Mathematical Modelling and Numerical Simulation with Applications

Editor-in-Chief
Mehmet Yavuz, PhD

MATHEMATICAL MODELLING AND NUMERICAL SIMULATION WITH APPLICATIONS

- **Kumar, Sunil** - National Institute of Technology, India.

- **Lupulescu, Vasile** - Constantin Brâncuşi University of Târgu-Jiu, Romania.

- **Merdan, Hüseyin** - TOBB University of Economy and Technology, Department of Mathematics, Türkiye.

- **Mohammed S. Abdo** - Hodeidah University, Al-Hodeidah, Department of Mathematics, Yemen.

- **Muñoz-Pacheco, Jesus Manuel** - Faculty of Electronics Sciences at the Autonomous University of Puebla (BUAP), Mexico.

- **Noeiaghdam, Samad** - Irkutsk National Research Technical University, Russian Federation.

- **Owolabi, Kolade** - Federal University of Technology, Nigeria.

- **Otero-Espinar, Maria Victoria** - University of Santiago de Compostela, Spain.

- **Panigoro, Hasan S.** - Universitas Negeri Gorontalo, Indonesia

- **Povstenko, Yuriy** - Jan Dlugosz University in Czestochowa, Poland.

- **Qureshi, Sania** - Mehran University of Engineering and Technology, Pakistan

- **Sabatier, Jocelyn** - Bordeaux University, France

- **Safaei, Mohammad Reza** - Florida International University, USA

- **Salahshour, Soheil** - Bahçeşehir University, Türkiye

- **Sarı, Murat** - Yıldız Technical University, Türkiye

- **Singh, Jagdev** - JECRC University, India.

- **Valdés, Juan Eduardo Nápoles** - Universidad Nacional del Nordeste, Argentina

- **Veeresha, Pundikala** - Christ University, India

- **Weber, Gerhard-Wilhelm** - Poznan University of Technology, Poland

- **Xu, Changjin** - Guizhou University of Finance and Economics, China

- **Yang, Xiao-Jun** - China University of Mining and Technology, China

- **Yuan, Sanling** - University of Shanghai for Science and Technology, China.

Contents

RESEARCH PAPER

Optical solitons of the complex Ginzburg–Landau equation having dual power nonlinear form using φ^6-model expansion approach

Muhammad Abubakar Isah [1,*,‡] and **Asif Yokus** [1,2,‡]

[1]Department of Mathematics, Faculty of Science, Firat University, 23200 Elazig, Türkiye, [2]Application and Research Center Advisory Board Member, Istanbul Commerce University, Istanbul, Türkiye
*Corresponding Author
‡myphysics_09@hotmail.com (Muhammad Abubakar Isah); asfyokus@firat.edu.tr (Asif Yokus)

Abstract

This paper employs a novel φ^6-model expansion approach to get dark, bright, periodic, dark-bright, and singular soliton solutions to the complex Ginzburg-Landau equation with dual power-law nonlinearity. The dual-power law found in photovoltaic materials is used to explain nonlinearity in the refractive index. The results of this paper may assist in comprehending some of the physical effects of various nonlinear physics models. For example, the hyperbolic sine arises in the calculation of the Roche limit and the gravitational potential of a cylinder, the hyperbolic tangent arises in the calculation of the magnetic moment and the rapidity of special relativity, and the hyperbolic cotangent arises in the Langevin function for magnetic polarization. Frequency values, one of the soliton's internal dynamics, are used to examine the behavior of the traveling wave. Finally, some of the obtained solitons' three-, two-dimensional, and contour graphs are plotted.

Keywords: φ^6-model expansion method; complex Ginzburg–Landau equation; soliton solutions; dual power-law nonlinearity
AMS 2020 Classification: 35C07; 35C08; 35G20; 35C08

1 Introduction

The study of surfaces [1, 2] in geometry [3, 4] and a wide range of mechanical problems were the first implementations of partial differential equations. In the nineteenth century, eminent mathematicians from all over the world showed a significant interest in researching a variety of problems arising from partial differential equations [5]. Optical solitons have emerged as a key study issue in the physical and natural sciences. Solitons have been discovered to play an important role in several disciplines of research, including optical fibers [6, 7], plasma physics

[8], nonlinear optics, and many others [9, 10]. Solitons may spread across trans-continental and trans-oceanic distances in fiber optics. Solitons are solutions to the nonlinear partial differential equations that describe a single moving wave. Optical solitons, which are formed as the result of a perfect balance between dispersion (or diffraction) and nonlinearity in a nonlinear medium, are frequently used in telecommunications and electromagnetics. Optical fibers are solutions to the Nonlinear Schrodinger equations. Soliton solutions contain particle-like structures, such as magnetic monopoles, as well as extended structures, such as domain walls and cosmic strings, which have implications in the cosmology of the early universe. Periodic solutions, such as $cos(x + t)$, are periodic traveling wave solutions [11, 12]. Dark soliton refers to a solitary wave with lower intensity than the background, bright soliton refers to a solitary wave with higher peak intensity than the background, and singular soliton refers to a solitary wave with discontinuous derivatives; examples of such solitary waves include compactions, which have finite (compact) support and peakons, whose peaks have a discontinuous derivative, dark solitons are modeled by the $\tan h$ functions whilst bright solitons are modeled by the $\sec h$. Understanding the dynamics of solitons can lead to a better understanding of the physics of the phenomena in which they exist. As a result, a number of sophisticated mathematical techniques have been developed to generate soliton solutions for a wide range of physical models such as the Kadomtsev–Petviashvili equation [13], the Benjamin–Ono equation [14], the disturbance effect in intracellular calcium dynamic on fibroblast cells [15], the Fisher equation [16], the nonlinear Schrödinger equation [17, 18], the Sharma–Tasso–Olver equation [19], the Murnaghan model [20], the Kaup-Kupershmidt equation [21], Navier–Stokes equation [22], the Zakharov–Kuznetsov equation [23], the B-type Kadomtsev–Petviashvili–Boussinesq equation [24] and others [25–27]. Recent analytical methods for solving PDEs, such as the eMETEM method [28], the generalized exponential rational function method [29], the extended sinh-Gordon equation expansion method [30], the q-homotopy analysis transform technique [31], the new extended direct algebraic method [32], the direct method [33], the Kudryashov's new function method [34], the split-step Fourier transform [35], the new modified unified auxiliary equation method [36], the $\left(\frac{1}{G'}\right)$-expansion method [37–39], the Jacobi elliptic functions [40].

The Ginzburg-Landau equation GLE is one of the most prominent partial differential equations in mathematics and physics, it was brought into the study of superconducting phenomenology theory by Ginzburg and Landau in the twentieth century. The GLE is commonly used to describe the propagation of optical solitons across optical fibers over extended distances. As a result, it is critical to examine the dynamic behavior of the GLE. Many researchers have recently solved the Complex Ginzburg-Landau equation CGLE with dual power law nonlinearity, among them; Arshed [41] solved this equation with the help of The exp(- $\phi(\xi)$)-expansion method and received different forms of solitons such as hyperbolic, rational and trigonometric functions. Jacobi's elliptic function expansion method is used to obtain some dark and periodic soliton solutions by Abdou et al. [42]. Al-Ghafri and Khalil [43] used The relation between the Weierstrass elliptic function and hyperbolic functions to derive optical soliton and period waves, in [44, 45], the trial solutions approach, exp(- $\phi(\xi)$)-expansion method and $\left(\frac{G'}{G^2}\right)$-expansion method are used, the $\left(\frac{G'}{G}\right)$ −method is used to secure the soliton solutions by Li Zhao, et al. [46], the other methods includes GPRE method [47].

The main objective of this study is to develop new solitons for the CGLE with dual power nonlinearity using the recently developed φ^6−model expansion method [18, 48], in which, to our knowledge, ıt has not been studied yet using the proposed technique. These new solitons include dark, bright, singular, rational, combined periodic, combined singular and periodic solitary wave

solutions.

The following is the outline for this paper; the mathematical analysis of the model will be given in Section 2. In Section 3, we will present the description of the φ^6–model expansion method. In Section 4, the φ^6–model expansion method will be applied to the CGLE model with dual power nonlinearity to get new travelling wave solutions to the model. Additionally, the physical structure of the traveling wave solution is graphically displayed in the related 2D, 3D and contour graphs. The physical dynamics of the soliton solutions are explored in Section 5, while the conclusions will be drawn in Section 6.

2 Mathematical analysis of the model

One of the most well-known partial differential equations in mathematics and physics, the Ginzburg-Landau equation was developed in the 20th century by Ginzburg and Landau and used to examine the superconducting phenomenology hypothesis. The propagation of optical solitons over optical fibers over long distances is frequently described using the GLE. The authors [41, 47] give the dimensionless shape of CGLE that will be investigated in this article as

$$iQ_t + aQ_{xx} + bF(|Q|^2)Q = \frac{1}{|Q|^2 Q^\star}\left[\alpha |Q|^2 (|Q|^2)_{xx} - \beta \left\{(|Q|^2)_x\right\}^2\right] + \gamma Q, \tag{1}$$

where x is the non-dimensional distance along the fibers and t is time in dimensionless form; a, b, α, β and γ are valued constants. The coefficients a and b are determined by the group velocity dispersion (GVD) and nonlinearity respectively. The terms with α, β and γ result from perturbation effects, specifically detuning.

The F in Eq.(1) is a real-valued algebraic function that must be smooth. $F(|Q|^2)Q$ is continuously differentiable k times, implying that

$$F(|Q|^2)Q \in \cup_{m,n=1}^{\infty} C^k\left((-n,n) \times (-m,m)\,;R^2\right), \tag{2}$$

by setting

$$\alpha = 2\beta. \tag{3}$$

Eq. (1) turns to

$$iQ_t + aQ_{xx} + bF(|Q|^2)Q = \frac{\beta}{|Q|^2 Q^\star}\left[2|Q|^2 (|Q|^2)_{xx} - \left\{(|Q|^2)_x\right\}^2\right] + \gamma Q. \tag{4}$$

To solve Eq. (1), the standard decomposition into phase-amplitude components:

$$Q(x,t) = U(\zeta)e^{i(-kx+\omega t+\theta)}, \tag{5}$$

and the wave variable ζ is given by

$$\zeta = (x - vt), \tag{6}$$

the function U represents the pulse shape and v is the soliton's velocity. In the phase factor, k denotes the soliton frequency, ω the soliton wave number and the phase constant θ. Substituting the amplitude-phase decomposition into Eq. (4) results in the following couple of equations after

breaking into real and imaginary parts:

$$- \left(ak^2 + \gamma + \omega \right) U + bF \left(U^2 \right) U + (a - 4\beta) U^{''} = 0, \tag{7}$$

and

$$v = -2ka. \tag{8}$$

In the following part, Eq. (7) will be examined by using dual power law nonlinearity.

3   Description of the proposed technique

According to Zayed et al. [27, 41] the following are the key steps of a recent φ^6-model expansion method:

Step-1: Consider the following nonlinear evolution equation for $Q = Q(x, t)$.

$$G(Q, Q_x, Q_t, Q_{xx}, Q_{xt}, Q_{tt}, ...) = 0, \tag{9}$$

here G is a polynomial of $Q(x, t)$ and its highest order partial derivatives, including its nonlinear terms.

Step-2: Making use of the wave transformation

$$Q(x, t) = Q(\zeta), \quad \zeta = x - vt, \tag{10}$$

where v represents wave velocity and Eq. (9) can be converted into the nonlinear ordinary differential equation shown below

$$\Omega(Q, Q', QQ', Q'', ...) = 0, \tag{11}$$

where the derivatives with respect to ζ are represented by the prime. **Step-3**: Suppose that the formal solution to Eq. (11) exists:

$$Q(\zeta) = \sum_{i=0}^{2N} \alpha_i U^i(\zeta), \tag{12}$$

where $\alpha_i (i = 0, 1, 2, \ldots, N)$ are to be determined constants, N can be obtained using the balancing rule and $U(\zeta)$ satisfies the auxiliary NLODE;

$$U'^2(\zeta) = h_0 + h_2 U^2(\zeta) + h_4 U^4(\zeta) + h_6 U^6(\zeta), \tag{13}$$
$$U^{''}(\zeta) = h_2 U(\zeta) + 2h_4 U^3(\zeta) + 3h_6 U^5(\zeta),$$

where $h_i (i = 0, 2, 4, 6)$ are real constants that will be discovered later.

Step-4: It is well known that the solution to Eq. (13) is as follows;

$$U(\zeta) = \frac{P(\zeta)}{\sqrt{fP^2(\zeta) + g}}, \tag{14}$$

provided that $0 < fP^2(\zeta) + g$ and $P(\zeta)$ is the Jacobi elliptic equation solution

$$P'^2(\zeta) = l_0 + l_2 P^2(\zeta) + l_4 P^4(\zeta), \tag{15}$$

where $l_i(i = 0, 2, 4)$ are unknown constants to be determined, f and g are given by

$$f = \frac{h_4(l_2 - h_2)}{(l_2 - h_2)^2 + 3l_0 l_4 - 2l_2(l_2 - h_2)},$$

$$g = \frac{3l_0 h_4}{(l_2 - h_2)^2 + 3l_0 l_4 - 2l_2(l_2 - h_2)}, \tag{16}$$

under the restriction condition

$$h_4^2(l_2 - h_2)[9l_0 l_4 - (l_2 - h_2)(2l_2 + h_2)] + 3h_6[-l_2^2 + h_2^2 + 3l_0 l_4]^2 = 0. \tag{17}$$

Step-5: According to [18, 25–27, 41, 48] it is well known that the Jacobi elliptic solutions of Eq. (15) can be calculated when $0 < m < 1$. We can have the exact solutions of Eq. (9) by substituting Eqs. (14) and (15) into Eq. (12).

Function	$m \to 1$	$m \to 0$	Function	$m \to 1$	$m \to 0$
$sn(\zeta, m)$	$tanh(\zeta)$	$sin(\zeta)$	$ds(\zeta, m)$	$csch(\zeta)$	$csc(\zeta)$
$cn(\zeta, m)$	$sech(\zeta)$	$cos(\zeta)$	$sc(\zeta, m)$	$sinh(\zeta)$	$tan(\zeta)$
$dn(\zeta, m)$	$sech(\zeta)$	1	$sd(\zeta, m)$	$sinh(\zeta)$	$sin(\zeta)$
$ns(\zeta, m)$	$coth(\zeta)$	$csc(\zeta)$	$nc(\zeta, m)$	$cosh(\zeta)$	$sec(\zeta)$
$cs(\zeta, m)$	$csch(\zeta)$	$cot(\zeta)$	$cd(\zeta, m)$	1	$cos(\zeta)$

4 Application of the φ^6-model expansion method

The dual power law is found in photovoltaic materials and is used to explain non-linearity in the refractive index. The formula for this law is $F(U) = U^n + rU^{2n}$, where r is a constant. When $n = 1$, is the non-linearity of the parabolic law which is the subset of the dual-power law [41, 45, 49], when $r = 0$ and $n = 1$, the relationship becomes the Kerr law of non-linearity, which is formed from the fact that a light wave in an optical fiber undergoes nonlinear responses owing to non-harmonic electron mobility in the presence of an external electric field [41].
For this non-linearity, Eq. (7) is reduced to

$$-\left(ak^2 + \gamma + \omega\right)U + b\left(U^{2n+1} + rU^{4n+1}\right) + (a - 4\beta)U'' = 0, \tag{18}$$

where $N = \frac{1}{2n}$ is obtained by balancing the higher order derivative and the nonlinear term in the above equation. The following transformation is used to achieve closed-form solutions

$$U = p^{\frac{1}{2n}},$$

that will reduce Eq. (18) into the ODE

$$-4n^2\left(ak^2 + \gamma + \omega\right)p^2 + 4bn^2 p^2\left(p + rp^2\right) + (a - 4\beta)\left(2npp'' + (1 - 2n)\left(p'\right)^2\right) = 0, \tag{19}$$

by balancing the highest order derivative and the highest nonlinear term in the above equation

$4N = 2N + 2$ we get $N = 1$, we obtain the following by substituting $N = 1$ in Eq. (12)

$$p(\zeta) = \alpha_0 + \alpha_1 U(\zeta) + \alpha_2 U^2(\zeta), \tag{20}$$

where α_0, α_1 and α_2 are constants to be determined.

We obtain the following algebraic equations by substituting Eq. (20) along with Eq. (13) into Eq. (19) and setting the coefficients of all powers of $U^i(\zeta)$, $i = 0, 1, \ldots, 8$ to zero

$U^0(\zeta)$; $-4n^2 \left(ak^2 + \gamma + \omega \right) \alpha_0^2 + 4bn^2\alpha_0^3 (1 + r\alpha_0)$

$\qquad - (-1 + 2n)(a - 4\beta) h_0\alpha_1^2 + 4n(a - 4\beta) h_0\alpha_0\alpha_2 = 0,$

$U^1(\zeta)$; $-2\alpha_1 \left(n\alpha_0 \left(4n \left(ak^2 + \gamma + \omega \right) - (a - 4\beta) h_2 - 2bn\alpha_0 (3 + 4r\alpha_0) \right) + 2(-1 + n)(a - 4\beta) h_0\alpha_2 \right) = 0,$

$U^2(\zeta)$; $\left((a - 4\beta) h_2 - 4n^2 \left(ak^2 + \gamma + \omega - 3b\alpha_0 (1 + 2r\alpha_0) \right) \right) \alpha_1^2$

$\qquad + 4\alpha_2 \left(n\alpha_0 \left(-2n \left(ak^2 + \gamma + \omega \right) + 2(a - 4\beta) h_2 + bn\alpha_0 (3 + 4r\alpha_0) \right) - (-1 + n)(a - 4\beta) h_0\alpha_2 \right) = 0,$

$U^3(\zeta)$; $2\alpha_1^3 \left(2n(a - 4\beta) h_4\alpha_0 + 2bn^2 (1 + 4r\alpha_0) \right)$

$\qquad + 2\alpha_1\alpha_2 \left((2 + n)(a - 4\beta) h_2 - 4n^2 \left(ak^2 + \gamma + \omega - 3b\alpha_0 (1 + 2r\alpha_0) \right) \right) = 0,$

$U^4(\zeta)$; $4bn^2r\alpha_1^4 + 12bn^2 (1 + 4r\alpha_0) \alpha_1^2\alpha_2 + (a - 4\beta) h_4 \left((1 + 2n) \alpha_1^2 + 12n\alpha_0\alpha_2 \right)$

$\qquad - 4 \left(-(a - 4\beta) h_2 + n^2 \left(ak^2 + \gamma + \omega - 3b\alpha_0 (1 + 2r\alpha_0) \right) \right) \alpha_2^2 = 0,$

$U^5(\zeta)$; $6\alpha_1 n(a - 4\beta) h_6\alpha_0 + 4\alpha_1\alpha_2 \left((1 + 2n)(a - 4\beta) h_4 + bn^2 \left(4r\alpha_1^2 + 3(1 + 4r\alpha_0) \alpha_2 \right) \right) = 0,$

$U^6(\zeta)$; $(a - 4\beta) h_6 \left((1 + 4n) \alpha_1^2 + 16n\alpha_0\alpha_2 \right) + 4\alpha_2^2 \left((1 + n)(a - 4\beta) h_4 + bn^2 (6r\alpha_1^2 + \alpha_2 + 4r\alpha_0\alpha_2) \right) = 0,$

$U^7(\zeta)$; $2\alpha_1\alpha_2 \left((2 + 7n)(a - 4\beta) h_6 + 8bn^2r\alpha_2^2 \right) = 0,$

$U^8(\zeta)$; $4 \left((1 + 2n)(a - 4\beta) h_6\alpha_2^2 + bn^2r\alpha_2^4 \right) = 0.$

We get the following result after solving the resulting system:

$$\alpha_0 = \frac{-1 - 2n}{4(1 + n)r}, \quad \alpha_1 = \frac{\sqrt{(1 + 2n) h_4}\sqrt{-a + 4\beta}}{2n\sqrt{br}}, \quad \alpha_2 = 0, \tag{21}$$

$$h_0 = \frac{b^2n^4 (1 + 2n)^2}{16r^2h_4 (1 + n)^4 (a - 4\beta)^2}, \quad h_2 = \frac{bn^2 (1 + 2n)}{2r(1 + n)^2 (a - 4\beta)},$$

$$h_6 = 0, \quad \omega = -1 \left(ak^2 + \gamma + \frac{b(1 + 2n)}{4r(1 + n)^2} \right).$$

In view of Eqs. (14), (20) and (21) along with the Jacobi elliptic functions in the table above, we obtain the following exact solutions of Eq. (18).

1. If $l_0 = 1$, $l_2 = -(1 + m^2)$, $l_4 = m^2$, $0 < m < 1$, then $P(\zeta) = sn(\zeta, m)$ or $P(\zeta) = cd(\zeta, m)$, we have

$$Q_1(x, t) = \left[\frac{-1 - 2n}{4(1 + n)r} + \frac{\sqrt{(1 + 2n)h_4}\sqrt{-a + 4\beta}}{2n\sqrt{br}} \left(\frac{sn(\zeta, m)}{\sqrt{f(sn(\zeta, m))^2 + g}} \right) \right]^{\frac{1}{2n}} e^{i(-kx + \omega t + \theta)}, \tag{22}$$

or

$$Q_{1,0}(x,t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)h_4}\sqrt{-a+4\beta}}{2n\sqrt{br}} \left(\frac{cd(\zeta,m)}{\sqrt{f(cd(\zeta,m))^2+g}} \right) \right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (23)$$

such that $0 < 2n\sqrt{br}(4(1+n)r)$, $\zeta = x - vt$ and f and g in Eq. (16) are given by

$$f = \frac{(1+m^2+h_2)h_4}{1-m^2+m^4-h_2^2}, \quad g = \frac{-3h_4}{1-m^2+m^4-h_2^2},$$

under the restriction condition

$$-h_4^2\left(-1-m^2-h_2\right)\left(-1+2m^2-h_2\right)\left(-2+m^2+h_2\right) = 0.$$

If $m \to 1$, then the soliton solution is obtained

$$Q_{1,1}(x,t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)h_4}\sqrt{-a+4\beta}}{2n\sqrt{br}} \left(\frac{\tanh(\zeta)}{\sqrt{\frac{-h_4\left(-3+(2+h_2)\tanh^2(\zeta)\right)}{-1+h_2^2}}} \right) \right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (24)$$

such that

$$h_4^2(-2-h_2)[-1+h_2]^2 = 0.$$

If $m \to 0$, then the periodic wave solution is obtained

$$Q_{1,2}(x,t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}} \left(\frac{\sin(\zeta)}{\sqrt{\frac{-\left(-3+(1+h_2)\sin^2(\zeta)\right)h_4}{-1+h_2^2}}} \right) \right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (25)$$

such that

$$h_4^2(-1-h_2)[(-2+h_2)(1+h_2)] = 0.$$

2. If $l_0 = 1-m^2$, $l_2 = 2m^2-1$, $l_4 = -m^2$, $0 < m < 1$, then $P(\zeta) = cn(\zeta,m)$, therefore

$$Q_2(x,t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}} \left(\frac{cn(\zeta,m)}{\sqrt{f(cn(\zeta,m))^2+g}} \right) \right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (26)$$
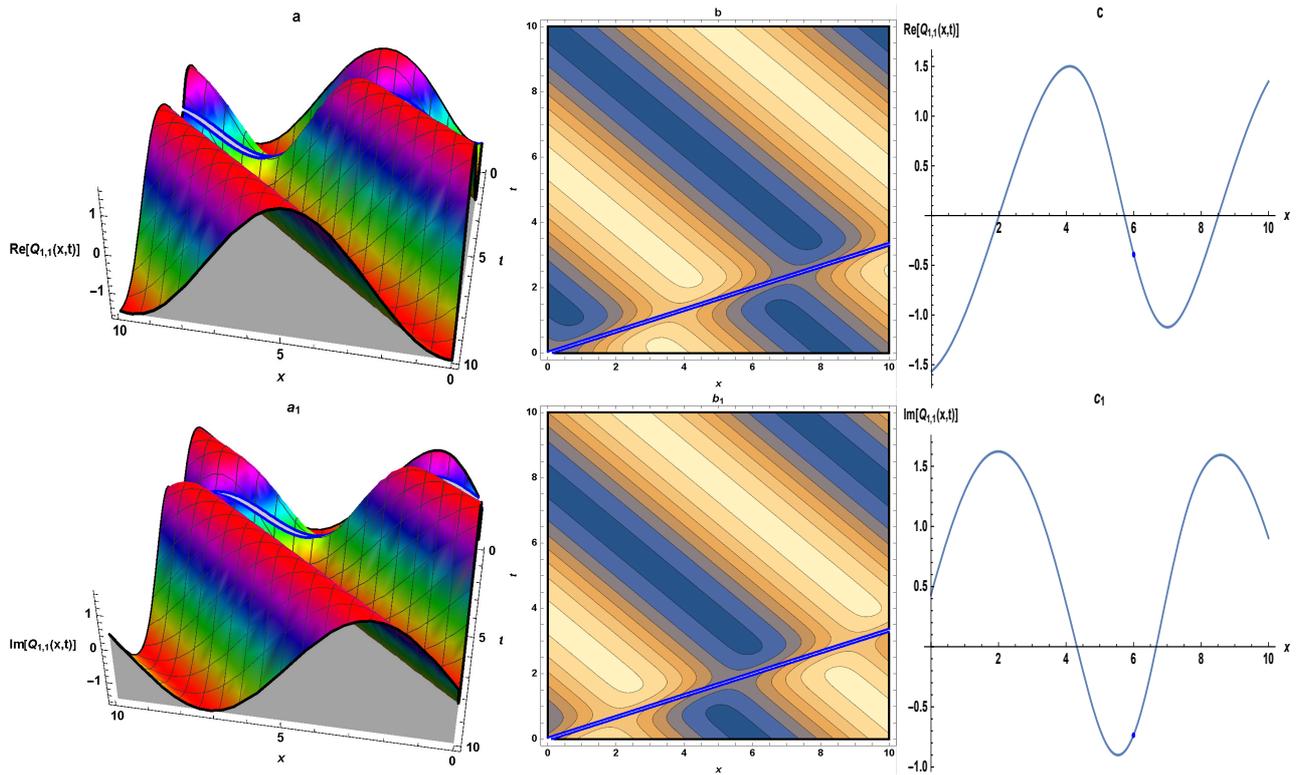
Figure 1. The 3Ds (a), (a_1), contours (b),(b_1) and 2Ds (c),(c_1) graphs of Eq. (24)

where f and g are determined by

$$f = -\frac{(-1 + 2m^2 - h_2)h_4}{1 - m^2 + m^4 - h_2^2}, \qquad g = \frac{3\left(-1 + m^2\right)h_4}{1 - m^2 + m^4 - h_2^2},$$

under the constraint condition

$$h_4^2\left(-1 + 2m^2 - h_2\right)\left[\left(-2 + m^2 + h_2\right)\left(1 + m^2 + h_2\right)\right] = 0.$$

If $m \to 1$, then the optical soliton is retrieved

$$Q_{2,1}\left(x,t\right) = \left[\frac{-1 - 2n}{4\left(1 + n\right)r} + \frac{\sqrt{\left(1 + 2n\right)\left(-a + 4\beta\right)h_4}}{2n\sqrt{br}}\left(\frac{sech(\zeta)}{\sqrt{\frac{-sech^2(\zeta)h_4}{1 + h_2}}}\right)\right]^{\frac{1}{2n}} e^{i(-kx + \omega t + \theta)}, \qquad (27)$$

provided that

$$h_4^2\left(1 - h_2\right)\left[h_2^2 + h_2 - 2\right] = 0.$$
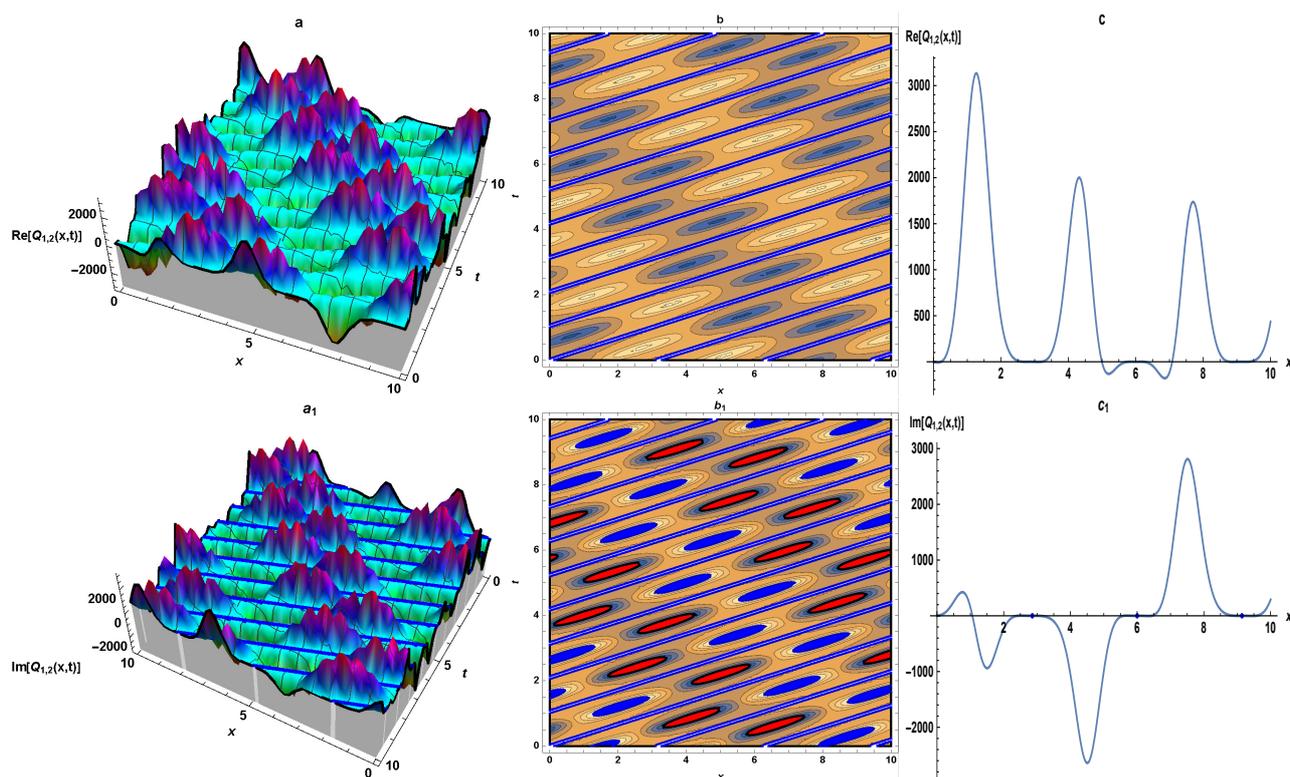
Figure 2. The 3Ds (a), (a_1), contours (b),(b_1) and 2Ds (c),(c_1) graphs of Eq. (25)

If $m \to 0$, then the periodic wave solution is obtained

$$Q_{2,2}(x,t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}} \left(\frac{\sin(\zeta)}{\sqrt{\frac{-\left(-3+(1+h_2)\sin^2(\zeta)\right)h_4}{-1+h_2^2}}} \right) \right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (28)$$

such that

$$h_4^2(-1-h_2)\left[(-2+h_2)(1+h_2)\right] = 0.$$

3. If $l_0 = m^2 - 1$, $l_2 = 2 - m^2$, $l_4 = -1$, $0 < m < 1$, then $P(\zeta) = dn(\zeta, m)$ which gives

$$Q_3(x,t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}} \left(\frac{dn(\zeta,m)}{\sqrt{f(dn(\zeta,m))^2+g}} \right) \right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (29)$$

where f and g are determined by

$$f = \frac{(-2+m^2+h_2)h_4}{1-m^2+m^4-h_2^2}, \qquad g = \frac{-3\left(-1+m^2\right)h_4}{1-m^2+m^4-h_2^2},$$
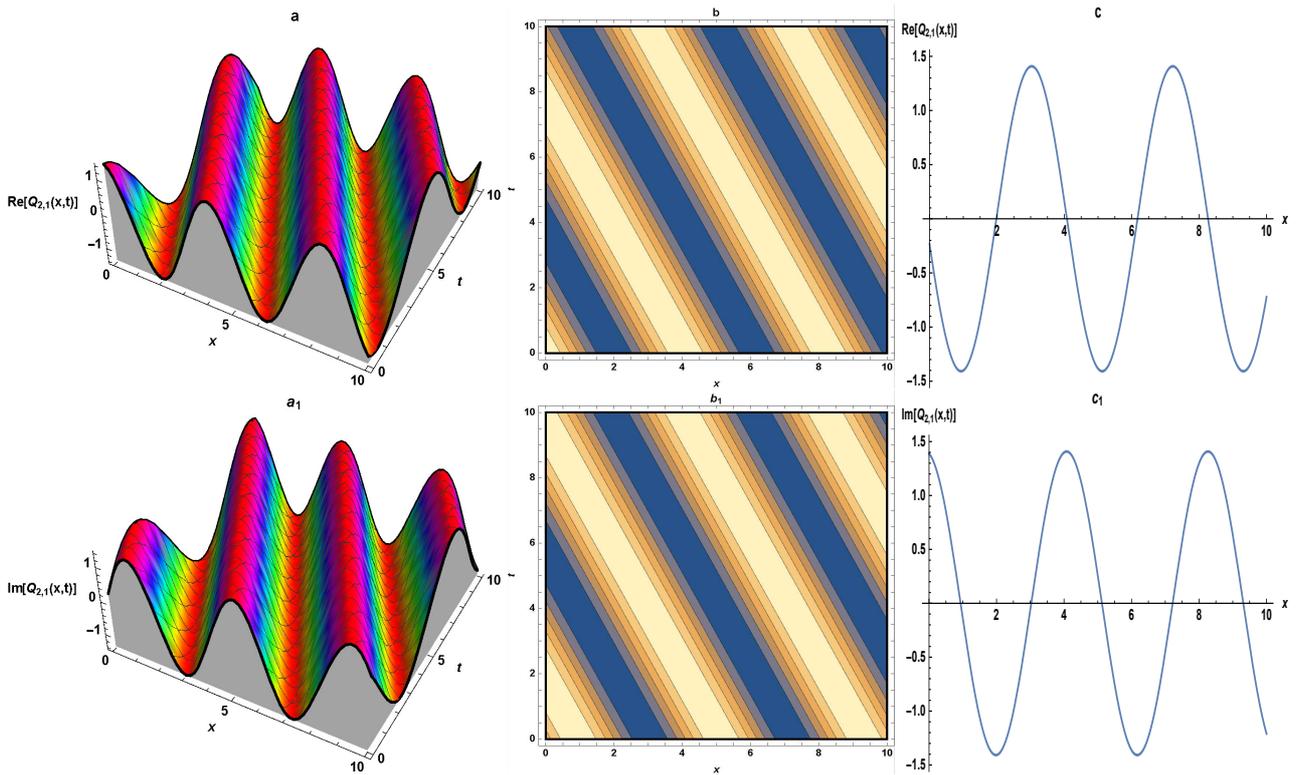
Figure 3. The 3Ds (a), (a_1), contours (b),(b_1) and 2Ds (c),(c_1) graphs of Eq. (27)

under the restriction condition

$$h_4^2 \left(2 - m^2 - h_2\right) \left[-\left(-1 + 2m^2 + h_2\right)\left(1 + m^2 + h_2\right)\right] = 0.$$

If $m \to 1$, then the soliton solution is obtained

$$Q_{3,1}\left(x,t\right) = \left[\frac{-1 - 2n}{4\left(1 + n\right)r} + \frac{\sqrt{\left(1 + 2n\right)\left(-a + 4\beta\right)h_4}}{2n\sqrt{br}}\left(\frac{sech(\zeta)}{\sqrt{\frac{-sech^2(\zeta)h_4}{1 + h_2}}}\right)\right]^{\frac{1}{2n}} e^{i\left(-kx + \omega t + \theta\right)}, \qquad (30)$$

provided that

$$h_4^2 \left(1 - h_2\right)\left[-2 + h_2 + h_2^2\right] = 0.$$

If $m \to 0$, then the rational solution is obtained

$$Q_{3,2}\left(x,t\right) = \left[\frac{-1 - 2n}{4\left(1 + n\right)r} + \frac{\sqrt{\left(1 + 2n\right)\left(-a + 4\beta\right)h_4}}{2n\sqrt{br}}\left(\frac{1}{\sqrt{\frac{-h_4}{-1 + h_2}}}\right)\right]^{\frac{1}{2n}} e^{i\left(-kx + \omega t + \theta\right)}, \qquad (31)$$

such that

$$h_4^2 \left(2 - h_2\right)\left[\left(1 + h_2\right)^2\right] = 0.$$
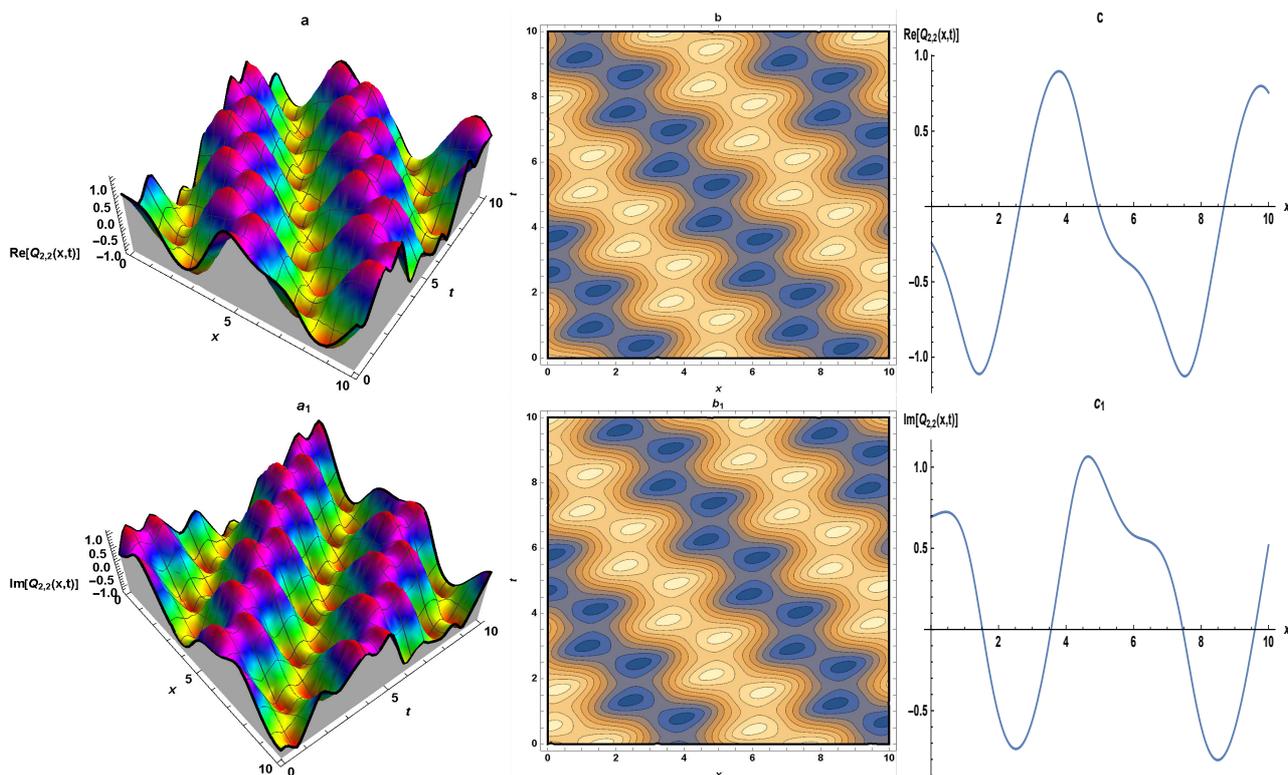
Figure 4. The 3Ds (a), (a_1), contours (b),(b_1) and 2Ds (c),(c_1) graphs of Eq. (28)

4. If $l_0 = m^2$, $l_2 = -(1 + m^2)$, $l_4 = 1$, $0 < m < 1$, then $P(\zeta) = ns(\zeta, m)$ or $P(\zeta) = dc(\zeta, m)$, then

$$Q_{4,1}(x,t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}} \left(\frac{ns(\zeta,m)}{\sqrt{f(ns(\zeta,m))^2+g}} \right) \right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (32)$$

or

$$Q_{4,2}(x,t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}} \left(\frac{dc(\zeta,m)}{\sqrt{f(dc(\zeta,m))^2+g}} \right) \right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (33)$$

where f and g are given by

$$f = \frac{(1+m^2+h_2)h_4}{1-m^2+m^4-h_2^2}, \qquad g = \frac{-3m^2h_4}{1-m^2+m^4-h_2^2},$$

under the constraint condition

$$h_4^2 \left(-1 - m^2 - h_2 \right) \left[-\left(-1 + 2m^2 - h_2 \right) \left(-2 + m^2 + h_2 \right) \right] = 0.$$
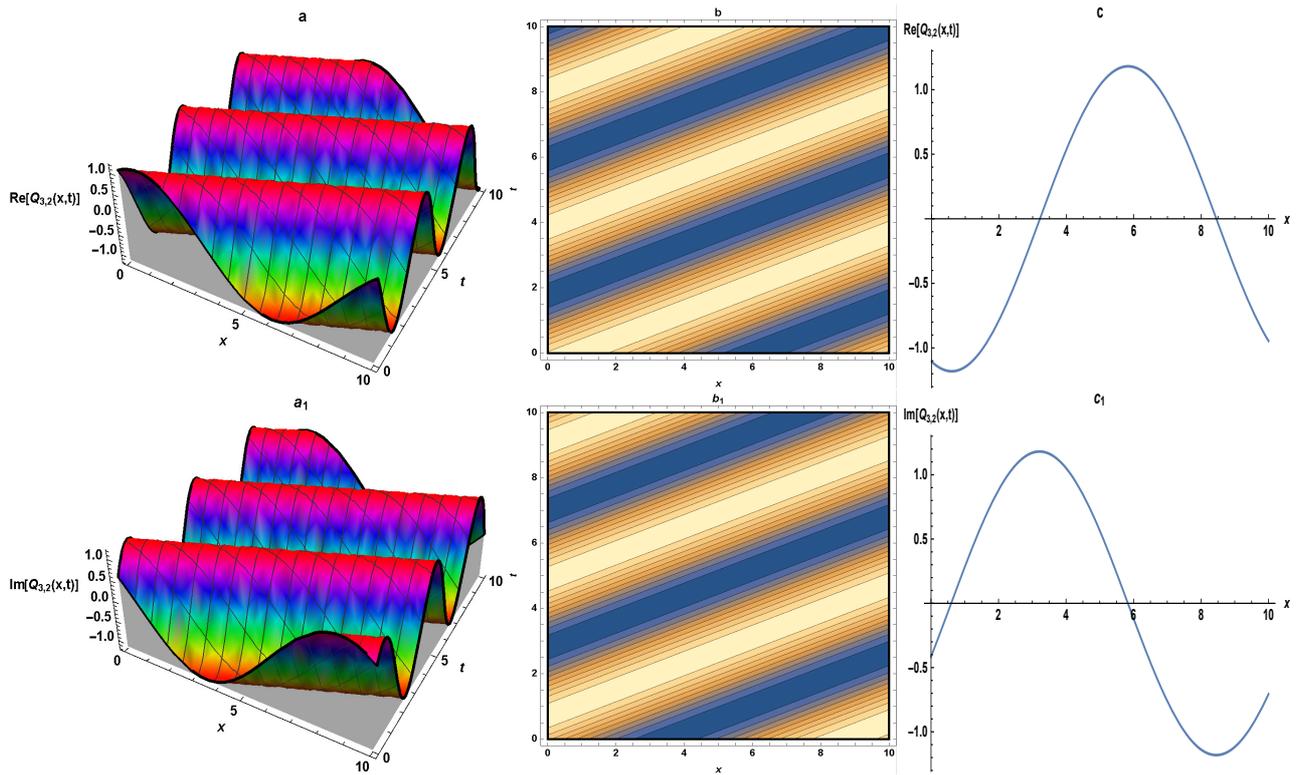
Figure 5. The 3Ds (a), (a_1), contours (b),(b_1) and 2Ds (c),(c_1) graphs of Eq. (31)

