithm is globally convergent. Numerical experiments on test problems and some applications, such as image restoration and sparse signal recovery, further show that our approach is efficient, consistently outperforming other approaches in terms of convergence speed and solution quality.

*Corresponding author

Received 26 September 2025; revised 24 November 2025; accepted 5 December 2025

Youcef Elhamam Hemici
Laboratory of Fundamental and Numerical Mathematics, Department of Mathematics, Faculty of Sciences, Setif1 University-Ferhat Abbas, 19000, Algeria. e-mail: youcefelhamam.hemici@univ-setif.dz

Samia Khelladi
Laboratory of Fundamental and Numerical Mathematics, Department of Mathematics, Faculty of Sciences, Setif1 University-Ferhat Abbas, 19000, Algeria. e-mail: samia.boukaroura@univ-setif.dz

Djamel Benterki
Laboratory of Fundamental and Numerical Mathematics, Department of Mathematics, Faculty of Sciences, Setif1 University-Ferhat Abbas, 19000, Algeria. e-mail: djbenterki@univ-setif.dz

How to cite this article
Hemici, Y.E., Khelladi, S. and Benterki, D., An efficient hybrid three-term conjugate gradient method with practical applications. *Iran. J. Numer. Anal. Optim.*, 2026; 16(1): 264-282. https://doi.org/10.22067/ijnao.2025.95643.1739

AMS subject classifications (2020): 90C26, 90C30, 65K05

Keywords: Unconstrained optimization, Conjugate gradient method, Descent direction, Image restoration, Signal recovery.

1 Introduction

Optimization techniques are widely used in science, engineering, economics, management, industry, and other areas. The most common optimization algorithms are the descent methods, namely gradient, Newton's, and conjugate gradient methods.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. Our goal is to solve the unconstrained optimization problem $\{\min f(x),\ x \in \mathbb{R}^n\}$. The principle of the conjugate gradient method is to construct a sequence

$$x_1 \in \mathbb{R}^n,\ x_{k+1} = x_k + \alpha_k d_k, \qquad \text{for } k \geq 1, \tag{1}$$

where x_1 is an initial point and x_k denotes the current iterate. The step size $\alpha_k > 0$ is computed using an exact or inexact line search, while the search direction d_k is defined as

$$d_1 = -g_1,\ d_{k+1} = -g_{k+1} + \beta_k d_k, \qquad \text{for } k \geq 1,$$

where $g_k = \nabla f(x_k)$ is the gradient of f at x_k and β_k is a scalar conjugacy coefficient.

The conjugate gradient method was first introduced by Hestenes and Stiefel (HS) in 1952 [19]. Later, different formulas of the parameter β_k were introduced, leading to various conjugate gradient directions d_k. Some of the most well-known choices of β_k include those proposed by Fletcher and Reeves (FR) [16], Polak–Ribière–Polyak (PRP) [29, 28], conjugate descent (CD) [6], Liu–Storey (LS) [22], and Dai–Yuan (DY) [15, 14]. For more details, see [18].

These methods are identical for a strongly convex quadratic function and the exact line search, but they behave differently for general nonlinear functions and the inexact line search. The FR, DY, and CD conjugate gradient methods are known for their strong convergence but poor practical behavior, as they are prone to taking short steps, which can lead to jamming. On the other hand, the LS, HS, and PRP conjugate gradient methods are well known for their good practical behavior due to their ability to switch to the steepest descent direction in order to avoid short steps that result in jamming [20], but they may not converge in general.

Based on the latter, many researchers have proposed modifications of the original schemes. Dai and Wen [8] proposed two modified conjugate gradient methods, denoted DPRP and DHS, given by

$$\beta_k^{DPRP} = \frac{\|g_{k+1}\|^2 - \frac{\|g_{k+1}\|}{\|g_k\|}\left|g_{k+1}^T g_k\right|}{\mu\left|g_{k+1}^T d_k\right| + \|g_{k+1}\|^2}, \qquad \mu > 1, \tag{2}$$

$$\beta_k^{DHS} = \frac{\|g_{k+1}\|^2 - \frac{\|g_{k+1}\|}{\|g_k\|}\left|g_{k+1}^T g_k\right|}{\mu\left|g_{k+1}^T d_k\right| + d_k^T y_k}, \qquad \mu > 1. \tag{3}$$

Additionally, inspired by (2) and (3), Zhang and Zheng proposed in [36] a modified LS method denoted as DLS, given by

$$\beta_k^{DLS} = \frac{\|g_{k+1}\|^2 - \frac{\|g_{k+1}\|}{\|g_k\|}\left|g_{k+1}^T g_k\right|}{\mu\left|g_{k+1}^T d_k\right| - d_k^T y_k}, \qquad \mu > 1. \tag{4}$$

In recent years, many researchers have introduced new hybrid conjugate gradient methods. The main concept behind hybridization is to combine classical conjugate gradient strategies in order to take advantage of their positive properties while reducing undesirable behaviors such as the jamming phenomenon. Depending on how the combination is formed, hybrid schemes can be either convex or nonconvex. Several examples of such approaches have been reported in the literature, including those proposed by Dalladji, Belloufi, and Sellami [9], Mtagulwa and Kaelo [26], Djordjević [11, 10, 12], Ben Hanachi, Sellami, and Belloufi [3], Yang, Luo, and Dai [35], Rivaie, Mustafa, and Abdelrhaman [30], Sellami and Chaib [31, 32], Sulaiman et al. [34], and Ouaoua, Khelladi, and Benterki [27].

In 2022, Jiang et al. [21], based on the hybrid conjugate gradient method and the convex combination technique, proposed a new family of hybrid three-term conjugate gradient methods:

$$d_{k+1} = \begin{cases} -g_{k+1} & \text{if} \quad k = 0, \\ -g_{k+1} + (1-\lambda_k)\beta_k^* d_k + \lambda_k \theta_k g_k & \text{if} \quad k > 0, \end{cases} \tag{5}$$

where $\beta_k^* = \max\{0, \min\{\beta_k, \beta_k^{DY}\}\}$, β_k is any conjugate parameter, $\lambda_k = \frac{\left|g_{k+1}^T d_k\right|}{\|g_{k+1}\|\|d_k\|}$ with $\lambda_k \in [0,1]$, and $\theta_k = -\eta\frac{g_{k+1}^T g_k}{\|g_k\|^2}$, $0 < \eta < 1$.

The objective of this work is to propose a new hybridization between LS and DLS using the formula (5) to calculate a new descent search direction.

The paper is structured as follows. In Section 2, we introduce the proposed hybrid conjugate gradient parameter together with the corresponding algorithm. Section 3 provides a thorough analysis of the descent property of the obtained direction and establishes the global convergence of

the algorithm under the strong Wolfe line search conditions. In Section 4, we assess the efficiency of the proposed method through numerical experiments on standard benchmark functions, as well as its application to image restoration and signal recovery. Finally, Section 5 presents the conclusion of this paper.

2 New hybrid conjugate gradient algorithm

Inspired by the works of Zhang and Zheng [36] and Jiang et al. [21], we propose a new hybrid parameter between LS and DLS (4), denoted by β_k^{MDLS}, defined by

$$\beta_k^{MDLS} = \begin{cases} \beta_k^{LS} = -\frac{g_{k+1}^T y_k}{g_k^T d_k} & \text{if} \quad \|g_{k+1}\|^2 \geq \left|g_{k+1}^T g_k\right|, \\ \beta_k^{DLS} = \frac{\|g_{k+1}\|^2 - \frac{\|g_{k+1}\|}{\|g_k\|}\left|g_{k+1}^T g_k\right|}{\mu\left|g_{k+1}^T d_k\right| - d_k^T y_k} & \text{otherwise}, \end{cases} \tag{6}$$

where $\mu > 1$.

The search direction is given as follows:

$$d_{k+1} = \begin{cases} -g_{k+1} & \text{if } k = 0 \text{or } \left|g_{k+1}^T g_k\right| \geq 0.5\|g_{k+1}\|^2, \\ -g_{k+1} + (1-\lambda_k)\beta_k^{MDLS} d_k + \lambda_k \theta_k g_k & \text{if } \ k > 0, \end{cases} \tag{7}$$

where $\lambda_k = \frac{\left|g_{k+1}^T d_k\right|}{\|g_{k+1}^T\|\|d_k\|}$ and $\theta_k = -\eta \frac{g_{k+1}^T g_k}{\|g_k\|^2}$. Note that $\lambda_k \in [0,1]$ and $0 < \eta < 1$.

We evaluate the behavior of our method, we use the strong Wolfe line search,that is, we find the stepsize α_k using the following conditions:

$$f(x_k + \alpha_k d_k) \leq f(x_k) + \delta \alpha_k g_k^T d_k, \tag{8}$$

$$\left|g_{k+1}^T d_k\right| \leq -\sigma g_k^T d_k, \tag{9}$$

where $0 < \delta < \sigma < 1$.

The corresponding algorithm is given bellow.

3 Convergence analysis

We introduce the following theorem to establish the sufficient descent condition.

Algorithm 1 MDLS

Step 0: Start by the initial point x_1 and a tolerance $\varepsilon > 0$.
Step 1: Set $k = 1$ and compute $d_1 = -g_1$.
Step 2: If $\|g_k\| \leq \varepsilon$, then **Stop**; else go to **Step 3**.
Step 3: Compute the stepsize $\alpha_k \in]0, 1]$, which satisfies strong Wolfe conditions (8), (9).
Step 4: Compute $x_{k+1} = x_k + \alpha_k d_k$.
Step 5: Compute $g_{k+1} = \nabla f(x_{k+1})$, $y_k = g_{k+1} - g_k$.
Step 6: Take $\beta_k = \beta_k^{MDLS}$ (6).
Step 7: Compute d_{k+1} using (7).
Step 8: Set $k = k + 1$ and return to **Step 2.**

Theorem 1. Let g_k and d_k denote the sequences produced by the $MDLS$ algorithm. Then, the corresponding search direction satisfies the sufficient descent condition given by

$$g_{k+1}^T d_{k+1} \leq -c\|g_{k+1}\|^2, \quad \text{for all } k \geq 0, \tag{10}$$

where c is a positive constant.

Proof. Multiplying (7) by g_{k+1}^T from the left gives

$$g_{k+1}^T d_{k+1}^{MDLS} = -\|g_{k+1}\|^2 + (1 - \lambda_k)\beta_k^{MDLS} g_{k+1}^T d_k + \lambda_k \theta_k g_{k+1}^T g_k. \tag{11}$$

We now analyze two cases.

Case 1. If $\|g_{k+1}\|^2 \geq \left|g_{k+1}^T g_k\right|$, then $\beta_k^{MDLS} = \beta_k^{LS}$ with

$$\beta_k^{LS} = \frac{g_{k+1}^T y_k}{-g_k^T d_k}, \qquad \theta_k = -\eta \frac{g_{k+1}^T g_k}{\|g_k\|^2}.$$

We have

$$\frac{(g_{k+1}^T g_k)^2}{\|g_k\|^2} = \|g_{k+1}\|^2 \cos^2 \omega. \tag{12}$$

Substituting these into (11) and using (12), we get

$$\begin{aligned} g_{k+1}^T d_{k+1}^{MDLS} &= -\|g_{k+1}\|^2 + (1 - \lambda_k)\frac{g_{k+1}^T y_k}{-g_k^T d_k} g_{k+1}^T d_k - \eta\lambda_k \frac{(g_{k+1}^T g_k)^2}{\|g_k\|^2} \\ &= -(1 + \eta\lambda_k \cos^2 \omega)\|g_{k+1}\|^2 + (1 - \lambda_k)\frac{g_{k+1}^T y_k}{-g_k^T d_k} g_{k+1}^T d_k, \end{aligned} \tag{13}$$

where ω is the angle between g_k and g_{k+1}.

Now, using the second condition of strong Wolfe (9), we get

$$\frac{g_{k+1}^T y_k}{-g_k^T d_k}\, g_{k+1}^T d_k \le \sigma g_{k+1}^T y_k.$$

Using this in (13) gives

$$\begin{aligned} g_{k+1}^T d_{k+1}^{MDLS} &\le -(1+\eta\lambda_k \cos^2\omega)\|g_{k+1}\|^2 + (1-\lambda_k)\sigma g_{k+1}^T y_k \\ &= -(1+\eta\lambda_k \cos^2\omega)\|g_{k+1}\|^2 + (1-\lambda_k)\sigma\|g_{k+1}\|^2 \\ &\quad -(1-\lambda_k)\sigma g_{k+1}^T g_k \\ &\le -(1+\eta\lambda_k \cos^2\omega)\|g_{k+1}\|^2 + 2(1-\lambda_k)\sigma\|g_{k+1}\|^2 \\ &= -(1-2\sigma+(2\sigma+\cos^2\omega)\lambda_k)\|g_{k+1}\|^2. \end{aligned}$$

Thus

$$g_{k+1}^T d_{k+1}^{MDLS} \le -c\|g_{k+1}\|^2,$$

where

$$c = 1 - 2\sigma + (2\sigma + \cos^2\omega)\lambda_k.$$

Since $0 < \sigma < \frac{1}{2}$ implies $1 - 2\sigma > 0$, and $\lambda_k \ge 0$, $\cos^2\omega \ge 0$, we conclude $c > 0$.

Case 2. If $\|g_{k+1}\|^2 < \left|g_{k+1}^T g_k\right|$, then $\beta_k^{MDLS} = \beta_k^{DLS}$. By Stanimirović [33], we know that

$$\beta_k^{DLS} \le -\frac{\|g_{k+1}\|^2}{g_k^T d_k}.$$

Proceeding as in Case 1, we obtain

$$\begin{aligned} g_{k+1}^T d_{k+1}^{MDLS} &\le -(1+\eta\lambda_k \cos^2\omega)\|g_{k+1}\|^2 + (1-\lambda_k)\sigma\|g_{k+1}\|^2 \\ &= -(1-\sigma+(\sigma+\cos^2\omega)\lambda_k)\|g_{k+1}\|^2. \end{aligned}$$

Hence

$$g_{k+1}^T d_{k+1}^{MDLS} \le -c\|g_{k+1}\|^2,$$

where

$$c = 1 - \sigma + (\sigma + \cos^2\omega)\lambda_k > 0.$$

In all cases, we have shown that

$$g_{k+1}^T d_{k+1}^{MDLS} \le -c\|g_{k+1}\|^2, \qquad c > 0,$$

which proves the sufficient descent property (10). □

To prove the global convergence of the algorithm, we make the following assumptions:

- *(i)* The level set $S = \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ is bounded; that is, there exists a constant $B > 0$ such that
$$\|x\| \leq B, \quad \text{for all } x \in S.$$
- *(ii)* In a neighborhood $\mathcal{N}$ of S, the gradient $\nabla f(x)$ is Lipschitz continuous; that is, there exists a constant $L > 0$ such that
$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \text{for all } x, y \in \mathcal{N}. \tag{14}$$

These assumptions imply the existence of a constant $\kappa \geq 0$, where

$$\|\nabla f(x)\| \leq \kappa, \qquad \text{for all } x \in S. \tag{15}$$

Lemma 1. [37] Assume that conditions (i) and (ii) are satisfied. Consider iterate (1), where d_k is the descent direction in (7), and α_k is chosen according to the strong Wolfe line search conditions (8) and (9). Then, the Zoutendijk condition

$$\sum_{k\geq 0} \frac{(g_k^T d_k)^2}{\|d_k\|^2} < \infty, \tag{16}$$

is satisfied.

Lemma 2. [12] Assume that assumptions (i) and (ii) hold and that $\alpha_k > 0$ is computed by the strong Wolfe line search (8) and (9). Then there exists $\alpha_* > 0$ such that

$$\alpha_k \geq \alpha_*, \qquad \text{for all } k \geq 1.$$

Theorem 2. Suppose that assumptions (i), (ii) and (10) hold. Then the $MDLS$ algorithm converges and satisfies

$$\lim_{k\to\infty} \inf \|g_k\| = 0. \tag{17}$$

Proof. Assume that $g_k \neq 0$, for all k. We need to prove (17).

Suppose that (17) does not hold. Then there exists a constant $r > 0$ such that

$$\|g_k\| \geq r, \qquad \text{for all } k. \tag{18}$$

Let D denote the diameter of the set S.

We have

$$|\beta_k^{DLS}| \leq -\frac{\|g_{k+1}\|^2}{g_k^T d_k} \leq \frac{\kappa^2}{cr^2},$$

$$|\beta_k^{LS}| = \frac{g_{k+1}^T y_k}{-g_k^T d_k} \leq \frac{\kappa LD}{cr^2}.$$

On the one hand, we have

$$\begin{aligned}\|d_{k+1}\| &= \|g_{k+1}\| + (1-\lambda_k)|\beta_k^{MDLS}|\|d_k\| + \lambda_k|\theta_k|\|g_k\| \\ &= \|g_{k+1}\| + (1-\lambda_k)|\beta_k^{MDLS}|\frac{\|s_k\|}{\alpha_k} + \lambda_k|\theta_k|\|g_k\|. \end{aligned} \tag{19}$$

From Lemma 2 and by using (14) and (15), we get two conditions:

1. If $\|g_{k+1}\|^2 \geq \left|g_{k+1}^T g_k\right|$, then (19) becomes

$$\|d_{k+1}\| \leq \kappa + \frac{\kappa LD}{\alpha^* cr^2} + \eta\kappa = C,$$

where C is a constant.

2. If $\|g_{k+1}\|^2 < \left|g_{k+1}^T g_k\right|$, but in this case we have $\left|g_{k+1}^T g_k\right| \geq \|g_{k+1}\|^2$, then (19) becomes

$$\|d_{k+1}\| \leq \kappa + \frac{\kappa^2}{\alpha^* cr^2} + 2\eta\kappa = C,$$

where C is a constant.

Now, we obtain

$$\sum_{k\geq 1} \frac{1}{\|d_{k+1}\|^2} = \infty. \tag{20}$$

Furthermore, combining inequalities (10) and (18) with the Zoutendijk condition (16), we obtain

$$c^2 t^4 \sum_{k\geq 0} \frac{1}{\|d_k\|^2} \leq \sum_{k\geq 0} \frac{c^2\|g_k\|^4}{\|d_k\|^2} \leq \sum_{k\geq 0} \frac{(g_k^T d_k)^2}{\|d_k\|^2} < \infty,$$

which contradicts condition (20). Therefore, assumption (18) must be not true.

Consequently, we conclude that

$$\lim_{k\to\infty} \inf \|g_k\| = 0.$$

□

4 Numerical experiments

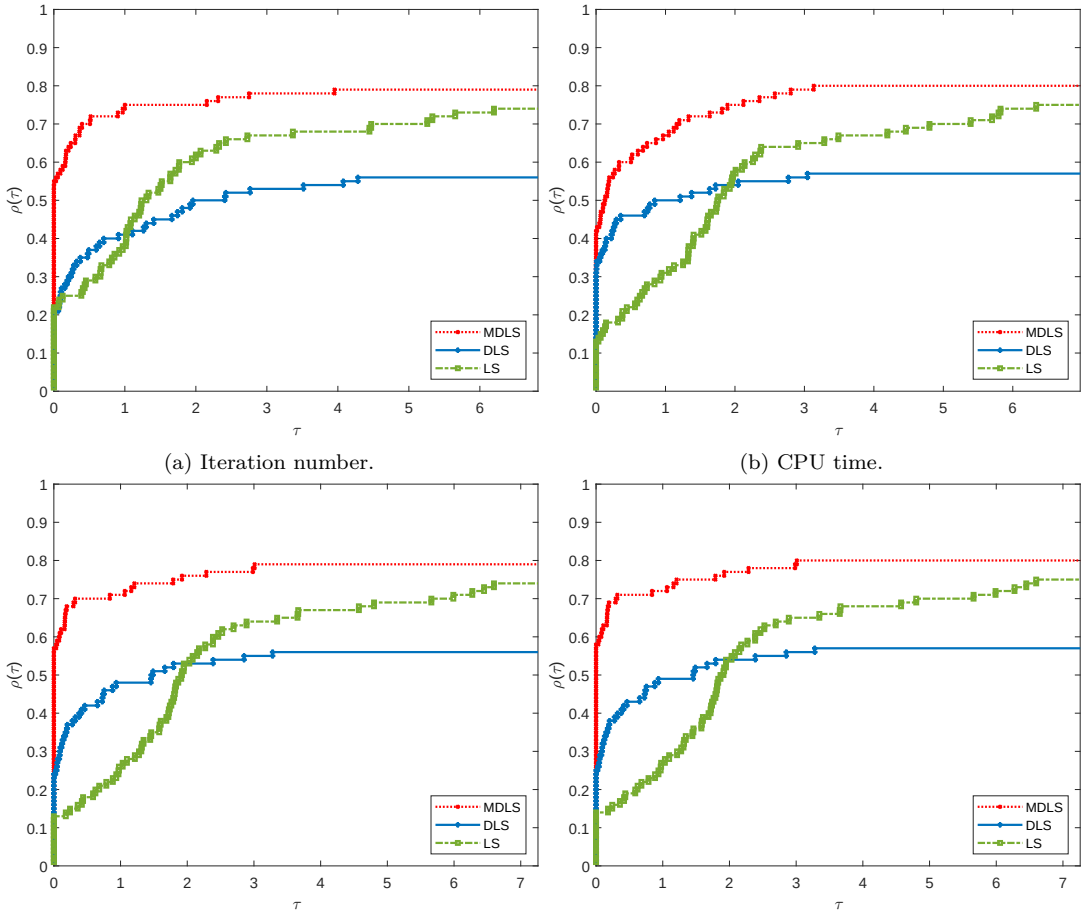
In this section, we present the performance of our new MDLS algorithm via some theoretical standard tests problems and practical applications.

4.1 Standard theoretical tests problems

We have selected various test functions given in Table 1, taken from the CUTE library of [1, 2, 17, 25] for different dimensions, from $n = 10$ to $n = 800000$. At the same time, we present a numerical comparison with other conjugate gradient algorithms (LS, DLS). We set the parameter $\eta = 0.4$, $\varepsilon = 10^{-6}$, and the stopping criterion is $\|g_k\| \leq \varepsilon$ or the iteration number > 4000. The stepsize α_k is calculated using the strong Wolfe line search with $\sigma = 0.01$, $\delta = 0.1$. In order to evaluate the data number of iterations, CPU time, gradient evaluation, and objective function evaluation, we use the performance profile of Dolan and Moré [13] presented in Figure 1.

Table 1: List of test functions and their dimensions.

Functions	**Dimensions**	**Functions**	**Dimensions**
Genrose	10000	Pen1	200, 1000
Osb2	10	Rosex	500, 1000
Eg2	100	Diagonal3	500, 2000
Freuroth	460	Singx	1000, 2000
Pen2	160	Diagonal1	800, 2000
Dixon3dq	150	Trid	500, 8000
Fletcbv3	100	Ie	500, 1500
Power1	150	Raydan1	500, 5000
Biggsb1	300	Lin	100, 1300
Dixmaani	360	Nonscomp	5000, 80000
Bv	2000, 20000	Dqrtic	5000, 150000
Liarwhd	6000, 30000	Dqdrtic	9000, 90000
Woods	150000, 200000	Dixmaana	6000, 90000
Quartc	80000, 50000	Dixmaanb	24000, 48000
Dixmaanc	2700, 27000	Dixmaand	12000, 90000
Dixmaane	2400, 48000	Dixmaanf	15000, 60000
Dixmaang	12000, 90000	Edensch	7000, 40000, 50000
Diagonal2	8000, 50000	Cosine	6000, 100000, 800000
Dixmaank	12000, 12000	Fletchcr	1000, 50000, 200000
Dixmaanl	2400, 24000	Exdenschnb	6000, 24000, 300000
Himmelbg	70000, 240000	Exdenschnf	90000, 280000, 600000
Dixmaanh	6000, 150000	Raydan2	2000, 20000, 500000
Dixmaanj	3000, 15000	Sine	100000, 250000, 500000



(a) Iteration number.

(b) CPU time.

(c) Objective function evaluation.

(d) Gradient evaluation.

Figure 1: Performance profile of MDLS, DLS and LS.

4.2 Practical applications

4.2.1 Image restoration

In this part, we evaluate the performance of the MDLS algorithm in comparison with the other conjugate gradient methods (LS and DLS) for image restoration problems affected by impulse noise. In this processing, we use the two-phase restoration scheme proposed by Chan, Ho, and Nikolova [7], which has proven effective in handling high-density of salt-and-pepper noise. This method was also used in [4, 18, 21, 23, 24], where it was shown that conjugate gradient-based

algorithms can successfully recover images even when the corruption ratio is extremely high, reaching up to 90%.

Let $X \in \mathbb{R}^{M\times N}$ represent the corrupted image, and let $\mathcal{A} = \{1,\dots,M\} \times \{1,\dots,N\}$ be the index set of all pixels. The restoration process is formulated as the following minimization problem:

$$\min_u F_\alpha(u) = \sum_{(i,j)\in\mathcal{B}} \left(\sum_{(m,n)\in\mathcal{V}_{i,j}\setminus\mathcal{B}} 2\varphi_\alpha(u_{i,j} - y_{m,n}) + \sum_{(m,n)\in\mathcal{V}_{i,j}\cap\mathcal{B}} \varphi_\alpha(u_{i,j} - u_{m,n}) \right),$$

where $\mathcal{B} \subset \mathcal{A}$ denotes the set of pixels identified as corrupted during the first phase, and the 4-neighborhood of a pixel located at position (i,j) is defined as

$$\mathcal{V}_{i,j} = \{(i,j-1),\ (i,j+1),\ (i-1,j),\ (i+1,j)\}.$$

The observed image is represented by the set $y = \{y_{i,j}\}$, where $y_{i,j}$ denotes the intensity value at pixel (i,j). The edge-preserving potential function is given by $\varphi_\alpha(t) = \sqrt{t^2+\alpha}$, where $\alpha = 100$ is used in our experiments.