If $m \to 1$, then the dark singular soliton solution is obtained

$$Q_{4,3}(x,t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}} \left(\frac{\coth(\zeta)}{\sqrt{\frac{-(-1+h_2+(2+h_2)csch^2(\zeta))h_4}{-1+h_2^2}}} \right) \right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (34)$$

such that

$$h_4^2 (-2-h_2) \left[(-1+h_2)^2 \right] = 0.$$

If $m \to 0$, then the periodic wave solution is obtained

$$p_{4,4}(x,t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}} \left(\frac{\csc(\zeta)}{\sqrt{\frac{-\csc^2(\zeta)h_4}{-1+h_2}}} \right) \right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (35)$$

such that

$$h_4^2 (-1-h_2) \left[(-2+h_2)(1+h_2) \right] = 0.$$

5. If $l_0 = -m^2, l_2 = 2m^2 - 1, l_4 = 1 - m^2, 0 < m < 1$, then $P(\zeta) = nc(\zeta, m)$, we have

$$Q_5(x,t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}} \left(\frac{nc(\zeta,m)}{\sqrt{f(nc(\zeta,m))^2 + g}} \right) \right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (36)$$

where f and g are given by

$$f = \frac{-(-1 + 2m^2 - h_2)h_4}{1 - m^2 + m^4 - h_2^2}, \qquad g = \frac{3m^2 h_4}{1 - m^2 + m^4 - h_2^2},$$

under the constraint condition

$$h_4^2 \left(-1 + 2m^2 - h_2 \right) \left[\left(-2 + m^2 + h_2 \right) \left(1 + m^2 + h_2 \right) \right] = 0.$$

If $m \to 1$, then the singular soliton solution is obtained

$$Q_{5,1}(x,t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}} \left(\frac{\cosh(\zeta)}{\sqrt{\frac{\left(-3+(1-h_2)\cosh^2(\zeta)\right)h_4}{-1+h_2^2}}} \right) \right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (37)$$

such that

$$h_4^2 (1 - h_2) \left[-2 + h_2 + h_2^2 \right] = 0.$$

If $m \to 0$, then the periodic wave solution is obtained

$$Q_{5,2}(x,t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}} \left(\frac{\sec(\zeta)}{\sqrt{\frac{-\sec^2(\zeta)h_4}{-1+h_2}}} \right) \right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (38)$$

such that

$$h_4^2 (-1 - h_2) \left[(-2 + h_2)(1 + h_2) \right] = 0.$$

6. If $l_0 = -1, l_2 = 2 - m^2, l_4 = -\left(1 - m^2\right), 0 < m < 1$, then $P(\zeta) = nd(\zeta, m)$ and we have

$$Q_6(x,t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}} \left(\frac{nd(\zeta,m)}{\sqrt{f(nd(\zeta,m))^2 + g}} \right) \right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (39)$$

where f and g are given by

$$f = \frac{(-2 + m^2 + h_2)h_4}{1 - m^2 + m^4 - h_2^2}, \qquad g = \frac{3h_4}{1 - m^2 + m^4 - h_2^2},$$

under the constraint condition

$$h_4^2 \left(2 - m^2 - h_2\right) \left[-\left(-1 + 2m^2 - h_2\right)\left(1 + m^2 + h_2\right)\right] = 0.$$

If $m \to 1$, then the singular soliton solution is obtained

$$Q_{6,1}(x,t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}}\left(\frac{\cosh(\zeta)}{\sqrt{\frac{\left(-3+(1-h_2)\cosh^2(\zeta)\right)h_4}{-1+h_2^2}}}\right)\right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (40)$$

such that

$$h_4^2\left(1 - h_2\right)\left[-\left(1 - h_2\right)\left(2 + h_2\right)\right] = 0.$$

If $m \to 0$, then the periodic wave solution is obtained

$$Q_{6,2}(x,t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}}\left(\frac{1}{\sqrt{\frac{-h_4}{-1+h_2}}}\right)\right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (41)$$

such that

$$h_4^2\left(2 - h_2\right)\left[(1 + h_2)^2\right] = 0.$$

7. If $l_0 = 1, l_2 = 2 - m^2, l_4 = 1 - m^2, 0 < m < 1$, then $P(\zeta) = sc(\zeta, m)$, we have

$$Q_7(x,t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}}\left(\frac{sc(\zeta,m)}{\sqrt{f(sc(\zeta,m))^2 + g}}\right)\right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (42)$$

where f and g are given by

$$f = \frac{(-2 + m^2 + h_2)h_4}{1 - m^2 + m^4 - h_2^2}, \qquad g = \frac{-3h_4}{1 - m^2 + m^4 - h_2^2},$$

under the constraint condition

$$h_4^2\left(2 - m^2 - h_2\right)\left[-\left(-1 + 2m^2 - h_2\right)\left(1 + m^2 + h_2\right)\right] = 0.$$

If $m \to 1$, then the singular soliton solution is obtained

$$Q_{7,1}(x,t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}}\left(\frac{\sinh(\zeta)}{\sqrt{\frac{\left(3+(1-h_2)\sinh^2(\zeta)\right)h_4}{-1+h_2^2}}}\right)\right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (43)$$

such that

$$h_4^2 (1 - h_2) \left[-2 + h_2 + h_2^2 \right] = 0.$$

If $m \to 0$, then the periodic wave solution is obtained
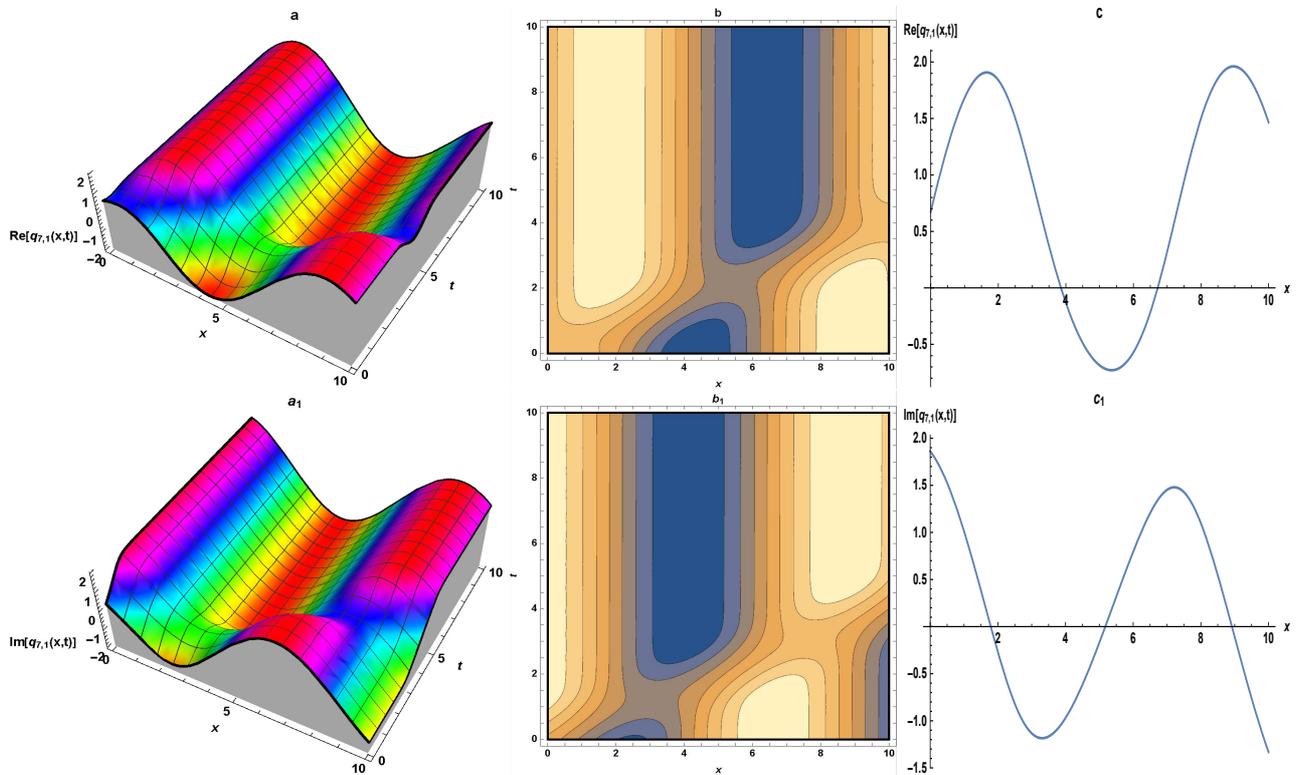
Figure 6. The 3Ds (a), (a_1), contours (b),(b_1) and 2Ds (c),(c_1) graphs of Eq. (43)

$$Q_{7,2} (x, t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}} \left(\frac{\tan(\zeta)}{\sqrt{\frac{(3-(-2+h_2)\tan^2(\zeta))h_4}{-1+h_2^2}}} \right) \right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (44)$$

such that

$$h_4^2 (2 - h_2) \left[(1 + h_2)^2 \right] = 0.$$

8. If $l_0 = 1, l_2 = 2m^2 - 1, l_4 = -m^2 (1 - m^2), 0 < m < 1$, then $P(\zeta) = sd(\zeta, m)$, we have

$$Q_8 (x, t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}} \left(\frac{sd(\zeta,m)}{\sqrt{f(sd(\zeta,m))^2+g}} \right) \right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (45)$$
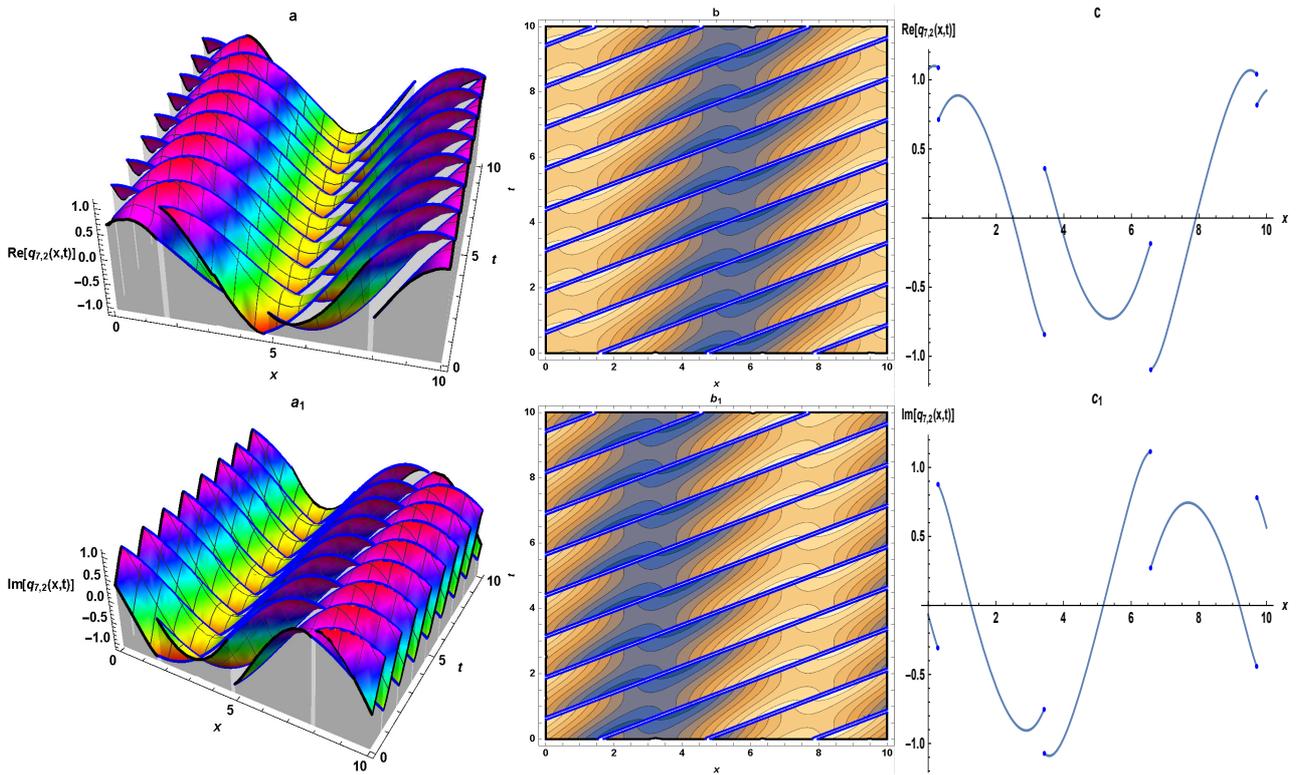
Figure 7. The 3Ds (a), (a_1), contours (b),(b_1) and 2Ds (c),(c_1) graphs of Eq. (44)

where f and g are given by

$$f = \frac{(-1 + 2m^2 - h_2)h_4}{1 - m^2 + m^4 - h_2^2}, \qquad g = \frac{-3h_4}{1 - m^2 + m^4 - h_2^2},$$

under the constraint condition

$$h_4^2 \left(-1 + 2m^2 - h_2\right) \left[\left(-2 + m^2 + h_2\right)\left(1 + m^2 + h_2\right)\right] = 0.$$

9. If $l_0 = 1 - m^2$, $l_2 = 2 - m^2$, $l_4 = 1$, $0 < m < 1$, then $P(\zeta) = cs(\zeta, m)$, we have

$$Q_9(x, t) = \left[\frac{-1 - 2n}{4(1 + n)r} + \frac{\sqrt{(1 + 2n)(-a + 4\beta)h_4}}{2n\sqrt{br}}\left(\frac{cs(\zeta, m)}{\sqrt{f(cs(\zeta, m))^2 + g}}\right)\right]^{\frac{1}{2n}} e^{i(-kx + \omega t + \theta)}, \quad (46)$$

where f and g are given by

$$f = \frac{(-2 + m^2 + h_2)h_4}{1 - m^2 + m^4 - h_2^2}, \qquad g = \frac{3(-1 + m^2)h_4}{1 - m^2 + m^4 - h_2^2},$$

under the constraint condition

$$h_4^2 \left(2 - m^2 - h_2\right) \left[-\left(-1 + 2m^2 - h_2\right)\left(1 + m^2 + h_2\right)\right] = 0.$$

If $m \to 1$, then the singular soliton solution is obtained

$$Q_{9,1}(x,t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}} \left(\frac{csch(\zeta)}{\sqrt{\frac{-csch^2(\zeta)h_4}{1+h_2}}} \right) \right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (47)$$

such that

$$h_4^2(1-h_2)\left[-2+h_2+h_2^2\right] = 0.$$

If $m \to 0$, then the periodic wave solution is obtained

$$Q_{9,2}(x,t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}} \left(\frac{\cot(\zeta)}{\sqrt{\frac{(3+(2-h_2)\cot^2(\zeta))h_4}{-1+h_2^2}}} \right) \right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (48)$$

such that

$$h_4^2(2-h_2)\left[(1+h_2)^2\right] = 0.$$

10. If $l_0 = -m^2(1-m^2), l_2 = 2m^2-1, l_4 = 1, 0 < m < 1$, then $P(\zeta) = ds(\zeta,m)$, we have

$$Q_{10}(x,t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}} \left(\frac{ds(\zeta,m)}{\sqrt{f(ds(\zeta,m))^2+g}} \right) \right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (49)$$

where f and g are given by

$$f = \frac{-(-1+2m^2-h_2)h_4}{1-m^2+m^4-h_2^2}, \qquad g = \frac{-3m^2(-1+m^2)h_4}{1-m^2+m^4-h_2^2},$$

under the constraint condition

$$h_4^2\left(-1+2m^2-h_2\right)\left[\left(-2+m^2+h_2\right)\left(1+m^2+h_2\right)\right] = 0.$$

11. If $l_0 = \frac{1-m^2}{4}, l_2 = \frac{1+m^2}{2}, l_4 = \frac{1-m^2}{4}, 0 < m < 1$, then $P(\zeta) = nc(\zeta,m) \pm sc(\zeta,m)$ or $P(\zeta) = \frac{cn(\zeta,m)}{1\pm sn(\zeta,m)}$, we have

$$Q_{11,1}(x,t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}} \left(\frac{nc(\zeta,m)\pm sc(\zeta,m)}{\sqrt{f(nc(\zeta,m)\pm sc(\zeta,m))^2+g}} \right) \right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (50)$$

or

$$Q_{11,2}(x,t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}} \left(\frac{\frac{cn(\zeta,m)}{1\pm sn(\zeta,m)}}{\sqrt{f\left(\frac{cn(\zeta,m)}{1\pm sn(\zeta,m)}\right)^2 + g}} \right) \right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (51)$$

where f and g are given by

$$f = \frac{-8(1+m^2-2h_2)h_4}{1+14m^2+m^4-16h_2^2}, \qquad g = \frac{12(-1+m^2)h_4}{1+14m^2+m^4-16h_2^2},$$

under the constraint condition

$$h_4^2 \left(\frac{1}{2}\left(1+m^2-2h_2\right) \right) \left[\frac{1}{16}\left(1+(-6+m)m+4h_2\right)\left(1+m(6+m)+4h_2\right) \right] = 0.$$

If $m \to 1$, then the combined singular soliton solution

$$Q_{11,3}(x,t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}} \left(\frac{\sinh(\zeta)+\cosh(\zeta)}{\sqrt{\frac{-(\sinh(\zeta)+\cosh(\zeta))^2 h_4}{1+h_2}}} \right) \right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (52)$$

or dark-bright optical soliton is obtained

$$Q_{11,4}(x,t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}} \left(\frac{\left(\frac{sech(\zeta)}{1+\tanh(\zeta)}\right)}{\sqrt{\frac{-\left(\frac{sech(\zeta)}{1+\tanh(\zeta)}\right)^2 h_4}{1+h_2}}} \right) \right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (53)$$

such that

$$h_4^2(1-h_2)\left[-2+h_2+h_2^2\right] = 0.$$

If $m \to 0$, then the periodic wave solution is obtained

$$Q_{11,5}(x,t) = \frac{1}{2} \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}} \left(\frac{\sec(\zeta)+\tan(\zeta)}{\sqrt{\frac{\left(3+2(1-2h_2)(\sec(\zeta)+\tan(\zeta))^2\right)h_4}{-1+16h_2^2}}} \right) \right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (54)$$
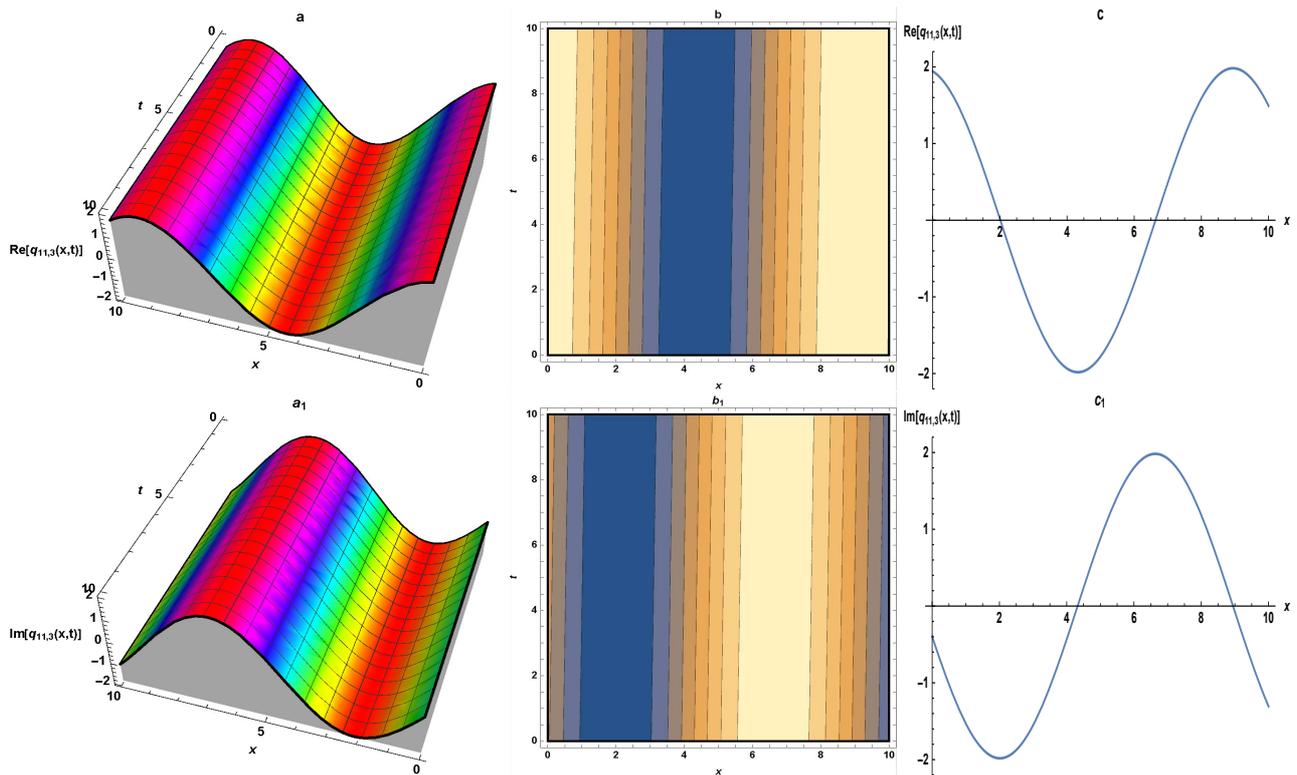
Figure 8. The 3Ds (a), (a_1), contours (b),(b_1) and 2Ds (c),(c_1) graphs of Eq. (52)

or

$$Q_{11,6}\left(x,t\right) = \frac{1}{2}\left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}}\left(\frac{\left(\frac{\cos(\zeta)}{1+\sin(\zeta)}\right)}{\sqrt{\frac{\left(3+2(1-2h_2)\left(\frac{\cos(\zeta)}{1+\sin(\zeta)}\right)^2\right)h_4}{-1+16h_2^2}}}\right)\right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (55)$$

such that

$$h_4^2\left(\frac{1}{2}-h_2\right)\left[\frac{1}{16}(1+4h_2)^2\right] = 0.$$

12. If $l_0 = \frac{-\left(1-m^2\right)^2}{4}$, $l_2 = \frac{1+m^2}{2}$, $l_4 = \frac{-1}{4}$, $0 < m < 1$, then $P(\zeta) = mcn(\zeta,m) \pm dn(\zeta,m)$, we have

$$Q_{12}\left(x,t\right) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}}\left(\frac{mcn(\zeta,m) \pm dn(\zeta,m)}{\sqrt{f(mcn(\zeta,m) \pm dn(\zeta,m))^2 + g}}\right)\right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (56)$$

where f and g are given by

$$f = \frac{-8(1+m^2-2h_2)h_4}{1+14m^2+m^4-16h_2^2}, \qquad g = \frac{12(-1+m^2)^2h_4}{1+14m^2+m^4-16h_2^2},$$
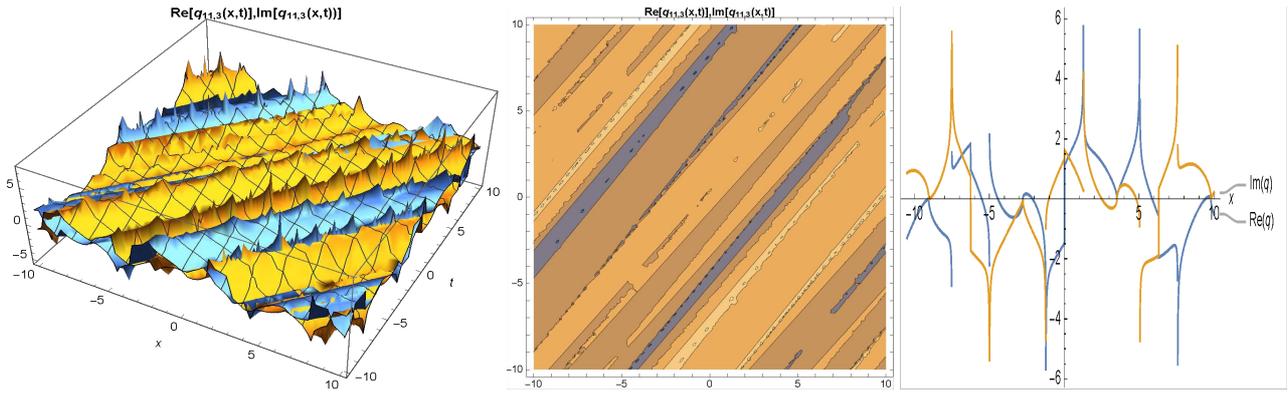
Figure 9. The 3D (a), contour (b) and 2D (c), graphs of Eq. (55)

under the constraint condition

$$h_4^2 \left(\frac{1}{2} \left(1 + m^2 - 2h_2 \right) \right) \left[\frac{1}{16} \left(1 + (-6+m)\, m + 4h_2 \right) \left(1 + m\,(6+m) + 4h_2 \right) \right] = 0.$$

13. If $l_0 = \frac{1}{4}$, $l_2 = \frac{1-2m^2}{2}$, $l_4 = \frac{1}{4}$, $0 < m < 1$, then $P(\zeta) = \frac{sn(\zeta,m)}{1 \pm cn(\zeta,m)}$, we have

$$Q_{13}\,(x,t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}} \left(\frac{\frac{sn(\zeta,m)}{1 \pm cn(\zeta,m)}}{\sqrt{f\left(\frac{sn(\zeta,m)}{1 \pm cn(\zeta,m)}\right)^2 + g}} \right) \right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (57)$$

where f and g are given by

$$f = \frac{8(-1+2m^2+2h_2)h_4}{1-16m^2+16m^4-16h_2^2}, \qquad g = \frac{-12h_4}{1-16m^2+16m^4-16h_2^2},$$

under the constraint condition

$$h_4^2 \left(\frac{1}{2} - m^2 - h_2 \right) \left[\frac{1}{16} + 2m^2 - 2m^4 + \left(\frac{1}{2} - m^2 \right) h_2 + h_2^2 \right] = 0.$$

If $m \to 1$, then the dark-bright optical soliton solution is obtained

$$Q_{13,1}\,(x,t) = \frac{1}{2} \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}} \left(\frac{\left(\frac{\tanh(\zeta)}{1+sech(\zeta)} \right)}{\sqrt{\frac{\left(3-2(1+2h_2)\left(\frac{\tanh(\zeta)}{1+sech(\zeta)} \right)^2 \right) h_4}{-1+16h_2^2}}} \right) \right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (58)$$

such that

$$h_4^2 \left(\frac{-1}{2} - h_2 \right) \left[\frac{1}{16} (1-4h_2)^2 \right] = 0.$$

If $m \to 0$, then the periodic wave solution is obtained

$$Q_{13,2}(x,t) = \frac{1}{2}\left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}}\left(\frac{\left(\frac{\sin(\zeta)}{1+\cos(\zeta)}\right)}{\sqrt{\frac{\left(3+2(1-2h_2)\left(\frac{\sin(\zeta)}{1+\cos(\zeta)}\right)^2\right)h_4}{-1+16h_2^2}}}\right)\right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (59)$$

such that

$$h_4^2\left(\frac{1}{2}-h_2\right)\left[\frac{1}{16}(1+4h_2)^2\right] = 0.$$

14. If $l_0 = \frac{1}{4}$, $l_2 = \frac{1+m^2}{2}$, $l_4 = \frac{(1-m^2)^2}{4}$, $0 < m < 1$, then $P(\zeta) = \frac{sn(\zeta,m)}{cn(\zeta,m)\pm dn(\zeta,m)}$, we have

$$Q_{14}(x,t) = \left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}}\left(\frac{\frac{sn(\zeta,m)}{cn(\zeta,m)\pm dn(\zeta,m)}}{\sqrt{f\left(\frac{sn(\zeta,m)}{cn(\zeta,m)\pm dn(\zeta,m)}\right)^2+g}}\right)\right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (60)$$

where f and g are given by

$$f = \frac{-8(1+m^2-2h_2)h_4}{1+14m^2+m^4-16h_2^2}, \qquad g = \frac{-12h_4}{1+14m^2+m^4-16h_2^2},$$

under the constraint condition

$$h_4^2\left(\frac{1}{2}\left(1+m^2-2h_2\right)\right)\left[\frac{1}{16}\left(1+(-6+m)\,m+4h_2\right)\left(1+m\,(6+m)+4h_2\right)\right] = 0.$$

If $m \to 1$, then the singular soliton solution is obtained

$$Q_{14,1}(x,t) = \left[\frac{-b_1}{3b_2} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}}\left(\frac{\sinh(\zeta)}{\sqrt{\frac{\left(3+(1-h_2)\sinh^2(\zeta)\right)h_4}{-1+h_2^2}}}\right)\right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (61)$$

such that $h_4^2\left(1-h_2\right)\left[-2+h_2+h_2^2\right] = 0$. If $m \to 0$, then the combined periodic wave solution is obtained

$$Q_{14,2}(x,t) = \frac{1}{2}\left[\frac{-1-2n}{4(1+n)r} + \frac{\sqrt{(1+2n)(-a+4\beta)h_4}}{2n\sqrt{br}}\left(\frac{\left(\frac{\sin(\zeta)}{1+\cos(\zeta)}\right)}{\sqrt{\frac{\left(3+2(1-2h_2)\left(\frac{\sin(\zeta)}{1+\cos(\zeta)}\right)^2\right)h_4}{-1+16h_2^2}}}\right)\right]^{\frac{1}{2n}} e^{i(-kx+\omega t+\theta)}, \quad (62)$$

such that $h_4^2\left(\frac{1}{2}-h_2\right)\left[\frac{1}{16}(1+4h_2)^2\right] = 0.$

5 Result and physical interpretations of the findings

The φ^6-model expansion technique is used as the integration algorithm. The bright, dark, singular, dark-bright, dark singular, and combined singular soliton solutions to the complex Ginzburg-Landau equation CGLE with dual power law nonlinearity were found. The dual-power law, found in photovoltaic materials, is used to explain nonlinearity in the refractive index. The traveling wave solutions discovered in this work are both physically and mathematically helpful. The constants in the computed wave propagation solutions must be given physical meaning in order to comprehend their physical significance.
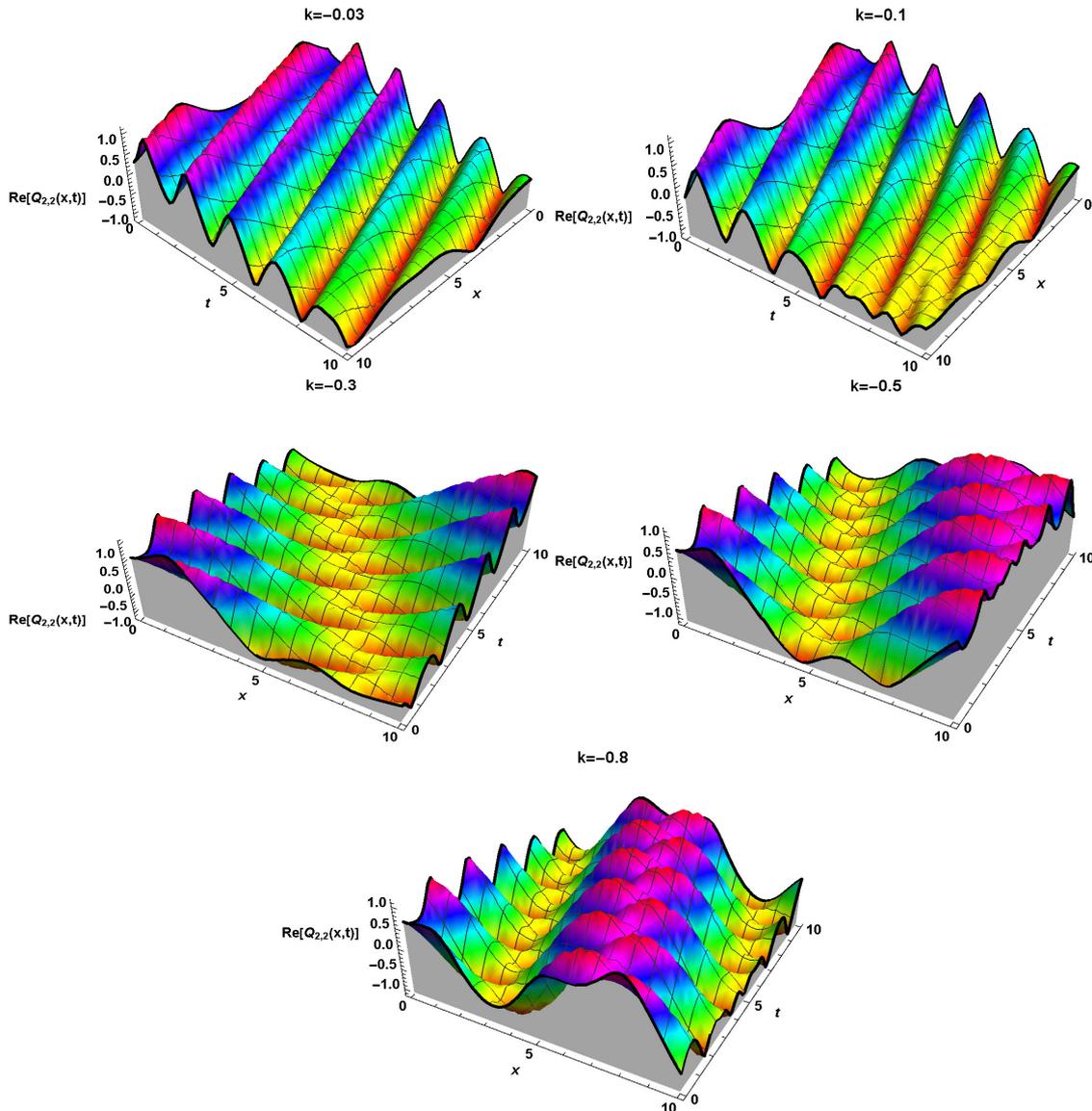
Figure 10. The 3D graphs of real part of Eq. (28) for
$\alpha_0 = 0.5, \alpha_1 = 2, h_1 = 1.2, h_2 = 0.9, h_3 = 2.6, h_4 = 0.5, w = 0.2, \theta = 0.5, n = 1, v = 3, \lambda = 0.5.$

In the physics literature, Eq. (5) represents the mathematical model that assumes the envelope of a forward-moving wave pulse evolves slowly in time and space related to a period or wavelength. $Q(x,t) = U(x - vt)e^{i(-kx + \omega t + \theta)}$, the function U represents the pulse shape and v is the soliton's velocity. In the phase factor, ω is the soliton wave number, θ is the phase constant and k denotes

the soliton frequency which is physically noteworthy in this study and whose various values will be studied for the moving wave's behavior. The parameter k is proportional to the velocity of the soliton as well as the length of the pulse. To evaluate the dynamical features and describe the evolution characteristic, we might pick appropriate values and functions for these parameters in Eq. (28).
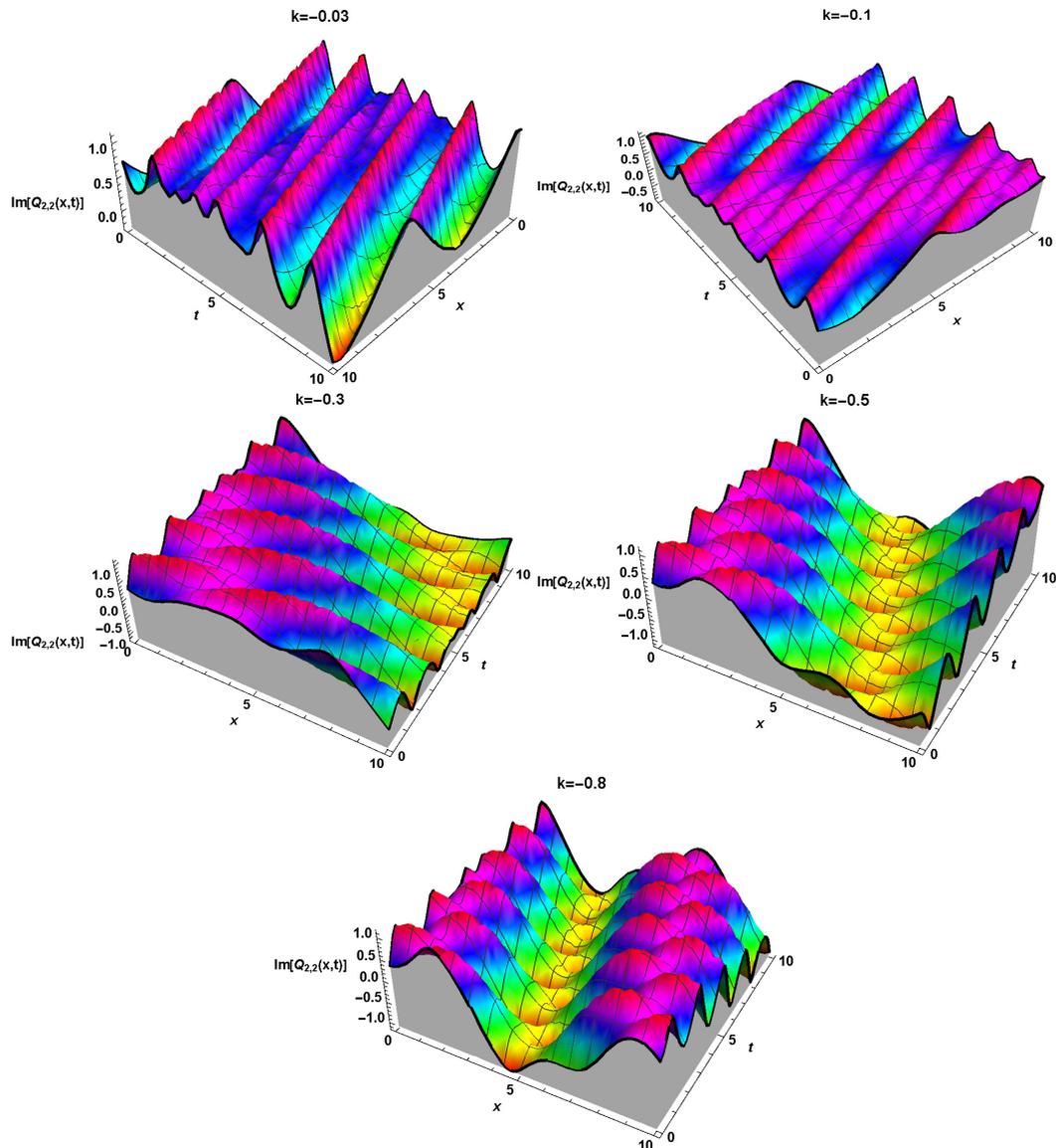
Figure 11. The 3D graphs of imaginary part of Eq. (28) for
$\alpha_0 = 0.5, \alpha_1 = 2, h_1 = 1.2, h_2 = 0.9, h_3 = 2.6, h_4 = 0.5, w = 0.2, \theta = 0.5, n = 1, v = 3, \lambda = 0.5.$

Figures 10 and 11 depict the behavior of a single wave at any given time, which is crucial in the transmission of energy from one location to another. In order to provide a new perspective to the topic let us investigate the physical implications of the parameters in the transformation, known as the classical wave transformation. The velocity of the propagating wave is proportional to the variable k. The frequency of the propagating wave is directly proportional to the velocity of the wave and inversely to its wavelength. The traveling wave shows diagonal wave behavior when k is increased. The frequency of a traveling wave is proportional to the wave number. The number of waves grows as the frequency increases.

6 Conclusion

Based on the research findings, it is important to point out that the selected scheme produces a wide range of innovative solutions that are both intriguing and valuable for the governing model. The obtained results in this work are believed to describe some of the CGLE's physical impacts. This research will be crucial to the understanding of the superconducting phenomenology theory, which is frequently used to explain the long-distance propagation of optical solitons via optical fibers and it will help in studying the photovoltaic materials. The φ^6-model expansion approach is helpful and efficient for constructing optical soliton solutions for most nonlinear physical phenomena. The behavior of a traveling wave solution chosen among fourteen different solutions generated for various values of frequency is investigated. When we compare our results in this paper to the results in [41–47], we conclude that our results are unique and have not been found elsewhere, the model will also be evaluated using fractional temporal evolution to account for slow-light pulses.

Declarations

Ethical approval

The authors state that this research complies with ethical standards. This research does not involve either human participants or animals.

Consent for publication

Not applicable.

Conflicts of interest

The authors declare that they have no conflict of interest.

Data availability statement

Data availability is not applicable to this article as no new data were created or analysed in this study.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author's contributions

M.A.I.: Methodology, Software, Conceptualization, Validation, Investigation, Writing - Original Draft, Visualization, Data Curation. A.Y.: Methodology, Conceptualization, Validation, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization, Supervision, Project Administration. All authors discussed the results and contributed to the final manuscript.

Acknowledgements

Not applicable.

References

[1] Isah, M.A. and Külahçı, M.A. A study on null cartan curve in Minkowski 3-space. *Applied Mathematics and Nonlinear Sciences,* 5(1), 413-424, (2020). [CrossRef]

[2] Isah, M.A. and Külahçı, M.A. Special curves according to bishop frame in Minkowski 3-space. *Applied Mathematics and Nonlinear Sciences,* 5(1), 237-248, (2020). [CrossRef]

[3] Isah, M.A., Isah, I., Hassan, T.L. and Usman, M. Some characterization of osculating curves according to darboux frame in three dimensional euclidean space. *International Journal of Advanced Academic Research,* 7(12), 47-56, (2021).

[4] Isah, I., Isah, M.A., Baba, M.U., Hassan, T.L. and Kabir, K.D. On integrability of silver riemannian structure. *International Journal of Advanced Academic Research,* 7(12), 2488-9849, (2021).

[5] Myint-U, T. and Lokenath, D. *Linear Partial Differential Equations for Scientists and Engineers.* Springer Science & Business Media: Berlin/Heidelberg, Germany, (2007).

[6] Ueda, T. and Kath, W.L. Dynamics of coupled solitons in nonlinear optical fibers. *Physical Review A,* 42(1), 563, (1990). [CrossRef]

[7] Hasegawa, A. and Matsumoto, M. Optical solitons in fibers. In *Springer Series in Photonics* (Vol 9) (pp. 41-59). Berlin, Heidelberg: Springer, (1989). [CrossRef]

[8] Musslimani, Z.H., Makris, K.G., El-Ganainy, R. and Christodoulides, D.N. Optical solitons in P T periodic potentials. *Physical Review Letters,* 100(3), 030402, (2008). [CrossRef]

[9] Kivshar, Y.S. and Agrawal, G.P. *Optical Solitons: from Fibers to Photonic Crystals.* Academic Press, (2003).

[10] Hasegawa, A. and Kodama, Y. Signal transmission by optical solitons in monomode fiber. *Proceedings of the IEEE,* 69(9), 1145-1150, (1981). [CrossRef]

[11] Manafian, J. and Heidari, S. Periodic and singular kink solutions of the Hamiltonian amplitude equation. *Advanced Mathematical Models & Applications,* 4(2), 134-149, (2019).

[12] Yokus, A. and Isah, M.A. Investigation of internal dynamics of soliton with the help of traveling wave soliton solution of Hamilton amplitude equation. *Optical and Quantum Electronics,* 54, 528, (2022). [CrossRef]

[13] Yokus, A. and Isah, M.A. Stability analysis and solutions of (2 + 1)-Kadomtsev–Petviashvili equation by homoclinic technique based on Hirota bilinear form. *Nonlinear Dynamics,* 109, 3029-3040, (2022). [CrossRef]

[14] Duran, S., Yokuş, A., Durur, H. and Kaya, D. Refraction simulation of internal solitary waves for the fractional Benjamin–Ono equation in fluid dynamics. *Modern Physics Letters B,* 35(26), 2150363, (2021). [CrossRef]

[15] Joshi, H., Yavuz, M. and Stamova, I. Analysis of the disturbance effect in intracellular calcium dynamic on fibroblast cells with an exponential kernel law. *Bulletin of Biomathematics,* 1(1), 24-39, (2023). [CrossRef]

[16] Yel, G., Kayhan, M. and Ciancio, A. A new analytical approach to the (1+ 1)-dimensional conformable Fisher equation. *Mathematical Modelling and Numerical Simulation with Applications,* 2(4), 211-220, (2022). [CrossRef]

[17] Iskenderoglu, G. and Kaya, D. Chirped self-similar pulses and envelope solutions for a nonlinear Schrödinger's in optical fibers using Lie group method. *Chaos, Solitons & Fractals,* 162, 112453, (2022). [CrossRef]

[18] Ghanbari, B. and Baleanu, D. A novel technique to construct exact solutions for nonlinear partial differential equations. The *European Physical Journal Plus,* 134(10), 506, (2019). [CrossRef]

[19] Kaya, D., Yokuş, A. and Demiroğlu, U. Comparison of exact and numerical solutions for the Sharma–Tasso–Olver equation. In *Numerical Solutions of Realistic Nonlinear Phenomena* (Vol 31) (pp. 53-65). Cham: Springer, (2020). [CrossRef]

[20] Duran, S., Durur, H., Yavuz, M. and Yokus, A. Discussion of numerical and analytical techniques for the emerging fractional order Murnaghan model in materials science. *Optical and Quantum Electronics,* 55(6), 571, (2023). [CrossRef]

[21] Durur, H., Yokuş, A. and Yavuz, M. (2022). Behavior analysis and asymptotic stability of the traveling wave solution of the Kaup-Kupershmidt equation for conformable derivative. In *Fractional Calculus: New Applications in Understanding Nonlinear Phenomena*, (Vol 3) (pp. 162-185). Bentham Science. [CrossRef]

[22] Shafqat, R., Niazi, A.U.K., Yavuz, M., Jeelani, M. . and Saleem, K. (2022). Mild solution for the time-fractional Navier–Stokes equation incorporating MHD effects. Fractal and Fractional, 6(10), 580. [CrossRef]

[23] Yokus, A., Kuzu, B. and Demiroğlu, U. Investigation of solitary wave solutions for the (3+1)-dimensional Zakharov–Kuznetsov equation. *International Journal of Modern Physics B,* 33(29), 1950350, (2019). [CrossRef]

[24] Yokuş, A., Aydın, M.E., Duran, S. and Durur, H. Simulation of hyperbolic type solitary waves based on velocity parameter for (3+1)-dimensional the B-type Kadomtsev–Petviashvili–Boussinesq equation. *Modern Physics Letters B,* 36(22), 2250110, (2022). [CrossRef]

[25] Zhou, Q., Xiong, X., Zhu, Q., Liu, Y., Yu, H., Yao, P. et al. Optical solitons with nonlinear dispersion in polynomial law medium. *Journal of Optoelectronics and Advanced Materials,* 17, 82-86, (2015). [CrossRef]

[26] Zayed, E.M. and Al-Nowehy, A.G. Many new exact solutions to the higher-order nonlinear Schrödinger equation with derivative non-Kerr nonlinear terms using three different techniques. *Optik,* 143, 84-103, (2017). [CrossRef]

[27] Zayed, E.M., Al-Nowehy, A.G. and Elshater, M.E. New-model expansion method and its applications to the resonant nonlinear Schrödinger equation with parabolic law nonlinearity. *The European Physical Journal Plus,* 133(10), 417, (2018). [CrossRef]

[28] Ozisik, M., Secer, A., Bayram, M., Yusuf, A. and Sulaiman, T.A. On the analytical optical soliton solutions of perturbed Radhakrishnan–Kundu–Lakshmanan model with Kerr law nonlinearity. *Optical and Quantum Electronics,* 54(6), 371, (2022). [CrossRef]

[29] Atas, S.S., Ali, K.K., Sulaiman, T.A. and Bulut, H. Invariant optical soliton solutions to the Coupled-Higgs equation. *Optical and Quantum Electronics,* 54(11), 754, (2022). [CrossRef]

[30] Sulaiman, T.A. Three-component coupled nonlinear Schrödinger equation: optical soliton and modulation instability analysis. *Physica Scripta,* 95(6), 065201, (2020). [CrossRef]

[31] Veeresha, P., Yavuz, M. and Baishya, C. A computational approach for shallow water forced Korteweg–De Vries equation on critical flow over a hole with three fractional operators. *An International Journal of Optimization and Control: Theories & Applications (IJOCTA),* 11(3), 52-67, (2021). [CrossRef]

[32] Younas, U., Bilal, M., Sulaiman, T.A., Ren, J. and Yusuf, A. On the exact soliton solutions and different wave structures to the double dispersive equation. *Optical and Quantum Electronics,* 54, 1-22, (2022). [CrossRef]

[33] Kudryashov, N.A. Exact solutions of the complex Ginzburg–Landau equation with law of

four powers of nonlinearity. *Optik,* 265, 169548, (2022). [CrossRef]

[34] El-Ganaini, S. and Al-Amr, M.O. New abundant solitary wave structures for a variety of some nonlinear models of surface wave propagation with their geometric interpretations. *Mathematical Methods in the Applied Sciences,* 45(11), 7200-7226, (2022). [CrossRef]

[35] Farag, N.G., Eltanboly, A.H., El-Azab, M.S. and Obayya, S.S.A. Pseudo-spectral approach for extracting optical solitons of the complex Ginzburg Landau equation with six nonlinearity forms. *Optik,* 254, 168662, (2022). [CrossRef]

[36] Tarla, S., Ali, K.K. and Yilmazer, R. Newly modified unified auxiliary equation method and its applications. *Optik,* 269, 169880, (2022). [CrossRef]

[37] Duran, S. Travelling wave solutions and simulation of the Lonngren wave equation for tunnel diode. *Optical and Quantum Electronics,* 53, 458, (2021). [CrossRef]

[38] Durur, H. Traveling wave solutions of the oskolkov equation arising in incompressible viscoelastic Kelvin–Voigt fluid. *Bilecik Şeyh Edebali Üniversitesi Fen Bilimleri Dergisi*, 9(2), 931-938, (2022). [CrossRef]

[39] Duran, S., Yokuş, A., Durur, H. and Kaya, D. Refraction simulation of internal solitary waves for the fractional Benjamin–Ono equation in fluid dynamics. *Modern Physics Letters B*, 35(26), 2150363, (2021). [CrossRef]

[40] Aslan, E.C. Solitary solutions and modulation instability analysis of the nonlinear Schrödinger equation with Cubic-Quartic nonlinearity. *Journal of Advanced Physics*, 6(4), 579-585, (2017). [CrossRef]

[41] Isah, M.A. and Yokuş, A. The investigation of several soliton solutions to the complex Ginzburg-Landau model with Kerr law nonlinearity. *Mathematical Modelling and Numerical Simulation with Applications*, 2(3), 147-163, (2022). [CrossRef]

[42] Abdou, M.A., Soliman, A.A., Biswas, A., Ekici, M., Zhou, Q. and Moshokoa, S.P. Dark-singular combo optical solitons with fractional complex Ginzburg-Landau equation. *Optik*, 171, 463-467, (2018). [CrossRef]

[43] Al-Ghafri, K.S. Soliton behaviours for the conformable space-time fractional complex Ginzburg-Landau equation in optical fibers. *Symmetry*, 12(2), 219, (2020). [CrossRef]

[44] Biswas, A., Yildirim, Y., Yasar, E., Triki, H., Alshomrani, A.S., Ullah, M.Z. et al. Optical soliton perturbation with complex Ginzburg–Landau equation using trial solution approach. *Optik*, 160, 44-60, (2018). [CrossRef]

[45] Arshed, S., Biswas, A., Mallawi, F. and Belic, M.R. Optical solitons with complex Ginzburg–Landau equation having three nonlinear forms. *Physics Letters A*, 383(36), 126026, (2019). [CrossRef]

[46] Li, Z. and Han, T. New exact traveling wave solutions of the time fractional complex Ginzburg-Landau equation via the conformable fractional derivative. *Advances in Mathematical Physics*, 2021, 1-12, (2021). [CrossRef]

[47] Arshed, S., Raza, N., Rahman, R.U., Butt, A.R. and Huang, W.H. Sensitive behavior and optical solitons of complex fractional Ginzburg–Landau equation: a comparative paradigm. *Results in Physics*, 28, 104533, (2021). [CrossRef]

[48] Sajid, N. and Akram, G. Novel solutions of Biswas-Arshed equation by newly φ^6-model expansion method. *Optik*, 211, 164564, (2020). [CrossRef]

[49] Biswas, A. Chirp-free bright optical solitons and conservation laws for complex Ginzburg-

Landau equation with three nonlinear forms. *Optik* 174, 207-215, (2018). [CrossRef]

Mathematical Modelling and Numerical Simulation with Applications (MMNSA)
(https://dergipark.org.tr/en/pub/mmnsa)

RESEARCH PAPER

A harmonic oscillator model of atmospheric dynamics using the Newton-Kepler planetary approach

Alexander Munson [ID] [1,*,‡]

[1]Department of Mathematics, Physics, and Computer Science, Wilson College, PA 17201, United States
*Corresponding Author
‡alexander.munson@wilson.edu (Alexander Munson)

Abstract

Projection of future meteorological patterns such as median temperature and precipitation is necessary for governments to facilitate civil aviation, forecast agricultural productions, and advise public energy policies. Various models are proposed based on historical data, such as the short-term 7-day forecast or the long-term Global Forecast System, to study climate change. The main contribution of the paper is that it gives a feasible, cost-effective model for median-term projections with statistically tested accuracies well within the accepted margins of the scientific community. This model is the starting point to provide general guidelines to governments to forecast levels of energy consumption for residential cooling in summer and heating in winter to provide energy subsidies for low-income populations and for organizations supporting countries needing energy assistance. Additionally, mid-term models are also useful during global energy disruptions. A theoretical model is derived based on orbital mechanics, planetary science, and astronomy using Newton's Law of Universal Gravitation and Kepler's Laws of Planetary Motions. The model is then optimized with historical data in a specific region. The model's predictions are then statistically compared with the actual observed temperature outside the training data. In sum, the current harmonic oscillator method can be beneficially utilized by governments to forecast natural phenomena in order to provide timely assistance to respective populations such as predicting extreme temperature fluctuations in the planning of agricultural productions.

Keywords: Harmonic oscillators; Newton's law of gravitation; Kepler's laws of planetary motion; orbital mechanics

AMS 2020 Classification: 35C07; 35A25; 76A05

1 Introduction

The study of atmospheric phenomena to project weather patterns by mathematical modelling has a long history. The traditional approach has been gathering vast historical and empirical data and

fitting them to various dynamical system frameworks consisting of covariant and interdependent explanatory variables. Although new models have been proposed throughout the years, certain challenges persist. A key obstacle is that the newer models frequently require ultrafast computing powers incurring increasingly steep and prohibitive costs to national and local governments. This challenge is particularly acute for developing countries in need of economic aid and disaster relief. Secondly, these new computation-heavy models may lack the atmospheric flexibilities unique to a region due to its specific topographies. Additionally, these costly models can be deficient at times in providing sufficient information that answers to the administrative tasks of energy consumption needs for its populations. Indeed, the cost-effective allocations of energy supplies in a region's smart energy grid are a perennial challenge to governments at all levels, national, regional, and local. Vast research has been directed to achieve efficiency in a region's energy grid given its specific climate patterns. For example, a recent research paper by Yilmaz studied the feasibility of generative adversarial networks (GANs) to improve energy efficiency by incorporating information and communication technologies in Türkiye energy grid [1]. The paper's author was able to demonstrate that trained synthetic load data are effective in reducing prediction errors in load when historical data and risk management calculations are combined. Putting these considerations together, the current paper examines the feasibility of harmonic oscillators as a basic principle in mid-term climate projections. The main objective and contribution of the current paper is that it gives a feasible, cost-effective model for median-term temperature projections with statistically tested accuracies well within the accepted margins of the scientific community. The paper is not the end, but the starting point in developing a cost-effective oscillation model with higher accuracy and provide certain general guidelines for future research in this direction.

The remainder of this section will review some popular regional climate models in detail. Section 2 will broaden the discussion to show the extensive applicability of harmonic oscillators in biology, chemical engineering, physics, and biochemistry. In Section 3, the paper gives the classical mechanics background and the basic atmospheric model. Section 4 furnishes the mathematical model with historical meteorological data in a region. Section 5 shows the statistical analyses of both the consistency and the predictive feasibility of the model, and finally, Section 6 gives an overview of future research directions.

Regional climate models are increasingly important in that they consider historical data unique to the locale and approach projections with the understanding that atmospheric behaviors at the global level cannot be reasonably construed as a summation of mesoscale regional ensembles. Different regional climate models coexist including Modele Atmospheric Regionale (MAR) developed by University of Grenoble in France, Advanced Regional Prediction System (ARPS) by University of Oklahoma, and Aire Limitée Adaptation Dynamique Développement International (ALADIN) developed and shared by 16 European and African countries. One particular advanced regional model that has gained notice is the Regional Atmospheric Modelling System (RAMS) by Colorado State University. With a projection scale of a few hundred square kilometers, RAMS utilizes the nonlinear finite-difference method fine-tuned by historical data to produce projections. The troposphere is divided into 3D grid intervals of equal volume Dq where q is a discretized stochastic variable. Time t is also considered as a discrete variable. Observations are made at discrete points in time $t_n = t_0 + nDt$ for $n = 0, 1, 2, 3, \ldots$, and Dt is the sampling time interval. The probability that the stochastic state variable q is in the volume grid i at time t_n is given by the relation $f_i(t_n)Dq$. Probability density function f at the next time interval t_{n+1} is given by the discretized sum of current states

$$f_i(t_{n+1}) = f_i(t_n) + \sum_{i'} D_i^{i'} f_{i'}(t_n). \tag{1}$$

In the above relation, $D_i^{i'}$ is the probability transition matrix from the i-th state to the i'-th state. The coefficients in the transition matrix satisfy $D_i^{i'} = W_i^{i'} - \delta_{ii'} \sum_p W_{i'}^p$, where $W_j^{k'}$ values are the transition coefficients updated in real-time by historical data in the recent past and normalized to conserved probability

$$\sum_{i'} W_i^{i'} = 1. \tag{2}$$

One challenge has been striking a balance between the desired resolution in the phase space on the one hand and the availability of computing resources and historical data on the other. Another challenge is that while the estimate of the transition improves with growing time series length, there is a threshold time series length beyond which forecast accuracies do not improve significantly. The RAMS climate projection system uses the HPE Cray supercomputer at 800 petaFlops to handle the ever-increasing need for vast computations. Other regional climate models such as MAR face even more challenges as different terrains and topographies across the Mediterranean require more variabilities to be incorporated into the computations, further straining government agencies' tight budgets in times of inflation, natural disasters, and energy supply disruptions. It is with these considerations, feasible and cost-effective mathematical models based on historical data become a focus of research.

2 Applicability of harmonic oscillators

A harmonic oscillator exemplifies the simplicity of the smooth transfer between kinetic energy and potential energy. The equation of the phase space of the harmonic oscillator is a second-order differential equation with initial values. The long-term asymptotic behaviors can be understood through the presence and the nature of damping in the system. Applications of harmonic oscillators abound in nearly all branches of natural, social, and behavioral science ranging from physics, biology, chemistry, ecology, economics, financial mathematics, and game theory to psychology. In ecology, Koshkin and Meyers showed recently in 2022 that a stressed predator-prey system can be modelled by harmonic oscillators where asymptotic limits can collapse into an attractor point in the Lotka-Volterra differential equation system [2]. Setting $x(t)$ and $y(t)$ to be the populations of the two species at time t respectively. The Lotka-Volterra system describes the interdependence between the prey population density $x(t)$ and the predator population density $y(t)$, respectively, as

$$\begin{aligned} \frac{dx}{dt} &= y(\alpha - \beta x), \\ \frac{dy}{dt} &= x(-\gamma + \delta y), \end{aligned} \tag{3}$$

where α, β, and γ are dynamical system constants and δ is a system parameter describing the species-specific covariance relation between prey's population and the growth rate of predator's population. Koshkin and Meyers defined the energy of the predator-prey harmonic oscillator system to be $V(x,y) \equiv x^2 + by^2$, where b is also a constant of the system. Each constant C represents an integral energy level of the predator-prey system

$$x^2 + by^2 = C. \tag{4}$$

The time derivatives of V along the trajectories of the system are given by

$$\dot{V}(x,y) \equiv \nabla V \bullet (\dot{x},\ \dot{y}) = \frac{\partial V}{\partial x}\dot{x} + \frac{\partial V}{\partial y}\dot{y}. \tag{5}$$

Denoting the point $(x(t),y(t))$ as $p(t)$ in the graph of two species, the Ω-limit is defined as the asymptotic behavior of the system

$$\omega_p \equiv \left\{ q \in \mathbb{R}^2 \mid p(t_k) \longrightarrow q \text{ as } t \to \infty \right\}. \tag{6}$$

Koshkin and Meyers found that if environmental stress expressed as a damping coefficient $a > 0$ is present in the system, the derivative of the energy in the long-term limit is negative

$$\frac{d}{dt}V(x,y) = \frac{\partial V}{\partial x}\dot{x} + \frac{\partial V}{\partial y}\dot{y} = -2ax^2 < 0, \tag{7}$$

and the damped predator-prey oscillator system enters into a closed trajectory of energy ellipses and eventually collapses to a single-point Ω-limit q [2]. Harmonic oscillation also has numerous applications in chemical engineering, industrial engineering, and material science. In chemical engineering, certain cyclic hydrocarbon compounds such as benzene exhibits strong chemical stability due to the perfect symmetry in its six carbon-to-carbon and carbon-to-hydrogen covalent bonds. This stability of the benzene molecule causes intermolecular stacking to occur due to resonance delocalization of its electrons in the outer p atomic orbitals. The stable benzene molecule forms a double-deck crystalline structure in this type of intermolecular stacking whose noncovalent binding strength is modelled excellently by harmonic oscillators that utilize the periodic growth and decay of positive electrostatic potential on one benzene molecule relative to the negative electrostatic potential on the other benzene molecule in the above-described stacked structure. The electrostatic potential on a benzene molecule with resonance delocalization is given by the quantum harmonic potential

$$U(x) = \frac{1}{2}m\omega^2 x^2. \tag{8}$$

Here m is the molecular mass of benzene and ω is the angular frequency of the harmonic oscillation exhibited by the periodic increase and decrease in electrostatic potentials. In particular, the chemical stability of benzene due to the symmetry of its intra-molecular covalent bonds results in a slow evaporative property manifested by its distinct sweet aroma which is termed aromaticity in chemical engineering. The aromaticity of most cyclic hydrocarbon compounds is thus described excellently by a harmonic oscillation model with benzene as a reliable benchmark in industries. In a recent study, Arpa et al. [3] refined the harmonic oscillator model of aromaticity by reparametrizing the oscillation parameters of the classic carbon-to-carbon bond strength in acyclic hydrocarbon compounds ethane and ethylene. The researchers achieved the reparametrization of the harmonic oscillator model by approximating the aromaticity parameter α given as

$$\alpha = \frac{2}{\left(R_s - R_{opt}\right)^2 + \left(R_d - R_{opt}\right)^2}. \tag{9}$$

In Equation (9), R_s is the mean length of the pure single carbon-to-carbon bonds and R_d is the mean length of the pure double carbon-to-carbon bonds in ethane and ethylene and R_{opt} is the optimal carbon-to-carbon bond length in fully aromatic inert benzene experimentally measured to be 1.398-angstrom using neutron diffraction at 15 Kelvin. Thus, harmonic oscillator models are also important in chemical and industrial engineering research because these aromatic hydrocarbons are the backbone of industrial solvents, adhesives, plastics, and petrochemicals.

In quantum physics, harmonic oscillators are ubiquitous and are used extensively to model quantum phenomena. For example, Özdemir et al. [4] compared two types of computational bases in quantum entanglement. Denoting the initial state of an entangled system as ρ and the final state as ρ', the two states are related by a quantum phase operation $\mathcal{E} : \rho \rightarrow \mathcal{E}(\rho)$. One example of the quantum phase operation is the Kraus formalism given by

$$\mathcal{E}(\rho) = \sum_\mu E_\mu \rho \, E_\mu^\dagger. \tag{10}$$

The operator E_μ acts on the Hilbert space H_S of the entangled system and satisfies the completeness relation $\sum_\mu E_\mu^\dagger E_\mu = I$. Since a quantum system is never isolated and always interacts with its environment, the interaction introduces noise in quantum computation and information processing. This interaction is represented at time $t = 0$ in the entangled state as

$$\rho(0) = \rho_S(0) \otimes \rho_E(0). \tag{11}$$

The evolution of the time-dependent entangled system is given by

$$\rho(t) = U(t)\rho_S(0) \otimes \rho_E(0)U^\dagger(t), \tag{12}$$

where $U = e^{-iHt/\hbar}$. The Kraus representation of an entangled system is one where energy dissipation occurs via a damping process as a system interacts with its environment. Özdemir et al. [4] discovered that the Kraus representation for a single qubit whose computation basis is defined by the bosonic number states $\{\|0\rangle, \|1\rangle\}$ are given by the operators

$$
\begin{aligned}
A_0(t) &= |0\rangle\langle 0| + \sqrt{\eta} \, |1\rangle\langle 1|, \\
A_1(t) &= \sqrt{1-\eta} \, |0\rangle\langle 1|,
\end{aligned}
\tag{13}
$$

where $\sqrt{1-\eta}$ is the probability of the entangled system losing one qubit up to time t. The study compared the fidelities of the qubit states for the no-photon-decay event, the one-photon-decay event, and the two-photon-decay event. For example, it was discovered that the energy dissipation between the event of a one-photon-decay and two-photon decay up to time t, given as a damped harmonic oscillator with the above operator basis $\{A_i(t)\}$ exhibits a long-term fidelity behavior proportional to polynomial growths

$$\lim_{t \to \infty} F_1(t) - F_2(t) \propto \eta + \mathcal{O}\left[\eta^2\right]. \tag{14}$$

The harmonic oscillator has also found applications in biochemistry. For example, it is well-known that insulin and insulin-like growth factors 1 and 2, denoted as IGF1 and IGF2 have similar chemical structures. Although their precise mechanism of action is still unclear, it is known that these hormones activate two closely related receptor tyrosine-kinases called the insulin

receptor and IGF1 receptor respectively. Differences in the chemical kinetics of ligand binding and activations of receptors by insulin and IGF1 are generally thought to be the determining factors of their specificity. However, the receptors' ligand binding mechanisms show complex allosteric properties such as dependence on receptor disassociation rates. As a result, a more robust theory of receptor kinetics proved challenging for decades.

Kiselyov et al. [5] used harmonic oscillators to provide a valuable mathematical model for receptor kinetics. Using the concept that the active configuration of the receptors has higher free energy compared to that of the inactive configuration, Kiselyov et al. [5] modeled the behavior of the receptor mechanism near the point of activation as a harmonic oscillator. The idea is that separate subunits of the receptor are rigid bodies connected by covalent bonds similar to an elastic spring on a harmonic oscillator. In the model, the energy of the receptor oscillates harmonically when the receptor is in thermal equilibrium with the buffer. The energy of the receptor, as a result of random collisions with the buffer molecules, has a probability distribution given by the Maxwell formula

$$\frac{dP}{dE} = \frac{1}{KT}e^{-E/KT}, \tag{15}$$

where P, E, and T are the probability, energy, and temperature of the ensemble, respectively, and k is the Boltzmann constant. The above distribution implies that the fraction of the insulin receptors having sufficient activation energy for binding is given by

$$\int_{E_{activation}}^{\infty} \frac{1}{KT}e^{-E/KT}dE. \tag{16}$$

By tagging select sites alternately with regular and radioactive binding ligands, De Meyts et al. [5] were able to theorize the binding reaction kinetics. For example, for sites 1 and 2 on the insulin receptors, the chemical equilibrium of the level of $r_{1\times2}$ is given by

$$r_{1\times2} = \frac{r_{tot}k_{cr}/\left(k_{cr}+d_2\right)}{1+\dfrac{d_1d_2}{a_1\left(k_{cr}+d_2\right)/L}}, \tag{17}$$

where L and r_{tot} are the insulin concentration and total receptor levels respectively and the d_i's are the percentages of sites 1 and 2 remaining unbounded after insulin dissociation, k_{cr} is the critical kinetic constant, and a_1 is the dependence rate factor of site 1. Numerous other models of harmonic oscillators exist in natural science, behavioral science, and social science.

This paper is assembling a harmonic oscillator model of climate cycles with classic variables such as temperature. A word of caution is in order. In mathematical modelling, the current paper is fully aware that there exist explanatory models other than harmonic oscillators that can account for these natural phenomena. For example, in a recent paper, Mishra et al. [6] were able to combine the finite volume method and the Newton-Raphson method to study the calcium distribution across the membrane of a cholangiocyte cell and found that the diffusion rate follows a cutoff inverse sigmoid function. In addition, temperature itself can act with more complexity in natural phenomena that cannot be accounted for by a simple system of differential operators.

For example, in another recent study, Shah et al. [7] showed that temperature can cause convective flows across stratified surfaces in autocatalysis reactions in microorganisms. Thus, at times, a model based on these well-known mathematical frameworks alone is insufficient to provide explanatory power in studying some natural phenomena such as the movement of chemicals or

macro-particles across membranes or layers. At both the micro-level and the macro-level, natural phenomena can indeed exhibit more complexity than any simple but elegant mechanism alone can aptly capture. For example, the above-mentioned sigmoid saturation model, in contrast to a model of oscillators that shows periodicity, was successfully utilized as the activation function in a study of artificial intelligence-assisted detection methods of melanoma performed by Orhan and Yavşan [8]. In the study, the researchers chose the sigmoid function as the activation function in testing six classification algorithms' accuracy in melanoma detection models built on the convoluted neural network or CNN-based deep learning AI's. Among the six algorithms thus trained, the MobileNet algorithm achieved an accuracy rate of 89.4% in cancer detection [8]. In another study, Joshi et al. [9] considered a mathematical model that is framed to investigate the role of buffer and calcium concentration on fibroblast cells. Thus, future research should keep in mind that mathematical modelling of these natural or social phenomena may involve a multi-dimensional approach.

3 The classical harmonic oscillator

Harmonic oscillators are in essence the idealization of the continuous and reversible process of the transfer of kinetic energy into potential energy in a mass-spring system. The classical harmonic oscillator is best reified by the mass-spring system. The system has a natural equilibrium position where the mass and the restorative force of the spring are at static rest. Displacement of the mass away from the natural equilibrium position by a distance of x causes a restorative force F_1 proportional to the magnitude of the displacement and opposite in direction given by Hooke's law $F_1 = -kx$ where k is the spring constant. The force F_1 causes a potential energy $U(x)$ to be stored in the mass-spring system given by

$$U(x) = -\int_0^x F_1 du = \frac{k}{2}x^2. \tag{18}$$

Potential energy is continuously transferred to kinetic energy as the mass is moving from the stretched or compressed position to the equilibrium position. Denote the velocity of the mass as v, kinetic energy F_2 in the system with mass m at velocity v is given by $\frac{1}{2}mv^2$. By Newton's second law of motion, one obtains $F_2 = ma$. Since $a = \frac{d^2x}{dt^2}$, one has

$$F_2 = m\frac{d^2x}{dt^2}. \tag{19}$$

The conservation of energy means the sum of these two forces should be equal in magnitude and opposite in direction,

$$m\frac{d^2x}{dt^2} = -kx. \tag{20}$$

Normalizing both side by the mass m and re-writing $\frac{k}{m} = \omega^2$, where ω is the angular frequency, one has a linear second-order differential equation

$$\frac{d^2x}{dt^2} + \omega^2 x = 0. \tag{21}$$

The general solution of Eq. (17) giving the position x of the mass as a function of time t is

$$x(t) = Ae^{i\omega t} + Be^{-i\omega t}. \tag{22}$$

Using Euler's identity $e^{i\pi} = -1$, the above solution $x(t)$ can be rewritten in trigonometric terms. The period of motion is given by $T = \frac{2\pi}{\omega}$ and the frequency is given by $f = \frac{1}{T} = \frac{\omega}{2\pi}$. Should the system be subjected to damping, a positive damping constant β appears in the linear term $\frac{dx}{dt}$

$$\frac{d^2 x}{dt^2} + \frac{\beta}{m}\frac{dx}{dt} + \omega^2 x = 0. \tag{23}$$

Setting $\frac{\beta}{m} = 2\lambda$, the damped case Eq. (19) further splits into three subcases according to the rates of deceleration given by the parity of $\lambda^2 - \omega^2$. For example, in the underdamped case where $\lambda^2 - \omega^2 < 0$, the solution becomes infinitely periodic and the oscillation flattens as $t \to \infty$

$$x(t) = e^{-\lambda t}\left[c_1 \sin\sqrt{\lambda^2 - \omega^2}t + c_2 \cos\sqrt{\lambda^2 - \omega^2}t\right]. \tag{24}$$

Overall, the free classical harmonic oscillator is an idealization of a perpetual motion machine in which energy exchanges forms without incurring any loss. In laboratories, retarding elements such as friction cause damping and eventually results in a cessation of periodic motion.

4 Basic Newton-Keplerian planetary model

The angular frequency ω in the harmonic oscillation relation Eq. (21) model is closely related to Kepler's laws of planetary motion and available astronomical data. Kepler's third law states that the ratio between the square of a planet's orbital period T and its semi-major axis a is constant within the solar system. Let M_s denote the mass of the Sun, m denote the mass of a planet, r be its mean distance from the star, and ω be the angular velocity. Since the gravitational force is balanced by the centrifugal force, Newton's law of universal gravitation means

$$mr\omega^2 = G\frac{M_s \bullet m}{r^2}. \tag{25}$$

Theorem 1 *The square of the period T is proportional to the cube of the mean distance from the star.*

Proof The angular frequency is defined as the mean angular speed measured in radians, thus one has

$$\omega = \frac{2\pi}{T}. \tag{26}$$

Squaring the above relation gives

$$\left[\frac{2\pi}{T}\right]^2 = G\frac{M_s}{r^3}. \tag{27}$$

Inverting the above relation, one has

$$\frac{r^3}{T^2} = G\frac{M_s}{4\pi^2}. \tag{28}$$

Since M_s and G are constants, one has the proportionality $T^2 \propto r^3$. This completes the proof. ∎

In most solar systems, the orbits of the planets around their stars are elliptical. Let a and b denote the semi-major and semi-minor axes of the two-body system respectively, then one has the following theorem:

Theorem 2 *The orbital period T is a function of both the semi-major and semi-minor axes a and b, respectively.*

Proof Since $M_s \gg m$, one can simplify the dynamics of the system as a two-body system with specific angular momentum $\mathbf{h}$ being the angular momentum $\mathbf{L}$ normalized by the mass of the planet, $\mathbf{h} = \frac{\mathbf{L}}{\mathbf{m}}$. Let $h = \|\mathbf{h}\|$. In a gravitationally bounded two-body system where the planet's displacement can be measured by angular movement, the infinitesimal operators of angle $d\theta$, time dt, and area dA are related as follows

$$
\begin{aligned}
dt &= \frac{r^2}{h}d\theta, \\
dA &= \frac{r^2}{2}d\theta.
\end{aligned}
\tag{29}
$$

Simplifying the differential operator relations above, one has $dt = \frac{2}{h}dA$. Further simplification leads to

$$
\frac{dA}{dt} = \frac{h}{2}.
\tag{30}
$$

Since h as the modulus of the mean normalized angular momentum specific to a planet-star system in question is a constant, this implies that the rate of change in the area swept out by the planet rotating around the star $\frac{dA}{dt}$ is also a constant, thus giving the Kepler's second law: the planet sweeps out equal areas in equal time. Performing a contour integration of the above identity over the full elliptical orbit of the planet to obtain the orbital period T, one has the following geometric identity relating the orbital period to the semi-major and minor axes and angular momentum:

$$
T = \oint dt = \oint \frac{2}{h}dA = \frac{2\pi ab}{h}.
$$

∎

In subsequent discussions, the mean angular displacement is denoted by $\overline{\omega}$ where $\overline{\omega} = \frac{T}{C}$ where the normalizing factor C is the cycle number in one complete orbital revolution. Additionally, the mean angular displacement is denoted by ω. With these notations, Kepler's second law states that ω is given by

$$
\omega = \frac{2\pi ab}{t\sqrt{p}\sqrt{M_s + m}},
\tag{31}
$$

where p is a parameter specific to the elliptical orbit of the two-body system related to the semi-major and semi-minor axes a and b by $p = \frac{b^2}{a}$ and t is the time for one orbit revolution. Normalizing the semi-major axis a to unity, the above relation reduces to $\frac{b}{\sqrt{p}} = 1$. Hence, Eq. (29) gives $\omega = \frac{2\pi}{t\sqrt{M_s+m}}$. Reducing the mass of the Sun to unity, one has $\omega = \frac{2\pi}{t\sqrt{1+m}}$. Given the periodic nature of harmonic oscillators described above and the numerous applications across disciplines discussed in the last section, it is natural to infer that the cycles of seasonal changes such as annual cycles of monthly median temperature or lengths of daylight durations can also

be modelled by harmonic oscillations. Absent any sudden cosmic changes that adversely affect meteorological conditions such as monthly temperature, seasonal precipitations, or atmospheric pressure in a region, most data variables follow a periodic pattern within a year whose anomalies can be corrected by parametric statistics such as goodness of fit tests or analysis of variance tests. To that end, a rational model for monthly median temperature T in a region can be reasonably given by the coefficient-free harmonic oscillator

$$T(m) = \frac{iA}{2}\left[e^{-i\omega(m-c_1)} - e^{i\omega(m-c_1)}\right] + \frac{B}{2}\left[e^{i\omega(m-c_2)} + e^{-i\omega(m-c_2)}\right] + D. \tag{32}$$

In the above model, m is time given in months, the amplitude given by $\sqrt{A^2 + B^2}$ is modeled as an approximation of the difference between the highest monthly median temperature and the lowest monthly median temperature, the period of the oscillation is given by $\frac{2\pi}{\omega}$, and the undetermined coefficients c_i's are phase shifts and D is the vertical shift. Certain observations about a time-dependent atmospheric dynamics model are worthy of mention. First, since meteorology variables in shorter time intervals exhibit more random behaviors, the longer interval of months is used. Hence, there arises the consideration of which method of time-keeping should be utilized, be it the lunar calendar or other types. Second, once a particular method of time-keeping is selected, there is the additional consideration of the uneven residue of the orbital period of the Earth around the Sun and how the residue is accounted for in the model. Third, there is also the consideration of the axial precession of the Earth.

Table 1. Monthly median temperatures of New York City from July 2017 to July 2021

2017		2018		2019		2020		2021	
Month	Temperature	Month	Temperature	Month	Temperature	Month	Temperature	Month	Temperature
		Jan	-1	Jan	0.5	Jan	3.5	Jan	2
		Feb	5.5	Feb	2.5	Feb	4.5	Feb	1
		Mar	5	Mar	6	Mar	8	Mar	7.5
		Apr	10	Apr	13.5	Apr	10	Apr	12.5
		May	19.5	May	17	May	16	May	17.5
		Jun	22.5	Jun	22	Jun	22.5	Jun	23
Jul	24	Jul	26.5	Jul	26.5	Jul	26.6	Jul	24
Aug	23	Aug	26	Aug	23.5	Aug	25		
Sept	20.5	Sept	21.5	Sept	21	Sept	20		
Oct	17.5	Oct	14	Oct	15.5	Oct	13.5		
Nov	8	Nov	7	Nov	6.5	Nov	11.7		
Dec	2	Dec	3.5	Dec	3.5	Dec	3.5		

The precession is caused by the westward movements of the equinoxes along the ecliptic relative to visually fixed stars in the heavens. Although the axial precession has a period of approximately 25,798 years, the choice of longer time variables can introduce anomalies in the model. There exist other considerations in addition to the ones mentioned above. The harmonic oscillator model at present is one among a whole set of mathematical models utilized to understand the complexity of atmospheric dynamics. Hence, the parameters of the model can and should be improved as more accurate data become available. In the present discussion, the monthly median temperature in New York City is utilized. These monthly median temperatures were collected from July 2017 to July 2021 available on public domains [10]. It should be stressed that different public domains may post slightly different monthly median temperatures for the same region. The above temperatures are monthly median temperatures. Observe that the median temperatures

above exhibit sinusoidal wave patterns with maximum between the months of June and August of each year and minimum between the months of December and February. A connected scatterplot in Excel was used to visualize these data points. To further visualize the data, the months were coded with the first month, July of 2017 as 7 and each subsequent month having an increment of 1, thus converting categorical data into ordinal-numeric data ready for modeling and parametric statistical analysis. Notice in the scatterplot, a rapid increase in median temperatures between
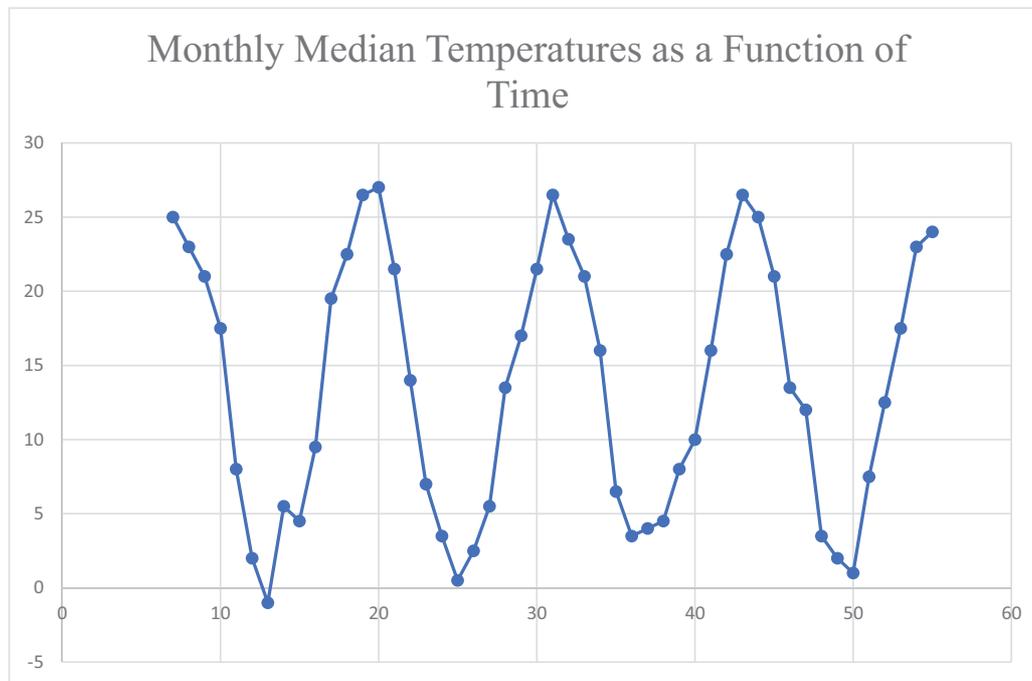
Figure 1. Connected scatterplot of monthly median temperatures of New York City

$5°$ and $20°$ spanned three months out of a year. These three months are typically March, April, and May in the spring. Conversely, a rapid decrease in median temperatures between $21°$ and $7°$ also spanned three months out of a year. These three months are typically September, October, and November in the fall. Also notice that the change in temperature between December of a year through February of the following year is usually less than $6°$ in general. This latter pattern was also exhibited in the three-month period between June and August of a year. For example, in the year 2019, the median temperatures in June, July and August were $21.5°$, $26.5°$, and $23.5°$ respectively. This indicates a temperature increase of only $5°$ in the three-month period in 2019.