To evaluate the effectiveness of the restoration process, we use the Peak Signal-to-Noise Ratio (PSNR), which is defined as

$$PSNR = 10\log_{10}\left[\frac{255^2}{\frac{1}{MN}\sum_{i,j}(x^r_{i,j} - x^*_{i,j})^2}\right],$$

where $x^r_{i,j}$ and $x^*_{i,j}$ represent the pixel values of the reconstructed and reference images, respectively.

The experiments are conducted on three gray-scale images of size 512×512 *Lady*, *Baboon*, and *Bridge*. The stopping criterion is defined as either the number of iterations exceeding 300 or the relative change in the objective function satisfying

$$\frac{|\Delta F_\alpha(u_k)|}{F_\alpha(u_k)} \le 10^{-4}.$$

We test the performance of the algorithms under various levels of salt-and-pepper noise: 30%, 50%, 70%, and 90%.

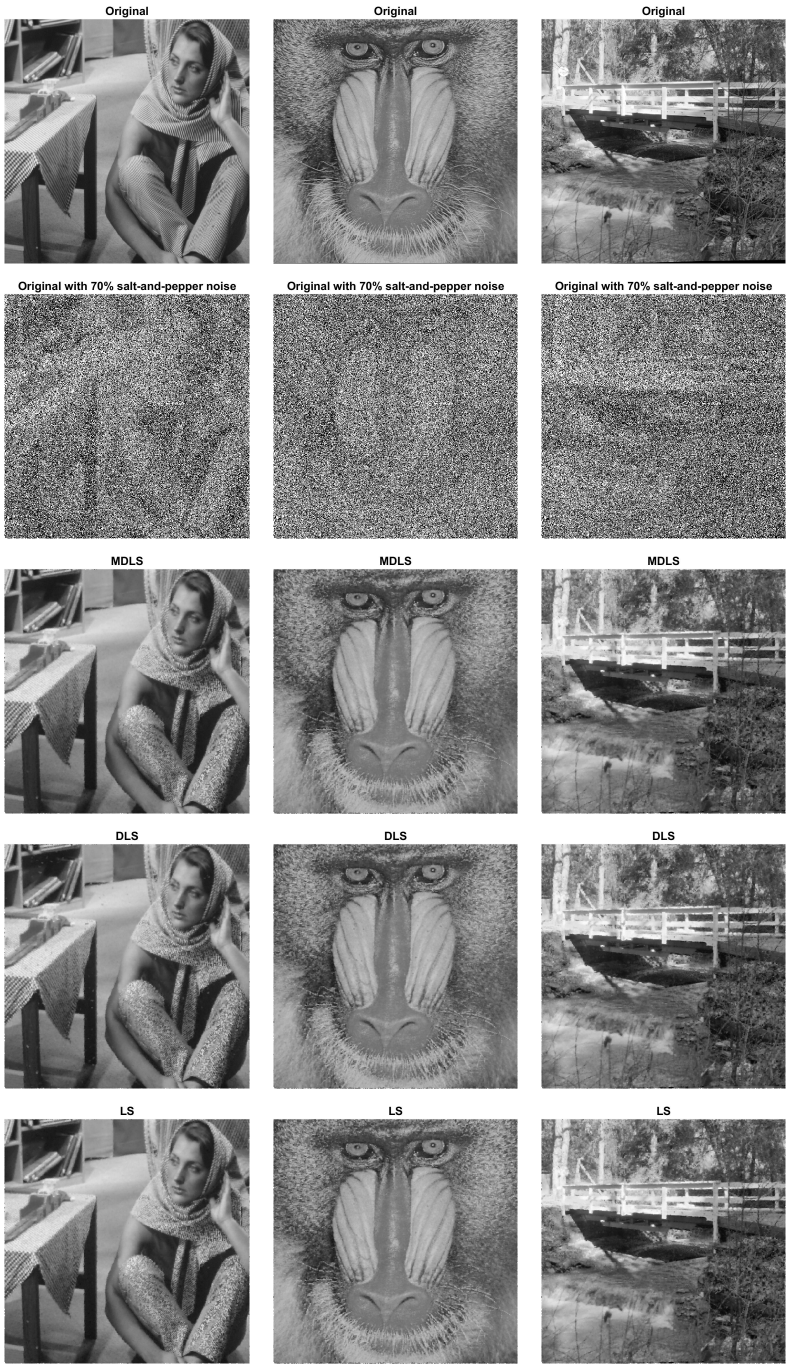


Figure 2: The noisy images with 70% salt-and-pepper noise and the restored images by MDLS, DLS, and LS.

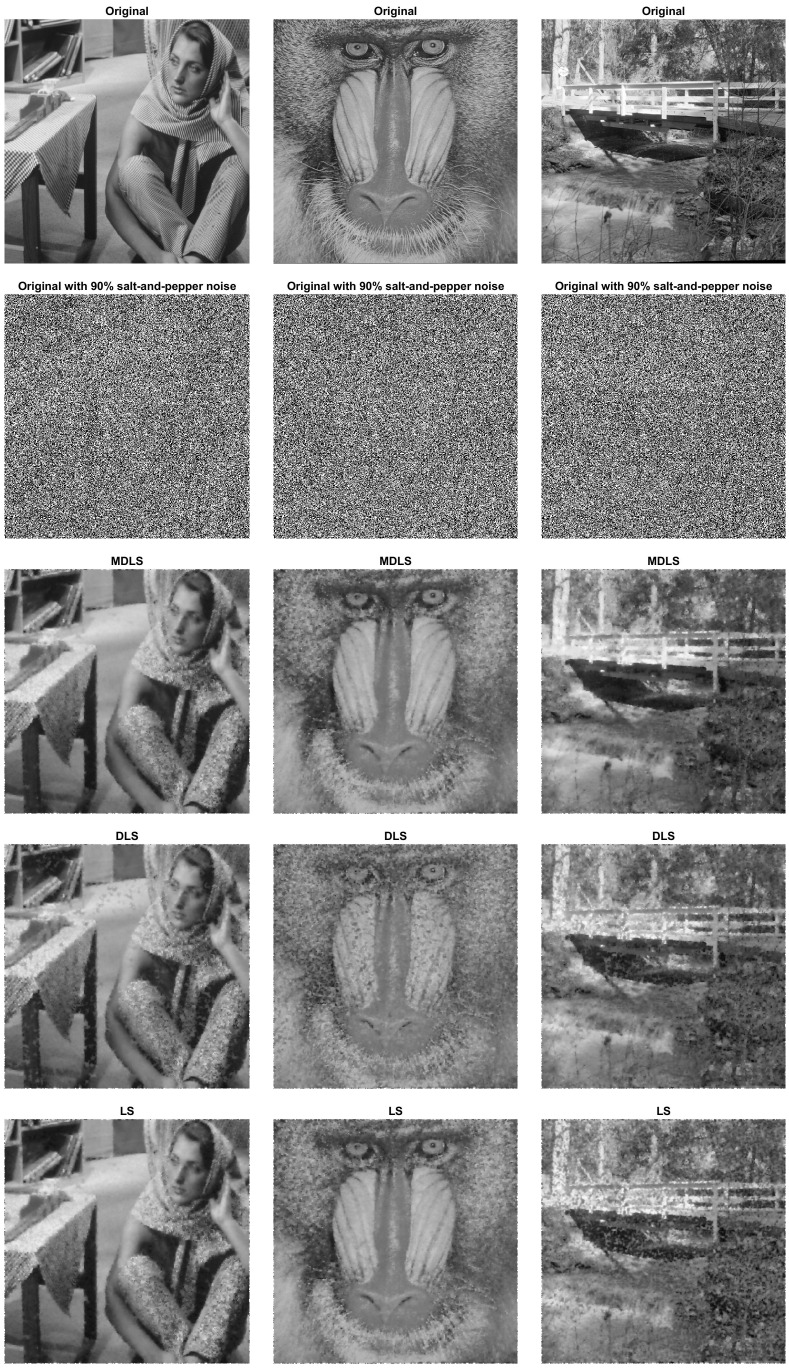


Figure 3: The noisy images with 90% salt-and-pepper noise and the restored images by MDLS, DLS, and LS.

Table 2: Performance results for different images and compression levels

Image/Noise	MDLS			DLS			LS		
	Iter	Time	PSNR	Iter	Time	PSNR	Iter	Time	PSNR
Barbara/0.30	16	8.73	28.39	55	13.71	28.40	15	14.40	28.27
Barbara/0.50	19	16.26	26.37	34	23.29	26.39	16	22.55	26.13
Barbara/0.70	16	21.43	24.36	56	37.22	24.34	24	43.75	24.37
Barbara/0.90	28	36.05	22.11	105	81.23	21.64	30	61.19	22.09
Baboon/0.30	18	7.52	27.91	39	11.95	27.89	15	9.11	27.84
Baboon/0.50	13	12.37	25.96	27	19.12	25.93	16	23.75	25.60
Baboon/0.70	18	28.81	23.80	75	38.74	23.75	16	25.47	23.75
Baboon/0.90	27	37.59	21.38	63	46.15	20.39	31	63.04	21.37
Bridge/0.30	15	8.54	28.54	31	11.35	28.39	14	8.23	28.46
Bridge/0.50	20	21.05	26.61	41	27.19	26.58	22	26.86	26.30
Bridge/0.70	24	23.80	24.45	88	55.35	24.36	26	46.82	24.42
Bridge/0.90	37	41.61	21.42	75	76.54	20.20	18	91.73	19.43

4.2.2 Sparse signal recovery

This part is dedicated to making a comparison of the three algorithms and analyzing the effectiveness of our proposed MDLS algorithm for solving sparse signal recovery problems. Sparse signal recovery is a critical problem in signal processing, where the objective is to recover a sparse signal from noisy or incomplete measurements. The optimization problem can be formulated as

$$\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{2}\|Ax-b\|^2 + \theta\|x\|_1,$$

where $A \in \mathbb{R}^{m\times n}$ (with $m < n$) is the measurement matrix, $b \in \mathbb{R}^m$ represents the observed data, $\|x\|_1 = \mathbf{e}_n^T x$ with $\mathbf{e}_n = (1, 1, \ldots, 1)^T \in \mathbb{R}^n$, and $\theta > 0$ is a regularization parameter controlling the trade-off between data fidelity and sparsity [5]. The effectiveness of signal recovery is typically evaluated using the mean squared error (MSE) given by

$$\text{MSE} = \frac{1}{n}\|x^* - x\|^2,$$

where x^* is the recovered signal and x is the original sparse signal. Note that a low MSE means that the estimated signal is very close to the true signal, indicating better performance. In our tests, N represents the sparsity level, which equals 32. The size of the problems is $n = 4096$ and $m = 1024$. The regularization parameter is $\theta = 2.9 \times 10^{-3}$.

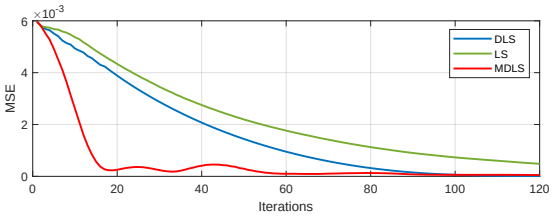


Figure 4: MSE in terms of iterations.

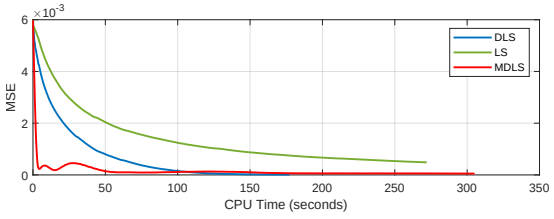


Figure 5: MSE in terms of CPU time.

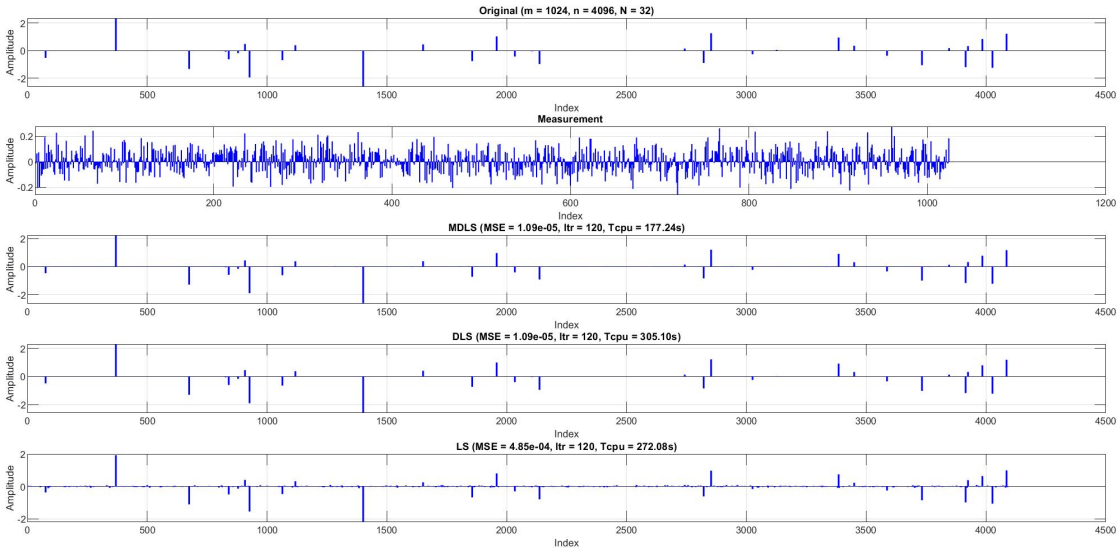


Figure 6: The original sparse signal compared to the restored signals.

Commentaries

The performance profile results, as depicted in Figures 1a, 1b, 1c, and 1d, show that our proposed conjugate gradient method MDLS, using β_k^{MDLS}, consistently outperforms methods based on β_k^{DLS} and β_k^{LS} for unconstrained optimization problems.

In the context of image restoration from Figures 2 and 3 and the results of Table 2, the MDLS method strikes an optimal balance between computational speed and image quality, particularly in scenarios with high levels of noise. It achieves superior PSNR values compared to other methods while requiring fewer iterations than the DLS and LS approaches.

For sparse signal restoration, Figures 4, 5, and 6 highlight the robustness and efficiency of the MDLS algorithm in addressing sparse signal recovery problems. It shows faster convergence

and reduced computational costs, offering a significant improvement over the other considered algorithms.

5 Conclusion

We proposed a new hybrid conjugate gradient method for solving unconstrained optimization problems. Our method, called MDLS, uses a new switching rule between the LS and DLS formulas. This rule automatically makes sure the search direction is a descent direction, without needing other extra safeguards that are often added to hybrid conjugate gradient methods.

We also used a three-term search direction to improve the algorithm's performance. This helps keep the value of β_k stable, avoids poor search directions, and makes the method more reliable. We proved that the MDLS method satisfies the sufficient descent condition and globally converges under the strong Wolfe line search.

Numerical tests showed that the MDLS method is more efficient compared to the standard LS and DLS methods. It consistently reduced the number of iterations and computing time while performing well in practical problems like image restoration and sparse signal recovery. This shows its potential for solving various optimization problems.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Andrei, N., *An unconstrained optimization test functions collection*, Adv. Model. Optim., 10 (2008), 147–161.

[2] Andrei, N., *Nonlinear conjugate gradient methods for unconstrained optimization*, Springer Optim. Appl., 158 (2020), 455–466.

[3] Ben Hanachi, S., Sellami, B. and Belloufi, M., *New iterative conjugate gradient method for nonlinear unconstrained optimization*, RAIRO Oper. Res., 56 (2022), 2315–2327.

[4] Cai, J., Chan, R. and Morini, B., *Minimization of an edge-preserving regularization functional by conjugate gradient type methods*, in *Image Processing Based on Partial Differential Equations*, Math. Vis., Springer, Berlin–Heidelberg, 2007.

[5] Candès, E., Romberg, J. and Tao, T., *Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Inf. Theory, 52(2) (2006), 489–509.

[6] Fletcher, R., *Practical methods of optimization: unconstrained optimization*, Vol. 1, Wiley, New York, 1987.

[7] Chan, R., Ho, C. and Nikolova, M., *Salt-and-pepper noise removal by median-type noise detectors and detail-preserving regularization*, IEEE Trans. Image Process., 14(10) (2005), 1479–1485.

[8] Dai, Z. and Wen, F., *Another improved Wei–Yao–Liu nonlinear conjugate gradient method with sufficient descent property*, Appl. Math. Comput., 218(13) (2012), 7421–7430.

[9] Delladji, S., Belloufi, M. and Sellami, B., *New hybrid conjugate gradient method as a convex combination of FR and BA methods*, J. Inf. Optim. Sci., 42(3) (2021), 591–602.

[10] Djordjević, S., *New hybrid conjugate gradient method as a convex combination of HS and FR methods*, J. Appl. Math. Comput., 2(9) (2018), 366–378.

[11] Djordjević, S., *New hybrid conjugate gradient method as a convex combination of LS and CD methods*, Filomat, 31(6) (2017), 1813–1825.

[12] Djordjević, S., *New hybrid conjugate gradient method as a convex combination of LS and FR methods*, Acta Math. Sci. Ser. B, 39(1) (2019), 214–228.

[13] Dolan, E. and Moré, J., *Benchmarking optimization software with performance profiles*, Math. Program., 91 (2002), 201–213.

[14] Dai, Y. and Yuan, Y., *An efficient hybrid conjugate gradient method for unconstrained optimization*, Ann. Oper. Res., 103 (2001), 33–47.

[15] Dai, Y. and Yuan, Y., *A nonlinear conjugate gradient method with a strong global convergence property*, SIAM J. Optim., 10(1) (1999), 177–182.

[16] Fletcher, R. and Reeves, C., *Function minimization by conjugate gradients*, Comput. J., 7(2) (1964), 149–154.

[17] Gould, N., Orban, D. and Toint, P., *CUTEr and SifDec: a constrained and unconstrained testing environment, revisited*, ACM Trans. Math. Softw., 29(4) (2003), 373–394.

[18] Hemici, Y. E., Khelladi, S. and Benterki, D., *New hybrid conjugate gradient method for nonlinear optimization with application to image restoration problems*, Kybernetika, 60(4) (2024), 535–552.

[19] Hestenes, M. and Stiefel, E., *Methods of conjugate gradients for solving linear systems*, J. Res. Natl. Bur. Stand., 49(6) (1952), 409–436.

[20] Hager, W. and Zhang, H., *A survey of nonlinear conjugate gradient methods*, Pac. J. Optim., 2 (2006), 35–58.

[21] Jiang, X., Liao, W., Yin, J. and Jian, J., *A new family of hybrid three-term conjugate gradient methods with applications in image restoration*, Numer. Algorithms, 91(1) (2022), 161–191.

[22] Liu, Y. and Storey, C., *Efficient generalized conjugate gradient algorithm. Part I: theory*, J. Optim. Theory Appl., 69(1) (1991), 129–137.

[23] Ma, G., Lin, H., Jin, W. and Han, D., *Two modified conjugate gradient methods for unconstrained optimization with applications in image restoration problems*, J. Appl. Math. Comput., 68 (2022), 4733–4758.

[24] Malik, M., Sulaiman, I., Abubakar, A., Ardaneswari, G. and Sukono, A., *A new family of hybrid three-term conjugate gradient method for unconstrained optimization with application to image restoration and portfolio selection*, AIMS Math., 8(1) (2023), 1–28.

[25] Moré, J., Garbow, B. and Hillstrom, K., *Testing unconstrained optimization software*, ACM Trans. Math. Softw., 7(1) (1981), 17–41.

[26] Mtagulwa, P. and Kaelo, P., *A convergent modified HS–DY hybrid conjugate gradient method for unconstrained optimization problems*, J. Inf. Optim. Sci., 40(1) (2019), 97–113.

[27] Ouaoua, M. L., Khelladi, S. and Benterki, D., *New parameter of conjugate gradient method for unconstrained nonlinear optimization*, Stat. Optim. Inf. Comput., 13(6) (2025), 2382–2390.

[28] Polyak, B., *The conjugate gradient method in extreme problems*, USSR Comput. Math. Math. Phys., 9 (1969), 94–112.

[29] Polak, E. and Ribière, G., *Note sur la convergence des méthodes de directions conjuguées*, Rev. Française Inform. Rech. Opér., 16 (1969), 35–43.

[30] Rivaie, M., Mustafa, M. and Abdelrhaman, A., *A new class of nonlinear conjugate gradient coefficients with exact and inexact line searches*, Appl. Math. Comput., 268 (2015), 1152–1163.

[31] Sellami, B. and Chaib, Y., *A new family of globally convergent conjugate gradient methods*, Ann. Oper. Res., 241 (2016), 497–513.

[32] Sellami, B. and Chaib, Y., *New conjugate gradient method for unconstrained optimization*, RAIRO Oper. Res., 50 (2016), 1013–1026.

[33] Stanimirović, P., Ivanov, B. and Mosić, D., *A survey of gradient methods for solving nonlinear optimization*, Electron. Res. Arch., 28(4) (2020), 1573–1624.

[34] Sulaiman, I. M., Kaelo, P., Khalid, R. and Nawawi, M. K. M., *A descent generalized RMIL spectral gradient algorithm for optimization problems*, Int. J. Appl. Math. Comput. Sci., 34(2) (2024), 225–233.

[35] Yang, X., Luo, Z. and Dai, X., *A global convergence of LS–CD hybrid conjugate gradient method*, Adv. Numer. Anal., 2013 (2013), Article ID 5.

[36] Zhang, Y. and Zheng, B., *Two new Dai–Liao-type conjugate gradient methods for unconstrained optimization problems*, J. Optim. Theory Appl., 175 (2017), 502–509.

[37] Zoutendijk, G., *Nonlinear programming, computational methods*, in *Integer and Nonlinear Programming*, North-Holland, Amsterdam, 1970, pp. 37–86.

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 1, 2026, pp 283–303
https://doi.org/10.22067/ijnao.2025.95542.1738
https://ijnao.um.ac.ir/

CC BY NC

Research Article

Dynamics and applications of quantum iterative methods

S.K. Nayak*, P.K. Parida and B. Mehta

Abstract

In this article, we introduce a new parametric family of the Quantum iterative method (PFQIM), constructed using the quantum derivative (q-derivative) to solve nonlinear equations. The stability and convergence analysis of the proposed family is analyzed through dynamical graphs. The dynamical and convergence planes of methods are examined using complex and real dynamics, respectively. In complex dynamics, we provide a scaling theorem for PFQIM to study affine conjugacy classes of iterative methods. Moreover, the influence of the q-derivative and the method parameter on stability and convergence is discussed in detail through real dynamics and real-world applications.

AMS subject classifications (2020): Primary 65H05; Secondary 26A33.

*Corresponding author

Received 20 September 2025; revised 21 October 2025; accepted 24 November 2025

Sapan Kumar Nayak
Department of Mathematics, Usha Martin University, Ranchi-835103, India. e-mail: sapannayak7@gmail.com

P. K. Parida
Department of Mathematics, Central University of Jharkhand, Ranchi-83522, India. e-mail: pkparida@cuj.ac.in

Babita Mehta
ICFAI University, Jharkhand, Ranchi-835222, India. e-mail: babita10625@gmail.com

How to cite this article
Nayak, S.K., Parida, P.K. and Mehta, B., Dynamics and applications of quantum iterative methods. *Iran. J. Numer. Anal. Optim.*, 2026; 16(1): 283-303. https://doi.org/10.22067/ijnao.2025.95542.1738

Keywords: Nonlinear equation, q-derivative, Basin of attraction, Scaling theorem, Dynamical plane, Convergence plane, $ACOC$.

1 Introduction

In many areas of science, mathematical models are very important. The billiard ball model of a gas, the Bohr model of an atom, the Lorenz model of the atmosphere, the Lotka-Volterra model of how predators and prey interact, the Schrödinger-type model [16], electrodynamics [1], and so on, are all examples of models based on mathematics. One of the most important tools of modern science is the mathematical model. When these models are applied to specific problems, they typically lead to nonlinear or differential equations, or even systems of equations. The difficulty is that finding exact solutions is extremely challenging if not impossible therefore numerical approximations using iterative methods become necessary. Let

$$S(z) = 0, \tag{1}$$

where $S : \mathbb{C} \to \mathbb{C}$ with $\mathbb{C}$ being the complex plane, and the most popular method is Newton's method, which produces a rational function with unknown dynamics. The dynamical properties of rational functions associated with iterative methods provide valuable information regarding their numerical stability and reliability. Many authors [4, 7, 8, 2, 15, 11, 20] have studied the complex dynamics of different types of methods and revealed their dynamical properties. In the last few years, complex dynamics have played an important role in mathematics, and the dynamical plane helps us to study complex variable rational functions. Later, in 2014, Magreñán [12] has provided a new tool, that is, a convergence plane, to easily understand the real dynamics of iterative methods. As a result, the real dynamics is different from complex dynamics, so many authors have explored the study of the real dynamics of several methods. In this literature, we use the quantum derivative (q-derivative) to study the dynamical behavior of a rational function obtained from the parametric family method (Chebyshev–Halley family method [8]) using complex dynamics as below:

$$\begin{cases} z_0 \text{given}, \\ z_{k+1} = z_k - \left(1 + \frac{0.5 L_S(z_k)}{(1-\delta * L_S(z_k))}\right) \frac{S(z_k)}{S'(z_k)}, \end{cases} \tag{2}$$

where $L_S(z_k) = \frac{S(z_k)S''(z_k)}{(S'(z_k))^2}$ and $\delta \in \mathbb{C}$ is the complex parameter. We obtain the super Halley method (SHM, $\delta = 1$), Halley method (HM, $\delta = 0.5$), Chebyshev method (CM, $\delta = 0$), and

Newton method (NM, $\delta = \infty$). Furthermore, using real dynamics for some real-world phenomena, interesting numerical results have been discussed.