This sequential pattern of rapid temperature increases between March, April, and May, followed by a period of minimal change, and the subsequent rapid decrease in September, October, and November which is turn followed by a period of minimal change in December, January, and February coincides well with that of a classical harmonic oscillator. Recall in the harmonic oscillator, the mass moves rapidly as it crosses the equilibrium position and moves slowly when the mass is close to the overstretched or over-compressed positions. Relation (26) gives a workable model of mean angular displacement per unit of time calculated in earth days as $\omega = 0.01720209895 \; radians/day$. In the current model, if one were to use months to minimize the effects of random meteorological anomalies, setting the uneven monthly residue to zero gives $\omega = 0.516062969 \; radians/month$. Allowing for 31 days in a month gives a higher monthly angular displacement of $\omega = 0.533365067 \; radians/month$. To optimize the monthly median temperatures specific to New York City from July, 2017 to July 2021, the intermediate value $\omega = 0.523598776$

was used for the model. In the next section, a visual proximity test was conducted and a parametric statistical test was also conducted to see how closely optimize the actual data. Thus, based on the available data, a harmonic oscillator model of the monthly median temperature T as a function of time measured in month m in New York City from July 2017 to July 2021 can be given as

$$T(m) = \frac{8.95}{2}\left[(1+i)e^{-0.523598776i(m-5.5)} + (1-i)e^{0.523598776i(m-5.5)}\right] + 13.4. \tag{33}$$

In this region-specific model, it is assumed that the phase angle $\phi = \tan^{-1}\left[\frac{A}{B}\right]$ is $\frac{\pi}{4}$ radians, thus setting $A = B$ in the coefficient-free oscillator in Eq. (30). The magnitude of the amplitude is modelled by trimmed minimum of the data set with the extreme cold month of January 2018 removed. Thus, the difference between the maximum and trimmed minimum is $26.6L' - 0.5L' = 26.1L'$ with an observed difference of $\frac{26.1L'}{2} \cong 13.1L'$. To fit the data sufficiently, the model utilized an amplitude of

$$\sqrt{A^2 + B^2} = \sqrt{8.95^2 + 8.95^2} \cong 12.7L' \tag{34}$$

whose difference is comfortably within 97% of the observed amplitude.

5 Numerical simulations and statistical analysis of the model

To test how closely the harmonic oscillator model approximates the actual data. A visual proximity test by superimposing the predicted temperatures and the actual data on the same graph shows the feasibility of the harmonic oscillator model trained on actual data when sufficient data were utilized as training data.
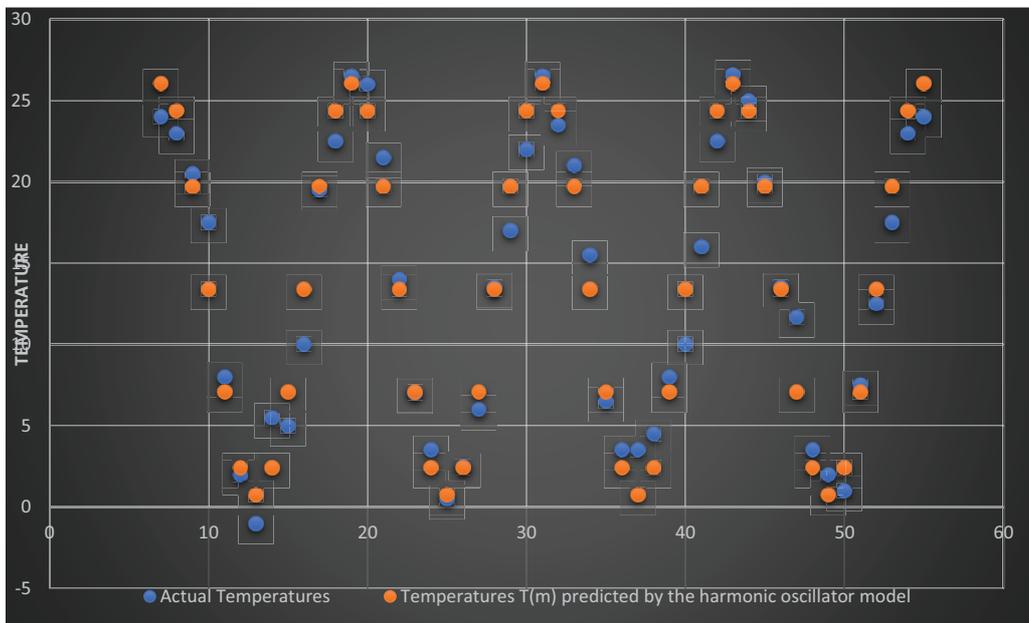
Figure 2. Comparison of monthly median temperatures of New York City compared to predicted by the harmonic oscillator model

Notice that in the overstretched and over-compressed regions in the visual proximity graph, the model's predicted temperatures $T(m)$ given as orange points approximate the actual temperature data given as blue dots well. In the equilibrium position, some variations can be observed.

For example, in October of 2017, the predicted median temperature was 13.399° and the actual recorded temperature was 17.5°. This pair constitutes the pair with the largest deviation. During the other months near the equilibrium point, the deviations were visibly smaller.

To complete the statistical analysis, a goodness of fit test was utilized to examine the predictive power of the harmonic oscillator model. The goodness of fit test is a parametric test that measures the deviation of the data points E expected by a hypothesized model and the actually observed field data O using the χ^2 distribution. The test statistic is given by

$$\chi^2 = \sum \frac{(O - E)^2}{E}. \tag{35}$$

For example, for the months between July 2017 and January 2018, the observed monthly median temperatures and the expected monthly median temperatures are given in the following table of partial results.

Table 2. Partial observed vs. predicted monthly temperatures of New York City

Year	Coded Month	Observed Temperature O	Expected Temperature E	O - E	$\frac{(O\text{-}E)^2}{E}$
2017	7	24	26.0572	-2.0572	0.1624
	8	23	24.3615	1.3615	0.0761
	9	20.5	19.7286	0.7714	0.0302
	10	17.5	13.4	4.1	1.2545
	11	8	7.0714	0.9286	0.1219
	12	2	2.4385	-0.4385	0.0789
2018	13	-1	0.7428	-1.7428	4.0891

Recall that the χ^2 distribution is a right-skewed distribution and that the goodness of fit test based on the χ^2 distribution is premised on the null hypothesis that the observed values are close to the expected values predicted by the model. Thus, at a preset significance level, one rejects the null hypothesis if the χ^2 the test statistic is high, implying significant deviations between the observed and the predicted values. Conversely, a low test statistic fails to reject, implying the observed values and the predicted values are statistically and sufficiently close not to warrant a rejection at the preset significance level.

With a degree of freedom at $k - 1 = 48$, at $\alpha = 0.10$ significance level, the χ^2 test fails to detect significant statistical differences between the actual monthly median temperatures and the predicted monthly median temperatures by the model at 90% confidence level demonstrating the consistency of the model. To further test the consistency of the model, the author put the model to the test with higher statistical rigor at the $\alpha = 0.05$ significance level. The test statistic is 34.7312 with a p-value above the null hypothesis rejection level again. Thus, both goodness of fit statistical tests failed to detect significant statistical differences between the actual monthly median temperatures and the predicted monthly median temperatures by the model at the 90% and 95% confidence levels. In addition to consistency, goodness of fit tests were also performed on historical data entirely outside of the training data used in the construction of the model in order to test the model's predictive power. Actual observed monthly median temperatures in New York City in the 12-month period from August, 2021 to July 2022 from the same public domain were statistically compared with model's predicted temperature at both the $\alpha = 0.05$ significance level and the $\alpha = 0.01$ significance level. At both significance levels, the goodness of fit test gave an F-ratio of 0.0605 and could not detect statistically significant differences between the

predicted median monthly temperature and the observed median temperature thereby supporting the predictive power of the cost-effective harmonic oscillation model.

6 Discussion and further research directions

The current research gives a vantage point to investigate the feasibility of using harmonic oscillators to model atmospheric dynamics, thus opening the possibilities in forecasting energy consumption and other areas of public policy related to climate science. Natural questions include but are not limited to in what ways can the model be improved; should damping be present in the model; if so, what are the causes of damping; what mechanism should be used in the model to account for axial precession of Earth, and can the same approach be utilized to analyze periodic cycles of annual precipitation. Numerous other questions and considerations exist in this direction. As such, the current paper is a starting point. This section will first discuss certain climate-specific research directions in the future, then it will broaden the discussion to include certain philosophical implications in the generalized harmonic oscillator model going forward. Based on the research presented, a possible area of future research to refine the current harmonic oscillator model is to fine-tune the mid-term monthly model by incorporating fluid dynamics with fractional calculus. In many areas of applied mathematics, fractional calculus has increasingly proved to hold explanatory power. For example, Ahmed et al. [11] have recently discovered a fractional calculus model using Caputo-Fabrizio fractional-order PDE to study the evolution of the cholera epidemic. The model thus derived contained a system composed of four fractional differential equations with input variables such as size of susceptible population and symptomatic infected population. Moreover, a theorem of the uniqueness and existence of solutions to the Caputo-Fabrizio fractional-order cholera model was also found. In atmospheric science, it is well-established that shorter-term atmospheric behaviors such as weekly median temperature or weekly precipitation in extreme weather conditions such as hurricanes can be modeled more accurately by hydrodynamic ensembles. For example, a hurricane is a moving frame of low hydrodynamic depressions with high atmospheric convection and an enclosed equatorial atmospheric circulation spreading to higher latitudes. In the analysis of these lower-than-mid-term projections, turbulent fluids in the hurricane column such as air or water vapors observe fractional conservation of mass in hydrodynamic behaviors given by

$$-\frac{\partial^\alpha \rho q_i}{\partial x_i^\alpha} = \frac{\Gamma(\alpha+1)}{\acute{x}^{\alpha+2}} \frac{\partial}{\partial t} \left(\acute{x}^3 n\rho \right). \tag{36}$$

In the above equation, the standard control volume's 3D lengths are given as $\acute{x}_i$ for $i = 1, 2, 3, \rho$ is the mean air/water vapor density, q_i is the specific hydrodynamic discharge passing through the i-th face of the standard control volume, $\Gamma(\alpha+1)$ is the gamma function, n is the porosity parameter of the air/water vapor mixture, and $\frac{\partial^\alpha}{\partial x_i^\alpha}$ is the α-th fractional derivative.

For example, Wheatcraft et al. [12] were able to utilize the above fractional calculus relation to study non-linear hydrodynamic flux in control volumes. The research showed that the hydrodynamic divergence term in the fractional mass conservation equation is the fractional convergence and that the scaling term in the fractional conservation of mass is in fact scale-invariant. Combining these latter considerations together with the research presented in the current paper, a more comprehensive mid-term or short-to-mid-term atmospheric model can be reasonably given by a system of both differential and fractional differential equations in which the master harmonic oscillator model first gives the overall projections, then the fractional derivative transient model modifies the above macro projections with more short-term refinement.

Another further research direction is to consider the philosophical nature of physical variables

present in the model. For example, is the meaning of mass m used in the definition of angular frequency ω limited to the total mass of the gravitationally bound two-body problem consisting of Sun and Earth with the negligible mass of the Moon absorbed into the approximation for computation purposes or are there more generalized epistemological ensembles that can account for harmonic oscillations in these vastly disparate natural and social phenomenological settings. This is a particularly important direction of future research based on the presented results. A review of the literature on the applications of the harmonic oscillator in natural sciences and social sciences shows that the meaning of the modelled variables can have discipline-specific interpretations in some cases while eludes satisfactory interpretations in others. For example, an application of the harmonic oscillator where the authors gave a plausible interpretation of the meaning of the modelled variables is a quantum spatial-periodic harmonic model of equity market with price limits by Meng et al. [13]. The authors of the study examined the price action movement of equities modelled as a fundamental particle moving, vibrating, and undergoing excitations in a quantum potential well with energy given as

$$V(x) = \frac{m\omega^2 x^2}{2}. \tag{37}$$

The authors reasoned that the energy band structures of the quantum harmonic oscillator model correspond to the non-linear market relations such as inter-band positive correlations and intra-band negative correlations between volatility and the transaction volume in unit time. The probability of locating the equity at price x modelled as the statistical location of the fundamental particle is given as the square modulus of the generalized Schrodinger wave function with a duly coupled Gaussian

$$|\varphi(x)|^2 = \sqrt{\frac{m\omega}{\pi\hbar}} e^{-m\omega x^2/\hbar}. \tag{38}$$

The authors posited that the modelled variables m, ω, $\hbar$ are interpreted as the mean total market capitalization of the equity, period of the equity fluctuation cycles, and most importantly the uncertainty of the irrational transaction volume of the equity respectively. To fit the model with a market circuit where trading is halted in the presence of extreme equity valuation volatilities, the authors further imposed a periodic boundary condition when the price exceeds \$d above the mean daily moving average or drops \$d below the identical mean daily moving average given as

$$\varphi(x) = e^{-ikd}\varphi(x+d). \tag{39}$$

where k is the one-dimensional Bloch wave number. With the market circuit thus modelled, the price of the equity is given as

$$\varphi(\xi) = e^{-\frac{\xi^2}{2}} \left[A \bullet H_1(\xi) + B \bullet H_2(\xi) \right]. \tag{40}$$

where H_i's are the first and the second Hermite polynomials and the parameters A and B are the constants specific to the equity in consideration. Thus modelled, the authors were able to obtain the solution of the equity wave function in Eq. (38) as

$$\left[H_1\left(\frac{-\beta d}{2}\right) - e^{-ikd}H_1\left(\frac{\beta d}{2}\right) \right] \bullet \left[H_2'\left(\frac{-\beta d}{2}\right) - e^{-ikd}H_2'\left(\frac{\beta d}{2}\right) + \beta d e^{-ikd}H_2\left(\frac{\beta d}{2}\right) \right], \tag{41}$$

where β is a re-parametrization constant coupling the price of the equity to ξ

$$\xi = \beta x = \sqrt{m\omega/\hbar \bullet} x. \tag{42}$$

It is worthy of consideration to note that the authors further inferred from the model that the reduced Planck constant $\hbar$ is the limit of irrational agency in equity trading, that is, in the perfectly rational market, $\hbar \to 0$.

One can ask: in the limit, is the efficient market hypothesis or EMH tenable? If financial information is made transparent, symmetric, and accessible to all market participants such as to all institutional investors and individual investors, can the rational market be a frictionless harmonic oscillator where the only uncertainty is the most natural quantum price fluctuations captured by a geometric Brownian motion X_t?

$$\varphi(\xi) \cong Ae^{i\xi t} + Be^{-i\xi t} + X_t. \tag{43}$$

If these conditions are tenable, then the logarithmic price of an equity P_t at time t is a simple linear sum of the expected value of the equity E_t and the stochastic discount factor M_t

$$\log P_{t+1} = \log M_t + E_t \bullet \log P_t. \tag{44}$$

A natural question is: what would a rational trading strategy be in the stated limit? These and other long-term considerations should be investigated by the mathematical modelling community.

Declarations

List of abbreviations

Not applicable.

Ethical approval

The author states that this research complies with ethical standards. This research does not involve either human participants or animals.

Consent for publication

Not applicable.

Conflicts of interest

The author confirms that there is no competing interest in this study.

Data availability statement

Data availability is not applicable to this article as no new data were created or analysed in this study.

Funding

Not applicable.

Author's contributions

The author has made substantial contributions to the conception, and design of the work, the acquisition, analysis, interpretation of data, and the creation of new software used in the work.

References

[1] Yilmaz, B. Generative adversarial network for load data generation: Tükiye energy market case. *Mathematical Modelling and Numerical Simulation with Applications*, 3(2),141-158, (2023). [CrossRef]

[2] Koshkin, S. and Meyers, I. Harmonic oscillators of mathematical biology: many faces of a predator-prey model. *Mathematics Magazine*, 95(3), 172-187, (2022). [CrossRef]

[3] Arpa, E.M. and Durbeej, B. HOMER: a reparameterization of the harmonic oscillator model of aromaticity (HOMA) for excited states. *Physical Chemistry and Chemical Physics*, 4(25), 95-127, (2023). [CrossRef]

[4] Özdemir, Ş.K., Liu Y., Miranowicz, A. and Imoto, N. Kraus representation of a damped harmonic oscillator and its application. *Physical Review A*, 70(4), 54-92, (2004). [CrossRef]

[5] Kiselyov, V.V., Versteyhe, S., Gauguin, L. and De Meyts, P. Harmonic oscillator model of the insulin and IGF1 receptors' allosteric binding and activation. *Molecular Systems Biology*, 5(243), 773-806, (2009). [CrossRef]

[6] Mishra, V., Nakul, N. and Adlakha, N. Finite volume simulation of calcium distribution in a cholangiocyte cell. *Mathematical Modelling and Numerical Simulation with Applications*, 3(1), 17-32, (2023). [CrossRef]

[7] Shah, N.A., Popoola, A.O., Oreyeni, T., Omokhuale, E. and Altine, M.M. A modelling of bioconvective flow existing with tiny particles and quartic autocatalysis reaction across stratified upper horizontal surface of a paraboloid of revolution. *Mathematical Modelling and Numerical Simulation with Applications*, 3(1), 74-100, (2023). [CrossRef]

[8] Orhan, H. and Yavsan, E. Artificial intelligence-assisted detection model for melanoma diagnosis using deep learning techniques. *Mathematical Modelling and Numerical Simulation with Applications*, 3(2),159-169, (2023). [CrossRef]

[9] Joshi, H., Yavuz, M. and Stamova, I. Analysis of the disturbance effect in intracellular calcium dynamic on fibroblast cells with an exponential kernel law. *Bulletin of Biomathematics*, 1(1), 24-39, (2023). [CrossRef]

[10] Current Results weather and science facts. New York City Temperatures: Averages by Month *https://www.currentresults.com/Weather/NewYork/Places/new-york-city-temperatures-by-month-average.php*, Access Date: 25th April 2023.

[11] Ahmed, I., Akgül, A., Jarad, F., Kumam, P., and Nonlaopon, K. A Caputo-Fabrizio fractional-order cholera model and its sensitivity analysis. *Mathematical Modelling and Numerical Simulations with Applications*, 3(2), 170-187, (2023). [CrossRef]

[12] Wheatcraft, S.W. and Meerschaert, M.M. Fractional conservation of mass. *Advances in Water Resources*, 31(10), 1377-1381, (2008). [CrossRef]

[13] Meng, X., Zhang, J.W., Xu, J. and Guo, H. Quantum spatial-periodic harmonic model for daily price-limited stock markets. *Physica A: Statistical Mechanics and its Applications*, 438, 154-160, (2015). [CrossRef]

Mathematical Modelling and Numerical Simulation with Applications (MMNSA)
(https://dergipark.org.tr/en/pub/mmnsa)

RESEARCH PAPER

Understanding the mathematical background of Generative Adversarial Networks (GANs)

Bilgi Yilmaz [ID] [1,2,*,‡] and **Ralf Korn** [ID] [1,3,‡]

[1]Department of Mathematics, Kaiserslautern Technical University, Kaiserslautern, Germany,
[2]Department of Mathematics, Faculty of Science, Kahramanmaraş Sütçü Imam University, Türkiye,
[3]Department of Financial Mathematics, Fraunhofer ITWM, Germany

*Corresponding Author
‡ yilmaz@mathematik.uni-kl.de (Bilgi Yilmaz); korn@mathematik.uni-kl.de (Ralf Korn)

Abstract

Generative Adversarial Networks (GANs) have gained widespread attention since their introduction, leading to numerous extensions and applications of the original GAN idea. A thorough understanding of GANs' mathematical foundations is necessary to use and build upon these techniques. However, most studies on GANs are presented from a computer science or engineering perspective, which can be challenging for beginners to understand fully. Therefore, this paper aims to provide an overview of the mathematical background of GANs, including detailed proofs of optimal solutions for vanilla GANs and boundaries for f-GANs that minimize a variational approximation of the f-divergence between two distributions. These contributions will enhance the understanding of GANs for those with a mathematical background and pave the way for future research.

Keywords: Generative adversarial networks; unsupervised learning; qualitative analysis
AMS 2020 Classification: 91G15; 91G20; 91B02; 62E99

1 Introduction

Generative Adversarial Networks (GANs) introduced by [1] consist of generative and discriminative neural network models that are usually denoted by letters G and D, respectively. To visualize GANs environment better in our mind, the generative model may be regarded as a counterfeiter who is attempting to produce a fraud Van Gogh's Starry Night painting and sell it without being noticed, whereas the discriminative model is equivalent to an expert who specializes in Van Gogh, trying to detect the counterfeit fraud painting. However, the counterfeiter does not care about producing images that are a variation of the original Starry Night painting. In the applications of GANs, the aim is not to present a new image identical to the original painting. Instead, it

aims to create a unique illustration of Starry Night that the Van Gogh expert recognizes as an unknown Van Gogh painting that is unprecedented anywhere before. As a result, a computation starts between the generator and discriminator over the fraud painting detection. The competition continues until the counterfeiter becomes intelligent enough to deceive the expert successfully. More precisely, the discriminator's role is to distinguish the real and fraud paintings, while the generator's role is to generate fraud paintings in such a way that it can mislead the discriminator, and the discriminator is unable to cope with rejecting the fraud paintings any longer (see Figure 1).
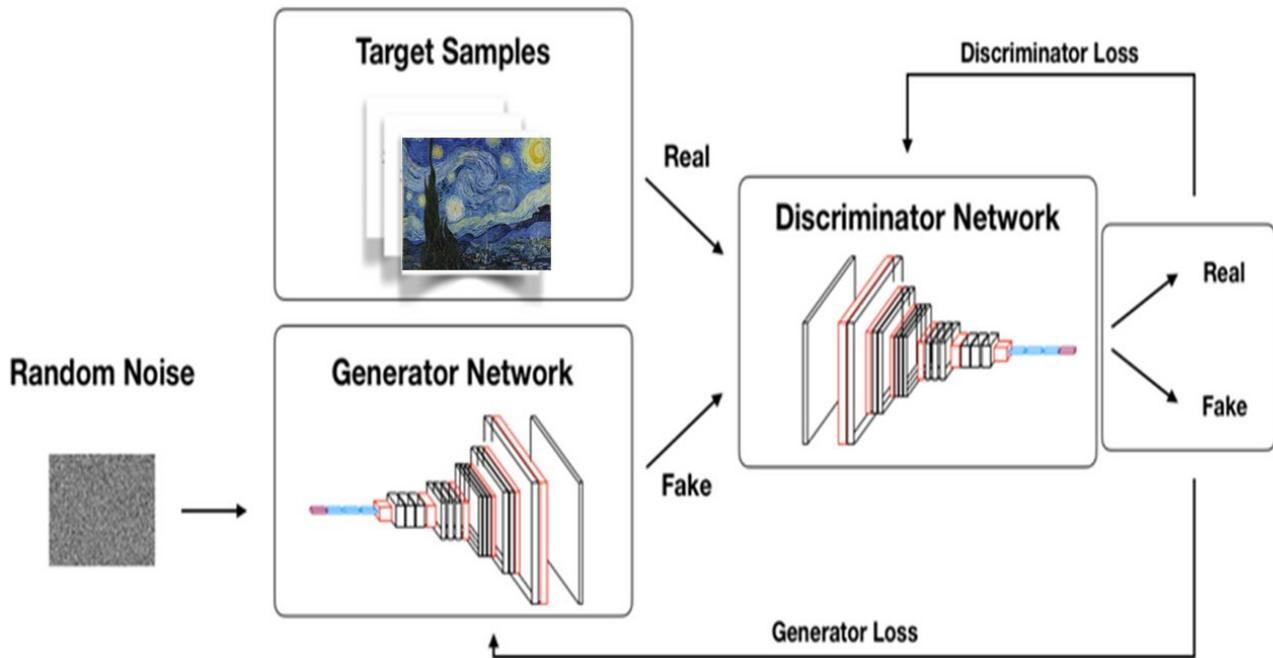
Figure 1. A visualization of the discriminator and generator networks as a counterfeiter and Van Gogh's painting expert

Figure 1 presents a visualization of the training process of Generative Adversarial Networks (GANs). The GAN training process involves two main components: the Generator and the Discriminator. The Generator takes random noise from the latent space as input and generates fake data, attempting to mimic the real data distribution. Initially, the generated data is random and typically of low quality. A batch of real data is sampled from the training dataset that serves as the ground truth for the Discriminator during training. The Discriminator is trained on both real and fake data. It is presented with the real data and the corresponding labels (1 for real) to learn to distinguish real data from fake data. It is then presented with the fake data generated by the Generator and the corresponding labels (0 for fake) to learn to identify the fake data. The Discriminator's performance is evaluated using a loss function, such as binary cross-entropy, which measures how well the Discriminator is differentiating the real data from the fake data. The Generator is trained to deceive the Discriminator by generating fake data that appears as realistic as possible. The Generator takes random noise as input and aims to generate data that the Discriminator labels as real. The Generator's performance is evaluated using the Discriminator's response to the fake data it generates. The Generator's loss function encourages it to generate data that deceives the Discriminator (i.e., the Discriminator's prediction closer to

1). The model parameters of both the Generator and Discriminator are updated using gradient descent or some variant, optimizing considering loss functions. The process continues iteratively, with the Generator getting better at generating realistic data, and the Discriminator becoming more skilled at differentiating real from fake data. The ideal state is reached when the Generator can create data that is indistinguishable from real data, and the Discriminator cannot confidently classify between the real and generated data. It is important to note that the Generator never uses the real data as input and trains solely with random noises.

To put this in a positive framework, we can say that the discriminator serves as a kind of quality control of the generated data. The better the discriminator performs, the better the benchmark for the generator. Then, the generator can finally beat the benchmark in a form in which the optimal strategy of the discriminator is essentially only guessing whether the generated data are fraud or real. Finally, the generator is ready to be used in synthetic data generation.

Some of the key issues that are critical in the applications of GANs are as follows:

- Quantifying "similar objects" is trickier than it sounds, and it carries the core to GANs. In the field of mathematics research, we have many alternative methods to quantify the similarity between any two objects, which can also lead us to different objectives in setting up GANs.
- In the application of GANs, we aim to generate original objects, which can be distant in which distance measure we consider to any objects at hand as a training dataset χ (i.e. we do not want to copy χ, but we feel the generated objects and χ belong to the same class).
- We do not care about generating a perturbation of the original painting. Instead, we want to produce a fake painting that the expert is going to consider like a unique painting that belongs to Van Gogh, which she has seen for the first time in her life.
- In this setting, the appropriate concept of similarity is distributional similarity. We call two objects similar if both are samplings from the same (or roughly same) probability distribution. This means that the two objects share similar characteristics and features that are determined by the underlying probability distribution. Therefore, we maintain a training dataset denoted with $\chi \subset \mathbb{R}^n$ that consists of samples gathered from μ. In this context, μ is a probability distribution, and its density is represented by $p(x)$. We want to arrive at a reasonable approximating probability distribution ν having a density $q(x)$ to μ. Then, we can obtain artificial or synthetic objects that are identical to objects in the training (real) dataset χ by sampling from ν.
- You may question, why we do not just consider the distributions as $\nu = \mu$ and obtain samples from the real data distribution μ.
  Unfortunately, such sampling is exactly the main problem of GANs since μ *is not known explicitly*. The only thing that we know is that we have a finite set of samples χ sampled from μ. Consequently, the actual issue is identifying the properties of μ by only using χ. In this sense, we must focus on specifying an appropriate probability distribution ν as an approximation process to μ.
- In addition to considering a distribution similar to μ in the sense of probability distances, one can also try to characterize μ by the empirical behavior of the data, their so-called *stylized facts*.
- Generally, the success of GANs depends on the sophistication of μ and the training dataset χ size.

The basic approach of GANs

The purpose of the study is to clarify the mathematical background of GANs. Therefore, it focuses on only theoretical aspects of GANs and contains any applications.

To approximate a given probability distribution μ, GANs require an initially defined probability distribution to start its training. Generally, the initial distribution, which we define as γ, is

introduced in space $\mathbb{R}^d$. Here, the space dimension d is not necessarily identical to the space dimension n (of $\mathbb{R}^n$). Now, suppose we have chosen the initial distribution γ to be the standard normal distribution, and we have denoted it with $N(0, \mathcal{I}_d)$. However, we are free to choose γ from other well-known probability distribution families (e.g., uniform). GANs utilize a technique to discover a mapping G, defined as $G : \mathbb{R}^d \longmapsto \mathbb{R}^n$. At this stage, consider a random variable $z \in \mathbb{R}^d$ sampled from initial distribution γ. Then, we can claim that the mapping $G(z)$ is from the same distribution family as μ. To emphasize, the probability distribution of $G(z)$ can be defined in the form of $\gamma \circ G^{-1}$. Here, G^{-1} denotes the inverse of G, and the inverse maps subsets of space $\mathbb{R}^n$ to subsets in space $\mathbb{R}^d$. Therefore, in the GANs modeling method, we desire to find a mapping $G(z)$ that satisfies $\gamma \circ G^{-1} = \mu$ or at least $\gamma \circ G^{-1}$ is a reasonable approximation of the real data distribution μ.

The vanilla GAN approach forms an adversarial system from which the generator receives updates on a continuous basis to increase output accuracy. More rigorously, the vanilla GAN presents a neural network called a discriminator, which attempts to label the observed samples as real, and generated samples as fake. From this perspective, the discriminator behaves like a classifier that attempts to distinguish real samples from fake samples. To this end, the discriminator assigns a probability $D(x) \in [0, 1]$ to each sample x for its probability of being a real sample. If samples $G(z_j)$ are outputs of the generator, the discriminator attempts to restrict them since they are fake samples.

In the early stage of training a GAN, restricting generated samples as fake should not be challenging since the generator is not elegant at generating realistic samples. However, after each attempt G fails to produce realistic samples to trick D, and G learns and adjusts itself with a refinement update. Thus, the improved G performs more reasonably compared to the one used at the early stages, and then it is the discriminator D's progression to revise for refinement. In an ideal case, through such an adversarial iterative process, we can eventually arrive at an equilibrium point; therefore, even the most reasonable D cannot perform more satisfactory labeling than a random guess. At this point, the samples generated by G become extremely identical to training samples χ in distribution. Consequently, the discriminator decision becomes completely random, and the probability of being real approximates 50%.

In GANs modelling approach, we have to define both the discriminator and generator by utilizing neural networks to understand the distributional properties of given data. Each neural network has its corresponding parameters ω and θ. These parameters are used in the training of the discriminator and generator and include the weights (also known as synaptic weights) of the neural network layers, as well as the biases of these layers. They are learned during training to optimize the performance of the GAN in generating realistic samples. Hence, we should register $D_\omega(x)$ for the discriminator and $G_\theta(z)$ for the generator, and we should denote $\nu_\theta := \gamma \circ G_\theta^{-1}$. Thus, it is clear that our task is to identify the desired generator $G_\theta(z)$ by adequately adjusting its parameter θ.

Building a GAN framework

As we mentioned above, there are two parties in GANs modeling method: generator $G_\theta(z)$ and a discriminator $D_\omega(x)$ who are in competition, and both parties have their own roles during the modeling process. More specifically,

The generator:

- The generator operates with a random vector whose length is fixed and, then, produces a fake sample in the corresponding domain.
- The vector is sampled from the Gaussian distribution (generally) and utilized to seed the

generator. After the training, points in the multidimensional vector space conform with points in the real data domain, forming a compact replica of the training data distribution.

- The vector space is called the latent space or equally vector space. It consists of some latent variables or some hidden variables, which are critical for the domain but cannot be observed directly.

The discriminator:

- The discriminator uses a sample from the domain as input (it may be either real or fake) and assigns a real or fake (generated) binary class label.
- The real sample directly comes from the original data, while fake samples are only outputs of the generator.
- The discriminator is a classifier model. When the training is finished, the discriminatory model is junked as we are curious about in the generator. Occasionally, the generator can be reset as it has learned to effectively determine characteristic from examples sampled from the problem domain. Some or all of the characteristics extraction layers can be utilized in transfer learning applications by utilizing the same or similar input data.

Both players in the min-max game are expressed by a corresponding function. Each function is differentiable concerning its inputs and parameters. As it introduced above, the discriminator is a differentiable function denoted by D that uses x as input and is allowed to use only the discriminator network weights ω as parameters. On the other hand, the generator is specified by G and uses the random vector z as the initial input and is only allowed to use the weights of the generator network θ as parameters [2].

In this setting, both players have their own loss functions. The loss functions are described with regard to parameters specific to players. The discriminator desires to minimize the problem $L^{(D)}(\omega, \theta)$ and it must accomplish the minimization by controlling only its parameters ω. On the other hand, the generator desires to minimize $L^{(G)}(\omega, \theta)$ and must accomplish the minimization by controlling only its parameters θ. Here, the discriminator and generator losses rely on the other player's parameters. However, both players are limited to controlling only their own parameters. Since each player's loss relies on the opposite player's parameters, despite each player being allowed to regulate its parameters and cannot control the opposite player's parameters, such a scenario is generally expressed as a game rather than a classical optimization problem [2].

As we mentioned already, generator G is a differentiable function. After we produce its random vector z from a well-known initial distribution called γ, G generates a fake sample x, which is implicitly sampled from the model distribution ($\mathbb{P}_{model} = \nu$). Commonly, a deep neural network is utilized to characterize the generator. However, we have some constraints on the configuration of the corresponding neural network. If we want $\mathbb{P}_{model}$ to have complete support on $\mathcal{X}$, the dimension of the generator should be at least as large as the dimension of $\mathcal{X}$ [2].

In a similar fashion, discriminator D is also a differentiable function, whose objective is to categorize samples accurately as real and fake. The discriminator is also naturally characterized by a deep neural network. Again, it has some restrictions on the configuration of its corresponding network. It has to use only real and fake samples as entries and assigns a probability score $D(x) \in [0, 1]$ for each x [2]. Here, notice that the generator never sees the real data and only uses random vector z as input, while the discriminator uses both real, and the generator's output.

A simple derivation of the loss functions

Before starting the definition of the loss functions, note that in the classical GANs architectures, the design of the discriminator loss functions $L^{(D)}$ always remains the same. They differ only by the cost function for the generator, $L^{(G)}$ [2]. The loss function introduced in the original study [1]

is obtained from the binary cross-entropy formula as follows

$$L(\hat{y}, y) = \left[y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y}) \right]. \tag{1}$$

Here, y and $\hat{y}$ correspond to the original and fake data, respectively.

In the training of the discriminator, the label of data assigned by the real data $\mu(x)$ is $y = 1$ (real/observed data) and $\hat{y} = D(x)$. Then, by substituting this into Eq. (1), we have

$$L(D(x), 1) = \log(D(x)), \tag{2}$$

and for the data sampled from the generator, the label is $y = 0$ (fake data) and $\hat{y} = D(G(z))$. Similarly, by substituting these into Eq. (1), we end up with

$$L(D(G(z)), 0) = \log(1 - D(G(z))).$$

In this setting, the goal of the discriminator is to accurately classify its input as fake or real. Therefore, the given loss functions for G and D have to be maximized. Then, the final loss function of D is denoted as

$$L^{(D)} = \max \left[\log(D(x)) + \log(1 - D(G(z))) \right]. \tag{3}$$

At this stage, it is important to remember that the generator is competing against the discriminator. Hence, the generator aims to minimize the optimization problem given in Eq. (3), and consequently, its loss function evolves to

$$L^{(G)} = \min \left[\log(D(x)) + \log(1 - D(G(z))) \right]. \tag{4}$$

Now, let us combine the loss functions (3) and (4). By combining these two equations, we obtain a min-max problem as

$$L = \min_{G} \max_{D} \left[\log(D(x)) + \log(1 - D(G(z))) \right]. \tag{5}$$

Here, it is worth emphasizing that the loss function in Eq. (5) is valid only for a single data point. Therefore, to consider the entire dataset, we need to consider the expectation of the combined loss function as

$$\min_{G} \max_{D} V(D, G) = \min_{G} \max_{D} \left[\mathbb{E}_{x \sim \mu}[\log(D(x))] + \mathbb{E}_{z \sim \gamma}[\log(1 - D(G(z)))] \right]. \tag{6}$$

The min-max formulation introduced in Eq. (6) is a concise one-liner function that intuitively captures the adversarial nature of the competition between the players G and D. However, in practice, individual loss functions are defined for both players since the gradient of $y = \log(x)$ is steeper around $x = 0$ than $y = \log(1 - x)$. This means that trying to maximize $\log(D(G(z)))$, or equivalently minimizing $-\log(D(G(z)))$ leads to quicker and more significant improvements in the generator performance than attempting to minimize $\log(1 - D(G(z)))$.

2 Mathematical description of vanilla GANs

The adversarial game introduced in the previous section can be expressed mathematically by a min-max task for a target function defined by the discriminator $D(x) : \mathbb{R} :\longmapsto [0,1]$ and generator $G : \mathbb{R}^d \longmapsto \mathbb{R}^n$. Here, it is clear that G transforms the random vector $z \in \mathbb{R}^d$ sampled from γ into generated (fake) samples $G(z)$. Then, D attempts to distinguish the generated samples from the training samples that are supposed to be sampled from μ while G attempts to generate new samples that are identical in distribution to the data that we use in the training of GANs [3].

In the original study [1], a target loss function is introduced as

$$V(D,G) := \mathbb{E}_{x\sim\mu}[\log(D(x))] + \mathbb{E}_{z\sim\gamma}[\log(1 - D(G(z)))],$$

where $\mathbb{E}$ represents the expectation concerning the distribution appointed in the subscript. We can avoid the subscript if there is no confusion.

The vanilla GAN solves the min-max problem given in Eq. (6). Heuristically, for a given G, the optimization problem $\max_D V(D,G)$ reveals the optimal D to reject outputs $G(z)$ by assigning higher probabilities to samples from μ and low probabilities to outputs $G(z)$. In contrast, for a given D, $\min_G V(D,G)$, the optimization problem reveals the optimal G, and therefore, the outputs $G(z)$ attempt to deceive D by assigning high probabilities for $G(z)$ [3].

Then, let us define $y = G(z) \in \mathbb{R}^n$ having a distribution defined as $\nu := \gamma \circ G^{-1}$, and the random vector $z \in \mathbb{R}^d$ is from the γ distribution family. Thus, we may rearrange $V(D,G)$ in terms of D and ν as follows

$$
\begin{aligned}
\tilde{V}(D,\nu) : \ &= \ \mathbb{E}_{x\sim\mu}[\log(D(x))] + \mathbb{E}_{z\sim\gamma}[\log(1 - D(G(z)))] \\
&= \ \mathbb{E}_{x\sim\mu}[\log(D(x))] + \mathbb{E}_{y\sim\nu}[\log(1 - D(G(y)))] \\
&= \ \int_{\mathbb{R}^n} \log(D(x)) d\mu(x) + \int_{\mathbb{R}^n} \log(1 - D(y)) d\nu(y).
\end{aligned}
\tag{7}
$$

Then, the min-max problem defined in Eq. (6) evolves to

$$\min_G \max_D V(D,G) = \min_G \max_D \left(\int_{\mathbb{R}^n} \log(D(x)) d\mu(x) + \int_{\mathbb{R}^n} \log(1 - D(y)) d\nu(y) \right). \tag{8}$$

Now, suppose that the distributions μ and ν have densities given as $p(x)$ and $q(x)$, respectively. Note that this can only happen under the condition of $d \geq n$. This condition is necessary for GANs to ensure that the discriminator is sufficiently powerful to distinguish real samples from generated ones. When $d \geq n$, the discriminator possesses a greater number of parameters compared to the sample size in the training dataset. Consequently, this asymmetry facilitates the discriminator's ability to effectively differentiate between real and generated samples. If $d < n$, the discriminator may not effectively learn to distinguish real from generated samples, resulting in poor-quality generated samples.

By using the densities, we obtain

$$V(D,\nu) = \int_{\mathbb{R}^n} \left(\log(D(x)) p(x) + \log(1 - D(x)) q(x) \right) dx.$$

With the help of the current evolution, the min-max problem given in Eq. (6) evolves to

$$\min_{G} \max_{D} V(D,G) = \min_{G} \max_{D} \int_{\mathbb{R}^n} \left(\log(D(x))p(x) + \log(1 - D(x))q(x) \right) dx.$$

From the evolved problem, notice that the equation is equal to $\min_{\nu} \max_{D} \tilde{V}(D, \nu)$ under the condition $\nu = \gamma \circ G^{-1}$ for some generator G.

Proposition 1 ([1]). *For distributions μ and ν on $\mathbb{R}^n$ having densities $p(x)$ and $q(x)$, respectively*

$$\max_{D} V(D, \nu) = \max_{D} \int_{\mathbb{R}^n} \left(\log(D(x))p(x) + \log(1 - D(x))q(x) \right) dx$$

is achieved by $D_{p,q}(x) = \frac{p(x)}{p(x)+q(x)}$ for $x \in supp(\mu) \cup supp(\nu)$.

Proof Let us define the integrand as

$$f(D(x)) = \log(D(x))p(x) + \log(1 - D(x))q(x).$$

To find the optimal solution, we look at the first order condition $\frac{df(D(x))}{dD(x)} = 0$ and second order condition $\frac{df^2(D(x))}{dD(x)^2} = 0$. Hence, let us start with

$$\frac{df(D(x))}{dD(x)} = \frac{p(x)}{D(x)} - \frac{q(x)}{1 - D(x)} = 0.$$

By solving this equality for $D(x)$ we find the critical point

$$D_{p,q}(x) = \frac{p(x)}{p(x) + q(x)}.$$

Now, let us compute the second derivative

$$\frac{d^2 f(D(x)}{dD(x)^2} = \frac{-p(x)}{D(x)^2} - \frac{q(x)}{(1 - D(x))^2}.$$

Then, it is obvious that the second derivative is strictly negative for at least one of $p(x)$ or $q(x)$ being positive. Therefore, we find the optimal solution $D_{p,q}(x)$ as

$$D_{p,q}(x) = \frac{p(x)}{p(x) + q(x)}.$$

∎

As a result of Proposition 1, we can give the following remark immediately.

Remark 1. *The discriminator optimal solution of the min-max problem satisfies $D_{p,q}(x) = \frac{p(x)}{p(x)+q(x)} \in [0, 1]$, and this is the requirement for the optimal discriminator.*

Note that the optimal solution makes the following sense intuitively:

- If some sample x is favorably actual, we may anticipate $p(x)$ to be close to one and $q(x)$ to converge at zero. Hence, the optimal D assigns one to such samples.
- For a generated sample $x = G(z)$, we anticipate the optimal D to assign zero since $p(G(z))$ has to be close to zero. When we train G to its optimal value, density $q(x)$ gets very close to density $p(x)$, i.e. we obtain $D_{p,q}(G(z)) \approx 0.5$.

As a consequence of Proposition 1, we can introduce the following theorem immediately.

Theorem 1. *Suppose $p(x)$ is a probability density function defined on space $\mathbb{R}^n$. Additionally, consider a probability distribution v having a density function denoted as $q(x)$ and a discriminator function $D : \mathbb{R}^n \longmapsto [0,1]$ as usual. Then, we have a min-max problem as follows [3],*

$$\min_{v} \max_{D} \tilde{V}(D,v) = \min_{v} \max_{D} \int_{\mathbb{R}^n} \left(\log(D(x))p(x) + \log(1 - D(x))q(x) \right) dx, \qquad (9)$$

and, we reach a solution with a special choice of $q(x) = p(x)$ and $D(x) = \frac{1}{2}, \forall x \in supp(p)$.