The q-calculus is a branch of mathematics in which the derivative of a function can be taken without limit. It is a combination of mathematics and physics and was first developed in 1908 by Jackson [9]. Then the authors, Jing and Fan [10] introduced the q-Taylor's formula with its q-remainder term. Singh, Mishra, and Pathak [23] explored the q-analogue of the iterative schemes and provided some interesting comparison results with classical schemes. Recently, many authors [21, 18, 22] derived some higher-order multi-step q-iterative methods to solve nonlinear equations and determined the convergence of the methods.

Let us recall some basic definitions and results of the q-calculus for $q \in (0,1)$, which we have used in our study.

In q-calculus, for $q \in (0,1)$, the q-integer can be defined as below:

$$\begin{cases} [p]_q = 1 + q + q^2 + \cdots + q^p - 1 = \frac{1-q^p}{1-q} & \text{for } p = 1, 2, 3, \ldots, \\ [p]_q = 1 & \text{for } q = 1. \end{cases} \tag{3}$$

The q-factorial and q-binomial are defined as

$$\begin{cases} [p]_q! = [p]_q[p-1]_q \cdots [1]q, \\ [0]_q! = 1, \\ \binom{p}{m}_q = \frac{[p]_q!}{[m]_q![p-m]_q!}. \end{cases} \tag{4}$$

Definition 1. Suppose that $S : \mathbb{C} \to \mathbb{C}$ is a continuous function. Then the q-derivative is defined as follows:

$$\begin{cases} D_q S(z) = \frac{S(qz)-S(z)}{qz-z}, & q \neq 1, \\ \frac{d}{dz} S(z) = \lim\limits_{h \to 0} \frac{S(z+h)-S(z)}{h}, & q = 1. \end{cases} \tag{5}$$

Definition 2. The product and quotient of the q-derivative of continuous $S(z)$ and $H(z)$ are

$$D_q(S(z)H(z)) = H(z)D_qS(qz) + S(qz)D_qH(z),$$

$$D_q \frac{S(z)}{H(z)} = \frac{H(z)D_qS(z) - S(z)D_qH(z)}{H(qz)H(z)}, \qquad H(z)H(qz) \neq 0.$$

Definition 3. The q-Taylor's series of a function $S(z)$ about the point z^* defined by Jackson

$$S(z) = \sum_{n=1}^{\infty} \frac{D_q^{(n)}(z - z^*)^n}{[n]!} \tag{6}$$

is
$(z - z^*)^0 = 1,\ (z - z^*) = \prod_{j=0}^{n-1}(z - qz^*)^j, \qquad$ for $n \in \mathbb{N}$,
and $D_q, D_q^{(2)}, D_q^{(3)}, \ldots$ are the q-derivatives.

Moreover, the rest of the manuscript is designed as follows: In section 2, we have shown the basic definition, order of convergence, and scaling theorem for the proposed method. Section 3 is devoted to the complex dynamics of the Quantum iterative method, with a brief discussion of its fixed and critical points. The dynamical plane of the method using a cubic polynomial has also been discussed in this section. In Section 4, the real dynamics of the proposed method using some real-world problems are discussed. Finally, section 5 ends with the conclusion.

2 Parametric family of Quantum iterative method (PFQIM) and its convergence analysis

To solve equation (1) iteratively, we introduce a parametric family of quantum iterative methods (PFQIM) defined as follows:

$$\begin{cases} z_0 \text{ given}, \\ z_{k+1} = Q_{S,\delta,q}(z_k) = z_k - \left(1 + \frac{0.5L_S(z_k)}{(1-\delta * L_S(z_k))}\right)\frac{S(z_k)}{D_q S(z_k)}, \end{cases} \tag{7}$$

where $L_S(z_k) = \frac{S(z_k)D_q^{(2)}S(z_k)}{(D_q S(z_k))^2}$ and $\delta \in \mathbb{C}$ is the complex parameter.

Theorem 1. Let $S(z)$ be an analytic function, and let $T(z) = \alpha z + \beta$, with $\alpha \neq 0$, be an affine map. Suppose $f(z) = \rho(S \circ T)(z)$, where ρ is a nonnull constant. Then $(T \circ Q_{S,\delta,q} \circ T^{-1})(z) = Q_{f,\delta,q}(z)$, that is, $Q_{S,\delta,q}$ and $Q_{f,\delta,q}$ are affine conjugate through T.

Proof. First, we have $T^{-1}(z) = \frac{1}{\alpha}(z - \beta)$. Moreover,

$$\begin{aligned} Q_{S,\delta,q} \circ T^{-1}(z) &= Q_{S,\delta,q}(T^{-1}(z)) \\ &= T^{-1}(z) - \left(1 + \frac{0.5L_S(T^{-1}(z))}{(1 - \delta L_S(T^{-1}(z)))}\right)\frac{S(T^{-1}(z))}{D_q S(T^{-1}(z))}. \end{aligned}$$

On the other side, as $f \circ T^{-1}(z) = S(z)$,

$$D_q(f \circ T^{-1})(z) = \frac{1}{\alpha} D_q f(T^{-1}(z))$$

and

$$D_q^{(2)}(f \circ T^{-1})(z) = \frac{1}{\alpha^2} D_q^{(2)} f(T^{-1}(z)).$$

From the above, $D_q f(T^{-1}(z)) = \alpha D_q S(z)$ and $D_q^{(2)} f(T^{-1}(z)) = \alpha^2 D_q^{(2)} S(z)$. So we have $L_f(T^{-1}(z)) = L_S(z)$ and $\frac{S(T^{-1}(z))}{D_q S(T^{-1}(z))} = \frac{1}{\alpha} \frac{f(z)}{D_q f(z)}$. Now, by substituting these expressions and using the definition of $Q_{S,\delta,q} \circ T^{-1}(z)$, we have

$$\begin{aligned} T \circ Q_{S,\delta,q} \circ T^{-1}(z) &= T(Q_{S,\delta,q}(T^{-1}(z))) \\ &= \alpha . Q_{S,\delta,q}(T^{-1}(z)) + \beta \\ &= \alpha . \left[T^{-1}(z) - \left(1 + \frac{0.5 L_S(T^{-1}(z))}{(1 - \delta L_S(T^{-1}(z)))} \right) \frac{S(T^{-1}(z))}{D_q S(T^{-1}(z))} \right] \\ &\quad + \beta \\ &= z - \left(1 + \frac{0.5 L_f(z)}{(1 - \delta L_f(z))} \right) \frac{f(z)}{D_q f(z)} = Q_{f,\delta,q}. \end{aligned}$$

□

The scaling theorem enables the study of iterative methods through simpler function families after an appropriate coordinate transformation.

Now, we have presented the order of convergence of the proposed method in the next theorem.

Theorem 2. Let $S : D \subseteq \mathbb{C} \to \mathbb{C}$ be a q-differentiable function in the open disk $D \subseteq \mathbb{C}$. If $S(z) = 0$ has a simple root $z^* \in D$ and z_0 is the initial guess sufficiently close to z^*, then the above method defined in (7) has cubic order convergence denoted as $[3; q]$, where q indicates the q-derivative, and the error equation of the proposed method is given below:

$$e_{k+1} = (2P_2^2 - P_3 - 2\delta P_2^2) e_k^3 + O(e_k^4). \tag{8}$$

Proof. The q-Taylor series of $S(z_k)$ and its first order q-derivative are

$$S(z_k) = D_q S(z^*)(e_k + P_2 e_k^2 + P_3 e_k^3) + O(e_k^4), \tag{9}$$

$$D_q(z_k) = D_q S(z^*)(1 + 2P_2 e_k + 3P_3 e_k^2) + O(e_k^3), \tag{10}$$

where $e_k = z_k - z^*$ and $P_n = \frac{D_q^{(n)} S(z^*)}{[n]! D_q S(z^*)}$, $n = 2, 3, \ldots$. The second order q-derivative of q-Taylor's series can be given as

$$D_q^{(2)} S(z_k) = D_q S(z^*)(2P_2 + 6P_3 e_k) + O(e_k^2). \tag{11}$$

From (9) and (10), we have

$$\frac{S(z_k)}{D_q S(z_k)} = e_k - P_2 e_k^2 - (2P_3 - 2P_2^2)e_k^3 + O(e_k^4)$$

and

$$L_S(z_k) = \frac{S(z_k) D_q^{(2)} S(z_k)}{(D_q S(z_k))^2} = 2P_2 e_k + (6P_3 - 6P_2^2)e_k^2 + (16P_2^3 - 28P_2 P_3)e_k^3 + O(e_k^4).$$

Next, using (12) and (2) implies

$$\left(1 + \frac{0.5 L_S(T^{-1}(z))}{(1 - \delta L_S(T^{-1}(z)))}\right) \frac{S(z_k)}{D_q S(z_k)} = e_k + (2P_2^2 - P_3 - 2\delta P_2^2)e_k^3 + O(e_k^4). \tag{12}$$

Finally, substituting (12) into (7), and simplifying a little, the desired error equation will be

$$e_{k+1} = (2P_2^2 - P_3 - 2\delta P_2^2)e_k^3 + O(e_k^4).$$

□

Lemma 1. With the local error equation is given in (8) of Theorem 2, we define the asymptotic error constant as

$$K(\delta) = 2P_2^2 - P_3 - 2\delta P_2^2. \tag{13}$$

The term $K(\delta)$ linearly dependent on the parameter δ, and therefore the parameter can be chosen to minimize or even annihilate this coefficient, so we have the following special case.

Optimal choice of δ.

Setting $K(\delta) = 0$ gives

$$\delta^\star = \frac{2P_2^2 - P_3}{2P_2^2} = 1 - \frac{P_3}{2P_2^2}, \qquad (P_2 \neq 0).$$

If $\delta = \delta^\star$, then the cubic term vanishes and the error equation becomes

$$e_{k+1} = O(e_k^4),$$

indicating that the rate of convergence increases from cubic to at least quartic. Hence, $\delta^\star$ gives the optimal value minimizing the asymptotic error constant $|K(\delta)|$.

3 Complex dynamics PFQIM

It is required to recall some basic dynamical ideas in order to understand the dynamical behavior of an iterative method when applied to a polynomial $S(z)$. Let $R : \mathbb{C}^* \to \mathbb{C}^*$ be the rational function defined on the Riemann sphere. Then the orbit of a point $z_0 \in \mathbb{C}^*$ is the set of composition functions by $\{z_0, R(z_0), R^2(z_0), \ldots, R^n(z_0), \ldots\}$. A point $z_0 \in \mathbb{C}^*$ is a fixed point of R if $R(z_0) = z_0$, and the behavior of fixed points is attracting, repelling, or neutral if $|D_q R(z_0)| \leq d < 1$, $|D_q R(z_0)| > 1$ or $|D_q R(z_0)| = 1$, respectively. Moreover, if $|D_q R(z_0)| = 0$, then the fixed points are called as super attracting.

The fixed points that do not belong to the roots z^* of $S(z)$ are called strange fixed points. If z^* is an attracting fixed point, then the basin of attraction of the fixed point is the set of preimages $\mathcal{A}(z^*) = \{z_0 \in \mathbb{C}^* : R^n(z_0) \to z^*, n \to \infty\}$. The Fatou set, $\mathcal{F}(R)$, is the set of points whose orbits tend to an attractive fixed point z^*. The complement of $\mathcal{F}(R)$ is the Julia set $\mathcal{J}(R)$ that consists of the repelling fixed points.

We now apply PFQIM (7) to a quadratic polynomial $S(z) = z^2 - a$, where $a \in \mathbb{C}^*$ and the Möbius transformation $\mathcal{M}(z) = \frac{z-a}{z+a}$ with the inverse $[\mathcal{M}(z)]^{-1} = \frac{-a(z+1)}{z-1}$ to produce the rational map relating to the quantum iterative method (7) as follows:

$$Q_{\mathcal{M}}(z,\delta,q) = \frac{z(-1+q^2-5z+2qz+3q^2z+z^2+4qz^2+3q^2z^2+z^3+2qz^3+q^2z^3+4z\delta-4qz\delta-4z^2\delta-4qz^2\delta)}{1+2q+q^2+z+4qz+3q^2z-5z^2+2qz^2+3q^2z^2-z^3+q^2z^3-4z\delta-4qz\delta+4z^2\delta-4qz^2\delta}.$$

Also, the operator obtained from the quantum Newton method when $\delta \to \infty$ is

$$Q_{\mathcal{NM}}(z,\delta,q) = \frac{z(q+z+qz-1)}{q-z+qz+1}.$$

3.1 Study of fixed points and their stability

The fixed and strange fixed points of the rational operator $Q_{\mathcal{M}}(z,\delta,q)$ are obtained by solving the equation $Q_{\mathcal{M}}(z,\delta,q) = z$. Note that $z = 0$, $z = 1$ and $z = \infty$ are fixed points of $Q_{\mathcal{M}}(z,\delta,q)$ out of which the fixed points $z = 0$ and $z = \infty$ are super-attracting. The strange fixed points are
$z = \frac{2\delta-\sqrt{4\delta^2-8\delta-4\delta q+2q+3}-q-2}{q+1}$ and $z = \frac{2\delta+\sqrt{4\delta^2-8\delta-4\delta q+2q+3}-q-2}{q+1}$.

Next, we observe that the parameter δ affects not only the number of fixed points but also their stability. The q-derivative rational operator expression, which is required for studying the stability of fixed points and determining critical points, is

$$
\begin{aligned}
D_qQ_{\mathcal{M}}(z,\delta,q) = &-\frac{1}{(q-1)}\Big[q^2z^3+3q^2z^2+3q^2z+q^2+2qz^3-4\delta qz^2+4qz^2\\
&-4\delta qz+2qz+z^3-4\delta z^2+z^2+4\delta z5z-1\Big]\frac{1}{z(q-1)}\\
&\Big[(q^5z^3+3q^4z^2-q^3z^3-4\delta q^3z^2+2q^3z^2+3q^3z-q^2z^3\\
&+4\delta q^2z^2)+(-8q^2z^2-4\delta q^2z+q^2z+4\delta qz^2-2qz^2-3qz\\
&+z^3-4\delta z^2+5z^2+4\delta z-z)-(q^2z^3+3q^2z^2+3q^2z+q^2\\
&-4\delta qz^2+2qz^2-4\delta qz+4qz+2q-z^3+4\delta z^2-5z^2)\\
&-\frac{(4\delta z+z+1)}{(q-1)z}\Big(z(q^2z^3+3q^2z^2+3q^2z+q^2+2qz^3-4\delta qz^2\\
&+4qz^2-4\delta qz+2qz+z^3-4\delta z^2+z^2+4\delta z-5z-1)\\
&-qz(q^5z^3+2q^4z^3+3q^4z^2+q^3z^3-4\delta q^3z^2+4q^3z^2+3q^3z\\
&-4\delta q^2z^2+q^2z^2-4\delta q^2z+2q^2z+q^2+4\delta qz-5qz-1)\Big)\Big].
\end{aligned}
$$

Table 1 presents the critical points of the operators for different values of δ and q.

Table 1: Critical points of operators $Q_{\mathcal{M}}(z,\delta,q)$ and $Q_{\mathcal{NM}}(z,\delta,q)$(when $\delta=\infty$) for δ and $q = 0.1,\ 0.5$, and 0.999

q, δ	Critical points
$q=0.1$, $\delta=0$	-53.691, -3.721, -2.687, -0.186, 1.834, 5.451
$q=0.1$, $\delta=0.5$	27.998, -10, -1, -0.357, 2.864, 3.490
$q=0.1$, $\delta=1$	-9.229-17.935i, -9.229+17.935i, -0.226-0.440i, -0.226+0.440i, 2.955-1.123i, 2.955+1.123i
$q=0.1$, $\delta=\infty$	0, 0.7903, 12.6542
$q=0.5$, $\delta=0$	-13.806, -1.391-0.252i, -1.391+0.252i, -0.144, 0.647, 3.086
$q=0.5$, $\delta=0.5$	-8.772, -2, -1, -0.227, 1, 2
$q=0.5$, $\delta=1$	-4.665-1.082i, -4.665+1.082i, -0.406-0.0943i, -0.406+0.094i, 1.372-0.340i, 1.372+ 0.340i
$q=0.5$, $\delta=\infty$	0, 0.2280, 8.7720
$q=0.999$, $\delta=0$	-15.071, -1.053-0.031i, -1.053+0.031i, -0.073, 0.129, 8.566
$q=0.999$, $\delta=0.5$	-10.396, -1.111, -1, -0.106, 0.182, 6.098
$q=0.999$, $\delta=1$	-3.559-4.271i, -3.559+4.271i, -0.127-0.153i, -0.127+0.153i, 0.385, 2.878
$q=0.999$, $\delta=\infty$	0, 2.5025e-04, 4.0000e+03

The dynamical graphs of the proposed method in this paper are plotted with a mesh of 1500×1500 real points using a Lenovo Ideapad flex 5, 1.19 GHz, Intel(R) Core(TM) i5-1035G1 CPU. The stopping criteria of the convergence plane have a tolerance of 10^{-3} and a maximum of

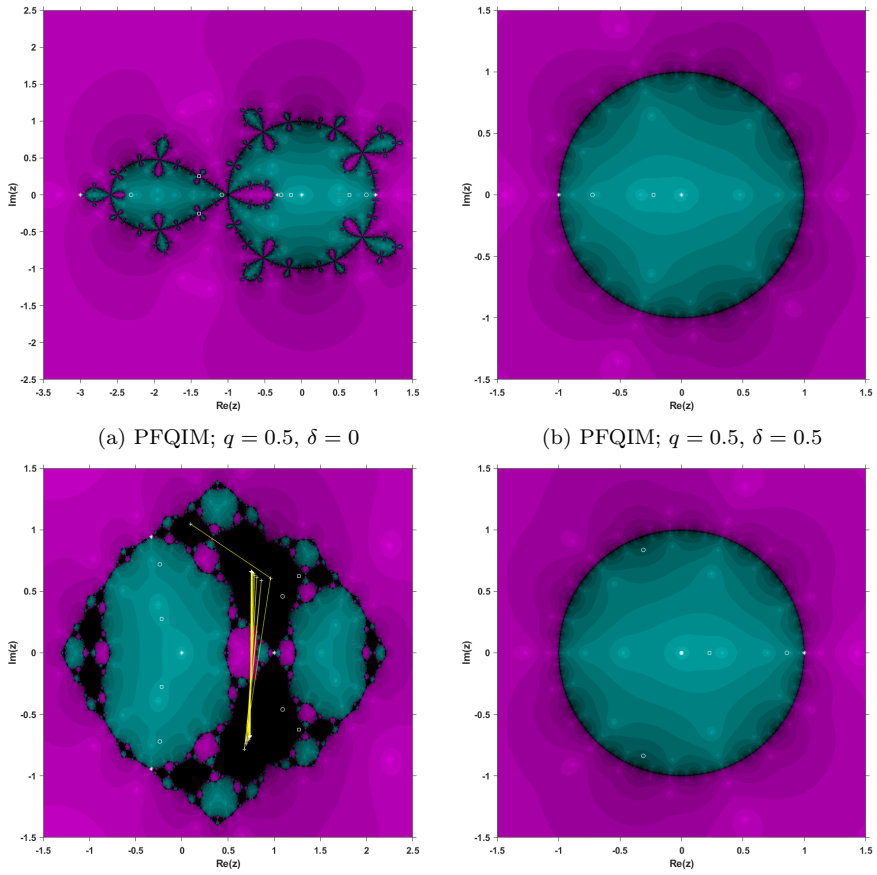
20 iterations. We have represented the convergence basins of each iterative procedure using the MATLAB program described in [5]. If the iterative method converges to the fixed point zero, then it is painted in cyan, and for ∞, it is painted in magenta, and if no root is found with a maximum of 20 iterations, the dynamical plane is painted in black. The attracting fixed points are marked in the figures by white stars, the parabolic fixed points by white circles, and the critical points by white squares. When the number of iterations decreases, the color will be more intense. If the method reaches the maximum number of iterations, then the point will become black.

We have plotted the dynamical planes of PFQIM for different values of δ with different values of q, such as $q = 0.5$ and $q = 0.999$. It can be observed that when the value of q is close to 1, the parametric plane is stable, excluding the boundary of the Fatou set. The star marks and square marks are denoted as the fixed and critical points, while circle points are taken as the parabolic points of the corresponding operator $Q_{\mathcal{M}}(z, \delta, q)$.

In particular, Figure 1 shows that in case (a), the method is stable with more fixed points, critical points, and iterations; in case (b), the number of fixed and critical points are reduced, and the Julia set is connected in the circumference; in case (c), the method is stable with more Julia sets; $z = 0$ is the repelling fixed point, and an orbit with 20 iterations is shown; and in case (d), the method is stable, but there are no Julia sets. In Figure 2, the dynamical planes are brighter than the previous one, meaning the method requires fewer iterations to converge on a root. In Figure 3, the stable behavior of a fixed point is seen at $q = 0.999$ and $\delta = 3.9 + 0.1i$, surrounding a complicated Julia set.

3.2 Dynamics plane on cubic polynomials

In this part, we have represented some dynamical planes of cubic polynomials when applied to PFQIM. Figure 4 illustrates the behavior of the family (7) on the cubic polynomial $z^3 - 1$. For different values of δ, we have called iterative methods like the Quantum Newton method (QNM, for $\delta = \infty$), Quantum Chebyshev method (QCM, for $\delta = 0$), Quantum Halley method (QHM, for $\delta = 0.5$), and the Quantum super-Halley method (QSCM, for $\delta = 1$). In Figure 4: (a), (b), (c), and (d), the dynamical planes of PFQIM are shown for $q = 0.5$, where the iterative family of method's behavior is unstable; in Figure 4: (e), (f), (g), and (h), the basin of attraction towards the divergent region is reduced, and the more stable zone is obtained; the intensity of the color is more intense, i.e., the convergence is faster. Moreover, the percentage of convergence of each dynamic plane is given in the figures, which gives an idea of the stability of the methods.



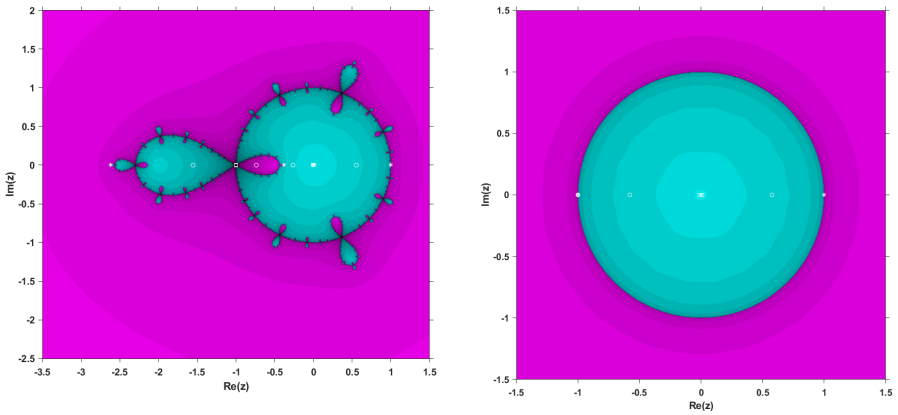
(a) PFQIM; $q = 0.5$, $\delta = 0$

(b) PFQIM; $q = 0.5$, $\delta = 0.5$

(c) PFQIM; $q = 0.5$, $\delta = 1$

(d) QNM; $q = 0.5$, $\delta = \infty$

Figure 1: Dynamical planes of $Q_{\mathcal{M}}(z, \delta, q)$ and $Q_{\mathcal{NM}}(z, \delta, q)$



(a) PFQIM; $q = 0.999$, $\delta = 0$

(b) PFQIM; $q = 0.999$, $\delta = 0.5$

(c) PFQIM; $q = 0.999$, $\delta = 1$

(d) QNM; $q = 0.999$, $\delta = \infty$

Figure 2: Dynamical planes of $Q_{\mathcal{M}}(z, \delta, q)$ and $Q_{\mathcal{NM}}(z, \delta, q)$

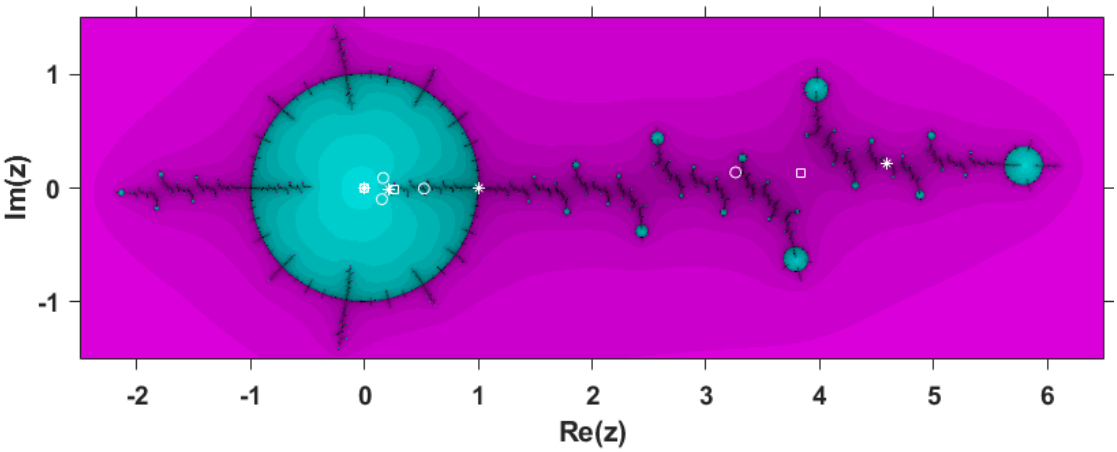
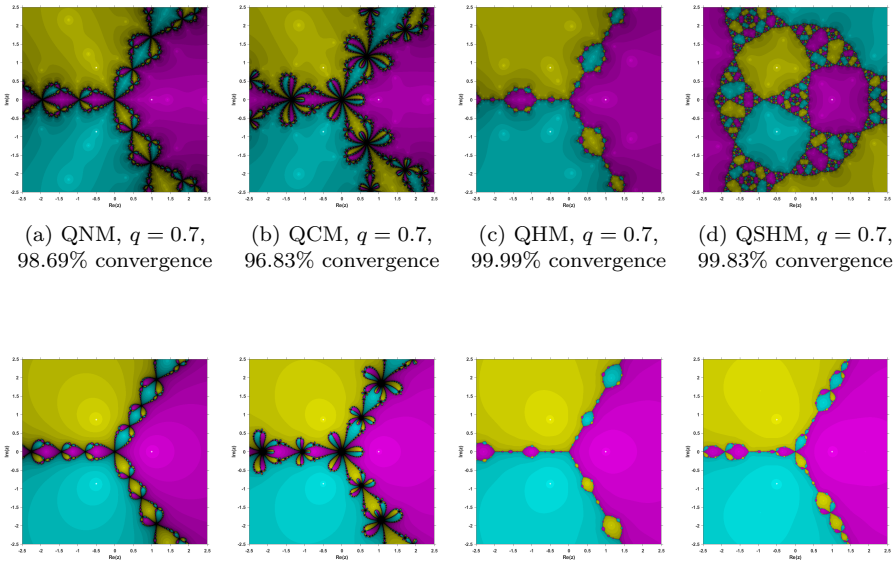


Figure 3: Dynamical plane of operator at $q = 0.999$ and $\delta = 3.9 + 0.1i$



(a) QNM, $q = 0.7$, 98.69% convergence (b) QCM, $q = 0.7$, 96.83% convergence (c) QHM, $q = 0.7$, 99.99% convergence (d) QSHM, $q = 0.7$, 99.83% convergence

(e) QNM, $q = 0.999$, 99.53% convergence (f) QCM, $q = 0.999$, 98.55% convergence (g) QHM, $q = 0.999$, 99.99% convergence (h) QSHM, $q = 0.999$, 100% convergence

Figure 4: Dynamic planes of cubic polynomial $z^3 - 1$ on PFQIM

4 Real dynamics of PFQIM

In this part, we have visualized the real dynamics (for more details readers can see the articles, [14, 13]) of the proposed method using some real-world applications. Here, the value of the q-derivative is taken as $q \in (0, 2]$. Additionally, the approximate computational order of convergence ($ACOC$) [6], is used to confirm that the theoretical order of convergence is also obtained in practice for single variable equations, can be calculated by the formula given below:

$$ACOC = \frac{\ln |(x_{k+1} - x_k|/|x_k - x_{k-1}|)}{\ln(|x_k - x_{k-1}|/|x_{k-1} - x_{k-2}|)}, \qquad k = 2, 3, 4, \ldots.$$

In the numerical table, we have shown the root of the equation, error, functional value at the nth iteration, ACOC, and CPU time in seconds. The convergence planes are potted using Matlab R2022b by modifying the program made in [12]. The term stable points refers to the collection of all starting points for which the iterative process generates a convergence region. Initial guesses and the parameter q both play a major role in determining whether iterative methods are stable. We illustrate the stability and dynamical properties of the method using convergence planes.

4.1 Numerical results

Example 4.1 (Kepler's laws of planetary motion)**.** In astronomy, Kepler's laws of planetary motion [17] are used to locate the position of a planet in the elliptical path that moving around the Sun. Kepler's equation applies to the motion of any object in an elliptical orbit. For example, Earth or Mars orbits around the Sun, or any moon orbiting a planet, or a rocket or satellite in orbit about any planet or moon. So our first test example is Kepler's equation, which can be written as $S_1(x) = x - e\sin(x) - M$, where $M$is the mean anomaly and e is the eccentricity of an object in an elliptical orbit. Suppose $M = \frac{\pi}{6}$ and $e = 0.6$; then the equation reduces to $S_1(x) = x - 0.6\ \sin(x) - \frac{\pi}{6}$ with root $x = 1.04149$. The numerical tables contain q as the order of derivative, and x^* is the approximate root. Tables 2, 3, 5, and 6 show that the speed of convergence increases as the value of q approaches one. Furthermore, in Tables 4 and 7, the numerical results of corresponding classical iterative are presented (for initial guesses $x_0 = 2$ and $x_0 = 2.5$), where we can observed that PFQIM for different values of $q = 0.999$, $q = 0.999999$, and $q = 0.9999999$ performf quite similar to classical methods, although in some cases it show slightly variations in the $ACOC$.

Table 2: Results of QSHM, QHM for $S_1(x)$ with initial guess $x_0 = 2$

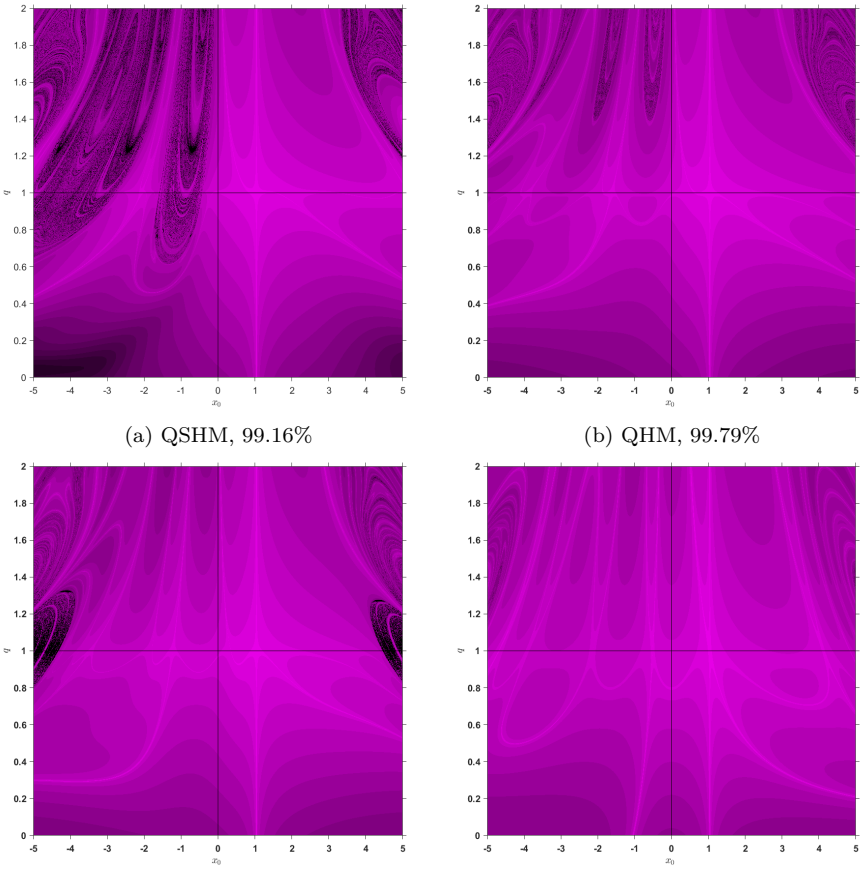
		QSHM						QHM				
q	x^*	$\lvert x_{k+1} - x_k\rvert$	$\lvert S_1(x_{k+1})\rvert$	ACOC	CPU	Iter	x^*	$\lvert x_{k+1} - x_k\rvert$	$\lvert S_1(x_{k+1})\rvert$	ACOC	CPU	Iter
0.7	x_1	5.41206e-06	4.0778e-07	1.01	0.05	07	x_1	3.39205e-06	2.55579e-07	1.0	0.51	07
0.8	x_1	5.024231e-06	2.590377e-07	0.99	0.06	06	x_1	2.434003e-06	1.254917e-07	0.99	0.49	06
0.999	x_1	1.56993e-05	4.23311e-09	2.3	0.04	03	x_1	4.42721e-06	1.19374e-09	3.7	0.057	03
0.999999	x_1	2.934109e-06	7.91589e-13	2.84	0.076	03	x_1	2.39488e-05	6.42819e-12	3.11	0.069	03
0.9999999	x_1	2.03524e-05	1.18538e-12	2.22	0.074	03	x_1	2.65808e-05	9.35374e-12	3.06	0.066	03

Table 3: Results of QCM, QNM for $S_1(x)$ with initial guess $x_0 = 2$

		QCM						QNM				
q	x^*	$\lvert x_{k+1} - x_k\rvert$	$\lvert S_1(x_{k+1})\rvert$	ACOC	CPU	Iter	x^*	$\lvert x_{k+1} - x_k\rvert$	$\lvert S_1(x_{k+1})\rvert$	ACOC	CPU	Iter
0.7	x_1	2.22229e-06	1.674418e-07	1.0	0.05	07	x_1	2.15963e-06	1.62721e-07	1.0	0.04	07
0.8	x_1	1.04465e-05	5.38613e-07	1.0	0.04	05	x_1	2.78538e-06	1.43606e-07	0.99	0.04	06
0.999	x_1	1.16708e-04	3.14705e-08	2.9	0.05	03	x_1	8.202927e-05	2.03785e-08	2.04	0.05	04
0.999999	x_1	1.5812e-04	4.2838e-11	2.84	0.074	03	x_1	8.9510e-05	2.0506e-09	2.02	0.78	04
0.9999999	x_1	1.4927e-04	9.0351e-11	2.86	0.65	03	x_1	8.9517e-05	2.0726e-09	2.02	0.65	04

Table 4: Results of classical SHM, HM, CM, and NM for $S_1(x)$ with initial guess $x_0 = 2$

Methods	x^*	$\lvert x_{k+1} - x_k\rvert$	$\lvert S_1(x_{k+1})\rvert$	ACOC	CPU	Iter
SHM	x_1	2.7000004e-06	1.11022302e-16	2.85	0.045	03
HM	x_1	2.363779e-05	6.661338e-16	3.12	0.045	03
CM	x_1	1.5668661e-04	5.4578563e-13	2.85	0.045	03
NM	x_1	8.95180e-05	2.075144e-09	2.02	0.07	04



(a) QSHM, 99.16%

(b) QHM, 99.79%

(c) QCM, 98.81%

(d) QNM, 99.86%

Figure 5: Convergence planes of $S_1(x)$, $x_0 \in \mathbb{R}$

Table 5: Results of QSHM, QHM for $S_1(x)$ with initial guess $x_0 = 2.5$

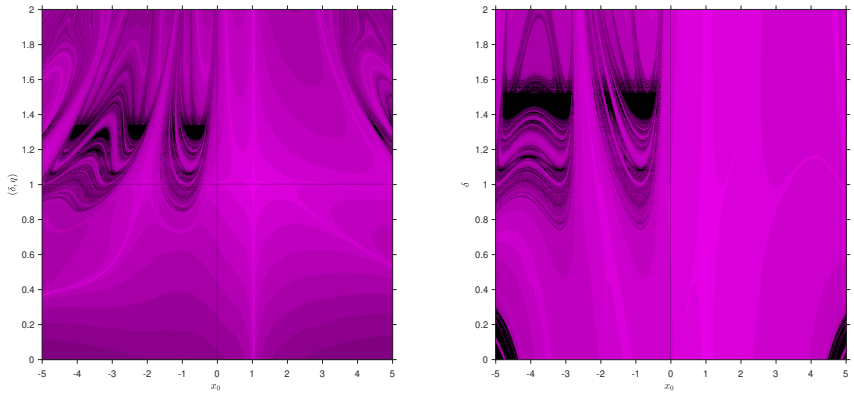
	QSHM						QHM					
q	x^*	$\lvert x_{k+1} - x_k\rvert$	$\lvert S_1(x_{k+1})\rvert$	ACOC	CPU	Iter	x^*	$\lvert x_{k+1} - x_k\rvert$	$\lvert S_1(x_{k+1})\rvert$	ACOC	CPU	Iter
0.7	x_1	8.043468e-06	6.06048e-07	1.00	0.067	07	x_1	3.79807e-06	2.86172e-07	1.0	0.71	07
0.8	x_1	5.309870e-06	2.737646e-07	0.99	0.066	06	x_1	1.145934e-05	5.90820e-07	1.00	0.66	05
0.999	x_1	2.99318e-04	8.06974e-08	3.11	0.057	03	x_1	5.40366e-04	1.45734e-07	3.18	0.064	03
0.999999	x_1	2.2857e-04	6.17301e-11	3.25	0.058	03	x_1	6.36395e-04	1.01844e-10	3.11	0.063	03
0.9999999	x_1	3.62815e-04	5.76748e-10	3.044	0.063	03	x_1	5.61367e-04	4.05469e-09	3.19	0.061	03

Table 6: Results of QCM, QNM for $S_1(x)$ with initial guess $x_0 = 2.5$

	QCM						QNM					
q	x^*	$\lvert x_{k+1} - x_k\rvert$	$\lvert S_1(x_{k+1})\rvert$	ACOC	CPU	Iter	x^*	$\lvert x_{k+1} - x_k\rvert$	$\lvert S_1(x_{k+1})\rvert$	ACOC	CPU	Iter
0.7	x_1	1.32199e-05	9.96074e-07	0.99	0.066	06	x_1	2.15963e-06	1.62721e-07	1.0	0.04	07
0.8	x_1	1.8032e-06	9.29707e-08	1.00	0.070	06	x_1	2.78538e-06	1.43606e-07	0.99	0.04	06
0.999	x_1	0.00189	5.106281e-07	2.74	0.060	03	x_1	6.41092e-04	6.6580e-08	2.04	0.054	04
0.999999	x_1	0.00223	5.54425e-09	2.66	0.062	03	x_1	6.64269e-04	1.1411e-07	2.02	0.062	04
0.9999999	x_1	0.00201	3.7527e-10	2.71	0.062	03	x_1	6.6429e-04	1.1428e-07	2.02	0.055	04

Table 7: Results of classical SHM, HM, CM, and NM for $S_1(x)$ with initial guess $x_0 = 2.5$

Methods	x^*	$\lvert x_{k+1} - x_k\rvert$	$\lvert S_1(x_{k+1})\rvert$	ACOC	CPU	Iter
SHM	x_1	2.32205e-04	6.32383e-13	3.25	0.065	03
HM	x_1	6.362408e-04	1.17829e-11	3.11	0.062	03
CM	x_1	0.00201	1.158377e-09	2.71	0.059	03
NM	x_1	6.64292e-04	1.142994e-07	2.029	0.054	04



(a) PFQIM, for parameters (δ, q), 95.72% (b) PFQIM, for parameters (δ), 95.36%

Figure 6: Convergence planes of $S_1(x)$ for different parameters

The convergence plane of each iterative method and its percentage of convergence are described in Figure 5. The root x_1 is painted in the magenta color and the faster convergence of the method reaches the root is brighter in the convergence plane. As given in the figure, QSHM (99.16%), QHM (99.79%), QCM (98.81%), and QNM (99.86%) among which the cubic order convergence method gives faster convergence then QNM. It can be also observed that at the order of derivative $q = 1$, the method is not converging to the root. Although the qualitative behavior of the proposed method on S_1 is illustrated for different parameters in Figure 6

(a) shows the combination of parameters pairs (δ, q) along $y - axis$, while Figure 6 (b) represents the variation of the parameter δ at $q = 0.999$ to study the stability of method in depth way.

Our second test function is the polynomial obtained from the Schrödinger wave equation, which is given in the example below.

Example 4.2 (Schrödinger wave equation for a hydrogen atom)**.** In quantum physics, the solution of the Schrödinger wave equation [19] for a charged particle moving in a Coulomb potential is linked to a probability distribution for the electron's location relative to the core. The standard Schrödinger equation for a particle with mass m traveling in a central potential is given by

$$-\frac{\hbar^2}{2m}\frac{\partial \Phi}{\partial t} - K\frac{e^2}{r}\Phi = E\Phi,$$

where r is the distance of the electron from the core and E is the energy. Also, the equation has the following representation in spherical coordinates:

$$-\frac{\hbar^2}{2m}\left[\frac{1}{r^2}\frac{\partial(r^2\frac{\partial \Phi}{\partial r})}{\partial r} + \frac{1}{r^2\sin\theta}\frac{\partial(\sin\theta\frac{\partial \Phi}{\partial \theta})}{\partial \theta} + \frac{1}{r^2\sin^2\theta}\frac{\partial^2\Phi}{\partial \psi^2}\right] + \frac{e^2\Phi}{r} = E\Phi.$$

The standard technique can be used to separate the final equation into angular and radial parts. The angular equation can also be written as a system of two equations, one of which is the Legendre equation [3].

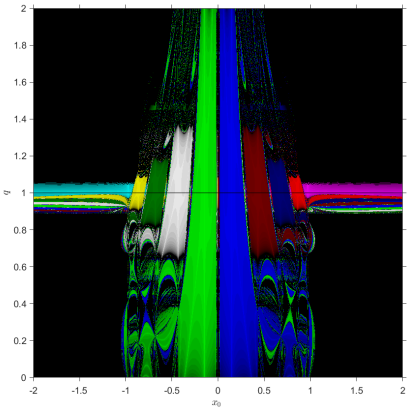
$$(1-x^2)\frac{d^2g}{dx^2} - 2x\frac{dg}{dx} + \left(n(n+1) - \frac{m^2}{1-x^2}\right)g(x) = 0.$$

The equation is simplified to Legendre polynomials when $m = 0$, corresponding to the azimuthal symmetry situation. We use the following form of the Legendre equation that we have derived:

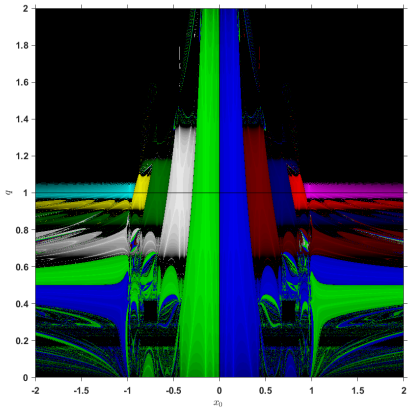
$$S_2(x) = 46189x^{10} - 109395x^8 + 90090x^6 - 30030x^4 + 3465x^2 - 63$$

with the roots, $x_1 = -0.9739$ (cyan), $x_2 = 0.9739$ (magenta), $x_3 = -0.8651$ (yellow), $x_4 = 0.8651$ (red), $x_5 = -0.6794$ (green), $x_6 = 0.6794$ (blue), $x_7 = -0.4334$ (white), $x_8 = 0.4334$ (dark red), $x_9 = -0.1489$ (dark green), and $x_{10} = 0.1489$ (dark blue) are painted in the respective colors given in Figure 7.

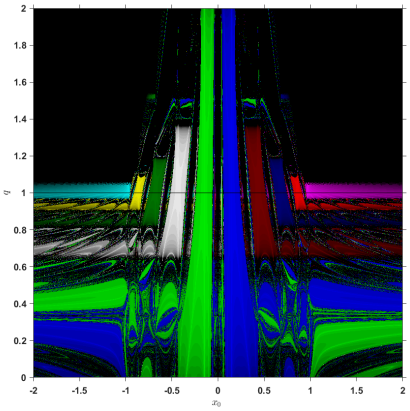
Moreover, the numerical results of $S_2(x)$ are shown in Tables 8 and 9 and compared with corresponding classical methods as given in Table 10, where "dgt" denotes divergence and "–" in $ACOC$ denotes an extremely large or undefined value.



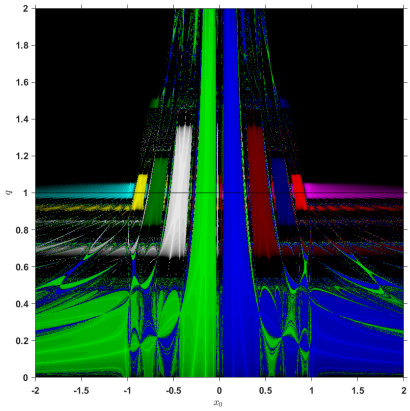
(a) QSHM, 34.95%



(b) QHM, 52.44%

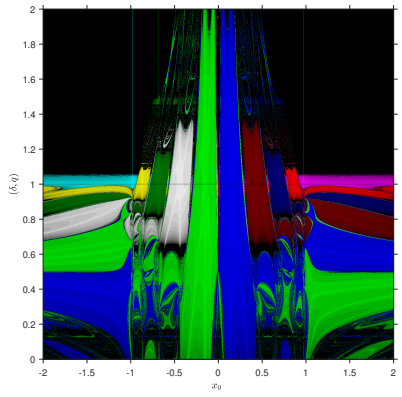


(c) QCM, 49.43%

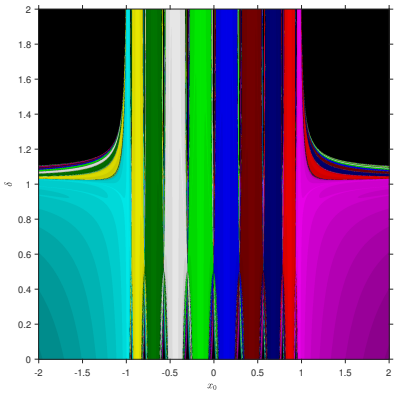


(d) QNM, 51.63%

Figure 7: Convergence planes of $S_2(x)$, $x_0 \in \mathbb{R}$



(a) PFQIM„ for parameters (δ, q) 55.13%



(b) PFQIM, for parameters (δ), 77.58%

Figure 8: Convergence planes of $S_2(x)$ for different parameters

Table 8: Results of QSHM, QHM for $S_2(x)$ with initial guess $x_0 = 1.5$

	QSHM						QHM					
q	x^*	$\lvert x_{k+1} - z_k\rvert$	$\lvert S_1(x_{k+1})\rvert$	ACOC	CPU	Iter	x^*	$\lvert x_{k+1} - x_k\rvert$	$\lvert S_1(x_{k+1})\rvert$	ACOC	CPU	Iter
0.7	x_8	5.75570e-07	1.61295e-04	1.0	0.08	21	x_8	5.7557e-07	1.61295e-04	1.0	0.05	21
0.8	dgt	0.03984	27.29704	–	0.05	500	dgt	0.03984	27.29704	–	0.05	500
0.999	x_2	5.0865e-08	5.74888e-06	0.7	0.06	04	x_2	5.08650e-08	5.74888e-06	0.7	0.05	12
0.999999	x_2	9.192652e-07	1.04597e-07	2.47	0.074	04	x_2	1.483287e-07	1.68633e-08	2.91	0.096	07
0.9999999	x_2	6.8850e-08	7.6988e-10	3.18	0.093	04	x_2	1.0325e-07	1.1996e-09	3.01	0.093	07

Table 9: Results of QCM, QNM for $S_2(x)$ with initial guess $x_0 = 1.5$

	QCM						QNM					
q	x^*	$\lvert x_{k+1} - x_k\rvert$	$\lvert S_1(x_{k+1})\rvert$	ACOC	CPU	Iter	x^*	$\lvert x_{k+1} - z_k\rvert$	$\lvert S_1(x_{k+1})\rvert$	ACOC	CPU	Iter
0.7	x_8	6.49974e-07	1.82144e-04	1.0	0.04	27	x_8	5.755708e-07	1.61295e-04	1.0	0.04	21
0.8	dgt	2.34549e-07	3.44064e-05	1.0	0.05	500	dgt	0.03984	27.29704	–	0.05	500
0.999	x_2	5.08650e-08	5.74888e-06	0.9	0.05	12	x_2	1.030859e-08	1.16369e-09	2.01	0.069	12
0.999999	x_2	2.10233e-06	2.3916e-07	2.79	0.063	08	x_2	1.03085e-08	1.16369e-09	2.01	0.064	12
0.9999999	x_2	2.05854e-06	1.913076e-08	2.81	0.073	08	x_2	1.07319e-08	1.15505e-10	2.00	0.066	12

Table 10: Results of classical SHM, HM, CM, and NM for $S_2(x)$ with initial guess $x_0 = 1.5$

Methods	x^*	$\lvert x_{k+1} - x_k\rvert$	$\lvert S_1(x_{k+1})\rvert$	ACOC	CPU	Iter
SHM	x_2	8.8108443e-07	3.63797880e-12	2.48	0.067	04
HM	x_2	1.63192236e-07	1.6825651e-11	2.87	0.07	07
CM	x_2	2.12884862e-06	4.2291503e-11	2.79	0.064	08
NM	x_2	1.0779137e-08	1.9554136e-11	2.00	0.046	12

The convergence planes of S_2 are illustrated in Figure 7, where distinct colors represent different roots of the polynomial, excluding black. The qualitative performance of the proposed method on S_2 is analyzed under different parameter choices. Figure 8 (a) displays the combinations of (δ, q) along the y-axis, while Figure 8 (b) illustrates how varying δ at $q = 0.999$ helps explore the stability of the method in greater detail.

5 Conclusion

We studied the complex and real dynamics of a newly designed Chebyshev–Halley family iterative method using the q-derivative. We proved a scaling theorem that enables PFQIM to analyze the dynamics of quadratic polynomials. It is observed that if the value of q is very close to one, then the method becomes more stable and converges faster. However, one limitation is that the method does not work for $q = 1$. Numerical experiments further demonstrated that PFQIM performs better than the corresponding classical iterative methods. The role of strange fixed points is clearly visualized in the dynamical planes. Both parameters q and δ showed strong stability behavior for various values. Moreover, since real dynamics differ from complex dynamics, they should be studied separately. The convergence plane proved to be a useful tool for visualizing the real dynamics of q-iterative methods.

Funding

This research received no external funding.

Conflicts of Interest

All authors jointly worked on the results, and they read and approved the final manuscript. Also, they declare that they have no competing interests.