Proof Let us now assume $p(x) = q(x)$ for all $x \in supp(p)$. Then, we have $\bar{D}(x) = 1/2$ and $\int_{\mathbb{R}^n} \log(1/2)p(x)dx = \int_{\mathbb{R}^n} \log(1/2)q(x)dx = -\log(2)$ as both p and q are probability densities. For this special choice of p, q, and D, we obtain

$$\tilde{V}(D,v) = -\log(4).$$

Note further that by the definition of the Jensen-Shannon divergence, we have

$$
\begin{aligned}
0 \leq JS(p\|q) &= 0.5(KL(p\|0.5(p+q)) + KL(q\|0.5(p+q))) \\
&= 2\log(2) + \int_{\mathbb{R}^n} \left(p(x) \log \left(\frac{p(x)}{p(x)+q(x)} \right) + q(x) \log \left(\frac{q(x)}{p(x)+q(x)} \right) \right) dx \\
&= \tilde{V}(D,v) + log(4).
\end{aligned}
$$

Therefore, $\tilde{V}(D,v)$ cannot be smaller than $-\log(4)$. Thus, we have proved that $q(x) = p(x)$ – and thus $\bar{D}(x) = 1/2$ – yields the minimum possible value of $\tilde{V}(D,v)$ for any v for the given choice of $D(x) = p(x)/(p(x) + q(x))$. Consequently, we end up with the desired result. ∎

Theorem 1 reveals that the solution to the min-max problem given by Eq. (9) is the result we seek under the hypothesis of the distributions having the same densities. Theorem 1 holds for all distributions in general.

Theorem 2. *Suppose that μ again is a probability distribution function given on space $\mathbb{R}^n$ as in Theorem 1. Then, for a probability distribution v and a discriminator $D : \mathbb{R}^n \longmapsto [0,1]$, we can introduce a min-max problem as follows [3]*

$$\min_{v} \max_{D} \tilde{V}(D,v) = \min_{v} \max_{D} \int_{\mathbb{R}^n} \left(\log(D(x))d\mu(x) + \log(1 - D(x))dv(x) \right), \qquad (10)$$

whose solution is achieved with the special choice $v = \mu$ and $D(x) = \frac{1}{2} \mu-a.e.$

Proof We first show that with the special choice of $v = \mu$ and $D(x) = \frac{1}{2}$ μ-almost everywhere, the min-max problem in Equation (10) is solved.

First, let's consider the objective function $\tilde{V}(D, \nu)$:

$$\tilde{V}(D, \nu) = \int_{\mathbb{R}^n} \left(\log(D(x)) d\mu(x) + \log(1 - D(x)) d\nu(x) \right).$$

Substituting $\nu = \mu$ and $D(x) = \frac{1}{2}$, we have:

$$
\begin{aligned}
\tilde{V}(D, \nu) &= \int_{\mathbb{R}^n} \left(\log\left(\frac{1}{2}\right) \mu(x) + \log\left(1 - \frac{1}{2}\right) \mu(x) \right) \\
&= \int_{\mathbb{R}^n} \left(-\log(2) \mu(x) - \log\left(\frac{1}{2}\right) \mu(x) \right) \\
&= -\log(2) \int_{\mathbb{R}^n} \mu(x) dx + \log\left(\frac{1}{2}\right) \int_{\mathbb{R}^n} \mu(x) dx.
\end{aligned}
$$

Since μ is a probability distribution function, the integral $\int_{\mathbb{R}^n} d\mu(x)$ is equal to 1. Therefore, the objective function simplifies to:

$$\tilde{V}(D, \nu) = -\log(2) + \log\left(\frac{1}{2}\right) = -2\log(2).$$

Hence, with the choice of $\nu = \mu$ and $D(x) = \frac{1}{2}$ μ-almost everywhere, the objective function $\tilde{V}(D, \nu)$ is minimized. To complete the proof, we need to show that for any other choice of ν and D, the objective function $\tilde{V}(D, \nu)$ is not smaller than 0. Let us consider an arbitrary choice of ν' and D' (where $\nu' \neq \mu$ or $D' \neq \frac{1}{2}$ μ-almost everywhere). Without loss of generality, assume that there exists a set $A \subset \mathbb{R}^n$ with positive measure such that $D'(x) \neq \frac{1}{2}$ for all $x \in A$. Since μ is a probability distribution function, we have $\mu(A) > 0$. Therefore, we can rewrite the objective function as:

$$
\begin{aligned}
\tilde{V}(D', \nu') &= \int_{\mathbb{R}^n} \left(\log(D'(x)) d\mu(x) + \log(1 - D'(x)) d\nu'(x) \right) \\
&\geq \int_A \left(\log\left(D'(x)\right) d\mu(x) + \log\left(1 - D'(x)\right) d\nu'(x) \right).
\end{aligned}
$$

Now, consider the term $\log(D'(x)) d\mu(x)$ for $x \in A$. Since $D'(x) \neq \frac{1}{2}$ for all $x \in A$, we have $\log(D'(x)) < 0$ for all $x \in A$. Therefore, $\log(D'(x)) d\mu(x) < 0$ for $x \in A$. On the other hand, consider the term $\log(1 - D'(x)) d\nu'(x)$ for $x \in A$. Since $D'(x) \neq \frac{1}{2}$ for all $x \in A$, we have $1 - D'(x) \neq \frac{1}{2}$ for all $x \in A$. Therefore, $\log(1 - D'(x)) < 0$ for all $x \in A$. Since ν' is a probability distribution, $d\nu'(x) \geq 0$ for all x. Hence, $\log(1 - D'(x)) d\nu'(x) \leq 0$ for $x \in A$. Combining these results, we have $\log(D'(x)) d\mu(x) + \log(1 - D'(x)) d\nu'(x) < 0$ for $x \in A$. Therefore, $\tilde{V}(D', \nu') < 0$. Since ν' and D' were chosen arbitrarily, we can conclude that for any other choice of ν and D, the objective function $\tilde{V}(D, \nu)$ is not smaller than 0. Hence, the solution to the min-max problem in Equation (10) is achieved with the special choice $\nu = \mu$ and $D(x) = \frac{1}{2}$ μ-almost everywhere. This completes the proof. ∎

Like many min-max problems, we may utilize the alternative optimization algorithm to find an optimal solution to the problem introduced by Eq. (9) that alternates by updating the discriminator and density q. Here, the updating process contains first updating the discriminator for density q, and second, updating density q with recently updated D. Notice that updating density q

means updating the generator. This process is repeated until we find an equilibrium point for the optimization.

Proposition 2. *If in each step of the training process, D is qualified to achieve an optimum point given $q(x)$, which is pursued by an update of approximating density $q(x)$ to further develop the criterion of minimization given as*

$$\min_q \int_{\mathbb{R}^n} \left(\log(D(x))p(x) + \log(1 - D(x))q(x) \right) dx.$$

At this stage, the approximating density q converges to the target density p.

Proof First, we show that if the discriminator D is qualified to achieve an optimum point given $q(x)$ in each step of the training process, then the approximating density q converges to the target density p. Let us consider the objective function to be minimized:

$$\min_q \int_{\mathbb{R}^n} \left(\log(D(x))p(x) + \log(1 - D(x))q(x) \right) dx.$$

In each step of the training process, the discriminator D is qualified to achieve an optimum point given $q(x)$. This means that for a fixed $q(x)$, the discriminator D is updated to maximize the objective function with respect to D. Let's denote this updated discriminator as D_q^*. Now, let us consider the objective function with the updated discriminator D_q^*:

$$\min_{q^*} \int_{\mathbb{R}^n} \left(\log(D_q^*(x))p(x) + \log(1 - D_q^*(x))q(x) \right) dx.$$

Since the discriminator D_q^* is optimized for a fixed $q(x)$, the objective function becomes:

$$\min_q \int_{\mathbb{R}^n} \left(\log(D_q^*(x))p(x) + \log(1 - D_q^*(x))q(x) \right) dx$$
$$= \min_q \left(\int_{\mathbb{R}^n} \left(\log(D_q^*(x))p(x) \right) dx + \min \left(\int_{\mathbb{R}^n} \left(\log(1 - D_q(^*x))q(x) \right) dx \right) \right).$$

The first term $\min_q \int_{\mathbb{R}^n} \left(\log(D_q(x))p(x) \right) dx$ does not depend on $q(x)$ and can be treated as a constant. Therefore, minimizing this term is equivalent to maximizing $\int_{\mathbb{R}^n} \left(\log(D_q(x))p(x) \right) dx$. Similarly, the second term $\min_q \int_{\mathbb{R}^n} \left(\log(1 - D_q(x))q(x) \right) dx$ does not depend on $p(x)$ and can be treated as a constant.

Therefore, minimizing this term is equivalent to maximizing $\int_{\mathbb{R}^n} \left(\log(1 - D_q((x))q(x) \right) dx$. Since the objective function is the sum of these two terms, minimizing the objective function is equivalent to maximizing both $\int_{\mathbb{R}^n} \left(\log(D_q(x))p(x) \right) dx$ and $\int_{\mathbb{R}^n} \left(\log(1 - D_q(x))q(x) \right) dx$.

Now, let us consider the first term $\int_{\mathbb{R}^n} \left(\log(D_q(x))p(x) \right) dx$. Since $D_q(x)$ is optimized for a fixed $q(x)$, it can be considered as a constant with respect to $p(x)$. Therefore, maximizing this term is equivalent to maximizing $\int_{\mathbb{R}^n} p(x)dx$.

Similarly, let us consider the second term $\int_{\mathbb{R}^n} \left(\log(1 - D_q(x))q(x) \right) dx$. Since $D_q(x)$ is optimized for a fixed $q(x)$, it can be considered as a constant with respect to $q(x)$. Therefore, maximizing

this term is equivalent to maximizing $\int_{\mathbb{R}^n} q(x)dx$. Since the objective function is the sum of these two terms, maximizing the objective function is equivalent to maximizing both $\int_{\mathbb{R}^n} p(x)dx$ and $\int_{\mathbb{R}^n} q(x)dx$.

Now, let us consider the convergence of the approximating density q to the target density p. As we maximize the objective function, we aim to maximize both $\int_{\mathbb{R}^n} p(x)dx$ and $\int_{\mathbb{R}^n} q(x)dx$. To achieve this, the approximating density q needs to converge to the target density p. Therefore, if in each step of the training process, the discriminator D is qualified to achieve an optimum point given $q(x)$, then the approximating density q converges to the target density p.

This completes the proof. ∎

In each step of the process, first, we find the optimal discriminator $D^*(x)$ for the current density $q(x)$. Later, update density $q(x)$ given the currently updated discriminator $D(x)$ to improve the accuracy. Repeating such a process finally leads us to the desired solution. In practice, nevertheless, we infrequently focus on optimizing discriminator D for a provided generator G. Instead, we generally focus on updating D a little while ago swapping to update generator G.

It is worth emphasizing here that the unconstrained min-max problems given by Eqs. (9) and (10) are not the same as the original min-max problem introduced in Eq. (6) or the equivalent to Eq. (7), where the probability distribution ν is constrained to $\nu = \gamma \circ G^{-1}$. However, it is useful in applications to suppose Eqs. (6) and (7) exhibit identical properties introduced in Theorem 2 and Proposition 2. We can suppose the same, even after further restricting the discriminator and generator functions are neural networks defined as $D = D_\omega$ and $G = G_\theta$ as instead. Then, set $\nu_\theta = \gamma \circ G_\theta^{-1}$. Under this setting, the min-max problem becomes $\min_\theta \max_\omega V(D_\omega, G_\theta)$, where

$$
\begin{aligned}
V(D_\omega, G_\theta) &= \mathbb{E}_{x \sim \mu}[\log(D_\omega(x))] + \mathbb{E}_{z \sim \gamma}[\log(1 - D_\omega(G_\theta(z)))] \qquad (11) \\
&= \int_{\mathbb{R}^n} \left(\log(D_\omega(x))d\mu(x) + \log(1 - D_\omega(x))d\nu_\theta(x) \right).
\end{aligned}
$$

Eq. (11) is the key to executing the fundamental optimization problem. Here, since we do not know the explicit form of μ (target distribution), we should approximate the expectations through sample averages. Thus, Eq. (11) helps us to find an approximation to $V(D_\omega, G_\theta)$. More precisely, suppose a set $\mathcal{A}$ that is a subset of samples drawn from the training/original dataset χ (a minibatch) defined above and suppose a set $\mathcal{B}$ that is a minibatch of samples in space $\mathbb{R}^d$ sampled from γ. Under these assumptions, we can approximate as [3]

$$
\begin{aligned}
\mathbb{E}_{x \sim \mu}[\log(D_\omega(x))] &\approx \frac{1}{|\mathcal{A}|} \sum_{x \in \mathcal{A}} \log(D_\omega(x)), \\
\mathbb{E}_{z \sim \gamma}[\log(1 - D_\omega(G_\theta(z)))] &\approx \frac{1}{|\mathcal{B}|} \sum_{z \in \mathcal{B}} \log(1 - D_\omega(G_\theta(z))).
\end{aligned}
$$

Note that a minibatch in the GANs framework refers to a small subset of training examples fed to the network in each training iteration. The minibatch size is typically chosen to balance the computational efficiency of training and the quality of GANs. Smaller minibatches can lead to faster training; however, they may result in a noisier gradient estimate and slower convergence. On the other hand, larger minibatches can provide a more accurate gradient estimate; however, they may require more memory and computational resources to process. During training, generator and discriminator networks are trained simultaneously by optimizing their respective loss functions using backpropagation. The minibatches of real data samples and generated samples are used to compute the discriminator's loss, while the generator's loss is computed using the generated

samples only. By using minibatches in GANs, the networks can efficiently learn the complex distribution of the data and generate high-quality synthetic samples.

3 f-divergence and f-GAN concepts

Recall our motivating problem defined for GAN having a probability distribution μ, known simply for the training samples at hand. We want to find a distribution ν through an iterative process. By beginning with a probability distribution ν and iteratively updating ν, we approximate the target distribution μ with ν. To approximate μ, first, we need to measure the distance between distributions μ and ν. The vanilla GAN uses the discriminator to approximate target distribution μ. However, we can use other measures to identify the distance between distributions.

f-divergence

We can measure the dissimilarity between any two distributions, in our case target distribution μ and approximated distribution ν, with the Kullback-Leibler (KL) divergence. Let $p(x)$ and $q(x)$ be the corresponding probability density functions of μ and ν defined on $\mathbb{R}^n$. Then, the distance between densities p and q is defined in the following form

$$\mathbb{D}_{KL}(p\|q) := \int_{\mathbb{R}^n} \log\left(\frac{p(x)}{q(x)}\right)p(x)dx.$$

Here, notice that $\mathbb{D}_{KL}(p\|q)$ is finite only if $q(x) \neq 0$ on $supp(p)$ almost everywhere. At this stage, we can conclude the following results for KL-divergence [4]:

- If $p(x) > q(x)$, x is a point in the real data with a high probability. This case is the heart of the '*mode dropping*' phenomenon. It occurs when we have large regions having high values of p, whereas having small values in q. Here, it is important to remark that if $p(x) > 0$ and $q(x) \to 0$, the integrand of $\mathbb{D}_{KL}$ rises to infinity very quickly. This means that such a cost function sets an exceptionally elevated cost to the generator's distribution that does not cover some data parts.
- If $p(x) < q(x)$, x has a low chance of being a data point, instead of a high chance of being a generated point. It is faced when we observe the generator producing an unrealistic image.If we observe $p(x) \to 0$ and $q(x) > 0$ we find that the value inside the $\mathbb{D}_{KL}$ shifts to 0. This means that such a cost function pays an exceptionally low cost for generating fake samples.

Remark 2. *Regarding GANs, $\mathbb{D}_{KL}(p\|q)$ has a unique minimum at $p(x) = q(x)$. Furthermore, it does not require knowing the unknown density $p(x)$ to estimate. However, it is impressive to notice that $\mathbb{D}_{KL}(p\|q)$ is not symmetrical for $p(x)$ and $q(x)$ [3, 4].*

Even though KL-divergence is widely used in the applications of GAN, there are other measures to identify the dissimilarity between distributions. For instance, the Jensen-Shannon (JS) divergence is given as

$$\mathbb{D}_{JS}(p\|q) := \frac{1}{2}\mathbb{D}_{KL}(p\|M) + \frac{1}{2}\mathbb{D}_{KL}(q\|M),$$

where $M = \frac{p(x)+q(x)}{2}$ is a divergence measure derived from KL-divergence.

The most significant benefit of JS-divergence is that it is well-defined for any densities $p(x)$, and $q(x)$ and symmetric concerning the densities ($\mathbb{D}_{JS}(p\|q) = \mathbb{D}_{JS}(q\|p)$) while KL-divergence is not symmetric.

Following Proposition 1, the minimization part of the min-max problem in the context of the vanilla GAN is exactly the minimization over density q of $\mathbb{D}_{JS}(p\|q)$ for a given p. As things stand,

$\mathbb{D}_{KL}$ and $\mathbb{D}_{JS}$ divergences are both particular cases of the $f - divergence$ where a more general form is introduced in [5] for such divergence measures.

Consider a strictly convex function $f(x)$ with a domain $I \subseteq \mathbb{R}$ that satisfies $f(1) = 0$. Additionally, for computation purposes, we interiorize $f(x) = +\infty$, $\forall x \notin I$ convention. Then, we can introduce the f-divergence concept as introduced in [3].

Definition 1. *Consider two probability density functions $p(x)$ and $q(x)$ defined on space $\mathbb{R}^n$. Then, the $f - divergence$ between these two densities is*

$$\mathbb{D}_f(p\|q) = \mathbb{E}_{x \sim q}\left[f\left(\frac{p(x)}{q(x)}\right)\right] = \int_{\mathbb{R}^n} f\left(\frac{p(x)}{q(x)}\right)q(x)dx,$$

where we adopt $f(\frac{p(x)}{q(x)})q(x) = 0$ if $q(x) = 0$.

Remark 3. *Since the $f - divergence$ is not symmetric $(\mathbb{D}_f(p\|q) \neq \mathbb{D}_f(q\|p))$ in general, we can confuse which density divides and which density in the fraction. If we obey the original setting introduced in [5], then the definition of $\mathbb{D}_f(p\|q)$ will be our $\mathbb{D}_f(q\|p)$. In this study, we adopt the definition introduced in [7], where the f-GAN concept is first introduced.*

Proposition 3. *Suppose $f(\cdot)$ is a strictly convex function defined on $I \subseteq \mathbb{R}$ and $f(1) = 0$. Further, suppose either $supp(p) \subseteq supp(q)$ (equivalent to $p \ll q$) or $f(x) > 0$ for $x \in [0,1]$. Then, for $\mathbb{D}_f(p\|q) \geq 0$ and $\mathbb{D}_f(p\|q) = 0$, the necessary and sufficient condition is $p(x) = q(x)$.*

Proof Using the convexity property of function f and Jensen's inequality, we have

$$\mathbb{D}_f(p\|q) = \mathbb{E}_{x \sim q}\left[f\left(\frac{p(x)}{q(x)}\right)\right] \geq f\left(\mathbb{E}_{x \sim q}\left[\frac{p(x)}{q(x)}\right]\right) = f\left(\int_{supp(q)} p(x)dx\right) = f(r),$$

where the equality holds if and only if the ratio $q(x)/p(x)$ is a constant or function f is linear on the range of the ratio $p(x)/q(x)$. The range of $p(x)/q(x)$ depends on the probability distributions $p(x)$ and $q(x)$ being considered. In general, the ratio $p(x)/q(x)$ can take any positive value, zero, or infinity, depending on the values of $p(x)$ and $q(x)$ for a given x. However, in the context of importance sampling, it is common to consider the ratio $p(x)/q(x)$ as a weighting function for sampling from the target distribution $p(x)$. In this case, the $p(x)/q(x)$ range is typically restricted to a finite interval to ensure that the importance weights are bounded and can be effectively used for sampling.

Function f is a strictly convex function, so it may only be previous or for that matter, we should have $p(x) = rq(x)$ on $supp(q)$ for the equality to hold. Suppose we have the $r \leq 1$ condition. If we have $supp(p) \subseteq supp(q)$, then we obtain $r = 1$, and hence, we have $\mathbb{D}_f(p\|q) \geq 0$. Such an equality holds if and only if we have $p = q$. Suppose $f(t) > 0$, $\forall t \in [0,1)$, then we also have $\mathbb{D}_f(p\|q) \geq f(r) \geq 0$. For $r < 1$, we have $\mathbb{D}_f(p\|q) \geq f(r) \geq 0$. Therefore, if $\mathbb{D}_f(p\|q) = 0$, the conditions $r = 1$ and $p = q$ hold. ∎

At this stage, we should note that $f - divergence$ can be specified for arbitrary probability distributions μ and ν on probability space Ω. Let τ be a third probability distribution that satisfies $\mu, \nu \ll \tau$, more specifically both μ and ν are absolutely continuous concerning the third probability distribution τ. For instance, suppose $\tau = \frac{1}{2}(\mu + \nu)$. Let $p = \frac{d\mu}{d\tau}$ and $q = \frac{d\nu}{d\tau}$ be Radon-Nikodym derivatives of p and q, respectively. We characterize the $f - divergence$ of probability distributions

μ and ν as [3]

$$\mathbb{D}_f(\mu\|\nu) := \int_\Omega f\left(\frac{p(x)}{q(x)}\right)q(x)d\tau = \mathbb{E}_{x\sim\nu}\left[f\left(\frac{p(x)}{q(x)}\right)\right]. \tag{12}$$

Here, once more we adopt the convention $f\left(\frac{p(x)}{q(x)}\right)q(x) = 0$ if $q(x) = 0$. Here, it is clear that this definition is free from the choice of the probability measure τ.

In the application of the $f - divergence$, the greatest difficulty is the unknown explicit expression of the target distribution denoted by μ. Hence, in the vanilla GAN setting, to calculate the $f - divergence$ ($\mathbb{D}_f(p\|q)$), we should express the divergence in terms of the average of samples. In [6], this problem is solved with the help of the convex conjugate of the convex function at hand.

Definition 2. *Suppose $f(\cdot)$ is a convex function on the interval defined as $I \subseteq \mathbb{R}$. The convex conjugate of f is simply a generalization of the celebrated Legendre transform. The convex conjugate $f^* : \mathbb{R} \longmapsto \mathbb{R} \cup \{\pm\infty\}$ is given as [3]*

$$f^*(y) = \sup_{t\in I}\left\{ty - f(t)\right\}.$$

We can introduce the following remark as an immediate result of the definition.

Remark 4. *The convex conjugate of convex functions is also called the Fenchel transform or Fenchel-Legendre transform.*

As we mentioned above, we may extend the convex conjugate f^* to $\mathbb{R}$ by defining $f(x) = +\infty$ for all $x \notin I$. Therefore, a more precise indication of f^* is illustrated in the following lemma.

Lemma 1. *Let $f(x)$ be a strictly convex and continuously differentiable function on $I \subseteq \mathbb{R}$, where $I^\circ = (a,b)$ with $a,b \in [-\infty, +\infty]$. Then [3],*

$$f^*(y) = \begin{cases} yf'^{-1}(y), & y \in f'(I^\circ) \\ \lim_{t\mapsto b^-}(ty - f(t)), & y \geq \lim_{t\mapsto b^-} f'(t) \\ \lim_{t\mapsto a^+}(ty - f(t)), & y \leq \lim_{t\mapsto a^+} f'(t). \end{cases}$$

Proof Define $g(t) = ty - f(t)$. Then, $g'(t) = y - f'(t)$ on $I \subseteq \mathbb{R}$, which is strictly decreasing since $f(t)$ is convex. Here, $g(t)$ is a function that is strictly concave on the domain defined with I. Note that, if $y = f'(t^*)$ for some $t^* \in I^\circ$, t^* is called a critical point of function g. Therefore, t^* has to be a global maximum of g. Therefore, $g(t)$ reaches its maximum at point $t = t^* = f'^{-1}(y)$. Now, suppose y is not in the range of f', in that case, $g'(t) > 0$ or $g'(t) < 0$ on I°. Suppose the case $g'(t) > 0\ \forall t \in I^\circ$. Here, it is clear that the supremum of function $g(t)$ is attained while $t \mapsto b^-$ because $g(t)$ is a monotonously increasing function. In a similar fashion, the second case $g'(t) < 0$, $\forall t \in I^\circ$ may be derived. ∎

Based on Lemma 1, we can give the following remark:

Remark 5. *Note that $+\infty$ is a potential f^* value. Hence, the domain of f^* ($Dom(f^*)$) is characterized as sets where f^* is finite.*

A result of Lemma 1, under the assumption that f is a continuously differentiable function, $\sup_{t\in I}\{ty - f(t)\}$ is achieved for some $t \in I$ if and only if, y is in the range of $f'(t)$. Such a result is

clear when $y \in f'(I^\circ)$, however, it is arguable relatively effortlessly for finite boundary points in domain I. More commonly, without the differentiability assumption, $\sup_{t \in I}\{ty - f(t)\}$ is achieved if and only if $y \in \partial f(t)$ for some $t \in I$ ($\partial f(t)$ is set of subderivatives). We summarize some of the important properties of the convex conjugate in the following proposition [3]:

Proposition 4. *Let $f(x)$ be a convex function defined on $\mathbb{R}$ having a range $\mathbb{R} \cup \{\pm\infty\}$. Then, its convex conjugate f^* is a convex and lower-semi continuous function. Moreover, if f is a lower-semi continuous function, f satisfies Fenchel duality $f = (f^*)^*$.*

Calculation of f-divergence using the convex dual

To calculate the $f - divergence$ from samples, [6] proposes using the convex dual of function f. Let μ and ν be probability two measures that satisfy $\mu, \nu \ll \tau$ for some probability measure τ, with $p = d\mu/d\tau$ and $q = d\nu/d\tau$. In the best scenario of $\mu \ll \nu$, by $f(x) = (f^*)^*(x)$, we retain

$$
\begin{aligned}
\mathbb{D}_f(\mu\|\nu) : \; & = \int_\Omega f\left(\frac{p(x)}{q(x)}\right)q(x)d\tau(x) \\
& = \int_\Omega \sup_t\{t\frac{p(x)}{q(x)} - f^*(t)\}q(x)d\tau(x) \\
& = \int_\Omega \sup_t\{tp(x) - f^*(t)q(x)\}d\tau(x) \\
& \geq \int_\Omega \left(T(x)p(x) - f^*(T(x))q(x)\right)d\tau(x) \\
& = \mathbb{E}_{x\sim\mu}[T(x)] - \mathbb{E}_{x\sim\nu}[f^*(T(x))],
\end{aligned}
\tag{13}
$$

where $T(\cdot)$ denotes any Borel function. Therefore, by considering T overall Borel functions, one obtains

$$
\mathbb{D}_f(\mu\|\nu) \geq \sup_T \left(\mathbb{E}_{x\sim\mu}[T(x)] - \mathbb{E}_{x\sim\nu}[f^*(T(x))]\right).
\tag{14}
$$

In addition, $\forall x$, $\sup_t\{t\frac{p(x)}{q(x)} - f^*(t)\}$ is achieved for some $t = T^*(x)$ if $\frac{p(x)}{q(x)}$ is in the f^* subderivatives range [6]. Hence, if it holds for $\forall x$, we obtain

$$
\mathbb{D}_f(\mu\|\nu) = \mathbb{E}_{x\sim\nu}[T^*(x)] - \mathbb{E}_{x\sim\mu}[f^*(T^*(x))].
$$

Such equality holds, generally under some light conditions.

Theorem 3. *Let $f(\cdot)$ be a strictly convex and continuously differentiable function on the domain $I \subseteq \mathbb{R}$ and let μ and ν be Borel two probability distributions on space $\mathbb{R}^n$ that satisfy $\mu \ll \nu$. Then, we have [6]*

$$
\mathbb{D}_f(\mu\|\nu) = \sup_T \left(\mathbb{E}_{x\sim\mu}[T(x)] - \mathbb{E}_{x\sim\nu}[f^*(T(x))]\right),
\tag{15}
$$

where $\sup_T$ is considered an overall Borel functions defined as $T : \mathbb{R}^n \longmapsto Dom(f^*)$. In addition, if the probability measure p satisfies $p(x) \in I, \forall x$, $T^*(x) := f'(p(x))$ is an optimizer of Eq. (15).

Proof We have obtained the upper bound for the problem in Eq. (14) showing the lower bound part will finish the proof. Let $p(x) = d\mu(x)/d\nu(x)$. Let us analyze Eq. (13) in detail by assuming

$q(x) = 1$, and $\sup_t \left\{ tp(x) - f^*(t) \right\}$ for each x. Let us express $g_x(t) = tp(x) - f^*(t)$, $S = Dom(f^*)$ and suppose $S^\circ = (a, b)$ where $a, b \in \mathbb{R} \cup \{\pm\infty\}$. Then, we can introduce a sequence $T_k(x)$ as follows:

If density function $p(x)$ is in the range of $f^{*\prime}$, say for instance $p(x) = f^{*\prime}(t_x)$, we formed $T_k(x) = t_x \in S$. If $p(x) - f^{*\prime} > 0$ for all t, then, $g_x(t)$ is a strictly increasing function. Hence, the supremum of $g_x(t)$ is achieved at the upper boundary point b. Therefore, we assign $T_k(x) = b_k \in S$, where $b_k \mapsto b^-$. Here, if $p(x) - f^{*\prime}(t) < 0, \forall t$, $g_x(t)$ becomes a strictly decreasing function. Therefore, in this case, the supremum of $g_x(t)$ is achieved at the lower boundary point a, and we assign $T_k(x) = a_k \in S$, where $a_k \mapsto a^+$. By Lemma 1 and its proof, we know that

$$\lim_{k \mapsto \infty} \left(T_k(x)p(x) - f^*(T_k(x)) \right) = \sup_t \{ tp(x) - f^*(t) \}.$$

Thus,

$$\lim_{k \mapsto \infty} \left(\mathbb{E}_{x \sim \nu}[T_k(x)] - \mathbb{E}_{x \sim \mu}[f^*(T_k(x))] \right) = \mathbb{D}_f(\mu \| \nu).$$

To show the proof of the last, suppose $p(x) \in I$. Then, again by Lemma 1, define $s(t) = f'^{-1}(t)$ for t in the range of f', then we can write

$$f^{*\prime}(t) = \left(ts(t) - f(s(t)) \right)' = s(t) + ts'(t) - f'(s(t))s'(t) = s(t).$$

Hence, we have $g'_x(t) = p(x) - f^{*\prime}(t) = p(x) - f'^{-1}(t)$. Then, $g_x(t)$ has a maximum at $t = f'(p(x))$. This result proves that $T^* = f'(p(x))$ is an optimizer for Eq. (15). ∎

Note that Theorem 3 holds for only $\mu \ll \nu$. However, one may give the following theorem for other cases.

Theorem 4. *Let $f(t)$ be a convex function where the domain of f^* includes (a, ∞) for some $a \in \mathbb{R}$. Let μ and ν be two Borel probability measures on $\mathbb{R}^n$ that satisfy $\mu \not\ll \nu$. Then,*

$$\sup_T \left(\mathbb{E}_{x \sim \mu}[T(x)] - \mathbb{E}_{x \sim \nu}[f^*(T(x))] \right) = +\infty,$$

holds. Here, $\sup_T$ is considered an overall Borel function defined as $T : \mathbb{R}^n \longmapsto Dom(f^*)$.

Proof Consider a new distribution defined as $\tau = \frac{1}{2}(\mu + \nu)$. Then, these two densities satisfy $\mu, \nu \ll \tau$. Moreover, let $p = d\mu/d\tau$ and $q = d\nu/d\tau$ be the Radon-Nikodym derivatives of the given densities. Here, we know that $\mu \not\ll \nu$. Therefore, we can find a set S_0 with $\mu(S_0) > 0$ on which $q(x) = 0$. Now, fix a point t_0 in the domain of f^*. Let us define $T_k(x) = k$ for $x \in S_0$, and $T_k(x) = t_0$ otherwise. Then we can introduce,

$$\mathbb{E}_{x \sim \mu}[T_k(x)] - \mathbb{E}_{x \sim \nu}[f^*(T_k(x))] \geq k\mu(S_0) - f^*(t_0)(1 - \nu(S_0)) \longmapsto +\infty$$

holds. This result leads us to the desired proof. ∎

At this stage, notice that the domain of f^* has no boundary from above, and Eq. (15) is not satisfied unless we have $\mu \ll \nu$. In many studies, we face a singular target distribution μ, as the training

data we are handling might have a lower-dimensional manifold. Hence, we can introduce the following theorem.

Theorem 5. *Consider a function $f(\cdot)$ that is a lower semicontinuous convex function and the domain I^* of f^* has $\sup I^* = b^* < +\infty$. Let μ and ν be two Borel probability measures on space $\mathbb{R}^n$, and $\mu = \mu_s + \mu_{ab}$, where $\mu_s \perp \nu$ and $\mu_{ab} \ll \nu$. Then [3],*

$$\sup_T \left(\mathbb{E}_{x\sim\mu}[T(x)] - \mathbb{E}_{x\sim\nu}[f^*(T(x))] \right) = \mathbb{D}_f(\mu\|\nu) + b^*\mu_s(\mathbb{R}^n),$$

where $\sup_T$ is carried over all Borel functions given as $T : \mathbb{R}^n \longmapsto Dom(f^*)$.

Proof First, let us define $\tau = \frac{1}{2}(\mu + \nu)$. Then, it is clear that $\mu, \nu \ll \tau$. Here, $\mu = \mu_{ab} + \mu_s$ decomposition is unique and assured by the celebrated *Lebesgue decomposition theorem*, where $\mu_{ab} \ll \nu$ and $\mu_s \perp \nu$. Furthermore, let $p_{ab} = d\mu_{ab}/d\tau$, $p_s = d\mu_s/d\tau$, and $q = d\nu/d\tau$ be the Radon-Nikodym derivatives of the densities. Here, we can divide $\mathbb{R}^n$ into $\mathbb{R}^n = \Omega \cup \Omega^c$, where $\Omega = supp(q)$. Then, we have $q(x) = p_{ab}(x) = 0$ for $x \in \Omega^c$ since we have $\mu_s \perp \nu$. Hence,

$$
\begin{aligned}
\sup_T \left(\mathbb{E}_{x\sim\mu}[T(x)] - \mathbb{E}[f^*(T_k(x))] \right) &= \sup_T \int_\Omega \left(T(x)p_{ab}(x) - f^*(T(x)))q(x) \right) d\tau \\
&\quad + \sup_T \int_{\Omega^c} T(x)p_{ab}(x)d\tau \\
&= \sup_T \int_\Omega \left(T(x)\frac{p_{ab}(x)}{q(x)} - f^*(T(x)) \right) q(x)d\tau + b^*\mu_s(\Omega^c) \\
&= \int_\Omega f\left(\frac{p_{ab}(x)}{q(x)} \right) q(x)d\tau + b^*\mu_s(\mathbb{R}^n) \\
&= \int_\Omega f\left(\frac{p(x)}{q(x)} \right) q(x)d\tau + b^*\mu_s(\mathbb{R}^n) \\
&= \mathbb{D}_f(\mu\|\nu) + b^*\mu_s(\mathbb{R}^n).
\end{aligned}
$$

∎

Variational divergence minimization (VDM) with f-GANs

It is possible to generalize the standard vanilla GAN with the help of $f-divergence$ measures. For a given probability distribution μ, f-GAN aims to minimize the distance between distributions via $\mathbb{D}_f(\mu\|\nu)$, concerning the probability distribution ν. Fulfilled in the sample space, f-GAN solves the min-max problem given as

$$\min_\nu \sup_T \left(\mathbb{E}_{x\sim\nu}[T(x)] - \mathbb{E}_{x\sim\mu}[f^*(T(x))] \right). \tag{16}$$

The f-GAN framework came on to stage primarily in [7], and the optimization problem given in Eq. (16) guides us to the (VDM).
Note that the *VDM* looks identical to the min-max problem given for the vanilla GAN. Here, the Borel function T is named a critic function, or shortly a critic. With the assumption $\mu \ll \nu$, by Theorem 3 it is equal to $\min_\nu \mathbb{D}_f(\mu\|\nu)$. As we mentioned earlier, one possible problem of the f-GAN is facing $\mu \not\ll \nu$ in Theorem 4. Then, Eq. (16) is generally not equal to $\min_\nu \mathbb{D}_f(\mu\|\nu)$. Luckily, some particularly selected f, such a case is not a problem anymore.

Theorem 6. *Suppose $f(t)$ is such a function that is lower semicontinuous and strictly convex, and the domain denoted as I^* of convex conjugate f^* satisfies $\sup I^* = b^* \in [0, \infty)$. Additionally, suppose that f is a continuously differentiable function on its domain and satisfies $f(t) > 0, \forall t \in (0,1)$, and let μ be Borel probability measures on space $\mathbb{R}^n$. Under these assumptions, we obtain our unique optimizer for [7]*

$$\inf_{\nu} \sup_{T} \left(\mathbb{E}_{x\sim\nu}[T(x)] - \mathbb{E}_{x\sim\mu}[f^*(T(x))] \right),$$

as $\nu = \mu$. Here, $\sup_T$ is assessed overall Borel functions $T : \mathbb{R}^n \longmapsto Dom(f^*)$ while $\inf_\nu$ is assessed overall potential Borel probability measures.

Proof From Theorem 5, for any Borel probability measure ν, we can write the following

$$\sup_{T} \left(\mathbb{E}_{x\sim\mu}[T(x)] - \mathbb{E}_{x\sim\nu}[f^*(T(x))] \right) = \mathbb{D}_f(\mu\|\nu) + b^*\mu_s(\mathbb{R}^n) \geq \mathbb{D}_f(\mu\|\nu).$$

By Proposition 3, such equality holds if and only if $\nu = \mu$. Consequently, $\nu = \mu$ becomes our unique optimizer for GANs. ∎

Some remarks on special solutions

Remark 6. *Suppose that both the density functions $p(x)$ and $q(x)$ satisfy $p(x) = q(x)$. Then, the optimal value becomes $D^*(x) = 1/2$. For such a special case, we have a loss function as [1]*

$$
\begin{aligned}
L(G^*, D^*) &= \int_{\mathbb{R}^n} \left(p(x)\log(D^*(x)) + q(x)\log(1 - D^*(x)) \right) dx \\
&= \log(\tfrac{1}{2}) \int_x p(x)dx + \log(\tfrac{1}{2}) \int_{\mathbb{R}^n} q(x)dx \\
&= -2log(2).
\end{aligned}
$$

Furthermore, if we calculate JS divergence, we have

$$
\begin{aligned}
\mathbb{D}_{JS}(\mu\|\nu) &= \frac{1}{2}\mathbb{D}_{KL}\left(\mu\|\frac{\mu+\nu}{2} \right) + \frac{1}{2}\mathbb{D}_{KL}\left(\nu\|\frac{\mu+\nu}{2} \right) \\
&= \frac{1}{2}\left(\log(2) + \int_{\mathbb{R}^n} p(x)\log\left(\frac{p(x)}{p(x)+q(x)} \right)dx \right) \\
&\quad + \frac{1}{2}\left(\log(2) + \int_{\mathbb{R}^n} q(x)\log\left(\frac{q(x)}{p(x)+q(x)} \right)dx \right) \\
&= \frac{1}{2}\left(\log(4) + L(G, D^*) \right). \quad (17)
\end{aligned}
$$

By rearranging Eq. (17), we find

$$L(G, D^*) = 2\mathbb{D}_{JS}(\mu\|\nu) - 2\log(2).$$

As an immediate result, we can also give the following remark.

Remark 7. *Under the assumptions given in the preceding remark, the followings hold [3]:*

- *Fundamentally, the objective of a GAN loss function is to quantify the similarity between the generated data distribution ν and the real sample distribution μ by using $\mathbb{D}_{JS}$ under the optimal discriminator D*

condition. The best generator (G^*) imitates the distribution of real data, which leads us to the minimum given as $L(G^*, D^*) = -2\log 2$.

- *If we train the discriminator D until it convergences, its error approximates 0. This indicates that the $\mathbb{D}_{JS}$ between the distributions has reached its maximum (it is easy to see that $0 \leq \mathbb{D}_J S(\mu\|\nu) \leq \ln(2)$). We can find it only if their distributions are not continuous (meaning: their densities are not absolutely continuous functions) or the distributions have disjoint supports. One potential reason behind the noncontinuity of the distribution is if their supports rely on low-dimensional manifolds. For such a case, there is substantial empirical and theoretical evidence to believe that the generated data distribution ν is focused on a low-dimensional manifold for many datasets.*
- *If both μ and ν rest in low-dimensional manifolds, they are almost undoubtedly disjoint. If the distributions have disjoint supports, we can always find a perfect discriminator that divides real and fake samples 100% accurately.*

4   Concluding remarks

In this study, we discovered and explored the mathematical background of GANs to illustrate a deep understanding of them for further extensions. Hence, in this study, we took a detailed tour of the mathematics behind GANs. After the celebrated work of Goodfellow et al. [1], new adversarial training objectives and techniques for generative modeling have been developed, such as Wasserstein GANs [8, 9]. Furthermore, GANs have been widely applied to new fields of research, including mathematical finance [10–12], time series generation [13, 14], audio synthesis [15], and fraud detection in financial datasets [16]. The underlying mathematics for these models are obviously different from what we have discussed above, but the study is a good starting point nonetheless.

Declarations

Ethical approval

The authors state that this research complies with ethical standards. This research does not involve either human participants or animals.

Consent for publication

Not applicable.

Conflicts of interest

The authors declare that they have no conflict of interest.

Funding

Author's contributions

B.Y.: Conceptualization, Methodology, Software, Data Curation, Writing-Original draft preparation. R.K.: Visualization, Investigation, Supervision, Validation, Writing-Reviewing and Editing. Both authors have made substantial contributions to the conception and design of the work. They have read and agreed to the published version of the manuscript.

Authors' information

- Dr. Bilgi Yilmaz holds an M.Sc. and Ph.D. degree in Financial Mathematics from the Institute of Applied Mathematics at Middle East Technical University, Türkiye. Currently, Dr. Yilmaz is a post-doctoral researcher at RPTU Kaiserslautern, Germany.
- Prof. Dr. Ralf Korn studied mathematics and business administration at the University of Mainz, Germany. After completing his doctorate and habilitation in Mathematics, Professor Korn received a call to the Technical University of Kaiserslautern in 1999, where he has been a professor of Financial Mathematics in the Mathematics Department ever since. One year later, he founded the Financial Mathematics department of the Fraunhofer Institute for Industrial Mathematics in Kaiserslautern, which he has headed ever since.