References

[1] Alawaideh, Y. M., Lupas, A. A., Al-Khamiseh, B. M., Yousif, M. A., Mohammed, P. O. and Hamed, Y. S., *Hamiltonian formulation for continuous systems with second-order derivatives: a study of Podolsky generalized electrodynamics*, Axioms, 13(10) (2024), 665.

[2] Babajee, D. K. R., Cordero, A. and Torregrosa, J. R., *Study of iterative methods through the Cayley quadratic test*, J. Comput. Appl. Math., 291 (2016), 358–369.

[3] Bronshtein, I. N., Musiol, G., Semendyayev, K. A. and Mühlig, H., *Handbook of mathematics*, Springer, Berlin, 2015.

[4] Campos, B., Cordero, A., Torregrosa, J. R. and Vindel, P., *Dynamics of the family of c-iterative methods*, Int. J. Comput. Math., 92 (2014), 1–11.

[5] Chicharro, F. I., Cordero, A. and Torregrosa, J. R., *Drawing dynamical and parameters planes of iterative families and methods*, Sci. World J., (2013), 23566140.

[6] Cordero, A. and Torregrosa, J. R., *Variants of Newton's method using fifth-order quadrature formulas*, Appl. Math. Comput., 190(1) (2007), 686–698.

[7] Cordero, A., Torregrosa, J. R. and Vindel, P., *Dynamics of a family of Chebyshev–Halley type methods*, Appl. Math. Comput., 219(16) (2013), 8568–8583.

[8] Hernández, M. A. and Salanova, M. A., *A family of Chebyshev–Halley type methods*, Int. J. Comput. Math., 47(1–2) (1993), 59–63.

[9] Jackson, F. H., *A q-form of Taylor's formula*, Messenger Math., 38 (1909), 62–64.

[10] Jing, S. C. and Fan, H. Y., *q-Taylor's formula with its q-remainder*, Commun. Theor. Phys., 23(1) (1995), 117.

[11] Lee, M. and Kim, Y. I., *The dynamical analysis of a uniparametric family of three-point optimal eighth-order multiple-root finders under the Möbius conjugacy map on the Riemann sphere*, Numer. Algorithms, 83(3) (2020), 1063–1090.

[12] Magreñán, Á. A., *A new tool to study real dynamics: the convergence plane*, Appl. Math. Comput., 248 (2014), 215–224.

[13] Magreñán, Á. A., Cordero, A., Gutiérrez, J. M. and Torregrosa, J. R., *Real qualitative behavior of a fourth-order family of iterative methods by using the convergence plane*, Math. Comput. Simul., 105 (2014), 49–61.

[14] Magreñán, Á. A. and Gutiérrez, J. M., *Real dynamics for damped Newton's method applied to cubic polynomials*, J. Comput. Appl. Math., 275 (2015), 527–538.

[15] Maroju, P., Magreñán, Á. A., Sarría, Í. and Kumar, A., *Local convergence of fourth- and fifth-order parametric family of iterative methods in Banach spaces*, J. Math. Chem., 58(3) (2020), 1572–1589.

[16] Muhammad, B., Alawaideh, Y. M., Rehman, S. U., Yousif, M. A., Younas, U., Baleanu, D. and Mohammed, P. O., *Gross–Pitaevskii systems of fractional order with respect to multicomponent solitary wave dynamics*, Sci. Rep., 15(1) (2025), 24337.

[17] Nievergelt, Y., *Computing the distance from a point to a helix and solving Kepler's equation*, Nucl. Instrum. Methods Phys. Res. A, 598(3) (2009), 788–794.

[18] Nonlaopon, K., Khan, A. G., Ameen, F., Awan, M. U. and Cesarano, C., *Multi-step quantum numerical techniques for finding the solutions of nonlinear equations*, Mathematics, 10(15) (2022), 2227.

[19] Pavkov, T. M., Kabadzhov, V. G., Ivanov, I. K. and Ivanov, S. I., *Local convergence analysis of a one-parameter family of simultaneous methods with applications to real-world problems*, Algorithms, 16(2) (2023), 103.

[20] Román, C. E. C., *A new approach to study the dynamics of the modified Newton's method to multiple roots*, Bull. Math. Soc. Sci. Math. Roumanie, 62 (2019), 110.

[21] Sana, G., Mohammed, P. O., Shin, D. Y., Noor, M. A. and Oudat, M. S., *On iterative methods for solving nonlinear equations in quantum calculus*, Fractal Fract., 5(60) (2021).

[22] Sana, G., Noor, M. A. and Baleanu, D., *Novel higher-order iterative schemes based on the q-calculus for solving nonlinear equations*, AIMS Math., 7(3) (2021), 2473–2498.

[23] Singh, P., Mishra, P. K. and Pathak, R. S., *q-iterative methods*, IOSR J. Math., 9(1) (2013), 06–13.

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 1, 2026, pp 304–338
https://doi.org/10.22067/ijnao.2025.94317.1676
https://ijnao.um.ac.ir/

CC BY NC

Research Article

Mathematical modeling and optimal control strategies of the typhoid transmission with cost-effectiveness analysis

K. Oubouskour*,[1], S. Khassal[1], B. Khajji[2] and O. Balatif[1]

Abstract

This paper presents the development of a sophisticated mathematical model aimed at optimizing the control of typhoid fever in Taiwan. The model incorporates key factors influencing transmission dynamics, including human interactions, exposure to contaminated sources, and the effectiveness of preventive measures. It is analyzed using optimal control theory and rigorous numerical simulations, enabling a comprehensive assessment of various intervention strategies. The findings highlight that a combination of awareness campaigns and vaccination programs constitutes the most effective approach to reducing disease spread. This work underscores the critical role of mathematical modeling as a decision-support tool in public health planning, providing data-driven insights to guide the prevention and control of typhoid fever.

AMS subject classifications (2020): Primary 92D30; Secondary 49K15, 34H05, 93C15.

*Corresponding author

Received 9 July 2025; revised 30 September 2025; accepted 2 October 2025

Khadija Oubouskour, e-mail: oubouskour.khadija@ucd.ac.ma
Sofiane Khassal, e-mail: sofiane.k@gmail.com
Bouchaib Khajji, e-mail: khajjibouchaib@gmail.com
Omar Balatif, e-mail: balatif.maths@gmail.com

[1]Department of Mathematics, Faculty of Sciences, Chouaib Doukkali University, El-Jadida, Morocco.
[2]Department of Mathematics and Computer Science, Hassan II University, Casablanca, Morocco.

How to cite this article
Oubouskour, K., Khassal, S., Khajji, B. and Balatif, O., Mathematical modeling and optimal control strategies of the typhoid transmission with cost-effectiveness analysis. *Iran. J. Numer. Anal. Optim.*, 2026; 16(1): 304-338. https://doi.org/10.22067/ijnao.2025.94317.1676

Keywords: Optimal control, Pontryagin's maximum principle, Typhoid fever, Hamiltonian, Cost-effectiveness.

1 Introduction

Typhoid fever, also referred to as enteric fever, is caused by Salmonella enterica serovars Typhi and Paratyphi A, B, and C.

S. Typhi is responsible for typhoid fever, whereas S. Paratyphi can also cause enteric fever [4]. Most infections from non-typhoidal Salmonella spp. lead to diarrhea and only rarely result in bloodstream infections. Individuals affected by typhoid fever, however, typically harbor the bacteria in both their bloodstream and intestinal tract. Notably, enteric fever is exclusive to humans and is primarily transmitted through contaminated food or water.

Medical research indicates that the duration of the incubation period for typhoid fever depends on the bacterial load in the body, which is influenced by the patient's overall health and stomach acidity [9]. The infectious dose of enteric fever is estimated at 10^5 bacilli. Older individuals and those with achlorhydria—characterized by reduced stomach acid secretion and impaired cell-mediated immunity—may acquire the infection after exposure to a lower infectious dose [17].

The incubation period for typhoid fever averages 10 to 20 days (range: 3–56 days), whereas for S. Paratyphi A, it typically ranges from 1 to 10 days [13]. During the first week, symptoms are often nonspecific and may include headache, fatigue, and fever. Constipation and a mild, non-productive cough are also frequently reported. In the second week, patients generally present with lethargy, persistent fever, mild abdominal distension, and splenomegaly. The appearance of rose-colored spots, usually 2–4 mm in diameter and blanching under pressure, is considered a classic sign of typhoid fever, although such spots are uncommon and difficult to detect in individuals with darker skin tones [13].

Confirmation of typhoid infection requires the isolation of S. Typhi from blood, bone marrow, or lesion samples. Importantly, asymptomatic carriers, individuals excreting S. Typhi in their stool or urine without showing symptoms can also contribute to transmission [40]. The most widely accepted diagnostic method is blood culture, as over 80% of infected individuals have detectable bacteria in their blood. Factors such as prior antibiotic use, the volume of blood cultured, and the timing of collection can explain why some cultures fail to isolate S. Typhi [41].

In 2017, an estimated 14.3 million people were infected with enteric fever, including 11 million S. Typhi cases and 120,000 deaths, more than half occurring in children under 15 years [37]. According to [34], these numbers decreased in 2021, with 9.3 million infections and 107,500 deaths annually. Although the disease has been controlled in developed countries, it remains

endemic in regions with poor hygiene. South Asia—including India, Pakistan, and Bangladesh—continues to show particularly high incidence rates [10, 24].

The emergence of bacterial resistance to common antibiotics and the existence of multiple Salmonella strains [6, 23] indicate that improvements in sanitation and hygiene alone are insufficient. Vaccination, therefore, remains a critical strategy for disease control, especially in epidemic areas. Currently, three types of licensed typhoid vaccines are available:

Typhoid conjugate vaccines (TCVs): New-generation TCVs include two licensed types, Typbar-TCV and PedaTyph™ (the latter licensed and marketed only in India). These are Vi-TT conjugate vaccines, produced with different carrier proteins. Additional TCV candidates are currently under regulatory review [42].

Vi capsular polysaccharide (ViPS) vaccine: Purified from the S. Typhi strain Ty2, this vaccine is used in ViPS-hepatitis A non-conjugate combination vaccines, mainly for travelers. Available data indicate that seroconversion against vaccine antigens is equivalent to that obtained with monovalent ViPS and hepatitis A vaccines [42].

Ty21a vaccine: Orally administered and prepared from a live attenuated Ty2 strain of S. Typhi, this vaccine has several genes modified by chemical mutagenesis, resulting in a strain that lacks the Vi antigen [42].

Mathematically, several models have been proposed to clarify transmission dynamics and evaluate the effectiveness of different preventive strategies against typhoid fever. These include models focusing on direct human-to-human transmission [27, 21], transmission via contaminated food and water [39, 29], and hybrid models combining both routes [16, 36]. Preventive strategies against typhoid epidemics have been studied through approaches involving treatment [28, 39], sanitation [3, 38], educational campaigns [33, 26], and media coverage [25]. The outcomes of such mathematical models contribute to the development of effective strategies for controlling and preventing typhoid fever.

For instance, Tilahun, Makinde, and Malonza [38] developed an epidemic model to investigate transmission dynamics and highlight the role of public health education in reducing disease spread in Taiwan and China. Musa et al. [26] proposed a model assessing the effects of vaccination, including herd immunity, and demonstrated that vaccination provides indirect short-term protection and reduces incidence. Alharbi, Alalhareth, and Ibrahim [3] designed a model to evaluate the role of direct and indirect transmission, incorporating vaccination coverage, treatment, and water sanitation. Their results revealed that reducing typhoid prevalence depends on implementing these interventions, minimizing contact with infected patients, and protecting water sources from fecal contamination. Similarly, Abboubakar et al. [1] developed a fractional derivative model incorporating vaccination, showing that immunization significantly reduces disease spread. Abboubakar and Racke [2] analyzed a standalone model including incomplete vaccina-

tion, hygiene measures, and treatment. Based on clinical data from Mbandjock, Cameroon, they concluded that large-scale vaccination campaigns, environmental sanitation, and therapeutic interventions must be combined to effectively control typhoid fever.

This body of research highlights the importance of mathematical models for understanding the transmission dynamics of typhoid fever and their relevance to public health interventions aimed at prevention and control.

Our main contribution lies in integrating several aspects often neglected in existing models. First, we consider the impact of infected immigrants on transmission dynamics by distributing them across the different compartments of the model. Second, we account for the role of population and hospital waste in contaminating water and food, as well as the effect of government sterilization efforts. Third, we include the therapeutic behaviors of infected individuals, assuming that patients may choose between drug-based and herbal treatments, which influences recovery rates. Some patients may also discontinue treatment due to financial constraints, lack of awareness, or premature relief of symptoms. Finally, we evaluate the effectiveness of various typhoid control strategies using a cost-effectiveness framework, enabling identification of the most optimal interventions under resource constraints.

The study is organized as follows: Section 2 presents the mathematical model, including its description and fundamental properties, such as the positivity and boundedness of solutions. Section 3 formulates the optimal control problem, establishes the existence of an optimal solution, and characterizes the controls using Pontryagin's maximum principle. Section 4 provides numerical simulations along with cost-effectiveness analyses of the proposed strategies. Finally, Section 5 is devoted to interpreting the results in a public health context and concludes with perspectives for future research. This last section is structured into two subsections: A discussion of the findings in relation to existing literature, and a conclusion summarizing the main contributions of the study.

2 Mathematical model

2.1 Model description

In this paper, we propose an eight-compartment model to describe the transmission dynamics of typhoid fever. The compartment of susceptible individuals, denoted by $S(t)$, represents the population at risk of infection who have neither contracted the disease nor developed immunity. Individuals temporarily immune as a result of vaccination are grouped into the vaccinated-

susceptible compartment, $S_v(t)$ [42]. Exposed individuals in the incubation period, denoted by $E(t)$, are also considered. The incubation time varies with the infectious dose, which for enteric fever is typically around 10^5 bacilli ingested (and about 10^3 for non-typhoidal *Salmonella*) [17].

Asymptomatic infectious individuals are represented by $I_A(t)$ [40], whereas symptomatic infectious individuals are described by $I_S(t)$ [13]. Those receiving medical treatment (conventional rather than herbal), either at home or in hospital, are included in the treatment compartment $H(t)$. Recovered individuals, denoted by $R(t)$, are those who have fully recovered, either through successful medical treatment or via natural immunity, and have acquired temporary protection against typhoid fever [19].

At any time t, the total human population is defined as

$$N(t) = S(t) + S_v(t) + E(t) + I_A(t) + I_S(t) + H(t) + R(t).$$

Note that the total population is not assumed to be constant, as it evolves according to the inflow of newborns and immigrants, and the outflow due to natural and disease-related mortality. Finally, the environmental compartment $E_n(t)$ represents the density of *Salmonella* present in contaminated water and food.

The overall structure of the proposed model is illustrated in Figure 1, while its parameters are summarized in Table 1.

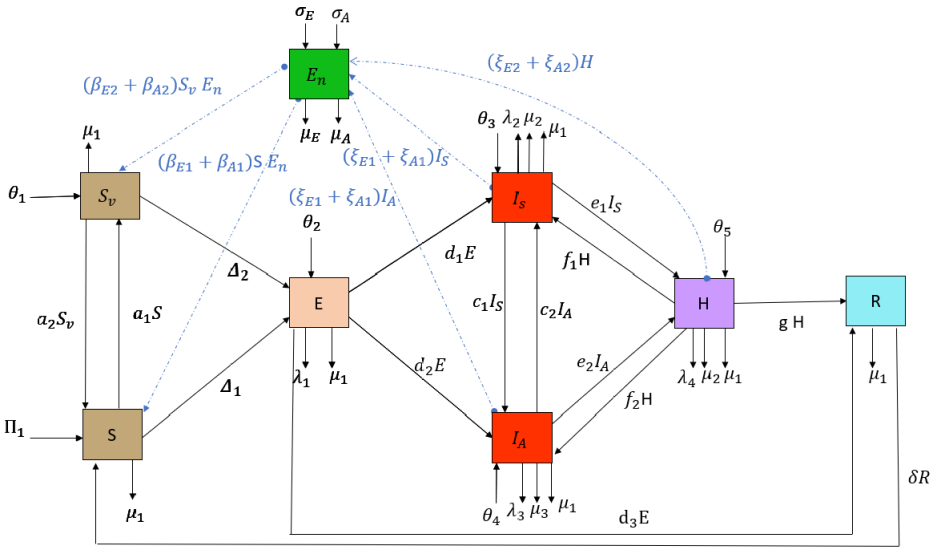


Figure 1: Model of typhoid fever

where

$$\Delta_1 = (\beta_{A1} + \beta_{E1})SE_n + (b_1 I_S + b_2 I_A)S.$$
$$\Delta_2 = (\beta_{A2} + \beta_{E2})S_v E_n + (b_1' I_S + b_2' I_A)S_v.$$

Table 1: Model parameters (person/day)

Symbol	Description
Π_1	Birth rate (newborns entering the population).
θ_1	Integration rate of vaccinated immigrants.
θ_2	Integration rate of immigrants carrying Salmonella bacteria.
θ_3	Integration rate of symptomatic infected immigrants.
θ_4	Integration rate of asymptomatic infected immigrants.
θ_5	Integration rate of immigrants based on treatment duration.
λ_1	Exit rate of immigrants carrying Salmonella.
λ_2	Discharge rate of symptomatic infectious immigrants.
λ_3	Discharge rate of asymptomatic infectious immigrants.
λ_4	Discharge rate of immigrants under treatment.
μ_1	Natural mortality rate.
μ_2	Disease-induced mortality rate for symptomatic infected individuals.
μ_3	Disease-induced mortality rate for asymptomatic infected individuals.
μ_E	Bacterial mortality rate in water.
μ_A	Bacterial mortality rate in food.
σ_E	Water disinfection rate.
σ_A	Food disinfection rate.
β_{A1}	Contamination rate via food.
β_{E1}	Contamination rate via water.
β_{A2}	Contamination rate of vaccinated susceptibles via food.
β_{E2}	Contamination rate of vaccinated susceptibles via water.
a_1	Vaccination rate.
a_2	Vaccine immunity waning rate.
b_1	Contamination rate of susceptibles by symptomatic infected.
b_2	Contamination rate of susceptibles by asymptomatic infected.
b_1'	Contamination rate of vaccinated individuals by symptomatic infected.
b_2'	Contamination rate of vaccinated individuals by asymptomatic infected.
d_1	Rate at which a Salmonella carrier becomes symptomatic.
d_2	Rate at which a Salmonella carrier becomes asymptomatic.
d_3	Natural immunity recovery rate.
e_1	Hospitalization rate of symptomatic infected individuals.
e_2	Hospitalization rate of asymptomatic infected individuals.
f_1	Rate of treatment cessation.
f_2	Rate of treatment discontinuation or resistance.
g	Recovery rate.
δ	Immunity loss rate.
ξ_{E1}	Bacterial shedding rate in water by infected individuals.
ξ_{A1}	Bacterial shedding rate in food by infected individuals.
c_1	Rate at which symptomatic individuals become asymptomatic.
c_2	Rate at which asymptomatic individuals become symptomatic.
ξ_{E2}	Bacterial shedding rate in water by hospitalized individuals.
ξ_{A2}	Bacterial shedding rate in food by hospitalized individuals.

As a result, the model of Figure 1 can be expressed by the following nonlinear system:

$$
\begin{cases}
\dfrac{dS(t)}{dt} &= \Pi_1 + a_2 S_v(t) + \delta R(t) - (\mu_1 + a_1) S(t) \\
&\quad -[b_1 I_S(t) + b_2 I_A(t) + (\beta_{E1} + \beta_{A1}) E_n(t)] S(t), \\
\dfrac{dS_v(t)}{dt} &= \theta_1 - (a_2 + \mu_1) S_v(t) + a_1 S(t) \\
&\quad -[b_1^{'} I_S(t) + b'_2 I_A(t) + (\beta_{E2} + \beta_{A2}) E_n(t)] S_v(t), \\
\dfrac{dE(t)}{dt} &= \theta_2 - (\mu_1 + \lambda_1 + d_1 + d_2 + d_3) E(t) \\
&\quad +[b_1 I_S(t) + b_2 I_A(t) + (\beta_{E1} + \beta_{A1}) E_n(t)] S(t) \\
&\quad +[b_1^{'} I_S(t) + b'_2 I_A(t) + (\beta_{E2} + \beta_{A2}) E_n(t)] S_v(t), \\
\dfrac{dI_S(t)}{dt} &= \theta_3 + d_1 E(t) + c_2 I_A(t) - (c_1 + e_1 + \lambda_2 + \mu_1 + \mu_2) I_S(t) + f_1 H(t), \\
\dfrac{dI_A(t)}{dt} &= \theta_4 + d_2 E(t) + c_1 I_S(t) - (c_2 + e_2 + \lambda_3 + \mu_1 + \mu_3) I_A(t) + f_2 H(t), \\
\dfrac{dH(t)}{dt} &= \theta_5 + e_1 I_S(t) + e_2 I_A(t) - (g + \lambda_4 + \mu_1 + \mu_2 + f_2 + f_1) H(t), \\
\dfrac{dR(t)}{dt} &= g H(t) + d_3 E(t) - (\mu_1 + \delta) R(t), \\
\dfrac{dE_n(t)}{dt} &= (-\sigma_E - \sigma_A - \mu_E - \mu_A) E_n(t) + (\xi_{E1} + \xi_{A1})(I_A(t) + I_S(t)) \\
&\quad +(\xi_{E2} + \xi_{A2}) H(t).
\end{cases}
\tag{1}
$$

Under the following initial conditions:

$S(0) = S^0 \geqslant 0$, $S_v(0) = S_v^0 \geqslant 0$, $E(0) = E^0 \geqslant 0$, $E_n(0) = E_n^0 \geqslant 0$, $I_A(0) = I_A^0 \geqslant 0$, $I_S(0) = I_S^0 \geqslant 0$, $H(0) = H^0 \geqslant 0$ and $R(0) = R^0 \geqslant 0$.

2.2 Model properties

The proposed model describes the evolution of typhoid fever within a population. To ensure the mathematical rigor of system (1), it is essential to investigate and establish fundamental properties, such as the existence, positivity, and boundedness of solutions. These properties guarantee that the model is well-posed and that the solutions remain biologically meaningful over time.

2.2.1 Positivity of solutions

Theorem 1. If $S^0 \geqslant 0$, $S_v^0 \geqslant 0$, $E^0 \geqslant 0$, $I_A^0 \geqslant 0$, $I_S^0 \geqslant 0$, $H^0 \geqslant 0$, $R^0 \geqslant 0$, $E_n^0 \geqslant 0$, then the solutions $S(t)$, $S_v(t)$, $E(t)$, $I_A(t)$, $I_S(t)$, $H(t)$, $R(t)$, and $E_n(t)$ of system (1) are positive for all $t \geqslant 0$.

Proof. It follows from the first equation of system (1) that

$$\frac{dS(t)}{dt} = \Pi_1 + a_2 S_v(t) + \delta R(t) - (\mu_1 + a_1 + (\beta_{A1} + \beta_{E1})E_n(t) + (b_1 I_S(t) + b_2 I_A(t))S(t).$$

Then

$$\frac{dS(t)}{dt} \geqslant -[\mu_1 + a_1 + (\beta_{A1} + \beta_{E1})E_n(t) + (b_1 I_S(t) + b_2 I_A(t))]S(t).$$

Let

$$F(t) = b_1 I_S(t) + b_2 I_A(t) + \mu_1 + a_1 + (\beta_{E1} + \beta_{A1})E_n(t).$$

Then

$$\frac{dS(t)}{dt} \geqslant -F(t)S(t).$$

By applying Gronwall's Lemma [12] and following the approach of Davarian et al. [11], we obtain

$$S(t) \geqslant S(0)\exp\left(-\int_0^t F(s)ds\right) \geqslant 0.$$

Similarly, we prove that $S_v(t) \geqslant 0, E(t) \geqslant 0, I_A(t) \geqslant 0, I_S(t) \geqslant 0, H(t) \geqslant 0,$ $R(t) \geqslant 0$ and $E_n(t) \geqslant 0$. □

2.2.2 Boundedness of the solutions

Theorem 2. The set $\Omega = \{(S, S_v, \ldots, R, E_n) \in \mathbb{R}^8_+ /$ $0 \leq S(t) + \cdots + R(t) \leq N(0) + \frac{Z}{\mu_1}$ and $0 \leq E_n(t) \leq E_n(0) + \frac{C}{q_{10}}\}$ is positively invariant under system with initial conditions $S^0 \geqslant 0$, $S_v^0 \geqslant 0$, $E^0 \geqslant 0$, $E_n^0 \geqslant 0$, $I_A^0 \geqslant 0$, $I_S^0 \geqslant 0$, $H^0 \geqslant 0$ and $R^0 \geqslant 0$,

where

$$\begin{aligned} Z &= \Pi_1 + \theta_1 + \theta_2 + \theta_3 + \theta_4 + \theta_5. \\ q_{10} &= \sigma_E + \sigma_A + \mu_E + \mu_A. \\ C &= \frac{Z}{\mu_1}(\xi_{E2} + \xi_{A2} + 2\xi_{E1} + 2\xi_{A1}). \end{aligned}$$

Proof. From system (1), the total population satisfies

$$\frac{dN(t)}{dt} = \frac{dS(t)}{dt} + \frac{dS_v(t)}{dt} + \cdots + \frac{dR(t)}{dt}.$$

We replace each derivative with its corresponding expression in system (1) to obtain

$$\frac{dN(t)}{dt} = Z - \mu_1 N(t) - \big[\lambda_1 E(t) + (\mu_2 + \lambda_2) I_S(t) + (\mu_3 + \lambda_3) I_A(t) + (\mu_2 + \lambda_3) H(t)\big].$$

Since all state variables $E(t), I_S(t), I_A(t), H(t)$ are non-negative, the last term is always non-negative. Hence,

$$\frac{dN(t)}{dt} \leq Z - \mu_1 N(t).$$

we rewrite

$$\frac{dN(t)}{dt} + \mu_1 N(t) \leq Z.$$

Multiplying by the integrating factor $e^{\mu_1 t}$ gives

$$\frac{d}{dt}\big(e^{\mu_1 t} N(t)\big) \leq Z e^{\mu_1 t}.$$

Integrating from 0 to t, we obtain

$$e^{\mu_1 t} N(t) - N(0) \leq Z \int_0^t e^{\mu_1 s}\, ds = \frac{Z}{\mu_1}\big(e^{\mu_1 t} - 1\big).$$

Thus,

$$N(t) \leq N(0) e^{-\mu_1 t} + \frac{Z}{\mu_1}\big(1 - e^{-\mu_1 t}\big).$$

The coefficients $e^{-\mu_1 t}$ and $1 - e^{-\mu_1 t}$ both lie in $[0, 1]$ and sum to 1. Therefore, the right-hand side is a convex combination of $N(0)$ and $\frac{Z}{\mu_1}$ [20], which implies

$$\min\left(N(0), \frac{Z}{\mu_1}\right) \leq N(t) \leq \max\left(N(0), \frac{Z}{\mu_1}\right), \qquad \text{for all } \ t \geq 0.$$

Since, by Theorem 1, $N(t) \geq 0$, for all $t \geq 0$, the total population satisfies the following boudedness condition:

$$0 \leq N(t) \leq \max\left(N(0), \tfrac{Z}{\mu_1}\right).$$

Finally, by the elementary inequality $\max(a, b) \leq a + b$ for $a, b \geq 0$, we deduce the weaker but simpler estimate

$$0 \leq N(t) \leq N(0) + \frac{Z}{\mu_1}, \qquad \text{for all } \ t \geq 0.$$

On the other hand, we have

$$\frac{dE_n(t)}{dt} = -(\sigma_E + \sigma_A + \mu_E + \mu_A) E_n(t) + (\xi_{E1} + \xi_{A1})(I_A(t) + I_S(t)) + (\xi_{E2} + \xi_{A2}) H(t).$$

Since $0 \leq I_A(t) \leq N(0) + \frac{Z}{\mu_1}, \quad 0 \leq I_S(t) \leq N(0) + \frac{Z}{\mu_1}$ and $0 \leq H(t) \leq N(0) + \frac{Z}{\mu_1}$.