  In addition, he has been one of the co-founders of the European Institute for Quality Management of Financial Mathematical Products and Processes EI-QFM in Kaiserslautern since 2010. Professor Korn is a member of the German Society for Insurance and Financial Mathematics (DGVFM), where he was a member of the board from 2003 to 2009 and has been deputy chairman of the board since 2011.

References

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S. et al. Generative adversarial nets. *Advances in Neural Information Processing Systems,* 27, (2014).

[2] de Meer Pardo, F. Enriching financial datasets with generative adversarial networks. *MS thesis, Delft University of Technology, The Netherlands*, (2019).

[3] Wang, Y. A mathematical introduction to generative adversarial nets (GAN). *ArXiv Prints, ArXiv:2009.00169,* (2020). [CrossRef]

[4] Arjovsky, M. and Léon, B. Towards principled methods for training generative adversarial Networks. *ArXiv Prints, arXiv:1701.04862,* (2017). [CrossRef]

[5] Syed A.M. and Samuel, D.S. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological),* 28(1), 131-142, (1996). [CrossRef]

[6] Nguyen, X., Wainwright, M.J. and Jordan M.I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. In Proceedings, *IEEE Transactions on Information Theory*, 56(11), pp. 5847-5861, (2010, October). [CrossRef]

[7] Nowozin, S., Cseke, B. and Tomioka, R. f-GAN: Training generative neural samplers using variational divergence minimization. In Proceedings, *Advances in Neural Information Processing Systems 29 (NIPS),* (2016, December).

[8] Arjovsky, M., Chintala, S. and Bottou, L. Wasserstein generative adversarial networks. In Proceedings *Proceedings of the 34th International Conference on Machine Learning (PMLR),* pp. 214-223, (2017, July).

[9] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. and Courville, A.C. Improved training of Wasserstein GANs. In Proceedings *Advances in neural information processing systems 30 (NIPS),* (2017, December).

[10] Ni, H., Szpruch, L., Wiese, M., Liao, S. and Xiao, B. Conditional sig-wasserstein gans for time series generation. *ArXiv Preprint, arXiv:2006.05421,* (2020). [CrossRef]

[11] Wiese, M., Bai, L., Wood, B. and Buehler, H. Deep hedging: learning to simulate equity option markets. *ArXiv Preprint, arXiv:1911.01700* (2019). [CrossRef]

[12] Wiese, M., Knobloch, R., Korn, R. and Kretschmer P. Quant GANs: deep generation of financial time series. *Quantitative Finance,* 20(9), 1419-1440, (2020).

[13] Ni, H., Szpruch, L., Sabate-Vidales, M., Xiao, B., Wiese, M. and Liao, S. Sig-Wasserstein GANs for time series generation. In Proceedings *of the Second ACM International Conference on AI in Finance (ICAIF),* pp. 1-8, (2021, November). [CrossRef]

[14] Yoon, J., Jarrett, D. and Van der Schaar, M. Time-series generative adversarial networks. *Advances in Neural Information Processing Systems 32 (NeurIPS),* (2019, December).

[15] Donahue, C., McAuley, J. and Puckette, M. Adversarial audio synthesis. *ArXiv Preprint, arXiv:1802.04208,* (2018). [CrossRef]

[16] Schreyer, M., Sattarov, T., Reimer, B. and Borth, D. Adversarial learning of deepfakes in accounting. *ArXiv Preprint, arXiv:1910.03810,* (2019). [CrossRef]

Mathematical Modelling and Numerical Simulation with Applications (MMNSA)
(https://dergipark.org.tr/en/pub/mmnsa)

RESEARCH PAPER

Mathematical dynamics for HIV infections with public awareness and viral load detectability

Umar Tasiu Mustapha [iD] [1,‡], **Abdurrahman Ado** [iD] [1,‡], **Abdullahi Yusuf** [iD] [1,*,‡],
Sania Qureshi [iD] [2,3,4,‡] **and Salihu Sabiu Musa** [iD] [5,6,‡]

[1]Department of Mathematics, Federal University Dutse, 7156 Jigawa, Nigeria, [2]Department of Basic Sciences and Related Studies, Mehran University of Engineering and Technology, Jamshoro 76062, Pakistan, [3]Department of Mathematics, Near East University, 99138 Mersin, Türkiye, [4]Department of Computer Science and Mathematics, Lebanese American University, Beirut, Lebanon, [5]Department of Applied Mathematics, Hong Kong Polytechnic University, Hong Kong SAR, China, [6]Department of Mathematics, Kano University of Science and Technology, Wudil, Nigeria

*Corresponding Author
‡umar.tasiu@fud.edu.ng (Umar Tasiu Mustapha); abdurrahmanado4@gmail.com (Abdurrahman Ado); yusufabdullahi@fud.edu.ng (Abdullahi Yusuf); sania.qureshi@faculty.muet.edu.pk (Sania Qureshi); salihu-sabiu.musa@connect.polyu.hk (Salihu Sabiu Musa)

Abstract

In this paper, we develop a nonlinear deterministic model that incorporates public awareness and treatment to describe the dynamics of HIV/AIDS in an infected population with detectable and undetectable viral load. The model undergoes backward bifurcation in which a stable disease-free equilibrium coexists with a stable endemic equilibrium. Numerical simulations carried out show the behavior of the state variables and the impact of public awareness in controlling the spread of HIV. The results show that public awareness will help in curtailing the spread of HIV infection, and when treatment is applied to infected individuals with detectable viral load can easily suppress their virus to become undetectable so that they cannot transmit HIV through sexual intercourse.

Keywords: HIV; viral load detectability; parameter calibration; bifurcation analysis; confidence interval
AMS 2020 Classification: 34C23; 62P10; 92B05; 92D25

1 Introduction

Human Immunodeficiency Virus (HIV) is a virus affecting the cells of the immune system (CD4$^+$ T) making the body vulnerable to other infectious diseases [1, 2]. CD4$^+$ T cells are orchestrators, regulators, and direct effectors of antiviral immunity [1]. If HIV is not treated, it leads to a severe

stage of HIV infection called Acquired Immunodeficiency Syndrome (AIDS) [3]. The virus is transmitted via direct contact with different kinds of body fluids such as blood, vaginal fluids, rectal fluids, semen, and breast milk of the infected individual or through mother to her child during the pregnancy period (i.e, vertical transmission) [3, 4]. HIV still remains one of the most severe global public health threats [5, 6]. Since its emergence, more than 79.3 million people became infected with HIV/AIDS, among which 36.3 million people have died from AIDS-related illness [5]. About 37.7 million people were living with HIV by 2020 [5]. As of 2020, out of the total HIV-infected individuals, about 20.6 million people (55%) with HIV were in Eastern and Southern Africa, 4.7 million people (13%) in Western and Central Africa, 5.7 million people (15%) in Asia and the Pacific region and 2.2 million people (6%) in Western and Central Europe and North America [5]. The majority of people infected with HIV are from low and middle-income countries.

As part of global commitment to decrease the transmission of HIV infection, the number of people accessing Antiretroviral Therapy (ART) has increased significantly from 7.8 million in 2010 to 27.5 million in 2020. New infection declined by 30% from 2.1 million in 2010 to 1.5 million people, and 84% of people living with HIV are aware of their status, while the remaining 16% needs access to be diagnosis and HIV test [5]. Before the introduction of ART, a minority of the people infected with HIV maintained normal $CD4+$ cell counted healthy range of $(450–1650\ cells/ml)$ and remained symptoms-free without treatment for several years and did not progress to AIDS stages [7]. Some of these individuals have low or non-detectable viral load and are referred to as non-progressors, classified as long-term non-progressors and controllers due to their resistance to viral replication in the absence of ART. Controllers are sub-divided into elite controllers (EC) with HIV RNA less than 50 $copies/ml$ and Viremic Controllers (VCs) with $50 - 2000\ copies/ml$ [7, 8]. HIV treatments aim at reducing the viral load until the virus is no longer detectable.

It was reported that taking a full dose of ART could effectively suppress the viral load (i.e., the amount of HIV in a person's blood) of infected individuals [4]. If the viral load is lower than 200 $copies/ml$ in blood, it is unlikely to be detected using a blood antibody test. In this scenario, an infected individual cannot transmit HIV through sexual intercourse [9]. Up till now, there is no cure for HIV infection. Still, the ART helps to suppress viral replication within the patient's body and allows immune system recovery to strengthen and regain the ability to fight new infection [4]. WHO endorsed ART regardless of $CD4+$ cell count to all people with HIV in 2016. Also, ART should be offered simultaneously with diagnosis among individuals that are ready for treatment. In June 2021, 187 countries adopted the first recommendation, and 82 low- and middle-income countries implemented the second policy [4].

Many models have been formulated to study the dynamics of HIV/AIDS infection with different control strategies incorporated in the model. The study by [10] showed that public health awareness on risk behaviours could help in decreasing the persistence of HIV/AIDS. A mathematical model for assessing the impact of condom usage, ART and voluntary testing in decreasing the spread of HIV was proposed by [6]. The study revealed that the hope of controlling HIV transmission using the intervention listed was highly remarkable. Furthermore, a study by [11] highlighted that the rate of vertical transmission which leads to an increase in the pre-AIDS and AIDS population is proportional to the infective population. The spread of the diseases can be reduced significantly by controlling the vertical population. A study by Bhunu et al. [12] suggested that even in the absence of ART, effective guidance and testing have tremendous effects on the prevention and control of the epidemic. [13] analyzed HIV/AIDS dynamics for a situation when only HIV-infected individuals who did not develop AIDS symptoms and are not under ART transmit the HIV virus. Recently, nonlinear fractional order models are also used to describe HIV/AIDS transmission dynamics. [14] proposed a dynamical fractional first-order HIV-1 with

Caputo derivative that studies the infection between cancer cells, healthy CD4$^+$ T lymphocytes and virus-infected CD4$^+$ T lymphocytes. The result revealed that fractional order derivatives have a significant effect on the dynamics process. The fractional order model of viral kinetics for primary infection of HIV-1 with immune control and treatment was analyzed by [15]. A nonlinear fractional-order HIV epidemic model solved via the $L1$ scheme was proposed by [16]. The result showed that the homotopy analysis method applied has effectiveness and strength in solving a compartmental model. Naik et al. [17] proposed a nonlinear fractional order model for HIV transmission dynamics with optimal control. The study recommended that to decrease the spread of HIV infections there is a need for personal precaution and periodic monitoring by researchers and medical professionals.

For further references to the related studies mentioned above, one can visit [6, 10, 17–21]. In this research work, we have formulated a new mathematical model of HIV/AIDS dynamics considering infected individuals with detectable viral load and infected individuals with undetectable viral load.

The significance of this research is to assess the impact of viral load detectability on HIV/AIDS transmission when some of the infected individuals with low viral load (undetectable viral load) can not transmit HIV sexually, and also assess the impact of public education and treatment on uninfected and infected population, respectively.

The paper is organized as follows: the model is developed and analyzed in Sections 2 and 3, respectively. Basic reproduction number, equilibria, and their stabilities and bifurcation analysis are given in 4. Model fitting, parameter estimation, and sensitivity analysis are conducted and presented in Section 5. Numerical simulation of the model is presented in Section 6. Discussion of the results and conclusion are provided in Section 7.

2 Model description

A nonlinear mathematical model is developed to study the transmission dynamics of HIV/AIDS to assess the impact of public awareness and treatment on the overall dynamics. The total population at time t, denoted by $N(t)$ is divided into the following disjoint compartments: uneducated susceptible $S_u(t)$ (individuals that are unaware on how to prevent HIV infection), educated susceptible $S_e(t)$ (individuals that are aware on how to prevent HIV infection), newly infected individuals $I_1(t)$, infected with detectable viral load $I_2(t)$ (infected individuals with higher viral load (> 200 copies/ml) that can be detected using a blood test and they can transmit HIV through sexual intercourse), infected with undetectable viral load $I_3(t)$ (infected individuals with a low level of HIV virus in their blood (< 200 copies/ml) that can not be detected using blood test and they cannot transmit the disease through sexual intercourse), infected individuals receiving treatment $I_t(t)$, AIDS patients(infected individuals with higher viral load and developed certain opportunistic infections). The risk behaviours that can lead to HIV/AIDS infection include unprotected sex, sharing of drugs and injection needles, lack or absence of blood tests for couples before getting married, or lack of condom usage during sex. We considered sexual transmission as the only mode of transmission, as such I_3 are considered non-infectious since they can not transmit HIV through sex, and assumed AIDS patients to be sexually inactive because their immune system is weak and unable to fight infections which cause several opportunistic diseases to them.

Recruitment of new individuals into the susceptible population occurs at a rate π (which are assumed to be sexually active and susceptible to HIV infection). p is the fraction of recruited individuals that are educated. Uneducated susceptible individuals become educated through public awareness campaigns on HIV infection at a rate τ. Susceptible uneducated and educated individuals become infected when in contact with the infected individuals at a rate λ and $\alpha\lambda$, respectively. It is assumed that susceptible educated are avoiding risk behaviour which reduces

their rate of HIV infection by α, with $0 < \alpha < 1$. Newly infected individuals become either infected with detectable viral load at a rate $\epsilon\theta$ or infected controllers which are undetectable at a rate $(1-\epsilon)\theta$ where θ is the progression rate from the newly infected compartment, while ϵ is the fraction of newly infected with detectable viral load. Infected individuals with detectable viral load move to treatment at a rate ϕ while some progress to AIDS at the rate ρ. Infected individuals under treatment when taking a full dose of ART their viral load will be undetectable through a blood test and assumed to move into the infected undetectable viral load class at the rate γ. Infected with the undetectable viral load when their viral load becomes detected through blood test moves to infected detectable class at the rate ω.

AIDS patients move to treatment class at a rate σ. Natural death of individuals occurs at a rate μ. δ_1 and δ_2 are the disease mortality rates of infected individuals at AIDS and treatment compartment, respectively, where $\delta_2 < \delta_1$ (we assume that individuals under treatment die at a rate less than AIDS patients that refused to go for treatment). Thus, we have $N(t) = S_u(t) + S_e(t) + I_1(t) + I_2(t) + I_3(t) + I_t(t) + A(t)$.
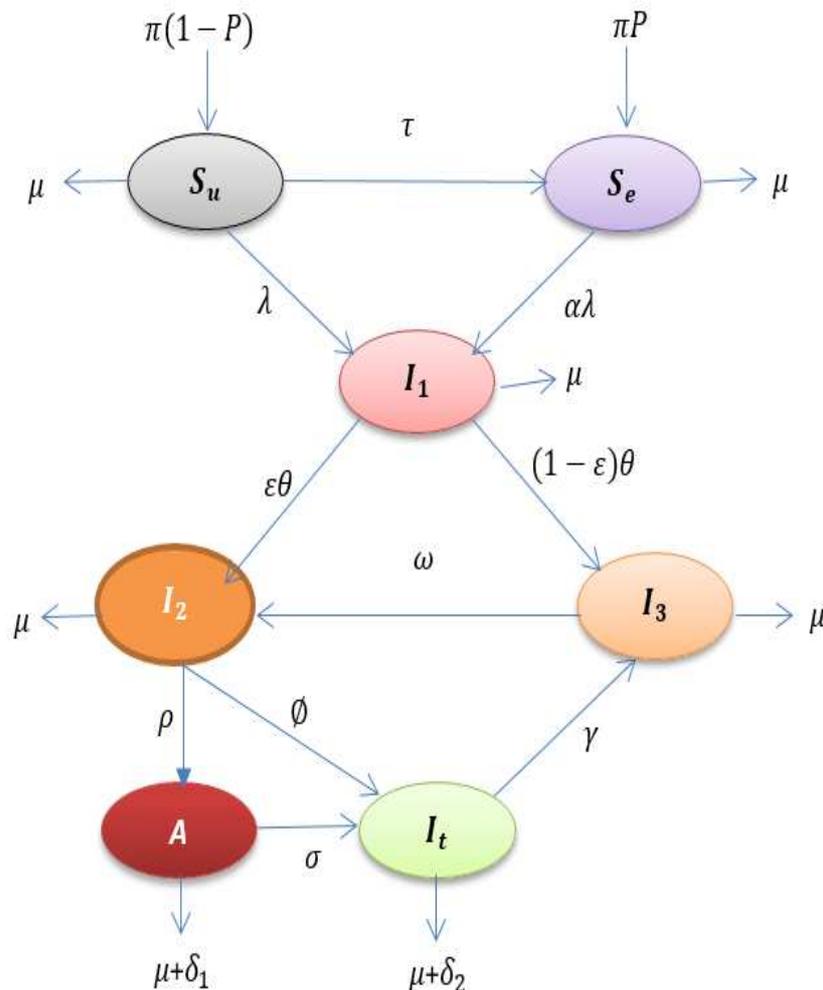
Figure 1. Schematic diagram of model (1). Solid arrows indicate transitions and expressions next to arrows show the *per capita* flow rate between compartments.

By considering the explanations of the model parameters and compartments, we obtain the

following system:

$$
\begin{aligned}
\frac{dS_u}{dt} &= \pi(1-p) - (\mu + \tau + \lambda)S_u, \\
\frac{dS_e}{dt} &= \pi p + \tau S_u - (\mu + \alpha\lambda)S_e, \\
\frac{dI_1}{dt} &= \lambda(S_u + \alpha S_e) - (\mu + \theta)I_1, \\
\frac{dI_2}{dt} &= \epsilon\theta I_1 + \omega I_3 - (\mu + \rho + \phi)I_2, \\
\frac{dI_3}{dt} &= (1-\epsilon)\theta I_1 + \gamma I_t - (\mu + \omega)I_3, \\
\frac{dI_t}{dt} &= \phi I_2 + \sigma A - (\mu + \gamma + \delta_2)I_t, \\
\frac{dA}{dt} &= \rho I_2 - (\mu + \sigma + \delta_1)A,
\end{aligned}
\tag{1}
$$

where, the force of infection for the transmission of HIV in this model is given by, $\lambda = \beta\left(\frac{\eta_1 I_1 + \eta_2 I_2 + I_t}{N}\right)$ and β is the effective contact rate that may result in HIV/AIDS infection, η_1, and η_2 ($\eta_1 > \eta_2$) denote an increase in infectiousness for newly infected individuals and infected individuals with higher viral load, respectively.

The description of the variables and parameters of the model are shown in Table 1.

Table 1. Interpretation and definitions of the state variables and parameters used in model (1).

Variable	Description
N	Total human population
S_u	Uneducated susceptible individuals
S_e	Educated susceptible individuals
I_1	Newly infected individuals
I_2	Infected individuals with detectable viral load
I_3	Infected individuals with undetectable viral load
I_t	Infected individuals under treatment
A	AIDS patients
λ	Force of infection
Parameter	
π	Recruitment rate of susceptible individuals
μ	Natural mortality rate
δ_1, δ_2	Death rate due to disease
p	Proportion of π that are educated
τ	Rate at which uneducated susceptible become educated about HIV infection
β	Effective contact rate
α	Parameter for decrease of infectiousness in S_e
θ	Rate of movement from infectious class
ϵ	Fraction of the rate of movement from infectious compartment
ϕ	Movement rate of infected with detectable viral load to treatment
ρ	Progression rate to AIDS compartment
γ	Rate at which treated individuals become undetectable viral load
ω	Rate at which infected individuals with undetectable viral load become detectable
σ	rate of movement of AIDS patients to treatment class at a rate
η_1, η_2	Infectiousness factor for newly infected and individuals with higher viral load

3 Analysis of the model

Boundedness and positivity of solutions

The model deals with the human population, each of its parameters, and the state variables are non-negative for all $t \geq 0$. We can now prove that each of the state variables of model (1) are non-negative for all $t \geq 0$.

Theorem 1 *System (1) defines a dynamical system in the closed set, given by,*

$$\Omega = \left\{ (S_u(t), S_e(t), I_1(t), I_2(t), I_3(t), I_t(t), A(t)) \, \epsilon \, R_+^7 : N \leq \frac{\pi}{\mu} \right\}.$$

Proof We are to show that R_+^7 is positively invariant, that is all solution of system (1) initiated in Ω do not leave Ω see (Theorem 2.1.5) of [22]. Let $S_u(0) > 0, S_e(0) > 0, I_1(0) > 0, I_2(0) > 0, I_3(0) > 0, I_t(0) > 0$, and $A(0)) > 0$. Suppose $S_u(0)$ and $S_e(0)$ are not positive, then there exists a time $\tilde{t} > 0$, such that $S_u(t) > 0$ and $S_e(t) > 0$ for $t \in [0, \tilde{t})$ and $S_u(\tilde{t}) = S_e(\tilde{t}) = 0$.

From the third, fourth, and fifth equation of model (1), we obtain,

$$\frac{dI_1(t)}{dt} \geq -(\theta + \mu)I_1(t) \quad \text{for } t \in [0, \tilde{t}),$$

$$\frac{dI_2(t)}{dt} \geq -(\rho + \Phi + \mu)I_2(t) \quad \text{for } t \in [0, \tilde{t}),$$

$$\frac{dI_3(t)}{dt} \geq -(\mu + \omega)I_3(t) \quad \text{for } t \in [0, \tilde{t}), \tag{2}$$

$$\frac{dI_t(t)}{dt} \geq -(\gamma + \mu + \delta_2)I_t(t) \quad \text{for } t \in [0, \tilde{t}),$$

$$\frac{dA(t)}{dt} \geq -(\mu + \sigma + \delta_1)A(t) \quad \text{for } t \in [0, \tilde{t}).$$

It follows that $I_1(0) > 0, I_2(0) > 0, I_3(0) > 0, I_t(0) > 0$ and $A(0) > 0$ for $t \in [0, \tilde{t})$. Thus, from the first and second equations of the system, we have

$$\frac{dS_u(t)}{dt} \geq -(\tau + \mu + \lambda)S_u(t) \quad \text{for } t \in [0, \tilde{t}),$$

$$\frac{dS_e(t)}{dt} \geq -(\mu + \alpha\lambda)S_e(t) \quad \text{for } t \in [0, \tilde{t}).$$

One can clearly see that, $S_u(0) > 0$ and $S_e(0) > 0$ which contradict our assumption of $S_u(\tilde{t}) = S_e(\tilde{t}) = 0$. Hence $S_u(t)$ and $S_e(t)$ are positive. Similarly, the positivity of the remaining state variable of the model can be seen from subsystem of (1) excluding the first and second equation written in matrix form as follows:

$$\frac{dX(t)}{dt} = \mathcal{M}Y(t) + B(t), \tag{3}$$

with

$$Y(t) = \begin{pmatrix} I_1, & I_2, & I_3, & I_t, & A \end{pmatrix}^T,$$

$$\mathcal{M} = \begin{pmatrix} \eta_1 K - K_1 & \eta_2 K & 0 & K & 0 \\ \epsilon\theta & -K_2 & \omega & 0 & 0 \\ K_6\theta & 0 & -K_3 & \gamma & 0 \\ 0 & \Phi & 0 & -K_4 & \sigma \\ 0 & \rho & 0 & 0 & -K_5 \end{pmatrix}, \tag{4}$$

$$B(t) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \end{pmatrix}^T,$$

where, $K = \beta\frac{S_u + \alpha S_e}{N}$, $K_1 = \theta + \mu$, $K_2 = \rho + \Phi + \mu$, $K_3 = \omega + \mu$, $K_4 = \gamma + \mu + \delta_2$, $K_5 = \sigma + \mu + \delta_1$ and $K_6 = (1 - \epsilon)$. Clearly, $\mathcal{M}$ is a Metzler matrix for the fact that both $S_u(t)$ and $S_e(t)$ are non-negative. Which shows subsystem (3) is a monotone system [23]. Thus, $\mathbb{R}^5_+$ is invariant under the flow of subsystem (3). Therefore, $\mathbb{R}^7_+$ is positively invariant under the flow of system (1). ∎

System (1) has a disease-free equilibrium which is determined by setting its right-hand sides to zero.

$$\epsilon^0 = (S_u^0, S_e^0, I_1^0, I_2^0, I_3^0, I_t, A) = \left(\frac{\pi(1 - P)}{\tau + \mu}, \frac{\pi(\tau + \mu p)}{\mu(\tau + \mu)}, 0, 0, 0, 0, 0 \right).$$

4 Basic reproduction number

The basic reproduction number (denoted by $R_0 = \rho(FV^{-1})$, where ρ is the spectral radius of the next generation matrix, (FV^{-1}) of model (1) is the number of new infections produced by HIV infected individuals with a detectable viral load when interacted with the fully susceptible population in the absence of awareness and treatment. It is determined using the next generation matrix approach [24] to establish the stability of the equilibrium. The matrix F represents the new infection terms and V for the remaining transition terms are respectively given by, R_0

$$F = \begin{bmatrix} \frac{\eta_1\beta\left(S_u^0 + \alpha S_e^0\right)}{N^0} & \frac{\eta_2\beta\left(S_u^0 + \alpha S_e^0\right)}{N^0} & 0 & \frac{\beta\left(S_u^0 + \alpha S_e^0\right)}{N^0} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{5}$$

$$V = \begin{bmatrix} K_1 & 0 & 0 & 0 & 0 \\ -\epsilon\theta & K_2 & -\omega & 0 & 0 \\ -K_6\theta & 0 & K_3 & -\gamma & 0 \\ 0 & -\phi & 0 & K_4 & -\sigma \\ 0 & -\rho & 0 & 0 & K_5 \end{bmatrix}. \tag{6}$$

The basic reproduction number R_0 is obtained as,

$$R_0 = \frac{\beta\left((K_4 K_6 \eta_2 \theta + \phi\left(-\gamma\,\eta_1 + K_6\theta\right)\right)\omega + \left((\varepsilon\theta\,\eta_2 + \eta_1 K_2\right)K_4 + \phi\,\varepsilon\theta\right)K_3\right)K_5\left(\mu(1-p) + \alpha(\tau + \mu p)\right)}{k_1\left((-\gamma\,\omega\,\phi + K_2 K_3 K_4\right)K_5 - \gamma\,\omega\,\rho\,\sigma\right)(\tau + \mu)}$$

$$+ \frac{\beta\rho\,\sigma\left((-\gamma\,\eta_1 + K_6\theta\right)\omega + K_3\varepsilon\theta\right)\left(\mu(1-p) + \alpha(\tau + \mu p)\right)}{k_1\left((-\gamma\,\omega\,\phi + K_2 K_3 K_4\right)K_5 - \gamma\,\omega\,\rho\,\sigma\right)(\tau + \mu)},$$

where $K_1 = \theta + \mu, K_2 = \rho + \phi + \mu, K_3 = \omega + \mu, K_4 = \gamma + \mu + \delta_2, K_5 = \sigma + \mu + \delta_1, K_6 = (1 - \epsilon)$ and are all positives.

Theorem 2 *The disease-free equilibrium (DFE) ϵ^0, of model (1), is locally-asymptotically stable (LAS) in Ω if $R_0 < 1$, and unstable if $R_0 > 1$.*

Global stability of disease-free equilibrium

Theorem 3 *The disease-free equilibrium (DFE) ϵ^0, of model (1) is globally-asymptotically stable (GAS) in Ω if $R_0 < 1$, and unstable if $R_0 > 1$.*

Proof To prove the GAS of DFE, the two axioms $[G_1]$ and $[G_2]$ must be satisfied for $R_0 < 1$ [25]. We re-write system (1) in the form:

$$\begin{aligned}\frac{dY_1}{dt} &= F(Y_1, Y_2),\\ \frac{dY_2}{dt} &= G(Y_1, Y_2) : G(Y_1, 0) = 0,\end{aligned} \tag{7}$$

where $Y_1 = (S_u^0, S_e^0)$ and $Y_2 = (I_1^0, I_2^0, I_3^0, I_t^0, A^0)$ with the elements of $Y_1 \in R_+^2$ representing the uninfected population and the elements of $Y_2 \in R_+^5$ representing the infected population. The DFE is now denoted as $E^0 = (Y_1^*, 0)$, where $Y_1^* = (N^0, 0)$. Now for the first condition, that is GAS of Y_1^*, gives,

$$\frac{dY_1}{dt} = F(Y_1, 0) = \begin{bmatrix} \pi(1 - P) - (\mu + \tau)S_u^0 \\ \pi p + \tau S_u^0 - \mu S_e^0 \end{bmatrix}.$$

Solving the linear differential equations gives,

$$S_u^0(t) = \frac{\pi(1-p)}{(\mu+\tau)} - \frac{\pi(1-p)}{(\mu+\tau)}e^{-(\mu+\tau)t} + S_u^0(0)e^{-(\mu+\tau)t},$$

$$S_e^0(t) = \frac{\pi p + \tau S_u^0}{\mu} - \frac{\pi p + \tau S_u^0}{\mu}e^{-\mu t} + S_e^0(0)e^{-\mu t}.$$

Now, it is easy to show that $S_u^0(t) + S_e^0(t) \to N^0(t)$ as $t \to \infty$ regardless of the value of $S_u^0(t)$ and $S_e^0(t)$. Thus, $Y_1^* = (N^0, 0)$ is globally asymptotically stable. Furthermore, for the second condition,

that is $\tilde{G}(Y_1, Y_2) = BY_2 - \tilde{G}(Y_1, Y_2)$ gives:

$$
B = \begin{pmatrix}
-(\mu + \theta) + \frac{\beta \eta_1 S_u^0}{N^0} + \frac{\alpha \beta \eta_1 S_e^0}{N^0} & \frac{\beta \eta_2 S_u^0}{N^0} + \frac{\beta \eta_2 S_e^0}{N^0} & 0 & \frac{\beta S_u^0 + \alpha \beta S_e^0}{N^0} & 0 \\
\epsilon \theta & -(\mu + \rho + \phi) & \omega & 0 & 0 \\
(1 - \epsilon)\theta & 0 & -(\mu + \omega) & \gamma & 0 \\
0 & \phi & 0 & -(\mu + \delta_2 + \gamma) & \sigma \\
0 & \rho & 0 & 0 & -(\mu + \delta_1 + \sigma)
\end{pmatrix}.
$$

$$\tag{8}$$

This is clearly Metziller matrix

$$
\tilde{G}(Y_1, Y_2) = \begin{pmatrix}
\lambda S_u^0 + \alpha \lambda S_e^0 - (\mu + \theta) I_1^0 \\
\epsilon \theta I_1^0 + \omega I_3^0 - (\mu + \rho + \phi) I_2^0 \\
(1 - \epsilon)\theta I_1^0 + \gamma I_t^0 - (\mu + \omega) I_3^0 \\
\phi I_2^0 + \sigma A^0 - (\mu + \gamma + \delta_2) I_t^0 \\
\rho I_2^0 - (\mu + \sigma + \delta_1) A^0
\end{pmatrix}.
$$

$$\tag{9}$$

Then,

$$
\tilde{G}(Y_1, Y_2) = BY_2 - \tilde{G}(Y_1, Y_2) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.
$$

Thus we have

$$
\tilde{G}(Y_1, Y_2) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T,
$$

It is clear that $\tilde{G}(Y_1, Y_2) = 0.$ ∎

Endemic equilibrium point

When HIV persists in the population, at least one of the infectious compartments of model (1) is not empty. As such, model (1) has endemic equilibrium which is obtained by setting the vector field of the model (1) to zero. Defined the endemic equilibrium state as;

$$
\epsilon^{**} = (S_u^{**}, S_e^{**}, I_1^{**}, I_2^{**}, I_3^{**}, I_t^{**}, A^{**}),
$$

where,

$$
\begin{aligned}
S_u^{**} &= \frac{\pi(1 - p)}{\lambda^{**} + \mu + \tau}, \\
S_e^{**} &= \frac{\pi(P\lambda^{**} + P\mu + P\tau + \tau)}{\alpha \lambda^{**2} + \alpha \lambda^{**} \mu + \alpha \lambda^{**} \tau + \lambda^{**} \mu + \mu^2 + \mu \tau}, \\
I_1^{**} &= \frac{\lambda^{**} \pi(P\alpha \lambda^{**} + P\alpha \mu + P\alpha \tau + \alpha \lambda^{**} + \alpha \tau + \mu)}{K_1 \left(\alpha \lambda^{**2} + \alpha \lambda^{**} \mu + \alpha \lambda^{**} \tau + \lambda^{**} \mu + \mu^2 + \mu \tau \right)},
\end{aligned}
$$

$$\tag{10}$$

$$I_2^{**} = -\frac{(\epsilon K_3 + \omega K_6)\lambda^{**}\left(((\lambda^{**}+\tau)\alpha+\mu)\pi(1-p)+P\pi\alpha(\lambda^{**}+\mu+\tau)\right)\theta K_5 K_4}{(\lambda^{**}+\mu+\tau)(\alpha\lambda^{**}+\mu)K_1(\gamma(\phi K_5+\rho\sigma)\omega-K_2 K_3 K_4 K_5)},$$

$$I_3^{**} = -\frac{\lambda^{**}\left(((\lambda^{**}+\tau)\alpha+\mu)\pi(1-p)+P\pi\alpha(\lambda^{**}+\mu+\tau)\right)\theta\left((\epsilon\gamma\phi+K_2 K_4 K_6)K_5+\gamma\rho\epsilon\sigma\right)}{(\lambda^{**}+\mu+\tau)(\alpha\lambda^{**}+\mu)K_1\left((\gamma\omega\phi-K_2 K_3 K_4)K_5+\gamma\omega\rho\sigma\right)},$$

$$I_t^{**} = -\frac{(\epsilon K_3 + \omega K_6)\lambda^{**}\left(((\lambda^{**}+\tau)\alpha+\mu)\pi(1-p)+P\pi\alpha(\lambda^{**}+\mu+\tau)\right)(\phi K_5+\rho\sigma)\theta}{\left((\gamma\omega\phi-K_2 K_3 K_4)K_5+\gamma\omega\rho\sigma\right)K_1(\alpha\lambda^{**}+\mu)(\lambda^{**}+\mu+\tau)},$$

$$A^{**} = -\frac{(\epsilon K_3 + \omega K_6)\lambda^{**}\left(((\lambda^{**}+\tau)\alpha+\mu)\pi(1-p)+P\pi\alpha(\lambda^{**}+\mu+\tau)\right)\theta\rho K_4}{(\lambda^{**}+\mu+\tau)(\alpha\lambda^{**}+\mu)K_1(\gamma(\phi K_5+\rho\sigma)\omega-K_2 K_3 K_4 K_5)}.$$

Existence of Endemic Equilibrium

Descartes rule of signs is applied in determining the existence of EEP.

Descartes rule of sign

Let $P(x)$ be a polynomial of degree n such that $n \geq 2$ with real coefficients. The number of positive roots or zeros of P is equal to the number of changes of sign of $P(x)$ or less by an even number. The force of infection at equilibrium of 1 is given by;

$$\lambda^{**} = \beta\left(\frac{\eta_1 I_1^{**} + \eta_2 I_2^{**} + I_t^{**}}{N^{**}}\right), \tag{11}$$

where,

$$N^{**} = S_u^{**} + S_1^{**} + I_2^{**} + I_3^{**} + I_t^{**} + A^{**}.$$

Substituting Eq. (10) into Eq. (11) gives $\lambda^{**} = 0$ and the quadratic equation in terms of λ^{**} stated as follows;

$$C_1 \lambda^{**^2} + C_2 \lambda^{**} + C_3 = 0, \tag{12}$$

where,

$$
\begin{aligned}
C_1 =\ & \alpha\big[(\phi+\rho)(\mu+\delta_2)\omega K_5+(\mu+\delta_1)\rho\gamma\omega+(K_2+\omega)\mu K_4 K_5 \\
& + \theta(\epsilon K_3+\omega K_6)(K_4 K_5+\phi K_5+\rho\sigma+\rho K_4)\big]+\alpha\left[(\epsilon\gamma\phi+K_2 K_4 K_6)\theta K_5+\gamma\rho\sigma\epsilon\theta\right], \\
C_2 =\ & [(\phi+\rho)(\mu+\delta_2)\omega K_5+(\mu+\delta_1)\rho\gamma\omega+(K_2+\omega)\mu K_4 K_5][K_1(\alpha(1-p)+p)+\alpha(\mu+\tau p)+\mu(1-p)] \\
& + (\epsilon K_3+\omega K_6)\theta(K_4 K_5+\phi K_5+\rho\sigma+\rho K_4)+(\epsilon\gamma\phi+K_2 K_4 K_6)\theta K_5+\gamma\rho\sigma\epsilon\theta \\
& - \alpha\beta\left[\eta_1\left((\phi+\rho)(\mu+\delta_2)\omega K_5+(\mu+\delta_1)\rho\gamma\omega+(K_2+\omega)\mu K_4 K_5\right)\right] \\
& + (\epsilon K_3+\omega K_6)\theta(\eta_2 K_4 K_5+\phi K_5+\rho\sigma), \\
C_3 =\ & K_1(\mu+\tau)\left[(\phi+\rho)(\mu+\delta_2)\omega K_5+(\mu+\delta_1)\rho\gamma\omega+(K_2+\omega)\mu K_4 K_5\right](1-R_o).
\end{aligned}
$$

The following results from the quadratic equation (12) are validated using the theorem below.

Theorem 4 *The endemic equilibrium (EE) of model* (1) *has a unique positive equilibrium whenever $R_0 > 1$.*

- (a) If $C_2 > 0$ and
  (i) $C_3 \geq 0$, model (1) has no positive equilibrium
  (ii) $C_3 < 0$, model (1) has a unique positive equilibrium
- (b) If $C_2 < 0$ and $C_3 > 0$ and
  (i) $C_2^2 - 4C_1 C_3 > 0$ model (1) has two positive equilibria

(ii) $C_2^2 - 4C_1C_3 = 0$ model (1) has a unique positive equilibrium

(iii) $C_2^2 - 4C_1C_3 < 0$ model (1) has no positive equilibrium

- (c) If $C_2 < 0$ and $C_3 \leqslant 0$, model (1) has a unique positive equilibrium.

Clearly, C_1 is always a positive real number, since all the parameters of the model are positive. The following cases are considered.

- case 1:if $C_2 > 0$ and $C_3 \geq 0 \iff R_0 \leqslant 1$
  The quadratic equation (12) has no positive real root, which implies the model has no positive equilibrium.
- Vase 2: if $C_2 > 0$ and $C_3 < 0 \iff R_0 > 1$
  The quadratic equation (12) has one positive real root, which implies the model has a unique positive equilibrium.
- Case 3: if $C_2 < 0$ and $C \leqslant 0 \iff R_0 \leqslant 1$
  The quadratic equation (12) has one positive real root, which implies the model has a unique positive equilibrium.
- Case 4: if $C_2 < 0$ and $C_3 > 0 \iff R_0 < 1$
  Quadratic equation (12) has either two, one or no positive real root depending on $C_2^2 - 4C_1C_3$, which implies the model has either two positive equilibria, unique positive equilibrium or no positive equilibrium.

From Case (4) when $R_0 < 1$, we have two equilibria exist. This implies the occurrence of backward bifurcation in model (1).

Bifurcation analysis

The HIV/AIDS dynamics model (1) exhibits backward (Subcritical) bifurcation near $R_o = 1$, that is the coexistence of disease-free equilibrium and endemic equilibrium when $R_0 < 1$. The epidemiological consequence of backward bifurcation is that $R_0 < 1$ will not guarantee the condition for disease control.

Centre manifold theorem stated in [25] is applied to model (1) for bifurcation analysis, to analyze the stability near disease-free equilibrium at $R_0 = 1$.

Let $\beta = \beta^{**}$ be the bifurcation parameter and $R_0 = 1$,

which implies;

$$\beta^{**} = \frac{k_1((-\gamma\omega\phi + K_2K_3K_4)K_5 - \gamma\omega\rho\sigma)(\tau+\mu)}{(((K_4K_6\eta_2\theta + \phi(-\gamma\eta_1 + K_6\theta))\omega + ((\varepsilon\theta\eta_2 + \eta_1K_2)K_4 + \phi\varepsilon\theta)K_3)K_5 + \rho\sigma((-\gamma\eta_1 + K_6\theta)\omega + K_3\varepsilon\theta))(\mu(1-p) + \alpha(\tau+\mu p))}.$$

The Theorem is applied by making the change of variables, let,

$S_u = x_1$, $S_e = x_2$, $I_1 = x_3$, $I_2 = x_4$, $I_3 = x_5$, $I_t = x_6$ and $A = x_7$ such that, $N = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7$.