Then

$$\begin{aligned} \frac{dE_n(t)}{dt} \leq & -(\sigma_E + \sigma_A + \mu_E + \mu_A)E_n(t) \\ & +(\xi_{E1} + \xi_{A1})(N(0) + \frac{Z}{\mu_1} + N(0) + \frac{Z}{\mu_1}) \\ & +(\xi_{E2} + \xi_{A2})(N(0) + \frac{Z}{\mu_1}). \end{aligned}$$

We put

$$\begin{aligned} q_{10} = & \quad \sigma_E + \sigma_A + \mu_E + \mu_A. \\ C = & \frac{Z}{\mu_1}(\xi_{E2} + \xi_{A2} + 2\xi_{E1} + 2\xi_{A1} + 2N(0)). \end{aligned}$$

Then

$$\frac{dE_n(t)}{dt} \leq -q_{10}E_n(t) + C,$$

Proceeding in the same way, using the proof of the previous inequality, we can derive the following result:

$$0 \leq E_n(t) \leq E_n(0) + \frac{C}{q_{10}}, \quad \text{for all } t \geq 0.$$

Hence, we prove that the region Ω is a positively invariant set for the system (1). □

2.2.3 Existence of solutions:

Theorem 3. The system (1) that satisfies a given initial condition $(S^0, S_v^0, \ldots, E_n^0)$ has an unique solution.

Proof. Let $Y = [S, S_v, \ldots, E_n]^T$ and $\varphi(Y) = \left(\frac{dS(t)}{dt}, \frac{dS_v(t)}{dt}, \ldots, \frac{dE_n(t)}{dt}\right)$. Then the system (1) can be rewritten in the following form:

$$\varphi(Y) = AY + B(Y),$$

where

$$A = \begin{pmatrix} -q_1 & a_2 & 0 & 0 & 0 & 0 & \delta & 0 \\ a_1 & -q_2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -q_3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & d_1 & -q_4 & c_2 & f_1 & 0 & 0 \\ 0 & 0 & d_2 & c_1 & -q_5 & f_2 & 0 & 0 \\ 0 & 0 & 0 & e_1 & e_2 & -q_6 & 0 & 0 \\ 0 & 0 & d_3 & 0 & 0 & g & -q_7 & 0 \\ 0 & 0 & 0 & q_8 & q_8 & q_9 & 0 & -q_{10} \end{pmatrix},$$

such that

$$\begin{aligned} q_1 &= a_1 + \mu_1, & q_2 &= a_2 + \mu_1, \\ q_3 &= \mu_1 + \lambda_1 + d_1 + d_2 + d_3, & q_4 &= c_1 + e_1 + \lambda_2 + \mu_1 + \mu_2, \\ q_5 &= c_2 + e_2 + \lambda_3 + \mu_1 + \mu_3, & q_6 &= g + \lambda_4 + \mu_1 + \mu_2 + f_1 + f_2, \\ q_7 &= \mu_1 + \delta, & q_8 &= \xi_{E1} + \xi_{A1}, \\ q_9 &= \xi_{E2} + \xi_{A2}, & q_{10} &= \sigma_E + \sigma_A + \mu_E + \mu_A, \end{aligned}$$

and

$$B(Y) = \begin{bmatrix} \Psi_1 & \Psi_2 & \Psi_3 & \theta_3 & \theta_4 & \theta_5 & 0 & 0 \end{bmatrix}^T,$$

where

$$\begin{aligned} \Psi_1 &= \Pi_1 - [(b_1 I_S(t) + b_2 I_A(t))S(t) + (\beta_{E1} + \beta_{A1})S(t)E_n(t)], \\ \Psi_2 &= \theta_1 - [(b_1^{'} I_S(t) + b'_2 I_A(t))S_v(t) + (\beta_{E2} + \beta_{A2})S_v(t)E_n(t)], \\ \Psi_3 &= \theta_2 + (b_1 I_S(t) + b_2 I_A(t))S(t) + (\beta_{E1} + \beta_{A1})S(t)E_n(t) \\ &\quad + (b_1^{'} I_S(t) + b'_2 I_A(t))S_v(t) + (\beta_{E2} + \beta_{A2})S_v(t)E_n(t). \end{aligned}$$

The second term of system (1) satisfies

$$\begin{aligned} |B(Y_1) - B(Y_2)| = 2|&(b_1 I_{S1}(t) + b_2 I_{A1}(t))S_1(t) + (\beta_{E1} + \beta_{A1})S_1(t)E_{n1}(t) \\ &+ (b_1^{'} I_{S1}(t) + b'_2 I_{A1}(t))S_{v1}(t) + (\beta_{E2} + \beta_{A2})S_{v1}(t)E_{n1}(t) \\ &- [(b_1 I_{S2}(t) + b_2 I_{A2}(t))S_2(t) + (\beta_{E1} + \beta_{A1})S_2(t)E_{n2}(t)] \\ &- [(b_1^{'} I_{S2}(t) + b'_2 I_{A2}(t))S_{v2}(t) + (\beta_{E2} + \beta_{A2})S_{v2}(t)E_{n2}(t)]|, \end{aligned}$$

then

$$\begin{aligned}|B(Y_1) - B(Y_2)| \leq 2|&b_1 I_{S1}(t)(S_1(t) - S_2(t)) + b_1 S_2(t)(I_{S1}(t) - I_{S2}(t))\\ &+b_2 I_{A1}(t)(S_1(t) - S_2(t)) + b_2 S_2(t)(I_{A1}(t) - I_{A2}(t))\\ &+(\beta_{E1} + \beta_{A1}) S_1(t)(E_{n1}(t) - E_{n2}(t))\\ &+(\beta_{E1} + \beta_{A1}) E_{n2}(t)(S_1(t) - S_2(t))\\ &+b_1^{'} I_{S1}(t)(S_{v1}(t) - S_{v2}(t)) + b'_1 S_{v2}(t)(I_{S1}(t) - I_{S2}(t))\\ &+b_2^{'} I_{A1}(t)(S_{v1}(t) - S_{v2}(t)) + b'_2 S_{v2}(t)(I_{A1}(t) - I_{A2}(t))\\ &+(\beta_{E2} + \beta_{A2}) S_{v1}(t)(E_{n1}(t) - E_{n2}(t))\\ &+(\beta_{E2} + \beta_{A2}) E_{n2}(t)(S_{v1}(t) - S_{v2}(t))|.\end{aligned}$$

We have $0 \leq N(t) \leq N(0) + Z^{'}$, where $Z^{'} = \dfrac{\Pi_1 + \theta_1 + \theta_2 + \theta_3 + \theta_4 + \theta_5}{\mu_1}$.
Then

$$\begin{aligned}|B(Y_1) - B(Y_2)| &\leq 2|(b_1 + b_2 + \beta_{E1} + \beta_{A1})(N(0) + Z^{'})(S_1(t) - S_2(t))\\ &\quad +(b_1 + b_1^{'})(N(0) + Z^{'})(I_{S1}(t) - I_{S2}(t))\\ &\quad +(b_2 + b_2^{'})(N(0) + Z^{'})(I_{A1}(t) - I_{A2}(t))\\ &\quad +(\beta_{E1} + \beta_{A1} + \beta_{E2} + \beta_{A2})(N(0) + Z^{'})(E_{n1}(t) - E_{n2}(t))\\ &\quad +(b_1^{'} + b_2^{'} + \beta_{E2} + \beta_{A2})(N(0) + Z^{'})(S_{v1}(t) - S_{v2}(t))|\\ &\leq K[|S_1(t) - S_2(t)| + |S_{v1}(t) - S_{v2}(t)| + |I_{S1}(t) - I_{S2}(t)|\\ &\quad +|I_{A1}(t) - I_{A2}(t)| + |E_{n1}(t) - E_{n2}(t)|],\end{aligned}$$

where

$$\begin{aligned}K = 2\max(&b_1 + b_2 + \beta_{E1} + \beta_{A1}, b_1^{'} + b_2^{'} + \beta_{E2} + \beta_{A2}, b_1 + b_1^{'}, b_2 + b_2^{'},\\ &\beta_{E1} + \beta_{A1} + \beta_{E2} + \beta_{A2}).\end{aligned}$$

Then $||\varphi(Y_1) - \varphi(Y_2)|| \leq W||Y_1 - Y_2||$, which $W = \max(K, ||A||) < \infty$.

Therefore the function φ is uniformly Lipschitz continuous, from the restriction on $S(t) \geqslant 0$, $S_v(t) \geqslant 0$, $E(t) \geqslant 0$, $I_A(t) \geqslant 0$, $I_S(t) \geqslant 0$, $H(t) \geqslant 0$, $R(t) \geqslant 0$ and $E_n(t) \geqslant 0$, we conclude the existence of the solution of the our system (1) [4, 22]. □

3 Optimal control problem

3.1 Problem formulation

Given the significant health risks posed by the spread of typhoid fever, particularly antimicrobial-resistant Salmonella Typhi, it is crucial to implement control strategies capable of limiting or

reducing its transmission. In this context, we have formulated an optimal control problem aimed at decreasing the incidence of symptomatic and asymptomatic infections, increasing recovery rates, and reducing the bacterial load in the environment. To achieve these objectives, six control measures are considered:

Control $\mathbf{u_1}$: Promotes awareness of hygiene practices, including the use of protective bibs, avoiding close contact with potentially contaminated objects, proper cleaning of supplies and water, and correct food packaging. This control targets the reduction of infection rates among susceptible individuals, both vaccinated and unvaccinated, as well as those already affected.

Control $\mathbf{u_2}$: Aims to reduce Salmonella contamination in environmental water sources by treating wastewater before it is discharged into the environment.

Control $\mathbf{u_3}$: Focuses on food sterilization to lower contamination rates and bacterial load in food, through proper packaging and the enforcement of sanitary protocols among food-handling personnel.

Control $\mathbf{u_4}$: Encourages booster vaccination among previously vaccinated individuals, emphasizing the importance of revaccination every three years to decrease contamination rates via person-to-person transmission, food, or water.

Control $\mathbf{u_5}$: Involves symptom screening of infected individuals to differentiate typhoid cases from other diseases such as malaria, helminthiases, leptospirosis, or murine typhoid [18, 30]. The primary objective is to reduce the environmental contribution of asymptomatic carriers, who remain contagious long after symptoms subside [37].

Control $\mathbf{u_6}$: Represents treatment-related efforts, including the provision of free medication, ambulance transport, patient care, monitoring during treatment, and isolation of infected patients in hospitals. This control aims to minimize treatment failures among hospitalized and infected individuals.

By integrating these six controls into the typhoid fever model (1), the resulting controlled system can be formally defined as follows:

$$
\begin{cases}
\frac{dS(t)}{dt} &= \Pi_1 + a_2 S_v(t) + \delta R(t) - (\mu_1 + a_1 + u_4(t))S(t) \\
&\quad -(1-u_1(t))[b_1 I_S(t) + b_2 I_A(t) + (\beta_{E1} + \beta_{A1})E_n(t)]S(t), \\
\frac{dS_v(t)}{dt} &= \theta_1 - (a_2 + \mu_1)S_v(t) + (a_1 + u_4(t))S(t) \\
&\quad -(1-u_1(t))[b_1^{'} I_S(t) + b'_2 I_A(t) + (\beta_{E2} + \beta_{A2})E_n(t)]S_v(t), \\
\frac{dE(t)}{dt} &= \theta_2 - (\mu_1 + \lambda_1 + d_1 + d_2 + d_3)E(t) \\
&\quad +(1-u_1(t))[(b_1 I_S(t) + b_2 I_A(t))S(t) + (\beta_{E1} + \beta_{A1})S(t)E_n(t) \\
&\quad +(b_1^{'} I_S(t) + b'_2 I_A(t))S_v(t) + (\beta_{E2} + \beta_{A2})S_v(t)E_n(t)], \\
\frac{dI_S(t)}{dt} &= \theta_3 + d_1 E(t) + c_2 I_A(t) + (1-u_6(t))f_1 H(t) \\
&\quad -[c_1 + e_1 + \lambda_2 + \mu_1 + \mu_2 + u_5(t) + u_6(t)]I_S(t), \\
\frac{dI_A(t)}{dt} &= \theta_4 + d_2 E(t) + c_1 I_S(t) - [c_2 + e_2 + \lambda_3 + \mu_1 + \mu_3]I_A(t) \\
&\quad +(1-u_6(t))f_2 H(t), \\
\frac{dH(t)}{dt} &= \theta_5 + (e_1 + u_5(t) + u_6(t))I_S(t) + e_2 I_A(t) \\
&\quad -[g + \lambda_4 + \mu_1 + \mu_2 + (1-u_6(t))(f_1 + f_2)]H(t), \\
\frac{dR(t)}{dt} &= gH(t) + d_3 E(t) - (\mu_1 + \delta)R(t), \\
\frac{dE_n(t)}{dt} &= [-\sigma_E - \sigma_A - (1+u_2(t))\mu_E - (1+u_3(t))\mu_A]E_n(t) \\
&\quad +(\xi_{E1} + \xi_{A1})(I_A(t) + I_S(t)) + (\xi_{E2} + \xi_{A2})H(t),
\end{cases}
\tag{2}
$$

where $S_v^0 \geqslant 0$, $S^0 \geqslant 0$, $E^0 \geqslant 0$, $E_n^0 \geqslant 0$, $I_A^0 \geqslant 0$, $I_S^0 \geqslant 0$, $H^0 \geqslant 0$, and $R^0 \geqslant 0$ are the given initial states.

Our aim is to minimize the number of carriers, both symptomatic and asymptomatic, while maintaining our environment on sanitation without bacteria. For that we search to minimize the following objective function:

$$
J(u) = E(T_e) + I_A(T_e) + I_S(T_e) + \int_0^{T_e} [E(t) + I_A(t) + I_S(t) + \sum_{i=1}^{6} \frac{A_i}{2} u_i^2(t)]dt. \tag{3}
$$

Here, A_i for $i = 1, 2, \ldots, 6$ represent the cost coefficients, which are chosen to evaluate the relative significance of $u = (u_1, u_2, \ldots, u_6)$ at time t, where T_e denotes the final time. In other words, our objective is to find the optimal control $u^* = (u_1^*, \ldots, u_6^*)$ such that

$$
J(u^*) = \min_{u \in U_{ad}} J(u),
$$

where U_{ad} is the set of admissible controls defined by

$$U_{ad} = \{u = (u_1, u_2, \ldots, u_6) : 0 \leq u_{i,\min} \leq u_i(t) \leq u_{i,\max} \leq 1 \text{ for } i \in \{1, 2, \ldots, 6\} \text{ and } t \in [0, T_e]\}.$$

3.2 Existence of an optimal control

Before studying the characterization of optimal control, we will use Fleming and Rishel [14, Corollary 4.1] to confirm its existence.

Theorem 4. Let us contemplate the optimal control problem stated in equations (2)–(3). It is established that there exists an optimal control $u^* \in U_{ad}$ for which

$$J(u^*) = \min_{u \in U_{ad}} J(u).$$

Proof. The existence of an optimal control $u^* \in U_{ad}$ can be established using in [14, Corollary 4.1]. To apply this result, we verify the following conditions:

1. The set of admissible controls and the corresponding state variables is nonempty.
2. The control set U_{ad} is convex and closed.
3. The right-hand side of the state system is bounded by a linear function of the state and control variables.
4. The integrand $L(E, I_A, I_S, u)$ of the objective functional is convex on U_{ad}, and there exist constants $\Upsilon_1, \Upsilon_2 \geq 0$ and $\rho \geq 1$ such that

$$L(E, I_A, I_S, u) \geq -\Upsilon_1 + \Upsilon_2 \left(|u_1|^2 + \cdots + |u_6|^2\right)^{\frac{\rho}{2}}.$$

Step 1: Non-emptiness of the control and state set. Using Theorem 7.1.1 of Boyce and DiPrima [8], consider $X = (S, S_v, \ldots, E_n)$ and the system

$$\dot{S} = G_S(t, X), \quad \dot{S}_v = G_{S_v}(t, X), \quad \ldots, \quad \dot{E}_n = G_{E_n}(t, X),$$

where $G_S, G_{S_v}, \ldots, G_{E_n}$ are the right-hand sides of system (2). Taking $u_i(t) = c_i$ for constants c_i, and noting that all parameters are constant and X is continuous, it follows that $G_S, \ldots, G_{E_n}$ are continuous along with their partial derivatives. By the Cauchy-Lipschitz Theorem, there exists a unique solution $X(t)$ satisfying the initial conditions. Therefore, the set of admissible controls and corresponding states is nonempty.

Step 2: Convexity and closure of the control set. By the definition, we have

$$0 \leq \Gamma v_1 + (1-\Gamma)v_2 \leq 1,$$

showing that U_{ad} is convex.

Step 3: Boundedness of the right-hand side.

For the total population $N(t)$ in system (2), as proved in Theorem 1, we have

$$0 \leq N(t) \leq N(0) + \frac{Z}{\mu_1}.$$

Writing the system in vector form $Y = [S, S_v, \ldots, E_n]^T$, we have

$$\varphi(Y) = AY + B(Y),$$

with

$$A = \begin{pmatrix} -q_1' & a_2 & 0 & 0 & 0 & 0 & \delta & 0 \\ q_2' & -q_2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -q_3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & d_1 & -q_4' & c_2 & q_5' & 0 & 0 \\ 0 & 0 & d_2 & c_1 & -q_5 & q_7' & 0 & 0 \\ 0 & 0 & 0 & q_8' & e_2 & q_9' & 0 & 0 \\ 0 & 0 & d_3 & 0 & 0 & g & -q_7 & 0 \\ 0 & 0 & 0 & q_8 & q_8 & q_9 & 0 & -q_{10}' \end{pmatrix},$$

where

$$\begin{aligned} q_1^{'} &= a_1 + \mu_1 + u_4(t) & q_2^{'} &= a_1 + u_4(t), \\ q_4^{'} &= c_1 + e_1 + \lambda_2 + \mu_1 + \mu_2 + u_5(t) + u_6(t), & q_2 &= a_2 + \mu_1, \\ q_3 &= \mu_1 + \lambda_1 + d_1 + d_2 + d_3, & q_5 &= c_2 + e_2 + \lambda_3 + \mu_1 + \mu_3, \\ q_5^{'} &= (1 - u_6(t))f_1, & q_7^{'} &= (1 - u_6(t))f_2, \\ q_7 &= \mu_1 + \delta, & q_8 &= \xi_{E1} + \xi_{A1}, \\ q_9^{'} &= -(g + \lambda_4 + \mu_1 + \mu_2 + (1 - u_6(t))(f_1 + f_2)), & q_9 &= \xi_{E2} + \xi_{A2}, \\ q_{10}^{'} &= \sigma_E + \sigma_A + (1 + u_2(t))\mu_E + (1 + u_3(t))\mu_A, & q_8^{'} &= e_1 + u_5(t) + u_6(t). \end{aligned}$$

and

$$B(Y) = \left[\Psi_1^{'}\ \Psi_2^{'}\ \Psi_3^{'}\ \theta_3\ \theta_4\ \theta_5\ 0\ 0\right]^t,$$

where

$$\begin{aligned}
\Psi_1^{'} &= \Pi_1 - (1-u_1(t))[(b_1 I_S(t) + b_2 I_A(t))S(t) + (\beta_{E1} + \beta_{A1})S(t)E_n(t)],\\
\Psi_2^{'} &= \theta_1 - (1-u_1(t))[(b_1^{'} I_S(t) + b'_2 I_A(t))S_v(t) + (\beta_{E2} + \beta_{A2})S_v(t)E_n(t),\\
\Psi_3^{'} &= \theta_2 + (1-u_1(t))[(b_1 I_S(t) + b_2 I_A(t))S(t) + (\beta_{E1} + \beta_{A1})S(t)E_n(t)]\\
&\quad +(1-u_1(t))[(b_1^{'} I_S(t) + b'_2 I_A(t))S_v(t) + (\beta_{E2} + \beta_{A2})S_v(t)E_n(t)].
\end{aligned}$$

Next, for any Y_1, Y_2, we estimate

$$|B(Y_1) - B(Y_2)| \leq K \sum |Y_1 - Y_2|,$$

where

$$K = 2|1-u_1(t)| \max\left(b_1 + b_2 + \beta_{E1} + \beta_{A1}, b'_1 + b'_2 + \beta_{E2} + \beta_{A2}, b_1 + b'_1, b_2 + b'_2, \beta_{E1} + \beta_{A1} + \beta_{E2} + \beta_{A2}\right).$$

Hence, φ is uniformly Lipschitz continuous and Condition 3 is satisfied.

Step 4: Convexity and growth of the integrand. The integrand of the objective functional is

$$L(E, I_S, I_A, u) = E(t) + I_A(t) + I_S(t) + \sum_{i=1}^{6} \frac{A_i}{2} u_i^2(t),$$

which is convex on U_{ad}. Since $A_i > 0$ and E, I_A, I_S are bounded, one can choose

$$\Upsilon_1 = 3 \inf_{t\in[0,T_e]} (E + I_S + I_A), \quad \Upsilon_2 = \inf_i \frac{A_i}{2}, \quad \rho = 2,$$

ensuring

$$L(E, I_S, I_A, u) \geq -\Upsilon_1 + \Upsilon_2 \sum_{i=1}^{6} |u_i|^2.$$

By steps 1–4 and Fleming and Rishel [14], there exists an optimal control u^* such that

$$J(u^*) = \min_{u\in U_{ad}} J(u).$$

□

3.3 Characterization of optimal controls

In this section, we apply Pontryagin's Maximum Principle [35] to derive the necessary conditions for optimality. To this end, we construct the Hamiltonian function, denoted by $\mathcal{H}$, which links the objective functional to the controlled dynamical system. This framework provides a rigorous characterization of the optimal controls.

The Hamiltonian function is defined at time t, by

$$\mathcal{H}(t) = E(t) + I_A(t) + I_S(t) + \sum_{i=1}^{6} \frac{A_i}{2} u_i^2(t) + \sum_{i=1}^{8} \phi_i g_i(X),$$

where g_i is the right side of the difference equation of the ith state variable and

$$X = (S, S_v, \ldots, E_n).$$

Theorem 5. Given the optimal control $u^* = (u_1^*, u_2^*, \ldots, u_6^*)$ and the solution S^*, S_v^*, E^*, I_a^*, I_s^*, H^*, R^* and E_n^*, of the corresponding state system (2), there exists adjoins variables ϕ_1, ϕ_2, ..., ϕ_8 satisfying

$$\begin{aligned}
\dot{\phi_1} &= -\phi_1[-\mu_1 - a_1 - u_4(t) - (1-u_1(t))(b_1 I_S(t) \\
&\quad + b_2 I_A(t) + (\beta_{E1} + \beta_{A1}) E_n(t))] \\
&\quad - \phi_2[a_1 + u_4(t)] - \phi_3[(1-u_1(t))(b_1 I_S(t) + b_2 I_A(t) + (\beta_{E1} + \beta_{A1}) E_n(t))] \\
&\quad + \phi_8(\beta_{E1} + \beta_{A1}), \\
\dot{\phi_2} &= -\phi_1[a_2] \\
&\quad - \phi_2[-a_2 - \mu_1 - (1-u_1(t))(b_1' I_S(t) + b_2' I_A(t) + (\beta_{E2} + \beta_{A2}) E_n(t))] \\
&\quad - \phi_3[(1-u_1(t))(b_1' I_S(t) + b_2' I_A(t) + (\beta_{E2} + \beta_{A2}) E_n(t))] - \phi_8[\beta_{E2} + \beta_{A2}], \\
\dot{\phi_3} &= -1 + \phi_3[\mu_1 + \lambda_1 + d_1 + d_2 + d_3] - \phi_4[d_1] - \phi_5[d_2] - \phi_7 d_3, \\
\dot{\phi_4} &= -1 + \phi_1[(1-u_1(t)) b_1 S(t)] + \phi_2[(1-u_1(t)) b_1' S_v(t)] - \phi_8[\xi_{E1} + \xi_{A1}] \\
&\quad - \phi_3[(1-u_1(t))(b_1 S(t) + b_1' S_v(t))] - \phi_5[c_1] - \phi_6[e_1 + u_5(t) + u_6(t)] \\
&\quad - \phi_4[-c_1 - e_1 - \lambda_2 - \mu_1 - \mu_2 - u_5(t) - u_6(t)],
\end{aligned}$$

$$
\begin{aligned}
\dot{\phi_5} &= -1 + \phi_1[(1-u_1(t))b_2S(t)] + \phi_2[(1-u_1(t))b_2^{'}S_v(t)] \\
&\quad -\phi_3[(1-u_1(t))(b_2S(t)+b_2^{'}S_v(t))] - \phi_4[c_2] \\
&\quad -\phi_5[-c_2-e_2-\lambda_3-\mu_1-\mu_3] - \phi_6[e_2] - \phi_8[\xi_{E1}+\xi_{A1}], \\
\dot{\phi_6} &= -\phi_4[(1-u_6(t))f_1] - \phi_5[(1-u_6(t))f_2] \\
&\quad -\phi_6[-g-\lambda_4-\mu_1-\mu_2-(1-u_6(t))(f_1+f_2)] - \phi_7[g] - \phi_8[\xi_{E2}+\xi_{A2}], \\
\dot{\phi_7} &= -\phi_1[\delta] - \phi_7[-\mu_1+\delta], \\
\dot{\phi_8} &= \phi_1[(1-u_1(t))(\beta_{E1}+\beta_{A1})S(t)] + \phi_2[(1-u_1(t))(\beta_{E2}+\beta_{A2})S_v(t)] \\
&\quad -\phi_3[(1-u_1(t))((\beta_{E1}+\beta_{A1})S(t)+(\beta_{E2}+\beta_{A2})S_v(t))] \\
&\quad -\phi_8[-\sigma_A-\sigma_E-(1+u_2(t))\mu_E-(1+u_3(t))\mu_A],
\end{aligned}
$$

with the transversal conditions at finally time T_e,

$$
\begin{aligned}
&\phi_1(T_e)=0, \quad \phi_3(T_e)=-1, \quad \phi_5(T_e)=-1, \quad \phi_7(T_e)=0, \\
&\phi_2(T_e)=0, \quad \phi_4(T_e)=-1, \quad \phi_6(T_e)=0, \quad \phi_8(T_e)=0.
\end{aligned}
$$

Furthermore, for $t \in [0, T_e]$, the optimal control $u^* = (u_1^*, u_2^*, \ldots, u_6^*)$ is given by

$$u_1^* = \min(1, \max(0, \frac{1}{A_1}[-\phi_1\Delta_1 - \phi_2\Delta_2 + \phi_3\Delta_3])), \tag{4}$$

$$u_2^* = \min(1, \max(0, \frac{1}{A_2}[\phi_8\mu_E E_n(t)])), \tag{5}$$

$$u_3^* = \min(1, \max(0, \frac{1}{A_3}[\phi_8\mu_A E_n(t)])), \tag{6}$$

$$u_4^* = \min(1, \max(0, \frac{1}{A_4}[\phi_1+\phi_2]S(t))), \tag{7}$$

$$u_5^* = \min(1, \max(0, \frac{1}{A_5}[\phi_4-\phi_6]I_S(t))), \tag{8}$$

$$u_6^* = \min(1, \max(0, \frac{1}{A_6}[\phi_4(I_S(t)+f_1H(t)) + \phi_5 f_2 H(t) - \phi_6((f_1+f_2)H(t)+I_S(t))])), \tag{9}$$

where

$$\Delta_1 = (b_1I_S(t) + b_2I_A(t))S(t) + (\beta_{E1}+\beta_{A1})S(t)E_n(t),$$

$\Delta_2 = (b_1^{'}I_S(t) + b'_2I_A(t))S_v(t) + (\beta_{E2}+\beta_{A2})S_v(t)E_n(t)$, and $\Delta_3 = \Delta_1 + \Delta_2$.

Proof. The Hamiltonian H_m is defined as follows:

$$\mathcal{H}(t) = [E(t) + I_A(t) + I_S(t) + \sum_{i=1}^{6} \frac{A_i}{2} u_i^2(t) + \sum_{i=1}^{8} \phi_i g_i(X),$$

where g_i is the right side of the difference equation of the ith state variable and

$$X = (S, S_v, \ldots, E_n).$$

For $t \in [0, T_e]$, we can obtain the adjoint equations and transversality conditions by using Pontryagin's maximum principle [35] such that

$$\begin{aligned}
\dot{\phi_1} = -\frac{\partial \mathcal{H}}{\partial S} &= -\phi_1[-\mu_1 - a_1 - u_4(t) \\
&\quad - (1-u_1(t))(b_1 I_S(t) + b_2 I_A(t) + (\beta_{E1} + \beta_{A1}) E_n(t))] \\
&\quad - \phi_2[a_1 + u_4(t)] - \phi_3[(1-u_1(t))(b_1 I_S(t) a + b_2 I_A(t) \\
&\quad + (\beta_{E1} + \beta_{A1}) E_n(t))] + \phi_8(\beta_{E1} + \beta_{A1}), \\
\dot{\phi_2} = -\frac{\partial \mathcal{H}}{\partial S_v} &= -\phi_1[a_2] - \phi_2[-a_2 - \mu_1 - (1-u_1(t))(b_1^{'} I_S(t) + b_2^{'} I_A(t) \\
&\quad + (\beta_{E2} + \beta_{A2}) E_n(t))] - \phi_3[(1-u_1(t))(b_1^{'} I_S(t) + b_2^{'} I_A(t) \\
&\quad + (\beta_{E2} + \beta_{A2}) E_n(t))] - \phi_8[\beta_{E2} + \beta_{A2}], \\
\dot{\phi_3} = -\frac{\partial \mathcal{H}}{\partial E} &= -1 + \phi_3[\mu_1 + \lambda_1 + d_1 + d_2 + d_3] - \phi_4[d_1] - \phi_5[d_2] - \phi_7 d_3, \\
\dot{\phi_4} = -\frac{\partial \mathcal{H}}{\partial I_S} &= -1 + \phi_1[(1-u_1(t)) b_1 S(t)] + \phi_2[(1-u_1(t)) b_1^{'} S_v(t)] \\
&\quad - \phi_8[\xi_{E1} + \xi_{A1}] - \phi_3[(1-u_1(t))(b_1 S(t) + b_1^{'} S_v(t))] - \phi_5[c_1] \\
&\quad - \phi_6[e_1 + u_5(t) + u_6(t)] \\
&\quad - \phi_4[-c_1 - e_1 - \lambda_2 - \mu_1 - \mu_2 - u_5(t) - u_6(t)], \\
\dot{\phi_5} = -\frac{\partial \mathcal{H}}{\partial I_A} &= -1 + \phi_1[(1-u_1(t)) b_2 S(t)] + \phi_2[(1-u_1(t)) b_2^{'} S_v(t)] \\
&\quad - \phi_4[c_2] - \phi_6[e_2] - \phi_3[(1-u_1(t))(b_2 S(t) + b_2^{'} S_v(t))] \\
&\quad - \phi_5[-c_2 - e_2 - \lambda_3 - \mu_1 - \mu_3] - \phi_8[\xi_{E1} + \xi_{A1}], \\
\dot{\phi_6} = -\frac{\partial \mathcal{H}}{\partial H} &= -\phi_4[(1-u_6(t)) f_1] - \phi_5[(1-u_6(t)) f_2] - \phi_7[g] - \phi_8[\xi_{E2} + \xi_{A2}] \\
&\quad - \phi_6[-g - \lambda_4 - \mu_1 - \mu_2 - (1-u_6(t))(f_1 + f_2)], \\
\dot{\phi_7} = -\frac{\partial \mathcal{H}}{\partial R} &= -\phi_1[\delta] - \phi_7[-\mu_1 + \delta],
\end{aligned}$$

$$\begin{aligned}\dot{\phi_8} = -\frac{\partial \mathcal{H}}{\partial E_n} &= \phi_1[(1-u_1(t))(\beta_{E1}+\beta_{A1})S(t)] \\ &+ \phi_2[(1-u_1(t))(\beta_{E2}+\beta_{A2})S_v(t)] \\ &- \phi_3[(1-u_1(t))((\beta_{E1}+\beta_{A1})S(t)+(\beta_{E2}+\beta_{A2})S_v(t))] \\ &- \phi_8[-\sigma_A-\sigma_E-(1+u_2(t))\mu_E-(1+u_3(t))\mu_A].\end{aligned}$$

So

$$\begin{aligned}\dot{\phi_1} &= -\frac{\partial \mathcal{H}}{\partial S}, \phi_1(T_e)=0, & \dot{\phi_5} &= -\frac{\partial \mathcal{H}}{\partial I_A}, \phi_5(T_e)=-1, \\ \dot{\phi_2} &= -\frac{\partial \mathcal{H}}{\partial S_v}, \phi_2(T_e)=0, & \dot{\phi_6} &= -\frac{\partial \mathcal{H}}{\partial H}, \phi_6(T_e)=0, \\ \dot{\phi_3} &= -\frac{\partial \mathcal{H}}{\partial E}, \phi_3(T_e)=-1, & \dot{\phi_7} &= -\frac{\partial \mathcal{H}}{\partial R}, \phi_7(T_e)=0, \\ \dot{\phi_4} &= -\frac{\partial \mathcal{H}}{\partial I_S}, \phi_4(T_e)=-1, & \dot{\phi_8} &= -\frac{\partial \mathcal{H}}{\partial E_n}, \phi_8(T_e)=0.\end{aligned}$$

For $t \in [0, T_e]$, the optimal control u^* can be solved from the optimality conditions,

$$\frac{\partial \mathcal{H}(t)}{\partial u_1(t)} = 0, \frac{\partial \mathcal{H}(t)}{\partial u_2(t)} = 0, \ \ldots, \ \frac{\partial \mathcal{H}(t)}{\partial u_6(t)} = 0.$$

With further simplification of the optimality condition and by taking account the bounds in U_{ad} of the controls, we can easily obtain $u_1^*, \ldots, u_6^*$ in the form (4)–(9). □

4 Numerical simulations and cost-effectiveness analysis

4.1 Numerical simulations

To illustrate the effectiveness of the various controls previously proposed to reduce the impact of typhoid fever in Taiwan, we present in this section the numerical simulation using MATLAB.

The optimality system of our control problem is solved iteratively by first solving the forward state system under the following initial conditions: $S^0 = 8{,}050{,}000$, $S_v^0 = 2{,}300{,}000$, $E^0 = 4{,}600{,}000$, $I_A^0 = 3{,}450{,}000$, $I_S^0 = 2{,}300{,}000$, $H^0 = 1{,}150{,}000$, $R^0 = 1{,}150{,}000$, $E_n^0 = 1{,}000$, then solving the backward adjoint system with the final conditions. Both state and adjoint systems are discretized using the explicit Euler method.

In contrast to approaches that monitor convergence via a tolerance criterion, our implementation uses a fixed number of iterations K, which is sufficient to ensure stable and accurate approximations of the states and controls. This avoids unnecessary computational overhead while guaranteeing the reliability of the results.

The numerical procedure can be summarized by the following pseudo-code:

Algorithm 1 Iterative scheme for solving the optimal control problem

```
Initialize parameters and initial state values.
Initialize all control variables u_i(t).
for k = 1 to K do
    Solve the state equations forward in time using the current controls (Euler method).
    Solve the adjoint equations backward in time with terminal conditions (Euler method).
    Update the controls u_i^{(k+1)}(t) using the explicit expressions from the Hamiltonian.
end for
```

Due to the limited availability of complete epidemiological data for typhoid fever in Taiwan, not all parameter values could be directly obtained from observational data. Where possible, we used reported values from published studies conducted in Taiwan. For parameters lacking local data, we selected values from the literature on similar settings or estimated reasonable ranges based on epidemiological knowledge and prior modeling studies. This approach ensures that the model remains biologically and epidemiologically plausible while enabling meaningful exploration of control strategies.

In this simulation, we used data and statistics mostly related to the Taiwan region: $f_1 = f_2 = 0.01$ [7], $\xi_{A1} = \xi_{E1} = 0.78$, $\delta = 0.054$, $c_2 = 0.0635$, $e_1 = 0.375$, $e_2 = 0.064$, $d_1 = 0.3402$, $d_2 = 0.0378$, $b_1 = b_2 = 10^{-8}$, $a_1 = 0.054$, $a_2 = 0.0292$, $\beta_{A1} = \beta_{E1} = 10^{-9}$, $\mu_2 = \mu_3 = 0.079$, $\mu_E = \mu_A = 0.1085$, and $\mu_1 = 3.42 \times 10^{-5}$ [3]. We also provided approximate values for the other parameters: $\Pi_1 = 1{,}034{,}602$, $\theta_1 = \theta_4 = 79{,}670$, $\theta_2 = 159{,}340$, $\theta_3 = \theta_5 = 39{,}835$, $\lambda_i = 0.001868$ for $i = 1, \ldots, 4$, $\sigma_E = \sigma_A = 0.01$, $\beta_{A2} = \beta_{E2} = b_1' = b_2' = 25 \times 10^{-9}$, $d_3 = 0.1$, $c_1 = 0.008$, $g = 0.0512$, $\xi_{A2} = \xi_{E2} = 0.05$.

Note that graphs with control are represented by dotted lines and graphs without control are represented by solid lines in future.

4.2 Strategy A: Sterilization of the water and food

In this strategy, we investigate the evolution of typhoid fever under the implementation of controlled water treatment and food sterilization measures. The results presented in Figure 2 indicate a reduction of approximately 134,000 carriers and 91,000 symptomatic individuals, while the number of asymptomatic individuals remains nearly unchanged. The bacterial load in the environment decreases by nearly 2 million germs by the end of the second week. The

decline in carriers and symptomatic cases is most pronounced in the first week, highlighting the rapid effect of sterilization. However, the stability of asymptomatic individuals suggests that additional interventions, such as screening or vaccination, may be required to fully control transmission. Overall, this strategy effectively reduces environmental contamination, which indirectly contributes to lowering infection rates in humans.

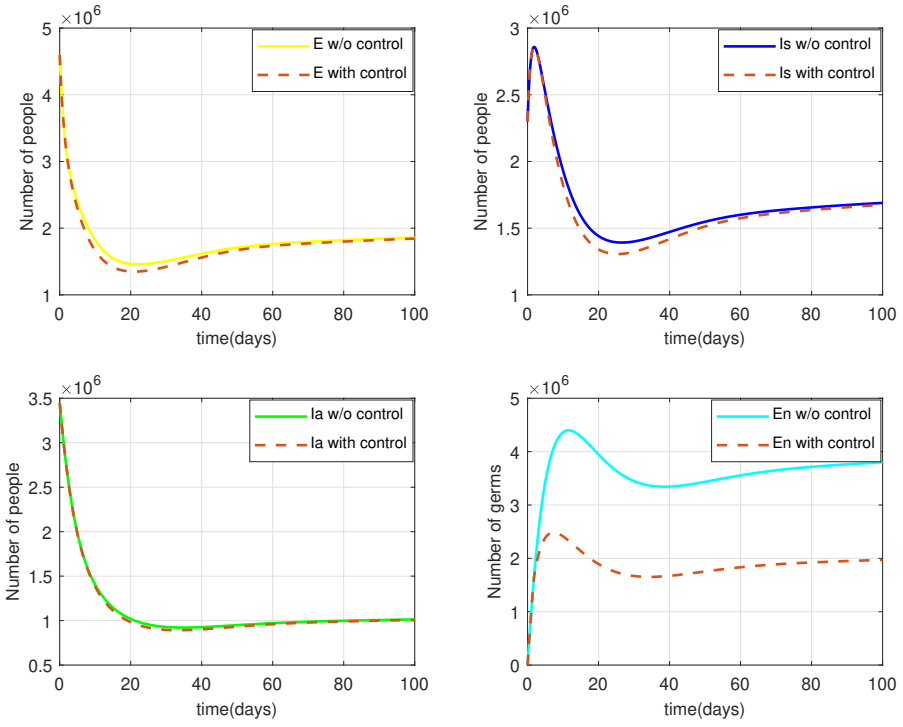


Figure 2: Illustration of the impact of sterilization of the water and food.

4.3 Strategy B: Screening and treatment

This strategy focuses on the impact of screening and treatment interventions (controls u_5 and u_6); see Figure 3. Starting from the first week, symptomatic infected individuals decrease significantly by approximately 2,308,000, while carriers decline by about 404,000. By the second week, approximately 333,300 carriers and 835,000 germs in the environment are reduced. These results indicate that active identification and treatment of infected individuals rapidly lowers

disease prevalence, particularly among symptomatic cases. Compared to Strategy A, this approach shows a more substantial and direct effect on reducing human infections, emphasizing the importance of timely screening and therapeutic management.

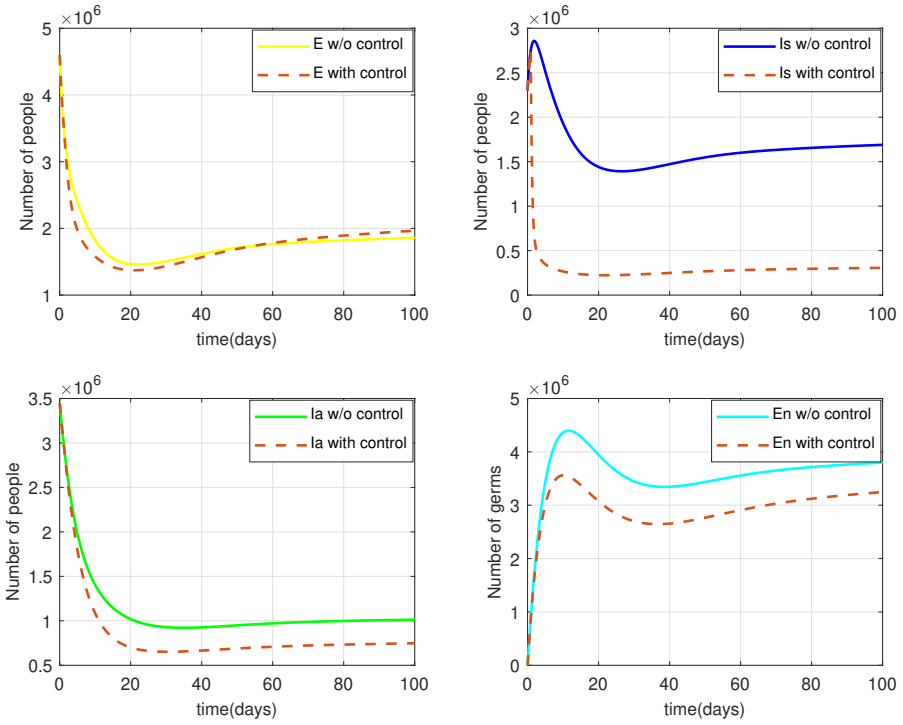


Figure 3: Illustration of the impact of treatment and symptom screening of infected.

4.4 Strategy C: Awareness campaign and vaccination program

In this strategy, we analyze the combined effect of awareness campaigns and vaccination programs; see Figure 4. During the first week, carriers decrease by approximately 1,531,800 individuals. In the second week, symptomatic cases are reduced by about 1,131,000, asymptomatic cases by 344.500, and environmental germs by 1,541,000. The results highlight that public education and vaccination contribute to a progressive reduction in both human infections and environmental contamination. The strategy shows a slower initial impact compared to direct treatment but

has a broader effect over multiple compartments, demonstrating the value of preventive measures in sustaining long-term control of typhoid fever.

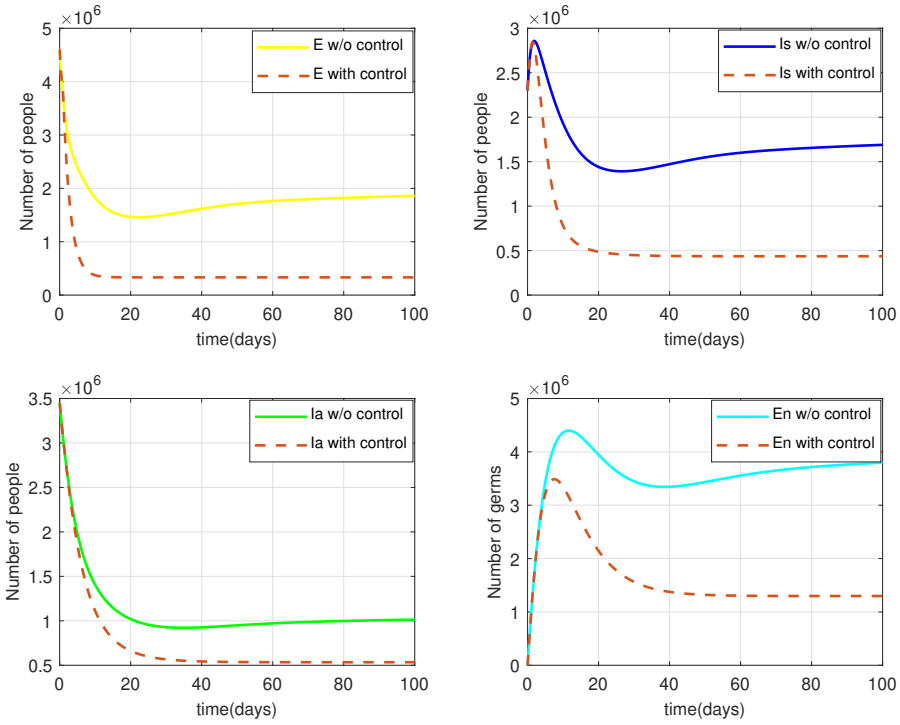


Figure 4: Illustration of the impact of awareness campaign with vaccination program.

4.5 Strategy D: Awareness program with sterilization of water and food

This strategy evaluates the combined effect of an awareness campaign and water/food sterilization; see Figure 5. The first week shows a reduction of approximately 1,626,400 carriers. By the second week, symptomatic cases decrease by about 1,183,400, asymptomatic cases by 329,500, and environmental germs by 2,289,000. Compared to previous strategies, this approach achieves substantial reductions across all compartments, demonstrating synergistic effects of preventive awareness and environmental control. The results suggest that this strategy is highly effective

in reducing both human infections and environmental contamination, making it a promising candidate for cost-effective intervention.

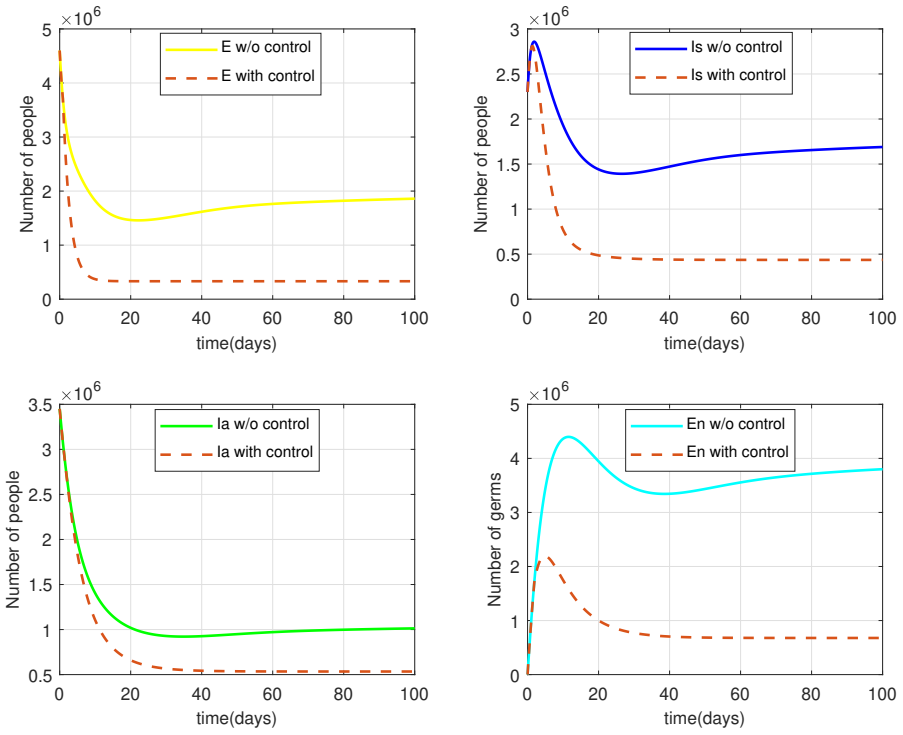


Figure 5: Illustration of the impact of awareness campaign with sterilization of the water and food.

4.6 Strategy E: Combination of all previous strategies

Here, we simultaneously implement all six controls to assess their collective impact; see Figure 6. From the first week, carriers decrease by approximately 1,578,800, symptomatic cases by 2,436,600, and environmental germs by 1,640,000. Asymptomatic individuals show a decline of about 520,500 from the second week. While this comprehensive strategy produces the greatest overall reduction in infections, the results indicate diminishing returns in certain compartments compared to Strategy D, particularly for carriers. These findings highlight that combining multiple interventions can maximize reductions but may not always proportionally improve efficiency. Strategy D remains highly effective while potentially being more cost-efficient.