Therefore; the equation of model (1) can be written in the form:

$$\frac{dX}{dt} = (f_1, f_2, f_3, f_4, f_5, f_6, f_7)^T,$$

such that

$$\frac{dX_1}{dt} = f_1 = \pi(1-p) - (\mu + \tau)x_1 - \frac{\beta^{**}x_1(\eta_1 x_3 + \eta_2 x_4 + x_6)}{N},$$

$$\frac{dX_2}{dt} = f_2 = \pi p + \tau x_1 - \mu x_2 + \frac{\alpha\beta^{**}x_2(\eta_1 x_3 + \eta_2 x_4 + x_6)}{N},$$

$$\frac{dX_3}{dt} = f_3 = \frac{\beta^{**}x_1(\eta_1 x_3 + \eta_2 x_4 + x_6)}{N} + \frac{\alpha\beta^{**}x_2(\eta_1 x_3 + \eta_2 x_4 + x_6)}{N} - (\mu + \theta)x_3,$$

$$\frac{dX_4}{dt} = f_4 = \epsilon\theta x_3 + \omega x_5 - (\mu + \rho + \phi)x_4, \tag{13}$$

$$\frac{dX_5}{dt} = f_5 = (1-\epsilon)\theta x_3 + \gamma x_6 - (\mu + \omega)x_5,$$

$$\frac{dX_6}{dt} = f_6 = \phi x_4 + \sigma x_7 - (\mu + \gamma + \delta_2)x_6,$$

$$\frac{dX_7}{dt} = f_7 = \rho x_4 - (\mu + \sigma + \delta_1)x_7.$$

Now, the Jacobian matrix of system (13) at disease-free equilibrium ϵ^0 is given by,

$$J(\epsilon^0) = \begin{bmatrix} -D_1 & 0 & -\eta_1 D_2 & -\eta_2 D_2 & 0 & D_2 & 0 \\ \tau & -\eta_1 & -\eta_1 D_3 & -\eta_2 D_3 & 0 & -D_3 & 0 \\ 0 & 0 & \eta_1(D_2 + D_3) - K_3 & \eta_2(D_2 + D_3) & 0 & (D_2 + D_3) & 0 \\ 0 & 0 & \epsilon\theta & -K_2 & \omega & 0 & 0 \\ 0 & 0 & K_6\theta & 0 & -K_3 & \gamma & 0 \\ 0 & 0 & 0 & \phi & 0 & -K_4 & \sigma \\ 0 & 0 & 0 & \rho & 0 & 0 & -K_5 \end{bmatrix}, \tag{14}$$

where, $D_1 = (\mu + \tau), D_2 = \frac{\beta^{**}\mu(1-p)}{\mu+\tau}, D_3 = \frac{\alpha\beta^{**}(\tau+\mu p)}{\mu+\tau}, K_1 = (\theta + \mu), K_2 = (\rho + \phi + \mu),$
$K_3 = (\omega + \mu), K_4 = (\gamma + \mu + \delta_2), K_5 = (\sigma + \mu + \delta_1)$ and $K_6 = (1-\epsilon).$
The linearized system (14) with $\beta = \beta^{**}$ has a zero eigenvalues. Now, let $V = [v_1, v_2, v_3, v_4, v_5, v_6, v_7]$ and $W = [w_1, w_2, w_3, w_4, w_5, w_6, w_7]^T$ be the corresponding left and right eigenvectors associated with the simple zero eigenvalues of the Jacobian Matrix of system (14), respectively.
Solving for the left eigenvectors W we have,

$$w_1 = \frac{\rho\mu K_4 w_2 + \eta_1\rho D_3 K_4 w_3 + D_3(\eta_2 K_4 K_5 + \rho\sigma + \phi K_5)w_7}{\rho K_4} > 0,$$

$$w_2 = w_2 > 0, w_3 = w_3 > 0, w_4 = \frac{K_5 w_7}{\rho} > 0,$$

$$w_5 = \frac{\rho\theta K_4 K_6 w_3 + \gamma(\rho\sigma + \phi K_5)w_7}{\rho K_3 K_4} > 0,$$

$$w_6 = \frac{(\rho\sigma + \phi K_5)w_7}{\rho K_4} > 0, w_7 = w_7 > 0.$$

Similarly, solving for the right eigenvectors V we have,

$$v_1 = v_2 = 0, v_3 = \frac{K_4 K_5 v_7 - \rho \gamma v_5}{\rho(D_2 + D_3)} > 0,$$

$$v_4 = \frac{K_3 v_5}{\omega} < 0, v_5 = v_5 < 0, v_6 = \frac{K_5 v_7}{\sigma} > 0, v_7 = v_7 > 0.$$

We used [25] as stated by [10] to find the direction of the bifurcation by computing a and b with,

$$a = \sum_{k,i,j=1}^{7} v_k w_i w_j \frac{\partial^2 f_k}{\partial x_i \partial x_j}(0,0), \quad b = \sum_{k,i=1}^{7} v_k w_i \frac{\partial^2 f_k}{\partial x_i \partial \beta^{**}}(0,0).$$

Now, computing the partial derivatives of system (13) which are non-zero. Since $v_1 = v_2 = 0$, and the second partial derivative of f_4, f_5, f_6 and f_7 are zeros, we only consider for $k = 3$ that is;

$$\frac{dX_3}{dt} = f_3 = \frac{\beta^{**} x_1(\eta_1 x_3 + \eta_2 x_4 + x_6)}{N} + \frac{\alpha \beta^{**} x_2(\eta_1 x_3 + \eta_2 x_4 + x_6)}{N} - (\mu + \theta)x_3.$$

We get;

$$\frac{\partial^2 f_3}{\partial x_1 \partial x_3} = \frac{\beta^{**} \eta_1}{N}, \frac{\partial^2 f_3}{\partial x_1 \partial x_4} = \frac{\beta^{**} \eta_2}{N}, \frac{\partial^2 f_3}{\partial x_1 \partial x_6} = \frac{\beta^{**}}{N},$$

$$\frac{\partial^2 f_3}{\partial x_2 \partial x_3} = \frac{\alpha \beta^{**} \eta_1}{N}, \frac{\partial^2 f_3}{\partial x_2 \partial x_4} = \frac{\alpha \beta^{**} \eta_2}{N}, \frac{\partial^2 f_3}{\partial x_2 \partial x_6} = \frac{\alpha \beta^{**}}{N}.$$

Therefore;

$$a = v_3 \sum_{i,j=3}^{7} w_i w_j \frac{\partial^2 f_3}{\partial x_i \partial x_j}(0,0) = \beta^{**} v_3(w_1 + \alpha w_2)(\eta_1 w_3 + \eta_2 w_4 + w_6) > 0.$$

Similarly;

$$b = v_3 \sum_{i=3}^{7} w_i \frac{\partial^2 f_3}{\partial x_i \partial \beta^{**}}(0,0) = \frac{v_3(\eta_1 w_3 + \eta_2 w_4 + w_6)(\mu(1 - p) + \alpha(\tau + \mu p))}{\mu + \tau} > 0.$$

Thus, we have $a > 0$ and $b > 0$. The following theorem holds:

Theorem 5 *When $\beta^{**} < 0$ the system is locally asymptotically stable and there exists a positive unstable equilibrium, while if $\beta^{**} > 0$ is unstable and there exists a negative and locally asymptotically stable equilibrium. Hence the requirement of having $R_0 < 1$ will not suffice the condition for the control of HIV/AIDS.*

Global stability of endemic equilibrium

Theorem 6 *The endemic equilibrium (EE), ϵ^{**} of model (1) is globally asymptotically stable (GAS) if $R_0 > 1$ and unstable if $R_0 < 1$.*

Proof We construct a Lyapunov function

$$
V = \left(S_u - S_u^{**} - S_u^{**}ln\left(\frac{S_u}{S_u^{**}}\right)\right) + \left(S_e - S_e^{**} - S_e^{**}ln\left(\frac{S_e}{S_e^{**}}\right)\right) + \left(I_1 - I_1^{**} - I_1^{**}ln\left(\frac{I_1}{I_1^{**}}\right)\right)
$$
$$
+ B_1\left(I_2 - I_2^{**} - I_2^{**}ln\left(\frac{I_2}{I_2^{**}}\right)\right) + B_2\left(I_3 - I_3^{**} - I_3^{**}ln\left(\frac{I_3}{I_3^{**}}\right)\right) + B_3\left(I_t - I_t^{**} - I_t^{**}ln\left(\frac{I_t}{I_t^{**}}\right)\right)
$$
$$
+ B_4\left(A - A^{**} - A^{**}ln\left(\frac{A}{A^{**}}\right)\right),
$$

(15)

where B_1, B_2, B_3 and B_4 are non negative constant. Differentiating V, we get

$$
V' = \left(1 - \frac{S_u^{**}}{S_u}\right)S_u' + \left(1 - \frac{S_e^{**}}{S_e}\right)S_e' + \left(1 - \frac{I_1^{**}}{I_1}\right)I_1' + B_1\left(1 - \frac{I_2^{**}}{I_2}\right)I_2'
$$
$$
+ B_2\left(1 - \frac{I_3^{**}}{I_3}\right)I_3' + B_3\left(1 - \frac{I_t^{**}}{I_t}\right)I_t' + B_4\left(1 - \frac{A^{**}}{A}\right)A'.
$$

(16)

Substituting (1) that is $S_u', S_e', I_1', I_2', I_3, I_t'$ and A' in to (16) we have,

$$
V' = \pi(1-p) - \lambda S_u - \mu S_u - \frac{S_u^{**}}{S_u}(\pi(1-p) - \lambda S_u - \mu S_u) + \pi p - \alpha\lambda S_e - \mu S_e
$$
$$
- \frac{S_e^{**}}{S_e}(\pi p - \alpha\lambda S_e - \mu S_e) + \lambda S_u + \alpha\lambda S_e - (\mu+\theta)I_1 - \frac{I_1^{**}}{I_1}(\lambda S_u + \alpha\lambda S_e - (\mu+\theta)I_1)
$$
$$
+ B_1(\omega I_3 - (\mu+\rho)I_2) - B_1\frac{I_2^{**}}{I_2}(\omega I_3 - (\mu+\rho)I_2) + B_2(\gamma I_t - (\mu+\omega)I_3)
$$
$$
- B_2\frac{I_3^{**}}{I_3}(\gamma I_t - (\mu+\omega)I_3) + B_3(\sigma A - (\mu+\gamma)I_t) - B_3\frac{I_t^{**}}{I_t}(\sigma A - (\mu+\gamma)I_t)
$$
$$
+ B_4(\rho I_2 - (\mu+\sigma)A) - B_4\frac{A^{**}}{A}(\rho I_2 - (\mu+\sigma)A).
$$

Setting the coefficient of I_2, I_3, I_t, and A to zero, the positive constant is determined as, $B_1 = (\mu+\omega), B_2 = \omega, B_3 = \frac{(\mu+\omega)(\mu+\rho)(\mu+\sigma)}{\rho\sigma}, B_4 = \frac{(\mu+\omega)(\mu+\rho)}{\rho}$. Now setting $\delta_1 = \delta_2 = 0$ in system (1), $N \to \frac{\pi}{\mu}$ as $t \to \infty$
Let $\tilde{\beta} = \frac{\mu\beta}{\pi}$,

$$
V' = \pi(1-p) - \lambda S_u - \mu S_u - \frac{S_u^{**}}{S_u}(\pi(1-p) - \lambda S_u - \mu S_u) + \pi p - \alpha\lambda S_e - \mu S_e
$$
$$
- \frac{S_e^{**}}{S_e}(\pi p - \alpha\lambda S_e - \mu S_e) + \lambda S_u + \alpha\lambda S_e - (\mu+\theta)I_1 - \frac{I_1^{**}}{I_1}(\lambda S_u + \alpha\lambda S_e - (\mu+\theta)I_1)
$$
$$
+ (\mu+\omega)(\omega I_3 - (\mu+\rho)I_2) - (\mu+\omega)\frac{I_2^{**}}{I_2}(\omega I_3 - (\mu+\rho)I_2)
$$
$$
+ \omega(\gamma I_t - (\mu+\omega)I_3) - \omega\frac{I_3^{**}}{I_3}(\gamma I_t - (\mu+\omega)I_3)
$$
$$
+ \frac{(\mu+\omega)(\mu+\rho)(\mu+\sigma)}{\rho\sigma}(\sigma A - (\mu+\gamma)I_t) - \frac{(\mu+\omega)(\mu+\rho)(\mu+\sigma)}{\rho\sigma}\frac{I_t^{**}}{I_t}(\sigma A - (\mu+\gamma)I_t)
$$
$$
+ \frac{(\mu+\omega)(\mu+\rho)}{\rho}(\rho I_2 - (\mu+\sigma)A) - \frac{(\mu+\omega)(\mu+\rho)}{\rho}\frac{A^{**}}{A}(\rho I_2 - (\mu+\sigma)A),
$$

which can be shown from model (1) that at steady state,

$$\pi(1-p) = \lambda S_u^{**} + \mu S_u^{**}, \pi p = \alpha \lambda S_e^{**} + \mu S_e^{**}, (\mu+\theta)I_1^{**} = \lambda S_u^{**} + \alpha \lambda S_e^{**},$$
$$\omega I_3^{**} = (\mu+\rho)I_2^{**}, \gamma I_t^{**} = (\mu+\omega)I_3^{**}, \sigma A^{**} = (\mu+\gamma)I_t^{**}, \rho I_2^{**} = (\mu+\sigma)A^{**}.$$

Using the above relations, we have,

$$\begin{aligned}
V' \leqslant & \left(2 - \frac{S_u}{S_u^{**}} - \frac{S_u^{**}}{S_u}\right) + \mu S_e^{**}\left(2 - \frac{S_e}{S_e^{**}} - \frac{S_e^{**}}{S_e}\right) + \lambda S_u^{**}\left(3 - \frac{S_u^{**}}{S_u} - \frac{I_1}{I_1^{**}} - \frac{I_1^{**}}{I_1}\right) \\
& + \alpha \lambda S_e^{**}\left(3 - \frac{S_e^{**}}{S_e} - \frac{I_1}{I_1^{**}} - \frac{I_1^{**}}{I_1}\right) + (\mu+\omega)(\mu+\rho)I_2^{**}\left(3 - \frac{A^{**}}{A} - \frac{I_2^{**}I_3}{I_2 I_3^{**}} - \frac{I_t I_3^{**}}{I_t^{**} I_3}\right) \\
& + \frac{(\mu+\omega)(\mu+\rho)(\mu+\sigma)A^{**}}{\rho}\left(2 - \frac{I_t}{I_t^{**}} - \frac{I_t^{**}}{I_t}\right).
\end{aligned}$$

From the fact that the arithmetic mean surpasses the geometric mean, the following, inequalities hold:

$$\left(2 - \frac{S_u}{S_u^{**}} - \frac{S_u^{**}}{S_u}\right) \leqslant 0, \left(2 - \frac{S_e}{S_e^{**}} - \frac{S_e^{**}}{S_e}\right) \leqslant 0, \left(3 - \frac{S_u^{**}}{S_u} - \frac{I_1}{I_1^{**}} - \frac{I_1^{**}}{I_1}\right) \leqslant 0,$$

$$\left(3 - \frac{S_e^{**}}{S_e} - \frac{I_1}{I_1^{**}} - \frac{I_1^{**}}{I_1}\right) \leqslant 0, \left(3 - \frac{A^{**}}{A} - \frac{I_2^{**}I_3}{I_2 I_3^{**}} - \frac{I_t I_3^{**}}{I_t^{**} I_3}\right) \leqslant 0, \left(2 - \frac{I_t}{I_t^{**}} - \frac{I_t^{**}}{I_t}\right) \leqslant 0.$$

Thus, we have that, $V' \leqslant 0$ for $R_0 > 1$. The equality condition $V' = 0$ will strictly hold only when $S_e = S_e^{**}, I_1 = I_1^{**}, I_2 = I_2^{**}, I_3 = I_3^{**}, I_t = I_t^{**}$ and $A = A^{**}$. Thus the endemic equilibrium ϵ^* is the only invariant set of model (1). Therefore, by applying the Lasalle invariance principle [26] the result follows. Therefore, the endemic equilibrium (EE) of model (1) is globally asymptotically stable (GAS). ∎

5 Model fitting and parameter calibration

After proposing an epidemiological model in terms of a nonlinear system of ordinary differential equations, it is of utmost importance to calibrate and estimate the suitable values of the biological parameters for the model to be of some use. It can be made possible only when one reaches authentic information about the actual data set for the epidemic over a certain period. This approach also helps one validate the proposed model for the disease under analysis. Several methods exist for calibrating and estimating such parameters, including the single shooting method, Gauss-Newton method, Nelder-Mead method, least squares, Monte Carlo sampling, and the local smoothing approach. Among the existing ones, the practice of least-squares is the most frequently used statistical approach for parameters' calibration in a nonlinear system of ordinary differential equations. The method minimizes discrepancies between actual data and the values predicted from the model's simulations for the infected class. The available data mainly presents the individuals infected with the disease. With the help of the least-squares method, the suitable estimated values of the parameters accompanying other essential information, including standard error, t-statistic, p-value, and the confidence interval, are computed in Table 2 wherein all p-values are < 0.05 with 95% confidence interval for each parameter. It may also be noted that the approximate value for the basic reproduction number is $R_0 = 3.85142$ while using the calibrated parameters given above and those estimated with the nonlinear least-squares method as shown in

Table 2. The most crucial statistical information, including minimum value, first quartile, second quartile, average, third quartile, maximum value, and standard deviation is collected in Table 3 for both real HIV cases and those predicted from the simulations of model (1). The values are in perfect agreement with each other for both cases. It is further ascertained in Figure 2 wherein the curve from the simulation of the newly infected cases (solid line) approaches the real HIV cases (solid dots) very well and thus in good alignment with the surveillance data. The residual plots of varying nature in Figure 3 are shown, with the x-axis being the period (1987-2014) and the y-axis for the residual values. These residuals seem to be equally and randomly spaced around the horizontal axis, making an ideal residual plot. Finally, a box and whisker plot is created in Figure 4 to obtain the additional detail for the analysis carried out in this section. The five statistics from the plot seem to have a good agreement, with the predicted one having an outlier.

Table 2. Values of fitted biological parameters including some important statistical measures obtained via least-squares non-linear curve fitting technique.

Par	Estimate	Standard error	t-statistic	p-value	Confidence interval
β	6.08348×10^{-1}	1.93409×10^{-2}	3.14541×10^{1}	1.26038×10^{-21}	$(5.68515 \times 10^{-1}, 6.48182 \times 10^{-1})$
ρ	7.14263	1.77252×10^{-1}	4.02965×10^{1}	2.8928×10^{-24}	$(6.77758, 7.50769)$
τ	5.51178×10^{-2}	6.80934×10^{-3}	8.09445	1.89584×10^{-8}	$(4.10937 \times 10^{-2}, 6.91419 \times 10^{-2})$

Table 3. Summary statistics for the real data, and the predicted data points obtained under simulations of model (1) for the newly infected individuals with HIV (I_1).

Data	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD
Real cases	6.1e+01	3.4750e+02	8.4600e+02	1.5013e+03	2.4720e+03	4.9460e+03	1.4877e+03
Predicted	6.1e+01	4.4027e+02	9.6770e+02	1.5474e+03	2.3224e+03	5.1929e+03	1.4394e+03

Table 4. Baseline values and ranges for parameters of model (1).

Parameters	Baseline (Range)	Units	Sources
N	$12,976,600$	Persons	Estimated
π	$969,907$	Year^{-1}	Estimated
μ	1.438×10^{-2}	Year^{-1}	[12]
δ_1	0.8	Year^{-1}	Estimated by [12]
δ_2	0.6	Year^{-1}	Estimated by [10]
p	0.02	Year^{-1}	fitted
τ	5.51178×10^{-2}	Year^{-1}	fitted
β	6.08348×10^{-1}	Year^{-1}	fitted
α	0.002	Year^{-1}	Estimated by [10]
θ	0.005	Year^{-1}	Estimated by [27]
ϵ	0.04	Year^{-1}	Estimated by [27]
ϕ	0.02	Year^{-1}	fitted
ρ	7.14263	Year^{-1}	fitted
γ	0.8	Year^{-1}	fitted
ω	0.93	Year^{-1}	fitted
η_1	0.6	Year^{-1}	fitted
η_2	0.5	Year^{-1}	fitted
σ	0.004	Year^{-1}	fitted

Figure 2. The best curve fitting for the real HIV cases [13] and the compartment of the newly infected cases from the proposed model given in model (1)
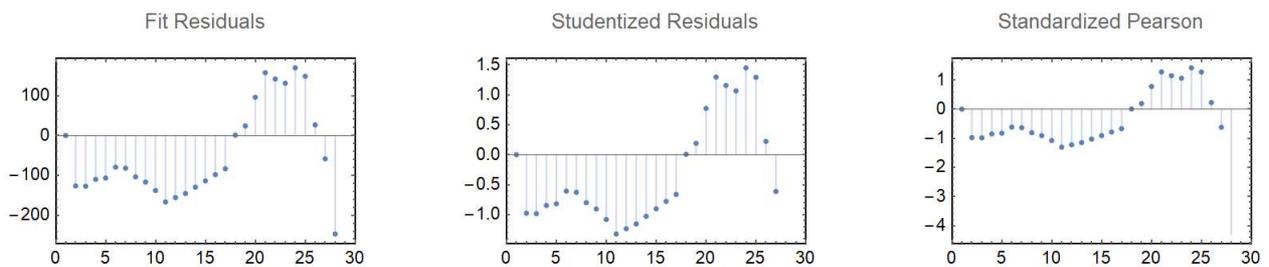
Figure 3. The residuals
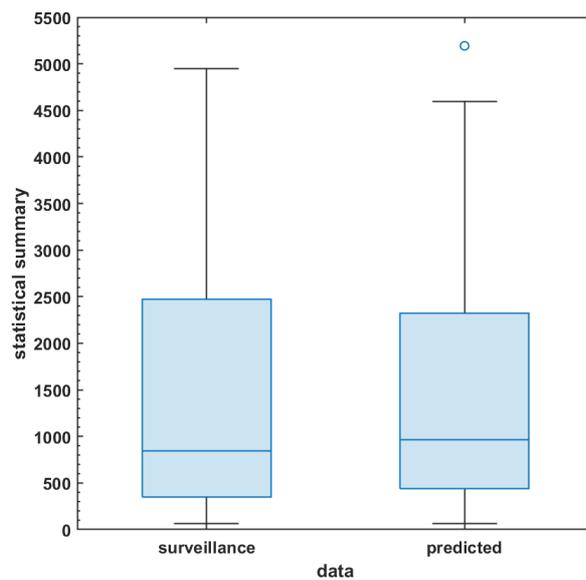
Figure 4. The BoxWhisker chart for each real surveillance HIV data value and those predicted from the proposed model (1)

Sensitivity analysis

In this section, we use the forward sensitivity index method to analyze the proposed HIV model in relation to the reproduction number R_0 with respect to the biological parameters used in the model. The method used to describe the sign of each parameter to determine the most sensitive parameters used in the model, those parameters with the negative sign are regarded as the most sensitive for decreasing the value of R_0 while parameters with positive values are sensitive for the increase of R_0 [28, 29]. The normalized local sensitivity index of R_0 with respect to the parameters is given by,

$$\chi_C^{R_0} = \frac{C}{R_0} \times \frac{\partial R_0}{\partial C}. \tag{17}$$

The indices for R_0 with respect to parameters are obtained as shown in Table 5.

Table 5. Forward normalized sensitivity indices

Parameters	Elasticity indices	Values of the elasticity index
θ	$\chi_\theta^{R_0}$	-0.2532
σ	$\chi_\sigma^{R_0}$	-0.01304
ω	$\chi_\omega^{R_0}$	0.0317
β	$\chi_\beta^{R_0}$	1.0000
δ_2	$\chi_{\delta_2}^{R_0}$	0.4789
α	$\chi_\alpha^{R_0}$	0.0096
τ	$\chi_\tau^{R_0}$	-0.0600
ρ	$\chi_{\eta_2}^{R_0}$	0.070
ϕ	$\chi_\phi^{R_0}$	-0.1493

Figure 5. Bar chart pictorial representation

The most sensitive epidemiological parameters that effectively determined the control of the spread of HIV infection are obtained and represented using a bar chart given in the forward normalized sensitivity indices Table 5.

6 Numerical scenarios and discussion

The transmission dynamics of the governing model may be efficiently investigated by utilizing numerical simulations using state variables of interest. This section looks at several forms of time-series graphs using the parameters determined by the nonlinear minimum-squares fitting approach. The transmission dynamics of the model have been simulated by using state variables and the parameters in Table 4. The behaviour of the state variables and pattern of movement from one compartment to another are examined.

(a)   **(b)**
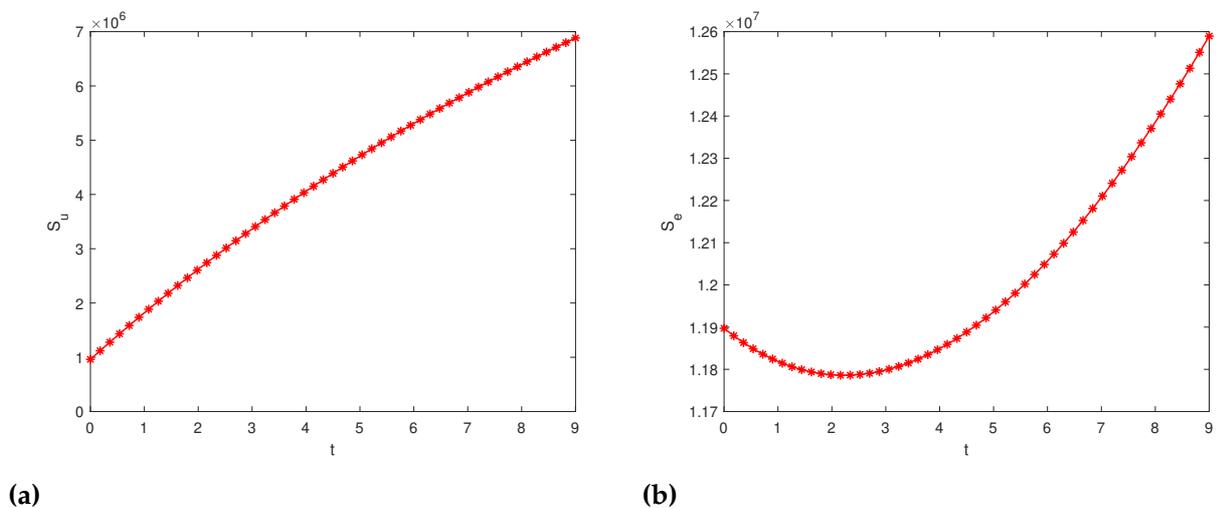
Figure 6. Behavior of the state variables (a) Susceptible Uneducated S_u, (b) Susceptible Educated S_e

(a)   **(b)**
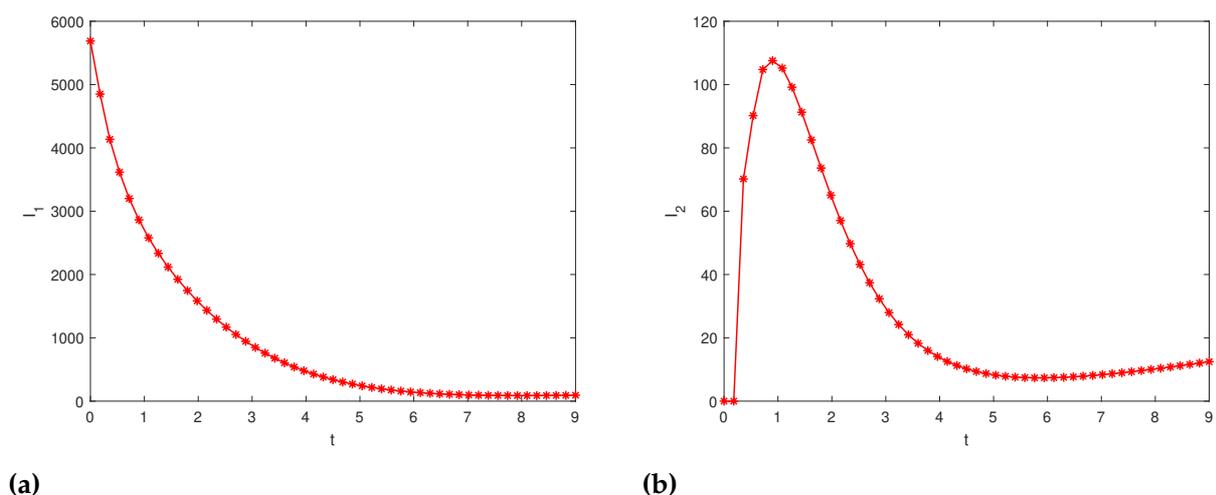
Figure 7. Behavior of the state variables (a) Newly infected individuals I_1, (b) Infected individuals with detectable viral load I_2

Figure 6(a) shows how the number of uneducated susceptible changes with time. The numbers

rise initially due to the recruitment of individuals and start decreasing due to the movement of individuals from the compartment to susceptible educated or newly infected compartment after being infected. Figure 6(b) shows how the number of susceptible educated rises with respect to time due to the movement of individuals from the uneducated compartment. Figure 7(a) shows how the newly infected individuals decrease persistently due to their movement into either infected individuals with undetectable viral load or infected individuals with detectable viral load. Figure 7(b) shows how the infected with detectable viral load raises due to the movement of newly infected individuals into the compartment, while decreases due to the movement of individuals into treatment or AIDS compartment.

(a)

(b)
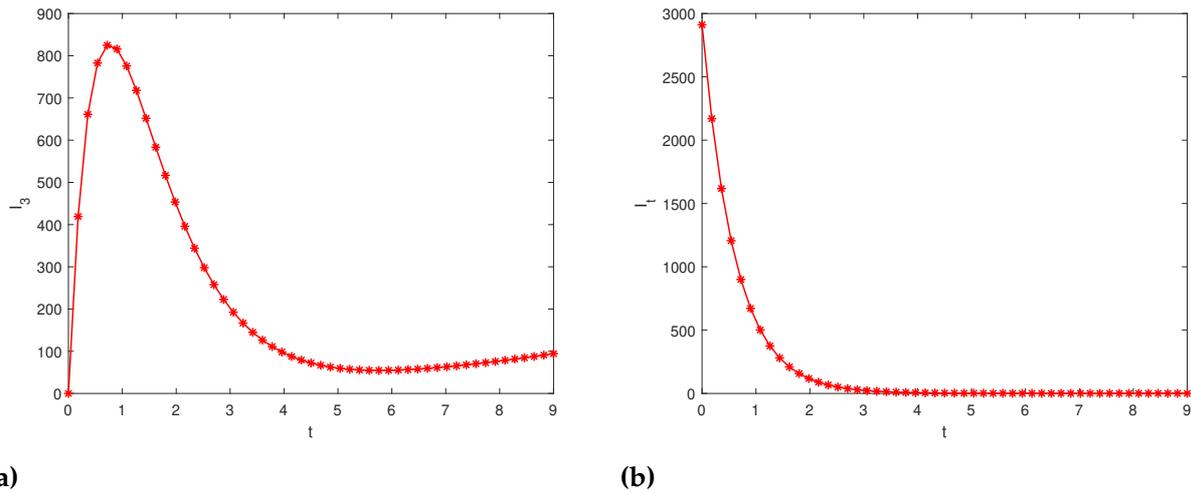
Figure 8. Behavior of the state variables (a) Infected individuals with undetectable viral load I_3, (b) Infected individuals under treatment I_t
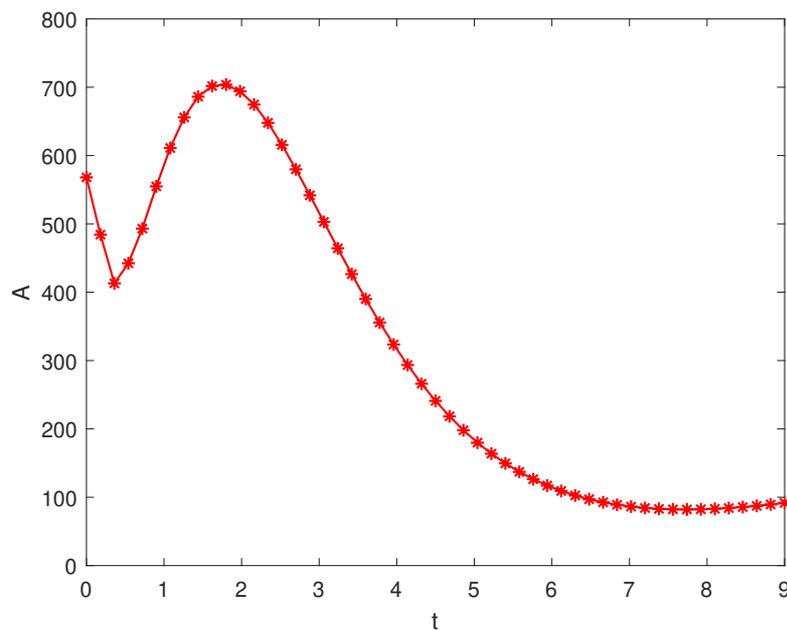
Figure 9. Behavior of the state variables AIDS patients A

(a)   **(b)**
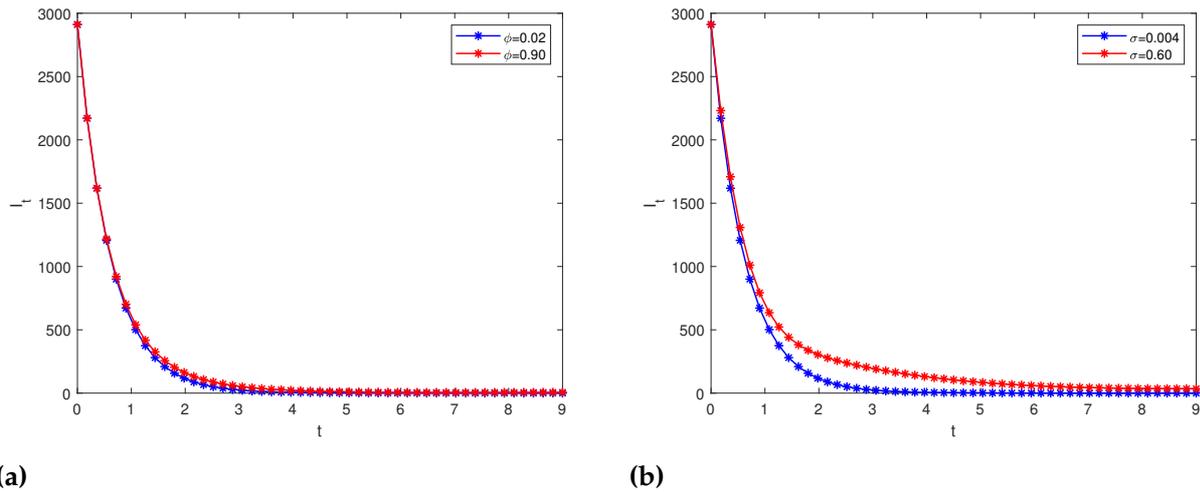
Figure 10. Patterns of I_t: (A) with different values of movement rate of infected with detectable viral load to treatment ϕ, (B) with different values of rate of movement of AIDS patients to treatment class at a rate σ

Figure 8(a) shows how infected individuals with undetectable viral load raises due to the movement of newly infected individuals into the compartment, while decreases after their viral load becomes detectable and moves into infected individuals with detectable viral load compartment. Figure 8(b) shows how infected individuals under treatment move out of the compartment to infected individuals with undetectable viral load after their viral load is suppressed.

(a)   **(b)**
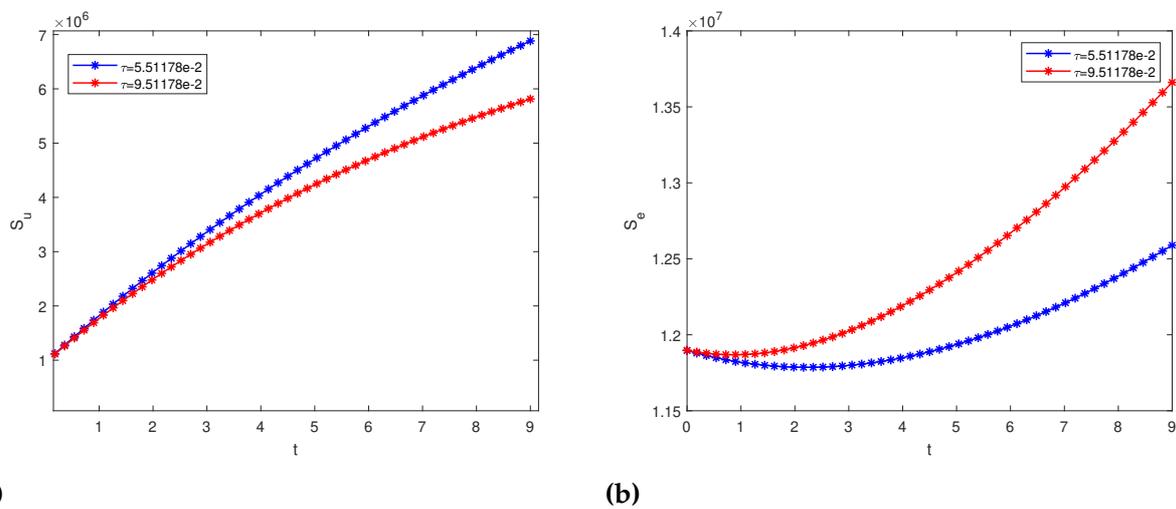
Figure 11. (a) Patterns of S_u with different values of rate at which uneducated susceptible become educated about HIV infection τ, (b) Patterns of S_e with different values of rate of movement of AIDS patients to treatment class at a rate τ

Figure 9 shows how AIDS patients decrease initially, possibly due to death, and later increase due to movement of infected individuals with detectable viral load, and decreases due to movement into treatment and death-related illness. Figures 10(a) and 10(b) show the pattern of infected individuals under treatment with different values of progression rate from undetectable viral load to treatment ϕ and progression rate from AIDS into treatment σ respectively. Despite the change in values of ϕ and σ, the graphs follow the exact pattern of I_t as in Figure 7(b). Change in values of ϕ and σ will be effective on the number of infected individuals with undetectable viral load and

AIDS-infected individuals, respectively. Figures 11(a) and 11(b) show the pattern of uneducated and educated susceptible individuals with different levels of education campaigns. The plots clearly show the impact of education parameter τ. As τ increases, the number of uneducated individuals decreases while educated individuals increase.

7   Summary and conclusion

In this paper, we have developed a nonlinear deterministic model that incorporates public awareness and treatment for the transmission dynamics of HIV/AIDS in an infected population with detectable and undetectable viral load. The analysis of the model reveals that the disease-free equilibrium is globally asymptotically stable whenever the associated reproduction number $R_0 < 1$ and unstable when $R_0 > 1$. Contrarily, the endemic equilibrium is globally asymptotically stable when the associated reproduction number is $R_0 > 1$ and unstable when $R_0 < 1$. Furthermore, the model undergoes the phenomenon of backward bifurcation in which a stable disease-free equilibrium coexists with a stable endemic equilibrium. The epidemiological implication of backward bifurcation is $R_0 < 1$ is a necessary but not sufficient condition for HIV control even when the classical requirement is satisfied, however the backward bifurcation analysis shows that when the bifurcation parameter $\beta^{**} < 0$ the system is locally asymptotically stable and there exists a positive unstable equilibrium, while if $\beta^{**} > 0$ is unstable and there exists a negative and locally asymptotically stable equilibrium. Hence, the requirement of having $R_0 < 1$ will not suffice the condition for the control of HIV/AIDS. The biological parameters of the model are fitted using the least square method with p-values < 0.05 and 95% confidence interval as shown in Table 2. The model was fitted with real HIV data cases on the newly infected compartment as shown in Figure 2. The most sensitive parameters for the control of the spread of HIV are identified using the forward sensitivity index method as shown in Figure 5, the most sensitive parameters that increase R_0 are β, ρ and ω, respectively. In addition, the numerical simulations carried out show the behavior of the state variables as shown in Figures 6,7,8, and 9. Similarly, Figure 11 shows the impact of public awareness. Finally, the results show that public awareness will help in curtailing the spread of HIV infection, and when treatment is applied to infected individuals with detectable viral load can easily suppress their virus to become undetectable so that they cannot transmit HIV through sexual intercourse. Future research should extend public awareness to infected individuals. In addition to this, a fractional order differential equation system can be used to describe HIV/AIDS transmission dynamics incorporating viral load detectability as the order has an effect on the dynamics and an optimal control problem can be applied to determine the optimal strategies for HIV eradication.

Declarations

Ethical approval

The authors state that this research complies with ethical standards. This research does not involve either human participants or animals.

Consent for publication

Not applicable.

Conflicts of interest

The authors declare that they have no conflict of interest.

References

[1] Swain, S.L., McKinstry, K.K. and Strutt, T.M. Expanding roles for CD4+ T cells in immunity to viruses. *Nature Reviews Immunology,* 12, 136-148, (2012). [CrossRef]

[2] Wilson, D.P., Law, M.G., Grulich, A.E., Cooper, D.A. and Kaldor, J.M. Relation between HIV viral load and infectiousness: a model-based analysis. *The Lancet,* 372(9635), 314-320, (2008). [CrossRef]

[3] Center for Disease Control and Prevention. About HIV *https://www.cdc.gov/hiv/basics/whatishiv.html,* (2021).

[4] World Health Organization. HIV and AIDS *www.who.int/news-room/fact-sheets/detail/hiv-aids,* (2021).

[5] HIV.gov. The Global HIV and AIDS Epidemic *https://www.hiv.gov/hiv-basics/overview/data-and-trends/global-statistics,* (2021).

[6] Podder, C.N., Sharomi, O., Gumel, A.B. and Strawbridge, E. Mathematical analysis of a model for assessing the impact of antiretroviral therapy, voluntary testing and condom use in curtailing the spread of HIV. *Differential Equations and Dynamical Systems,* 19, 283-302, (2011). [CrossRef]

[7] Gaardbo, J.C., Hartling, H.J., Gerstoft, J. and Nielsen, S.D. Thirty years with HIV infection—nonprogression is still puzzling: lessons to be learned from controllers and long-term nonprogressors. *AIDS Research and Treatment,* 2012, (2012). [CrossRef]

[8] Mandalia, S., Westrop, S.J., Beck, E.J., Nelson, M., Gazzard, B.G. and Imami, N. Are long-term non-progressors very slow progressors? Insights from the Chelsea and Westminster HIV cohort, 1988–2010. *PLoS One,* 7(2), e29844, (2012). [CrossRef]

[9] Nam aidsmap HIV & AIDS-sharing knowledge, changing lives. Viral load *https://www.aidsmap.com/about-hiv/viral-load,* (2017).

[10] Hussaini, N., Winter, M. and Gumel, A.B. Qualitative assessment of the role of public health education program on HIV transmission dynamics. *Mathematical Medicine and Biology: A Journal of the IMA,* 28(3), 245-270, (2011). [CrossRef]

[11] Waziri, A.S., Massawe, E.S. and Makinde, O.D. Mathematical modelling of HIV/AIDS

dynamics with treatment and vertical transmission. *Applied Mathematics,* 2(3), 77-89, (2012). [CrossRef]

[12] Bhunu, C.P., Mushayabasa, S., Kojouharov, H. and Tchuenche, J.M. Mathematical analysis of an HIV/AIDS model: impact of educational programs and abstinence in sub-Saharan Africa. *Journal of Mathematical Modelling and Algorithms,* 10, 31-55, (2011). [CrossRef]

[13] Silva, C.J. and Torres, D.F. A SICA compartmental model in epidemiology with application to HIV/AIDS in Cape Verde. *Ecological Complexity*, 30, 70-75, (2017). [CrossRef]

[14] Naik, P.A., Owolabi, K.M., Yavuz, M. and Zu, J. Chaotic dynamics of a fractional order HIV-1 model involving AIDS-related cancer cells. *Chaos, Solitons & Fractals,* 140, 110272, (2020). [CrossRef]

[15] Naik, P.A., Zu, J. and Owolabi, K.M. Modeling the mechanics of viral kinetics under immune control during primary infection of HIV-1 with treatment in fractional order. *Physica A: Statistical Mechanics and its Applications,* 545, 123816, (2020). [CrossRef]

[16] Naik, P.A., Yavuz, M. and Zu, J. The role of prostitution on HIV transmission with memory: A modeling approach. *Alexandria Engineering Journal,* 59(4), 2513-2531, (2020). [CrossRef]

[17] Naik, P.A., Zu, J. and Owolabi, K.M. Global dynamics of a fractional order model for the transmission of HIV epidemic with optimal control. *Chaos, Solitons & Fractals,* 138, 109826, (2020). [CrossRef]

[18] Ahmed, I., Akgül, A., Jarad, F., Kumam, P. and Nonlaopon, K. A Caputo-Fabrizio fractional-order cholera model and its sensitivity analysis. *Mathematical Modelling and Numerical Simulation with Applications,* 3(2), 170-187, (2023). [CrossRef]

[19] Ghosh, D., Santra, P.K. and Mahapatra, G.S. A three-component prey-predator system with interval number. *Mathematical Modelling and Numerical Simulation with Applications,* 3(1), 1-16, (2023). [CrossRef]

[20] Sabbar, Y. Asymptotic extinction and persistence of a perturbed epidemic model with different intervention measures and standard lévy jumps. *Bulletin of Biomathematics,* 1(1), 58-77, (2023). [CrossRef]

[21] Joshi, H., Yavuz, M. and Stamova, I. Analysis of the disturbance effect in intracellular calcium dynamic on fibroblast cells with an exponential kernel law. *Bulletin of Biomathematics,* 1(1), 24-39, (2023). [CrossRef]

[22] Moore, G. AM Stuart and AR Humphries Dynamical systems and numerical analysis (Cambridge Monographs on Applied and Computational Mathematics No. 2, Cambridge University Press, 1996). *Proceedings of the Edinburgh Mathematical Society*, 41(1), 213-216, (1998). [CrossRef]

[23] Smith, H.L. *Monotone dynamical systems: an introduction to the theory of competitive and cooperative systems: an introduction to the theory of competitive and cooperative systems* (Vol. 41). American Mathematical Society, (1995).