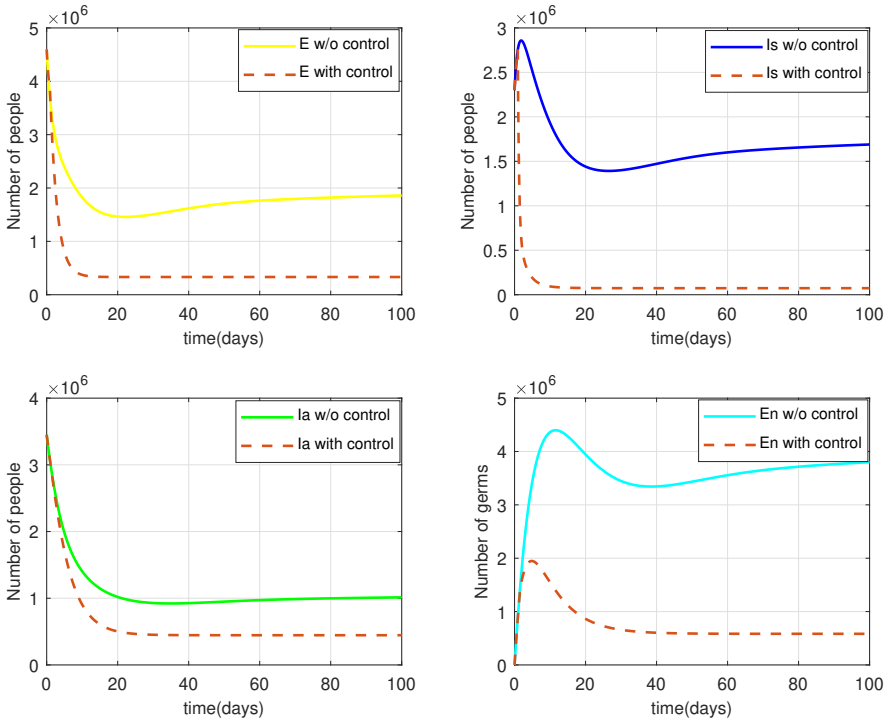


Figure 6: Illustration of the impact of combination of all strategies.

Remark 1. The six controls mentioned earlier can be used to create hundreds of different strategies. The selection of these five strategies was based not only on the outcomes of previous studies using these controls but also on practical criteria such as operational feasibility, estimated cost, and contextual relevance, ensuring that the analysis focuses on the most realistic and potentially effective interventions.

4.7 Cost-effectiveness analysis

In this section, we identify the most cost-effective strategy by comparing the cost-effectiveness ratios of the five intervention scenarios. To this end, we use the incremental cost-effectiveness ratio (ICER), a standard measure in health economics [15, 31, 32, 5].

The ICER is calculated as the ratio of the difference in costs to the difference in infections averted between two competing strategies i and j, where $i, j \in \{1, \ldots, 5\}$,

$$ICER(i,j) = \frac{C(i) - C(j)}{TA(i) - TA(j)},$$

with $C(\cdot)$ denoting the total cost and $TA(\cdot)$ the total infections averted. When comparing two strategies **M** and **N**, if $TA(N) > TA(M)$, then the ICER expresses the additional cost required to avert one extra infection by adopting **N** over **M**.

The total costs and infections averted for each strategy are summarized in Table 2.

Table 2: Total costs and infections averted for strategies S_1–S_5.

Strategy	Total averted infections (TA)	Total cost (TC)
$S_1 : u_2u_3$	3.6270×10^4	980
$S_2 : u_5u_6$	1.5420×10^6	989
$S_3 : u_1u_4$	3.2588×10^6	542.5536
$S_4 : u_1u_2u_3$	3.2650×10^6	1473.5
$S_5 : u_1u_2u_3u_4u_5u_6$	3.7103×10^6	2514.8

Step 1: S_1 **vs.** S_2

$$ICER(1) = \frac{TC(1)}{TA(1)} = 2.70 \times 10^{-2},$$

$$ICER(2) = \frac{TC(2) - TC(1)}{TA(2) - TA(1)} = 5.977 \times 10^{-6}.$$

Result: S_2 dominates S_1 (lower ICER, greater effectiveness), which implies that S_1 is excluded.

Step 2: S_2 **vs.** S_3

$$ICER(2) = \frac{TC(2)}{TA(2)} = 6.41 \times 10^{-4},$$

$$ICER(3) = \frac{TC(3) - TC(2)}{TA(3) - TA(2)} = -2.60 \times 10^{-4}.$$

Result: S_3 strongly dominates S_2 (negative ICER indicates lower cost with higher effectiveness), which implies that S_2 is excluded.

Step 3: S_3 **vs.** S_4

$$ICER(3) = \frac{TC(3)}{TA(3)} = 1.66 \times 10^{-4},$$

$$ICER(4) = \frac{TC(4) - TC(3)}{TA(4) - TA(3)} = 0.1501.$$

Result: S_3 is more cost-effective than S_4, since the additional cost of S_4 per infection averted is disproportionately high, which implies that S_4 is excluded.

Step 4: S_3 **vs.** S_5

$$ICER(3) = \frac{TC(3)}{TA(3)} = 1.66 \times 10^{-4},$$

$$ICER(5) = \frac{TC(5) - TC(3)}{TA(5) - TA(3)} = 4.36 \times 10^{-3}.$$

Result: Although S_5 averts the largest number of infections, its ICER is higher than that of S_3, which implies that S_5 is excluded.

Interpretation: Awareness campaigns combined with vaccination (S_3) offer the highest cost-effectiveness by simultaneously reducing transmission and increasing population immunity, effectively controlling typhoid fever.

5 Discussion and conclusion

5.1 Discussion

The inclusion of immigration in our model provides important insights into typhoid fever dynamics that are rarely captured in previous studies. Our simulations indicate that immigrants can substantially increase the number of exposed and infected individuals, as well as the environmental pathogen load, highlighting their potential role as a source of infection. Figures 7 and 8 illustrate these differences, emphasizing that immigration can amplify both the burden and the transmission potential of typhoid fever within the population.

Among the five control strategies proposed in this study, the combination of awareness campaigns with food and water sterilization is used here to illustrate the influence of immigration on disease dynamics. This example shows how such interventions can mitigate the effects of increased exposure and environmental contamination associated with population movement. These findings underline the importance of designing targeted preventive measures that address both resident and immigrant populations. Overall, incorporating immigration dynamics enhances the realism of the model and provides actionable guidance for public health planning.

5.2 Conclusion

In this study, we developed a deterministic mathematical model to minimize the spread of typhoid fever in Taiwan, explicitly incorporating the effects of immigrants from other endemic countries on local transmission dynamics, vaccine unreliability, assessment of government interventions, treatment of hospitalized patients, and contamination from population and hospital

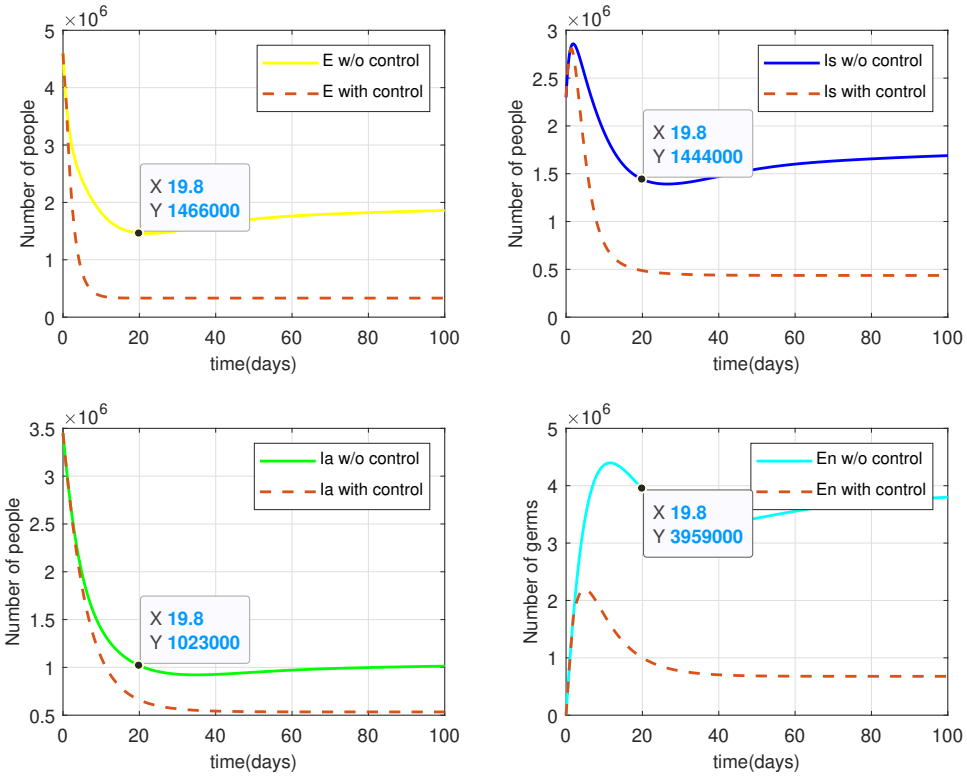


Figure 7: Projected dynamics of typhoid fever with immigration

waste. Importantly, we proposed six control measures: Increasing awareness, sterilizing food and water, promoting vaccinations, screening, and managing treatment.

Through the application of optimal control strategies, we successfully demonstrated a reduction in the disease burden, as evidenced by numerical simulations. Furthermore, using cost-effectiveness analysis, we identified that the most efficient strategy combines Awareness campaign and vaccination program.

Despite these insights, our model has inherent limitations, such as the assumption of homogeneous population mixing, constant parameter values, and the exclusion of additional environmental or spatial heterogeneities. Future research could address these limitations by integrating more detailed epidemiological data, exploring alternative combinations of control measures, considering spatial dynamics, or extending the modeling framework to other pathogens or regions. In particular, studying the spatial spread of typhoid fever may provide further insights into designing effective and targeted intervention strategies.

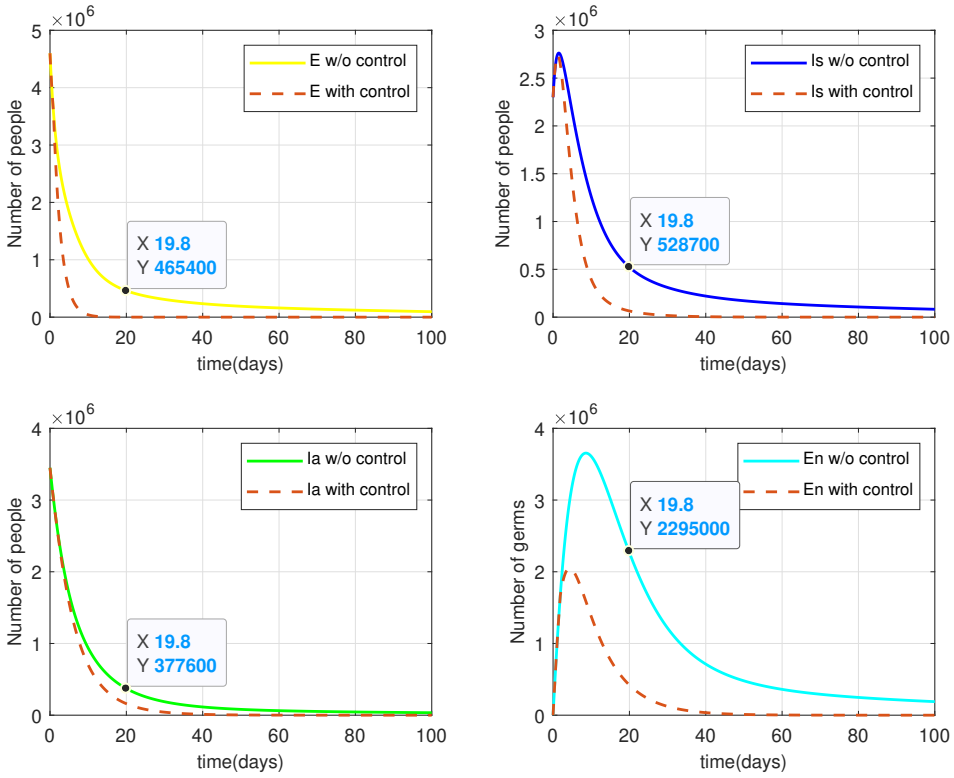


Figure 8: Projected dynamics of typhoid fever without immigration

Funding

This research received no external funding.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] Abboubakar, H., Kombou, L.K., Koko, A.D., Fouda, H.P.E., and Kumar, A. *Projections and fractional dynamics of the typhoid fever: A case study of Mbandjock in the Centre*

Region of Cameroon, Chaos, Soliton. Fract. 150, (2021) 111129.

[2] Abboubakar, H., and Racke, R. *Mathematical modeling, forecasting, and optimal control of typhoid fever transmission dynamics*, Chaos, Soliton. Fract. 149, (2021) 111074.

[3] Alharbi, M.H., Alalhareth, F.K., and Ibrahim, M.A. *Analyzing the Dynamics of a Periodic Typhoid Fever Transmission Model with Imperfect Vaccination*, Mathematics 11(15), (2023) 3298.

[4] Baig, U., Mehdi, S.M., and Iftikhar, N. *A pattern of antibiotic drug resistance of Salmonella Typhi and Salmonella Paratyphi among children with enteric fever in a tertiary care hospital in Lahore, Pakistan*, Croat. Med. J. 64(4), (2023) 256.

[5] Belhdid, S., Balatif, O., and Khajji, B. *Optimal control and cost-effectiveness analysis of scam rumor propagation over social networks*, Results Control Optim. 16, (2024) 100441.

[6] Bhutta, Z.A. *Current concepts in the diagnosis and treatment of typhoid fever*, BMJ 333(7558), (2006) 78–82.

[7] Birger, R., Antillón, M., Bilcke, J., Dolecek, C., Dougan, G., Pollard, A.J., Neuzil, K.M., Frost, I., Laxminarayan, R. and Pitzer, V.E. *Estimating the effect of vaccination on antimicrobial-resistant typhoid fever in 73 countries supported by Gavi: a mathematical modelling study*, Lancet Infect. Dis. 22(5), (2022) 679–691.

[8] Boyce, W.E., DiPrima, R.C., and Meade, D.B. *Elementary differential equations and boundary value problems*, John Wiley and Sons, 2021.

[9] Bronze, M.S., and Greenfield, R.A. *Biodefense: Principles and Pathogens*, Horizon Bioscience, Wymondham, Norfolk, England, 2005.

[10] Centers for Disease Control and Prevention. *Typhoid Fever and Paratyphoid Fever*, Available: https://www.cdc.gov/typhoid-fever/index.html, Accessed: 01-06-2023.

[11] Davarian, N., Heydari, A., Mehne, S.H.H. and Heydari, A. *The investigation of stability and optimal control of the multigroup Covid-19 epidemic model in Iran*, J. Math. Ext. 19, (2025).

[12] Encyclopedia of Mathematics. *Gronwall lemma*, Available: https://encyclopediaofmath.org/wiki/Gronwall_lemma, Accessed: 05-01-2024.

[13] Farrar, J., Hotez, P.J., Junghanss, T., Kang, G., Lalloo, D., White, N.J. and Garcia, P.J. (Eds.). *Manson's Tropical Diseases E-Book*, Elsevier Health Sciences, 2023.

[14] Fleming, W.H., and Rishel, R.W. *Deterministic and Stochastic Optimal Control*, Springer Science and Business Media, 2012.

[15] Folashade, B., and ELmojtaba Ibrahim, M. *Optimal control and cost-effective analysis of malaria/visceral leishmaniasis co-infection*, PLoS One 12(2), (2017) e0171102.

[16] González-Guzmán, J. *An epidemiological model for direct and indirect transmission of typhoid fever*, Math. Biosci. 96(1), (1989) 33–46.

[17] Government of Canada. *Salmonella enterica*, Available: https://www.canada.ca/fr/sante-publique/services/biosecurite-biosurete-laboratoire/fiches-techniques-sante-securite-agents-pathogenes\-evaluation\-risques/salmonella-enterica.html#note4, Accessed: 27-10-2023.

[18] Institut Pasteur. *Leptospirosis*, Available: https://www.pasteur.fr/fr/centre-medical/fiches-maladies/leptospirose, Accessed: 31-10-2023.

[19] Institut Pasteur. *Salmonellosis*, Available: https://www.pasteur.fr/fr/centre-medical/fiches-maladies/salmonellose, Accessed: 27-10-2023.

[20] Karloff, H. (n.d.). *Convex Functions and Applications*, University of South Carolina. Available: https://people.math.sc.edu/howard/Classes/555c/convex.pdf, Accessed: 30-09-2025.

[21] Karunditu, J.W., Kimathi, G., and Osman, S. *Mathematical modeling of typhoid fever disease incorporating unprotected humans in the spread dynamics*, J. Adv. Math. Comput. Sci. 32(3), (2019) 1–11.

[22] Khassal, S., Moumine, E.M., and Balatif, O. *Mathematical modeling, analysis, and optimal control of the cochineal insect impact on cacti plants*, Iran. J. Numer. Anal. Optim. 15(2), (2025).

[23] Mirza, S.H., Beechmg, N.J., and Hart, C.A. *Multi-drug resistant typhoid: a global problem*, J. Med. Microbiol. 44(5), (1996) 317–319.

[24] Mogasale, V., Maskery, B., Ochiai, R.L., Lee, J.S., Mogasale, V.V., Ramani, E., Kim, Y.E., Park, J.K. and Wierzba, T.F. *Burden of typhoid fever in low-income and middle-income*

countries: a systematic, literature-based update with risk-factor adjustment, Lancet Global Health 2(10), (2014) e570–e580.

[25] Mondal, J. *Influence of awareness programs by media in the typhoid fever: a study based on mathematical modeling*, J. Math. Model. 6(1), (2018) 1–26.

[26] Musa, S.S., Zhao, S., Hussaini, N., Usaini, S. and He, D. *Dynamics analysis of typhoid fever with public health education programs and final epidemic size relation*, Results Appl. Math. 10, (2021) 100153.

[27] Mushayabasa, S. *A simple epidemiological model for typhoid with saturated incidence rate and treatment effect*, Int. J. Math. Comput. Sci. 6(6), (2013) 688–695.

[28] Mushayabasa, S. *Modeling the impact of optimal screening on typhoid dynamics*, Int. J. Dyn. Control 4, (2016) 330–338.

[29] Mutua, J.M., Wang, F.B., and Vaidya, N.K. *Modeling malaria and typhoid fever co-infection dynamics*, Math. Biosci. 264, (2015) 128–144.

[30] Ogwa, M., Wenda, L., Lisimo, B., Lomboto, L., and Mokili, J.K.E.*Difficulté diagnostique et prise en charge de la fièvre typhoïde chez les enfants en milieu rural en Rd Congo (cas de l'HGR de Basoko)*, Int. J. Innov. Sci. Res. 31(2), (2017) 287–294.

[31] Okosun, K.O., Makinde, O.D., and Takaidza, I. *Impact of optimal control on the treatment of HIV/AIDS and screening of unaware infectives*, Appl. Math. Model. 37(6), (2013) 3802–3820.

[32] Okosun, K.O., Ouifki, R., and Marcus, N. *Optimal control strategies and cost-effectiveness analysis of a malaria model*, BioSystems 111(2), (2013) 83–101.

[33] Peter, O.J., Ibrahim, M.O., Edogbanya, H.O., Oguntolu, F.A., Oshinubi, K., Ibrahim, A.A., Ayoola, T.A. and Lawal, J.O. *Direct and indirect transmission of typhoid fever model with optimal control*, Results Phys. 27, (2021) 104463.

[34] Piovani, D., Figlioli, G., Nikolopoulos, G.K. and Bonovas, S. *The global burden of enteric fever, 2017–2021: a systematic analysis from the global burden of disease study 2021*, E. Clinical Medicine 77, (2024).

[35] Pontryagin. L.S. *Mathematical theory of optimal processes,* Routledge, 2018.

[36] Shaikh, A.S., and Nisar, K.S. *Transmission dynamics of fractional order Typhoid fever model using Caputo–Fabrizio operator*, Chaos, Soliton. Fract. 128, (2019) 355–365.

[37] Stanaway, J.D., Reiner, R.C., Blacker, B.F., Goldberg, E.M., Khalil, I.A., Troeger, C.E., Andrews, J.R., Bhutta, Z.A., Crump, J.A., Im, J. and Marks, F. *The global burden of typhoid and paratyphoid fevers: a systematic analysis for the Global Burden of Disease Study 2017*, Lancet Infect. Dis. 19(4), (2019) 369–381.

[38] Tilahun, G.T., Makinde, O.D., and Malonza, D. *Modelling and optimal control of typhoid fever disease with cost-effective strategies*, Comput. Math. Methods Med. 2017, (2017) 2324518.

[39] Tilahun, G.T., Makinde, O.D., and Malonza, D. *Co-dynamics of pneumonia and typhoid fever diseases with cost effective optimal control analysis*, Appl. Math. Comput. 316, (2018) 438–459.

[40] Watson, C.H., and Edmunds, W.J. *A review of typhoid fever transmission dynamic models and economic evaluations of vaccination*, Vaccine 33, (2015) C42–C54.

[41] World Health Organization. *Background document: the diagnosis, treatment and prevention of typhoid fever*, Background document: The diagnosis, treatment and prevention of typhoid fever (2003).

[42] World Health Organization. *Typhoid vaccines: WHO position paper–March 2018*, Wkly. Epidemiol. Rec. (2018) 153–172.

Aims and scope

Iranian Journal of Numerical Analysis and Optimization (IJNAO) is published by the Department of Applied Mathematics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad. Papers dealing with different aspects of numerical analysis and optimization, theories and their applications in engineering and industry are considered for publication.

Journal Policy

All submissions to IJNAO are first evaluated by the journal's Editor-in-Chief or one of the journal's Associate Editors for their appropriateness to the scope and objectives of IJNAO. If deemed appropriate, the paper is sent out for review using a single blind process. Manuscripts are reviewed simultaneously by reviewers who are experts in their respective fields. The first review of every manuscript is performed by at least two anonymous referees. Upon the receipt of the referee's reports, the paper is accepted, rejected, or sent back to the author(s) for revision. Revised papers are assigned to an Associate Editor who makes an evaluation of the acceptability of the revision. Based upon the Associate Editor's evaluation, the paper is accepted, rejected, or returned to the author(s) for another revision. The second revision is then evaluated by the Editor-in-Chief, possibly in consultation with the Associate Editor who handled the original paper and the first revision, for a usually final resolution.

The authors can track their submissions and the process of peer review via: http://ijnao.um.ac.ir

All manuscripts submitted to IJNAO are tracked by using "iThenticate" for possible plagiarism before acceptance.

Instruction for Authors

The Journal publishes all papers in the fields of numerical analysis and optimization. Articles must be written in English.
All submitted papers will be refereed and the authors may be asked to revise their manuscripts according to the referee's reports. The Editorial Board of the Journal keeps the right to accept or reject the papers for publication.
The papers with more than one authors, should determine the corresponding author. The e-mail address of the corresponding author must appear at the end of the manuscript or as a footnote of the first page.

It is strongly recommended to set up the manuscript by Latex or Tex, using the template provided in the web site of the Journal. Manuscripts should be typed double-spaced with wide margins to provide enough room for editorial remarks.

References should be arranged in alphabetical order by the surname of the first author as examples below:

[1] Brunner, H. *A survey of recent advances in the numerical treatment of Volterra integral and integro-differential equations*, J. Comput. Appl. Math. 8 (1982), 213-229.

[2] Stoer, J. and Bulirsch, R. *Introduction to Numerical Analysis*, Springer-verlag, New York, 2002.

Iranian Journal of Numerical Analysis and Optimization

CONTENTS Vol. 16, No. 1, pp 1-338

Unconditionally stable high-order scheme for the fractional generalized Burgers–Huxley equation 1
M. Hajipour, Y. Lakpour and D. Baleanu

A fractional order model for malaria control using optimal strategies and numerical simulations 19
A. Kouidere, A. El Bhih, I. Minifi and K. Adnaoui

Numerical exploration of pollutant transport using stochastic fractional diffusion and Karhunen–Loève expansion 38
Z. Moniri, A. Babaei and B. Parsa Moghaddam

A new numerical approach for tracing periodic solutions in nonlinear dynamical systems using CGM with PAC 61
K.F. Fazea

Fractional-order modeling and numerical analysis of double intermediate enzymatic reaction model using generalized Euler method 77
B. Koushick, N. Anbazhagan, R. Vignesh Raju, N. Jeeva

A novel three-term conjugate gradient approach for deep neural network training 102
F.S. Hosseini-Meighan and A. Ashrafi

Stability analysis of fast food consumption dynamics using a fractional-order framework 122
I. Sahib, M. Baroudi, H. Gourram, B. Khajji, A. Labzai and M. Belam

Measuring the efficiency of bi-level structured systems based on directional distance function and network DEA model 144
M. Habibpoor and M. Alirezaee

Qualitative and Numerical Analysis of a Time-Fractional Ebola Model in Population Dynamics 170
G. Saini, B. Ghosh and S. Chand

Transdermal drug delivery via a hollow microneedle considering the skin's mechanical properties: Analytical and numerical study 188
E. Azhdari and A. Emami

An improved search direction based on algebraic equivalent transformation technique for convex quadratic optimization 220
A. Kraria, B. Merikhi and D. Benterki

Quasilinearization method to solve several classes of nonlinear Lane–Emden equations by Gegenbauer polynomials 243
F. Sheikhi and B. Ghazanfari

An efficient hybrid three-term conjugate gradient method with practical applications 264
Y.E. Hemici, S. Khelladi and D. Benterki

Dynamics and applications of quantum iterative methods 283
S.K. Nayak, P.K. Parida and B. Mehta

Mathematical modeling and optimal control strategies of the typhoid transmission with cost-effectiveness analysis 304
K. Oubouskour, S. Khassal, B. Khajji and O. Balatif

web site: **https://ijnao.um.ac.ir**
Email: **ijnao@um.ac.ir**
ISSN-Print: **2423-6977**
ISSN-Online: **2423-6969**