[24] Van den Driessche P. and Wanmough, J. Reproduction numbers and sub-threshold endemic equilibria for compartimental models of disease transmition. *Mathematical Biosciences*, 180, 29-48, (2000). [CrossRef]

[25] Castillo-Charvez, C. and Song, B. Dynamical model of tuberculosis and their applications. *Mathematical Biosciences and Engineering*, 1(2), 361-404, (2004).

[26] Lasalle, J.P. The stability of dynamical systems. *Regional conference series in applied mathematics*, SIAM, Philadelphia, (1976).

[27] Berg, M.G., Olivo, A., Harris, B.J., Rodgers, M.A., James, L., Mampunza, S. et al. A high prevalence of potential HIV elite controllers identified over 30 years in Democratic Republic of Congo. *eBioMedicine*, 65, 103258, (2021). [CrossRef]

[28] Yusuf, A., Mustapha, U.T., Sulaiman, T.A., Hincal, E. and Bayram, M. Modeling the effect of horizontal and vertical transmissions of HIV infection with Caputo fractional derivative. *Chaos, Solitons & Fractals*, 145, 110794, (2021). [CrossRef]

[29] Mustapha, U.T. and Hincal, E. An optimal control of hookworm transmissions model with differential infectivity. *Physica A: Statistical Mechanics and its Applications*, 545, 123625, (2020). [CrossRef]

Mathematical Modelling and Numerical Simulation with Applications (MMNSA)
(https://dergipark.org.tr/en/pub/mmnsa)

RESEARCH PAPER

Two-dimensional Cattaneo-Hristov heat diffusion in the half-plane

Beyza Billur İskender Eroğlu [iD] [1,*,‡]

[1]Department of Mathematics, Balıkesir University, 10145 Balıkesir, Türkiye
*Corresponding Author
‡biskender@balikesir.edu.tr (Beyza Billur İskender Eroğlu)

Abstract

In this paper, Cattaneo-Hristov heat diffusion is discussed in the half plane for the first time, and solved under two different boundary conditions. For the solution purpose, the Laplace, and the sine- and exponential- Fourier transforms with respect to time and space variables are applied, respectively. Since the fractional term in the problem is the Caputo-Fabrizio derivative with the exponential kernel, the solutions are in terms of time-dependent exponential and spatial-dependent Bessel functions. Behaviors of the temperature functions due to the change of different parameters of the problem are interpreted by giving 2D and 3D graphics.

Keywords: Two-dimensional Cattaneo-Hristov equation; Laplace transform; sine-Fourier transform; exponential Fourier transform; Caputo-Fabrizio derivative

AMS 2020 Classification: 35R11; 35A22; 74A10

1 Introduction

Heat conduction, also called diffusion, is the exchange of thermal energy between different physical systems. The classical theory of heat conduction is based on Fourier's law, that is, almost 200 years ago. Fourier's law implies that infinitesimal heat changes propagate at an infinite speed. This result makes the law a paradox that cannot specifically represent microscopic heat distribution. The physical validity of this law is for heat transfer models in low dimensions and also in macroscopic scales.

To remove the inconsistency of Fourier's law for heat transfer occurring in non-homogeneous mediums or for microscopic scales, different non-local dependencies between heat flux and temperature gradient have been proposed. As a result, different types of heat conduction equations have emerged and this has led to the development of non-classical theories on heat conduction. In this sense, fractional operators with singular or non-singular kernels have played a significant

---

role in various types of real-world problems [1–6]. For instance, in a thin rectangular plate the non-local relation between the heat flux $q(t)$ and the temperature gradient $gradT = \left[\frac{\partial T}{\partial x} \frac{\partial T}{\partial y}\right]$ can be given by [7]

$$q(t) = -k \int_0^t K(t-\tau) \, gradT(x,y,\tau) \, d\tau, \tag{1}$$

where k is the coefficient of thermal conductivity. When this relation is combined with the laws of conservation of energy, it leads to the following generalized heat conduction equation [8]:

$$\frac{\partial T}{\partial t} = a \int_0^t K(t-\tau) \, \Delta T(x,y,\tau) \, d\tau. \tag{2}$$

in which a is the thermal diffusivity coefficient. The decisive factor here is the type of kernel function K which physically corresponds to the memory effects in heating systems. Some leading non-local laws with various types of kernel functions can be summarized as follows:

$$q(t) = -k \int_0^t gradT(x,y,\tau) \, d\tau \quad \text{(Full memory/without fading memory [9])}, \tag{3}$$

$$q(t) = \begin{cases} \frac{k}{\Gamma(\alpha)} \frac{\partial}{\partial t} \int_0^t (t-\tau)^{\alpha-1} \, gradT(x,y,\tau) \, d\tau, & 0 < \alpha \leq 1, \\ -\frac{k}{\Gamma(\alpha-1)} \int_0^t (t-\tau)^{\alpha-2} \, gradT(x,y,\tau) \, d\tau, & 1 < \alpha \leq 2, \end{cases} \quad \text{(long-tail memory [10])}, \tag{4}$$

$$q(t) = -\frac{k}{\xi} \int_0^t \exp\left(-\frac{t-\tau}{\xi}\right) gradT(x,y,\tau) \, d\tau \quad \text{(short-tail memory [11, 12])}, \tag{5}$$

where ξ denotes the finite relaxation time of the heating process. In Eq. (3), the kernel is constant $K = 1$ so there is no fading in memory. In Eq. (4), the relations between heat flux and temperature gradient have long-term memory power kernels $K = (t-\tau)^{\alpha-1}$ and $K = (t-\tau)^{\alpha-2}$. Thus, constitutive relations given in Eq. (4) led to the emergence of the heat equation with Caputo fractional derivative. Analytical solutions of these equations with different initial and boundary conditions and in different coordinate systems have been studied in detail by Povstenko [13–17]. Furthermore, the thermal stresses due to fractional heat conduction were researched [18–22], and even the optimal control problem of these thermal stresses was investigated later [23, 24].

The integro-differential equation with Jeffrey kernel $K = \exp\left(-\frac{t-\tau}{\xi}\right)$ based on the constitutive law stated in Eq. (5) was proposed for the damped heat diffusion in rigid conductors. A few years ago, Hristov conceived of relating the Jeffrey kernel in the Cattaneo model to the Caputo-Fabrizio fractional derivative that has a non-singular kernel [25]. The obtained model is called the Cattaneo-Hristov heat diffusion equation in the literature. This development shows that different constitutive equations can be reconstructed with non-singular fractional derivatives [26–28] which was detailed studied by researchers [29, 30]. In fact, this is a wise answer to understanding the physical background of fractional derivatives.

There are limited but undoubtedly valuable studies in the literature to find the analytical and

numerical solutions of the Cattaneo-Hristov heat equation [31–37] in one dimensional space. Although the Cattaneo-Hristov heat equation was constructed on the half-real line, it should be enlarged to the other coordinates according to the geometry of the medium heat conduction acting. In this manner, Avci [38] investigated the solution process of the Cattaneo-Hristov heat diffusion on an axial symmetrical finite cylinder and also analyzed the thermal stresses due to the heat sources applied from the boundaries of the cylinder. Motivated by this fact, the current study aims to represent the elastic heat diffusion in the half-plane. Therefore, this work focuses on solving the two-dimensional Cattaneo-Hristov equation with the Dirichlet boundary conditions by Laplace, sine- and exponential- Fourier integral transforms. To our knowledge, this is the first study on the two-dimensional Cattaneo-Hristov diffusion equation, and therefore it is possible that the work will contribute to the technological development of thermally elastic film materials.

On the other hand, we aim to investigate the harmonic temperature effect, which is a remarkable concept in the classical or fractional diffusion processes, for the Cattaneo-Hristov diffusion model. The behavior of the classical diffusion under a harmonic effect was first investigated by Ångström [39]. This physical phenomenon is referred to as "'oscillatory diffusion"' or "'diffusion waves"' in the literature. The harmonic effect on diffusion can be analyzed in two ways. In the first one, a harmonic source function is stipulated [40, 41]. On the other hand, it is considered that there is a harmonic effect at the boundary [42]. In [43], all possible harmonic effects are analyzed for a one-dimensional diffusion problem. The harmonic effect on the fractional diffusion models has been studied in the recent few years [44–47]. It should be noted that these fractional diffusion equations were described by Caputo derivative with the singular kernel. As far as is known, the current study is the first to examine the Cattaneo-Hristov diffusion process modeled with the Caputo-Fabrizio derivative under a harmonic boundary effect.

The paper is organized as follows: In Section 2, we give some preliminary definitions required for the formulation of problems. In Section 3, we obtain the fundamental solutions to the Dirichlet problem for the Dirac pulse and non-moving harmonic pulse, then evaluate the behavior of temperatures according to the change of order of the fractional derivative by the graphics. Moreover, we discuss the results from both the mathematical and physical perspectives in this section. Finally, we provide the concluding remarks in Section 4.

2 Preliminaries

The birth of fractional analysis occurred per se in the solution of Abel's tautochrone problem. In fact, Abel was unaware that he had found a new theory today known as the Riemann-Liouville fractional calculus [48]. This clearly shows us that fractional operators actually arise naturally when trying to understand physical phenomena. Fractional operators are particularly effective tools for understanding memory effects, clarifying hereditary properties, and modelling transport processes in complex environments. What is important is the accurate use and interpretation of fractional operators that differ depending on their kernel functions. As is known, the leading Riemann-Liouville and Caputo operators of conventional fractional calculus include singular kernels denoting long-tail memory. On the other hand, computational difficulties arising from the nature of these derivatives and their weakness in model processes complying with the exponential decay law have led to the emergence of the Caputo-Fabrizio and Atangana-Baleanu fractional derivatives with regular kernels.

Now, we remind the Caputo-Fabrizio fractional derivative, which also models the Cattaneo-Hristov heat diffusion discussed in the present study.

Definition 1 *[49] Let $f \in H^1(0, t)$ and $0 < \alpha < 1$, then the Caputo-Fabrizio fractional derivative is*

defined by

$$
{}^{CF}D_t^\alpha f(t) = \frac{\mathcal{N}(\alpha)}{1-\alpha} \int_0^t \frac{df(s)}{ds} \exp\left(-\frac{\alpha}{1-\alpha}(t-s)\right) ds,
\tag{6}
$$

where $\mathcal{N}(\alpha)$ denotes the normalization function satisfying $\mathcal{N}(0) = \mathcal{N}(1) = 1$.

The closed-form solution to the problem will be obtained using integral transforms. Since the Laplace transform is applied for the time variable, we indicate the Laplace transform property of the Caputo-Fabrizio derivative is as follows:

$$
\mathcal{L}\left\{{}^{CF}D_t^\alpha f(t)\right\}(s) = \frac{sf^*(s) - f(0)}{s + \alpha(1-s)}, \ 0 < \alpha \leq 1,
\tag{7}
$$

in which asterisk denotes the Laplace transform of the function. For the Dirichlet problems considered in the half plane, we apply the exponential Fourier transform via y variable [50]:

$$
\mathcal{F}\{f(y)\} = \overline{f}(\eta) = \int_{-\infty}^{\infty} f(y) e^{iy\eta} dy, \ -\infty < y < \infty,
\tag{8}
$$

with its inverse transform:

$$
\mathcal{F}^{-1}\left\{\overline{f}(\eta)\right\} = f(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \overline{f}(\eta) e^{-iy\eta} d\eta.
\tag{9}
$$

Also, the exponential-Fourier transform of the second-order derivative is reminded as

$$
\mathcal{F}\left\{\frac{d^2 f(y)}{dy^2}\right\} = -\eta^2 \overline{f}(\eta).
\tag{10}
$$

Then, we use the following sine-Fourier transform for the Dirichlet problem [50]:

$$
\mathcal{F}\{f(x)\} = \widetilde{f}(\xi) = \int_0^{\infty} f(x) \sin(x\xi) dx, \ 0 \leq x < \infty,
\tag{11}
$$

with the relevant inverse transform:

$$
\mathcal{F}^{-1}\left\{\widetilde{f}(\xi)\right\} = f(x) = \frac{2}{\pi} \int_0^{\infty} \widetilde{f}(\xi) \sin(x\xi) d\xi,
\tag{12}
$$

Since the sine-Fourier transform is used in the domain $0 \leq x < \infty$ for a prescribed Dirichlet boundary condition, we apply the following property

$$
\mathcal{F}\left\{\frac{d^2 f(x)}{dx^2}\right\} = -\xi^2 \widetilde{f}(\xi) + \xi f(0).
\tag{13}
$$

3 Statement of the problem

In this section, we aim to obtain the closed-form solutions to the Cattaneo-Hristov heat conduction problem in the half-plane. For this purpose, let us consider the initial boundary value problem defined as follows:

$$\frac{\partial T(x,y,t)}{\partial t} = a_1 \Delta T(x,y,t) + a_2(1-\alpha) \, {}^{CF}D_t^\alpha \Delta T(x,y,t),$$ (14)

$$0 < x < \infty, -\infty < y < \infty, \ 0 < t < \infty,$$

$$t = 0: \ T(x,y,0) = 0,$$ (15)

$$x = 0: \ T(0,y,t) = f(y,t),$$ (16)

where $\Delta T = \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2}$ is the Laplacian of temperature function. Firstly, applying the exponential-Fourier transform with respect to the spatial coordinate y and considering the property Eq. (10) lead to

$$\frac{\partial \overline{T}(x,\eta,t)}{\partial t} = a_1 \left(\frac{\partial^2 \overline{T}(x,\eta,t)}{\partial x^2} - \eta^2 \overline{T}(x,\eta,t) \right)$$

$$+ a_2(1-\alpha)^{CF}D_t^\alpha \left(\frac{\partial^2 \overline{T}(x,\eta,t)}{\partial x^2} - \eta^2 \overline{T}(x,\eta,t) \right).$$ (17)

with the transformed initial and boundary conditions:

$$t = 0: \ \overline{T}(x,\eta,0) = 0,$$

$$x = 0: \ \overline{T}(0,\eta,t) = \overline{f}(\eta,t).$$

Then, the sine-Fourier transform is applied according to the spatial coordinate x under the relation Eq. (13) and the result is obtained as follows:

$$\frac{\partial \widetilde{\overline{T}}(\xi,\eta,t)}{\partial t} = a_1 \left(-\left(\xi^2 + \eta^2 \right) \widetilde{\overline{T}}(\xi,\eta,t) + \xi \overline{f}(\eta,t) \right)$$

$$+ a_2(1-\alpha)^{CF}D_t^\alpha \left(-\left(\xi^2 + \eta^2 \right) \widetilde{\overline{T}}(\xi,\eta,t) + \xi \overline{f}(\eta,t) \right).$$ (18)

Finally, applying the Laplace transform to the time variable t gives the transformed solution:

$$\widetilde{\overline{T}}^*(\xi,\eta,s) = \xi \frac{(a\beta s + a_1\alpha)\overline{f}^*(\eta,s) - a_2\beta \overline{f}(\eta,0)}{\beta s^2 + [a\beta(\xi^2+\eta^2)+\alpha]s + a_1\alpha(\xi^2+\eta^2)},$$ (19)

where α is the order of the Caputo-Fabrizio derivative, a_1 and a_2 are some real constants such that

$$a_1 = \frac{k_1}{C_p\rho}, \ a_2 = \frac{k_2}{C_p\rho},$$ (20)

for effective thermal conductivity k_1 and elastic conductivity k_2. Also, C_p is the specific heat and ρ is the density of particles on the plate.

$$\beta = 1-\alpha, \ a = a_1 + a_2.$$ (21)

Inversion of the transformations gives the closed-form solution

$$T(x,y,t) = \frac{\sqrt{2}}{\pi\sqrt{\pi}} \int\limits_{-\infty}^{\infty} \int\limits_{0}^{\infty} \widetilde{\overline{T}}(\xi,\eta,t) e^{-iy\eta} \sin(x\xi)\, d\xi d\eta, \tag{22}$$

which can be arranged using Euler's formula as

$$T(x,y,t) = \frac{2\sqrt{2}}{\pi\sqrt{\pi}} \int\limits_{0}^{\infty} \int\limits_{0}^{\infty} \widetilde{\overline{T}}(\xi,\eta,t) \cos(y\eta) \sin(x\xi)\, d\xi d\eta. \tag{23}$$

To simulate the two-dimensional Cattaneo-Hristov diffusion equation, we consider two particular cases. In the 1st case, our the aim is to extend the original Cattaneo-Hristov heat diffusion problem considered for $x \in (0,\infty)$ to the half-real plane $(x,y) \in (0,\infty) \times (-\infty,\infty)$. In the 2nd case, we intend to examine the effect of the harmonic temperature function at the boundary on the Cattaneo-Hristov model, which has also an important effect on both classical and fractional heat conduction problems.

Case 1: Fundamental solution to two-dimensional Cattaneo-Hristov heat diffusion

Here, we consider the Dirac delta pulse at the boundary given by Eq. (16) for Cattaneo-Hristov heat diffusion equation:

$$T(0,y,t) = f(y,t) = \delta(y). \tag{24}$$

Substituting the exponential Fourier and Laplace transforms of this condition in Eq. (19) gives

$$\widetilde{\overline{T}}^{*}(\xi,\eta,s) = \xi \frac{a_1(\beta s + \alpha)}{\beta s^3 + [a\beta(\xi^2 + \eta^2) + \alpha]s^2 + a_1\alpha(\xi^2 + \eta^2)s}. \tag{25}$$

Next, inverting the Laplace transform reveals

$$\begin{aligned}
\widetilde{\overline{T}}(\xi,\eta,t) &= \frac{\xi}{\xi^2 + \eta^2} \left\{ \frac{1}{2}\left(\frac{C(\xi,\eta)}{B(\xi,\eta)} - 1\right) \exp\left(\frac{B(\xi,\eta) - A(\xi,\eta)}{2\beta}t\right) + 1 \right. \\
&\quad \left. - \frac{1}{2}\left(\frac{C(\xi,\eta)}{B(\xi,\eta)} + 1\right) \exp\left(\frac{-B(\xi,\eta) - A(\xi,\eta)}{2\beta}t\right) \right\},
\end{aligned} \tag{26}$$

where the notations defined in the following are used only for convenience

$$A(\xi,\eta) = a\beta\left(\xi^2 + \eta^2\right) + \alpha, \tag{27}$$

$$B(\xi,\eta) = \sqrt{a\beta^2\left(\xi^2 + \eta^2\right)^2 + 2(a_2 - a_1)a\beta\left(\xi^2 + \eta^2\right) + \alpha^2}, \tag{28}$$

$$C(\xi,\eta) = (a_2 - a_1)\beta\left(\xi^2 + \eta^2\right) - \alpha. \tag{29}$$

To make the solution suitable for numerical calculations, we need to reduce the double integral in Eq. (23) to a single integral by converting it to polar coordinates. For this purpose, we first suppose that

$$\xi = \rho \cos \theta, \eta = \rho \sin \theta, \tag{30}$$

and so we obtain

$$T(x,y,t) = \frac{2\sqrt{2}}{\pi\sqrt{\pi}} \int_0^\infty \int_0^{\frac{\pi}{2}} \widetilde{\overline{T}}_1(\rho,t) \cos(y\rho\sin\theta) \sin(x\rho\cos\theta) \rho^2 \cos\theta d\theta d\rho, \tag{31}$$

where $\widetilde{\overline{T}}_1 = \widetilde{\overline{T}}/\xi$ and since $\rho^2 = \xi^2 + \eta^2$ from Eq. (30), $\widetilde{\overline{T}}_1$ can be written as the function of (ρ, t) according to Eqs. (26)-(29). By using the change of the variable $v = \sin\theta$ and considering the following integral relation [51, 52]:

$$\int_0^1 \cos(y\rho v) \sin\left(x\rho\sqrt{1-v^2}\right) dv = \frac{\pi}{2} \frac{x}{\sqrt{x^2+y^2}} J_1\left(\sqrt{x^2+y^2}\right), \tag{32}$$

in which J_1 is the first kind Bessel function of order 1. Thereby, the closed-form solution is arrived at as

$$T(x,y,t) = \sqrt{\frac{2}{\pi}} \int_0^\infty \widetilde{\overline{T}}_1(\rho,t) \frac{x}{\sqrt{x^2+y^2}} J_1\left(\sqrt{x^2+y^2}\right) \rho^2 d\rho. \tag{33}$$

and solved by numerical computation of the improper integral. Then, the results are depicted in Figure 1.

In Figure 1(a), we aim to illustrate the dependence of heat diffusion on the variation of order of fractional derivatives. The 2D graphics show the cross-section of the temperature surface for the arbitrary values of $x = y = 0.5$. Note that the α parameter plays two critique roles in the discussed model, one as a coefficient and the other to determine the influence of fading memory. As α approaches 1, the damping memory effect weakens due to the coefficient role of α, and the temperature function tends to behave as in the classical heat equation. In the case of $\alpha = 1$, the elastic conductivity constant k_1 in the coefficient a_1 also loses its effect.

In Figure 1(b), the behavior of the temperature surface is shown for the arbitrary values of $\alpha = 0.6$ and $t = 0.5$. In this graph, the instantaneous Dirac heat pulse at the boundary of the region is clearly visible. For evolution equations such as heat conduction, examining the effects of instantaneous changes at the beginning or at the boundary is important both in obtaining fundamental solutions and in the sense of physical behavior. Due to this importance, the Dirac delta pulse effect is examined in different classical or fractional heat conduction models, as in the current study.
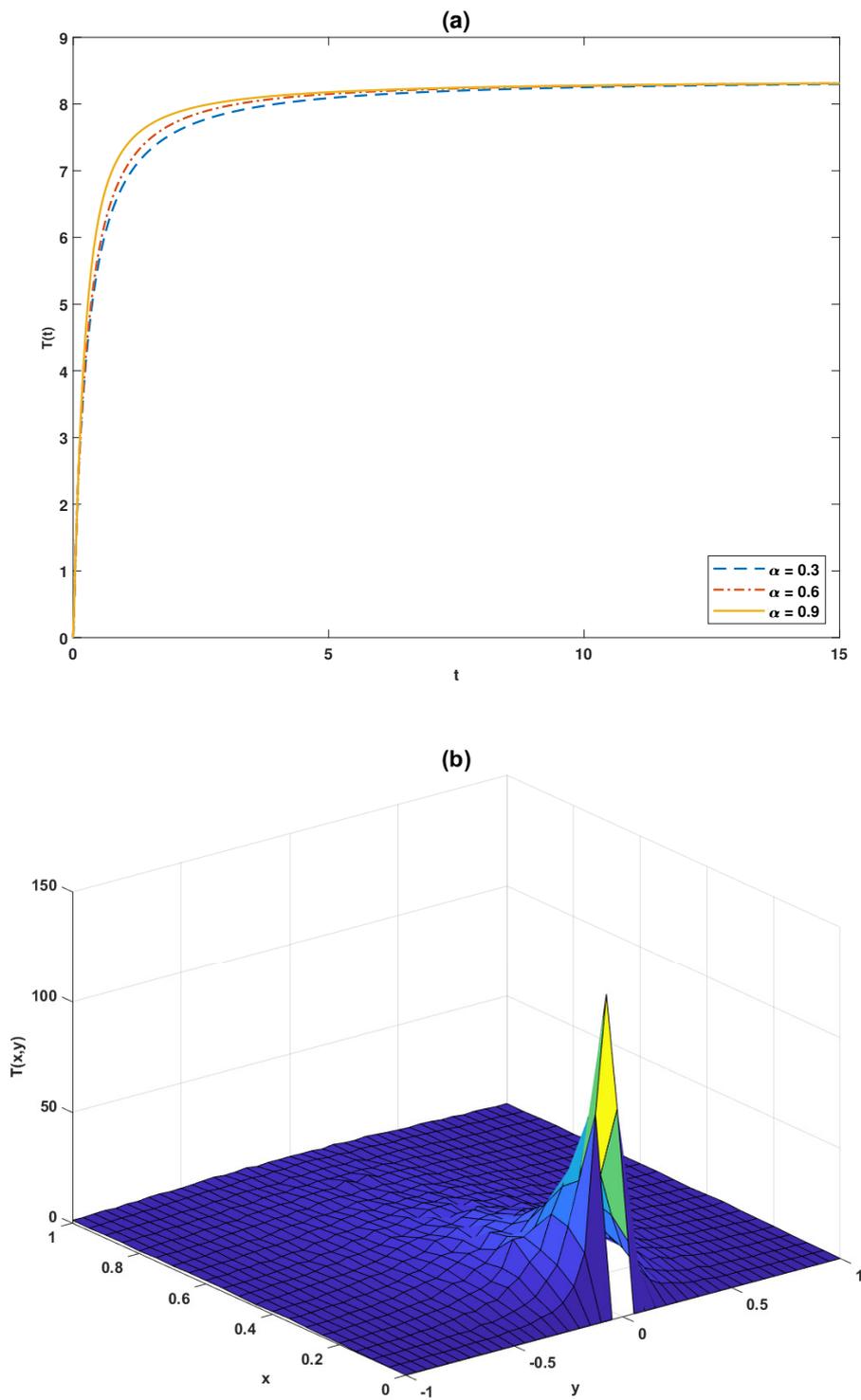
Figure 1. Temperature function for Dirac delta pulse at the boundary

Case 2: Non-moving harmonic temperature at the boundary

In this case, the behavior of the Cattaneo-Hristov heat diffusion is investigated under the effect of a time non-moving harmonic boundary temperature which is described by

$$T\left(0, y, t\right) = f\left(y, t\right) = \delta\left(y\right) \exp\left(i\omega t\right), \tag{34}$$

where ω denotes the angular frequency. Substituting integral transforms of the condition Eq. (34) into the transformed solution Eq. (19) and rearranging the results give

$$
\begin{aligned}
\widetilde{\overline{T}}^{*}(\xi,\eta,s) &= \xi\left\{\left[a_1\beta s^2 + a_1\alpha s - a_2\beta\omega^2\right] + i\left[a\beta\omega s + a_1\alpha\omega\right]\right\} \\
&\quad / \left\{\beta s^4 + \left[a\beta\left(\xi^2+\eta^2\right) + \alpha\right]s^3 + \left[\beta\omega^2 + a_1\alpha\left(\xi^2+\eta^2\right)\right]s^2 \right. \\
&\quad \left. + \left[a\beta\left(\xi^2+\eta^2\right) + \alpha\right]\omega^2 s + a_1\alpha\left(\xi^2+\eta^2\right)\omega^2\right\}.
\end{aligned}
\tag{35}
$$

For demonstration purposes, we focus on the real part of the transformed temperature function for the subsequent calculations. Inverting the Laplace transform of the real part of Eq. (35), one can obtain

$$
\begin{aligned}
\widetilde{\overline{T}}(\xi,\eta,t) &= \frac{\xi}{D(\xi,\eta)}\left\{\frac{E(\xi,\eta)}{2B(\xi,\eta)}\left[\exp\left(\frac{B(\xi,\eta)-A(\xi,\eta)}{2\beta}t\right) - \exp\left(\frac{-B(\xi,\eta)-A(\xi,\eta)}{2\beta}t\right)\right]\right. \\
&\quad + F(\xi,\eta)\sin\omega t + G(\xi,\eta)\cos\omega t \\
&\quad \left. - \frac{G(\xi,\eta)}{2}\left[\exp\left(\frac{B(\xi,\eta)-A(\xi,\eta)}{2\beta}t\right) - \exp\left(\frac{-B(\xi,\eta)-A(\xi,\eta)}{2\beta}t\right)\right]\right\},
\end{aligned}
\tag{36}
$$

where the notations $A(\xi,\eta)$ and $B(\xi,\eta)$ are given by Eqs. $(27)-(28)$ and the other abbreviations are as follows:

$$
\begin{aligned}
D(\xi,\eta) &= \left(a^2\beta^2\omega^2 + a_1^2\alpha^2\right)\left(\xi^2+\eta^2\right)^2 + 2a_2\alpha\beta\omega^2\left(\xi^2+\eta^2\right) + \beta^2\omega^4 + \alpha^2\omega^2, \\
E(\xi,\eta) &= \left[a^2(a_1-a_2)\beta^3\omega^2 - a_1^2 a_2\alpha^2\beta\right]\left(\xi^2+\eta^2\right)^2 \\
&\quad + \left[a(a_1-2a_2)\alpha\beta^2\omega^2 - 2a_1(a_1-a_2)\alpha\beta^2\omega^2 - a_1^2\alpha^3\right]\left(\xi^2+\eta^2\right) - 2a_2\beta^3\omega^4 - a_2\alpha^2\beta\omega^2, \\
F(\xi,\eta) &= a\beta^2\omega^3 + a_1\alpha^2\omega, \\
G(\xi,\eta) &= \left(a^2\beta^2\omega^2 + a_1^2\alpha^2\right)\left(\xi^2+\eta^2\right)^2 + a_2\alpha\beta\omega^2.
\end{aligned}
$$

Substituting the function $\widetilde{\overline{T}}(\xi,\eta,t)$ into Eq. (23) and using the same calculations in Eqs. (30)-(32) by indicating $\widetilde{\overline{T}}_2(\xi,\eta,t) = \widetilde{\overline{T}}(\xi,\eta,t)/\xi$ led to the closed-form solution as

$$
T(x,y,t) = \sqrt{\frac{2}{\pi}}\int_0^\infty \widetilde{\overline{T}}_2(\rho,t)\frac{x}{\sqrt{x^2+y^2}}J_1\left(\sqrt{x^2+y^2}\right)\rho^2 d\rho,
\tag{37}
$$

which is also depicted by calculating the improper integral numerically.
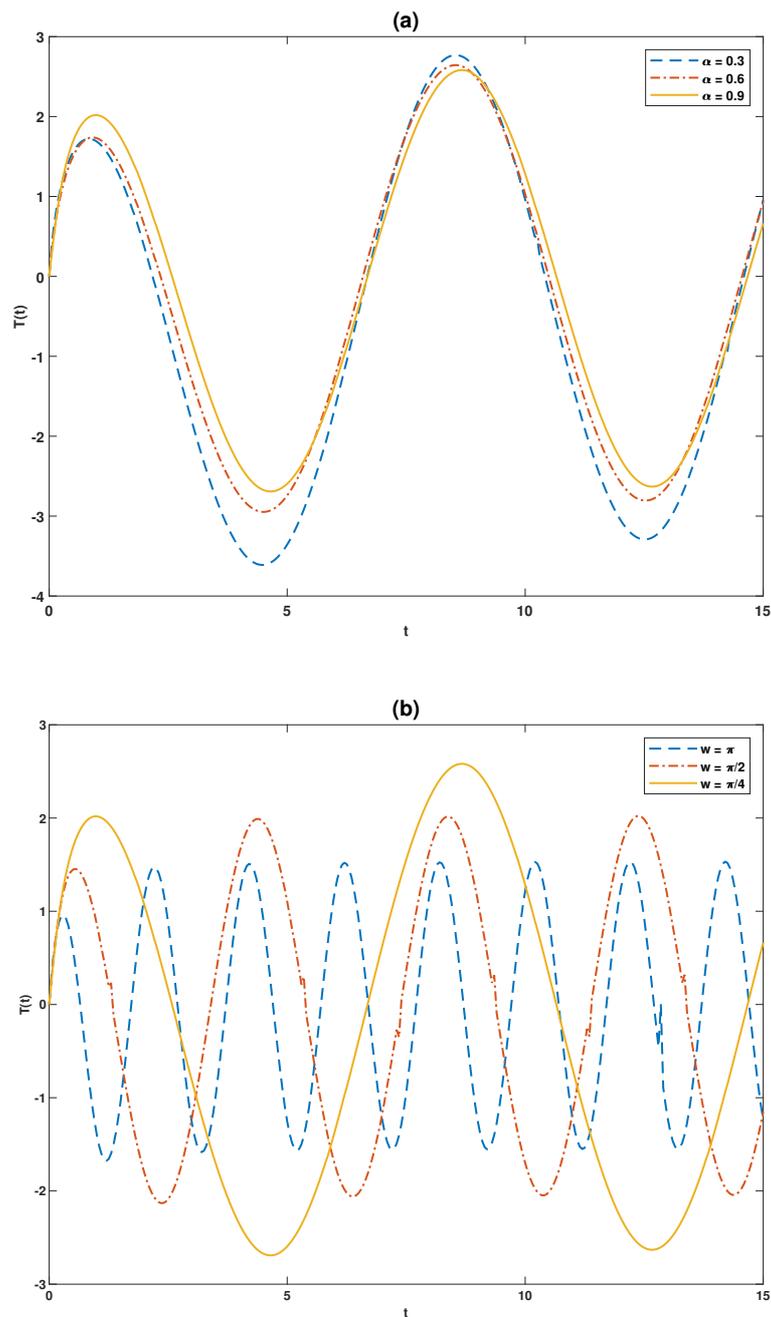
Figure 2. Time dependent temperature functions for non-moving harmonic boundary at $x = y = 0.5$ via variations of α and w, respectively.

Similar to Figure 1(a), Figure 2(a) shows also the dependence of temperature on the variation of α in the case of a non-moving harmonic temperature source at the boundary. Figure 2(b) shows the temperature response due to the change of angular frequency acting in the harmonic boundary temperature. As the angular frequency decreases, the wavelength of the temperature increases. It can be seen in both figures that temperature exhibits wave behavior similar to the boundary condition. This result clearly indicates Cattaneo's theory that wave phenomena may also occur in heat diffusion.

(a)

(b)
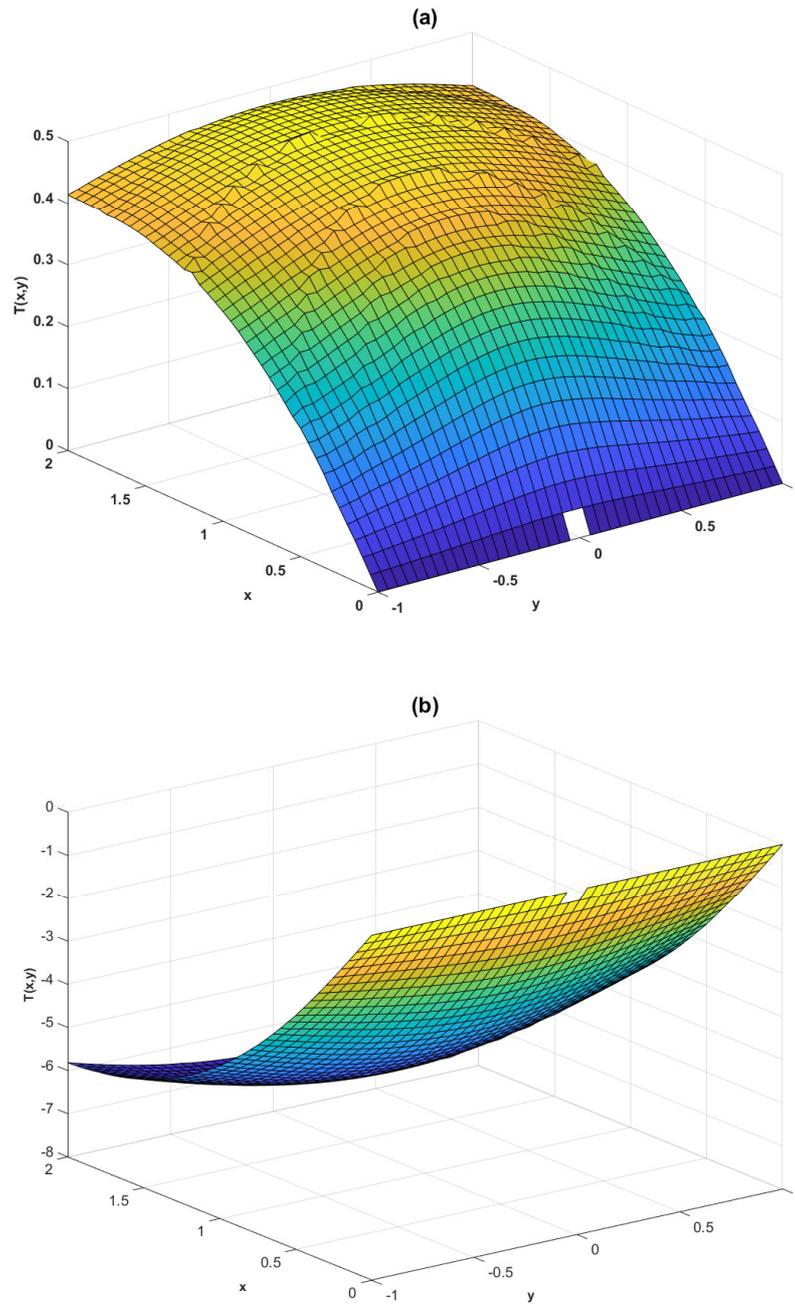
Figure 3. Temperature surfaces with non-moving harmonic boundary for $\alpha = 0.6$ at $t = 0.5$ and $t = 1.5$, respectively.

The time cross-section of the elastic heat diffusion at arbitrarily chosen times $t = 0.5$ and $t = 1.5$ is demonstrated in Figure 3. The wavelike temperature behaviour can be clearly seen for $\alpha = 0.6$ in both figures. This case can be similarly observed from the other α values depicted in Figure 2(a).

4 Conclusion

From the engineering point of view, it is important to know the mechanical and thermal behaviors of the materials under a heat force. These properties can be analyzed experimentally or with mathematical tools. In terms of mathematical analysis, it is crucial to exact modeling of the heat diffusion of the material. Although the Cattaneo-Hristov equation that models heat diffusion

with fading memory was constructed on the real line that physically corresponds to a wire, it is also significant to know the heat diffusion of a plate or a film with a fading memory effect. This situation can be generalized according to the geometry of materials, such as cylinders, spheres, cubes etc., which vary via the application area of the engineering problems. Therefore, this paper concerns the Cattaneo-Hristov diffusion equation in the half-plane. Two types of boundary conditions have been considered for the Dirichlet problem which are Dirac delta and non-moving harmonic temperatures, respectively. The closed-form solutions are arrived at by applying Fourier and Laplace integral transforms. The temperature functions have been illustrated under the variations of the model parameters using MATLAB software. These analyses performed for the two boundary temperatures can also be considered for different boundary conditions and different coordinate systems in future works.

Declarations

List of abbreviations

Not applicable.

Ethical approval

The authors state that this research complies with ethical standards. This research does not involve either human participants or animals.

Consent for publication

Not applicable.

Conflicts of interest

The author confirms that there is no competing interest in this study.

Data availability statement

Data availability is not applicable to this article as no new data were created or analysed in this study.

Funding

Not applicable.

Author's contributions

The author has made substantial contributions to the conception, design of the work, the acquisition, analysis, interpretation of data, and the creation of new software used in the work. Author has read and agreed to the published version of the manuscript.

Acknowledgements

References

[1] Yavuz, M. and Sene, N. Approximate solutions of the model describing fluid flow using generalized ρ-Laplace transform method and heat balance integral method. *Axioms*, 9(4), 123, (2020). [CrossRef]

[2] Hristov, J. Magnetic field diffusion in ferromagnetic materials: fractional calculus approaches. *An International Journal of Optimization and Control: Theories & Applications (IJOCTA)*, 11(3), 1–15, (2021). [CrossRef]

[3] Joshi, H. and Jha, B.K. Chaos of calcium diffusion in Parkinson's infectious disease model and treatment mechanism via Hilfer fractional derivative. *Mathematical Modelling and Numerical Simulation with Applications,* 1(2), 84-94, (2021). [CrossRef]

[4] Martinez-Farias F.J., Alvarado-Sanchez, A., Rangel-Cortes, E. and Hernandez-Hernandez, A. Bi-dimensional crime model based on anomalous diffusion with law enforcement effect. *Mathematical Modelling and Numerical Simulation with Applications,* 2(1), 26-40, (2022). [CrossRef]

[5] Sene, N. Second-grade fluid with Newtonian heating under Caputo fractional derivative: analytical investigations via Laplace transforms. *Mathematical Modelling and Numerical Simulation with Applications,* 2(1), 13-25, (2022). [CrossRef]

[6] Joshi, H., Yavuz M. and Stamova, I. Analysis of the disturbance effect in intracellular calcium dynamic on fibroblast cells with an exponential kernel law. *Bulletin of Biomathematics,* 1(1), 24-39, (2023). [CrossRef]

[7] Gurtin, M.E. and Pipkin, A.C. A general theory of heat conduction with finite wave speeds. *Archive for Rational Mechanics and Analysis,* 31, 113-126, (1968). [CrossRef]

[8] Nigmatullin, R.R. On the theory of relaxation for systems with "remnant" memory. *Physica Status Solidi (b),* 124(1), 389-393, (1984).
[CrossRef]

[9] Green, A.E. and Naghdi, P.M. Thermoelasticity without energy dissipation. *Journal of Elasticity,* 31, 189-208, (1993). [CrossRef]

[10] Gorenflo, R., Mainardi, F., Moretti D. and Paradisi, P. Time fractional diffusion: a discrete random walk approach. *Nonlinear Dynamics,* 29, 129-143, (2002). [CrossRef]

[11] Cattaneo, C. Sulla Conduzione del Calore. *Atti del Seminario Matematico e Fisico dell'Universita di Modena e Reggio Emilia,* 3, 83-101, (1948).

[12] Cattaneo, C. Sur une forme de l'equation de la chaleur eliminant la paradoxe d'une propagation instantantee. *Comptes Rendus de l'Académie des Sciences,* 247, 431-433, (1958).

[13] Povstenko, Y.Z. Fractional heat conduction equation and associated thermal stress. *Journal Thermal Stresses,* 28(1), 83-102, (2004). [CrossRef]

[14] Povstenko, Y.Z. Thermoelasticity that uses fractional heat conduction equation. *Journal of Mathematical Sciences,* 162, 296-305, (2009). [CrossRef]

[15] Povstenko, Y.Z. Time-fractional heat conduction in an infinite medium with a spherical hole under Robin boundary condition. *Fractional Calculus and Applied Analysis,* 16(2), 354-369, (2013). [CrossRef]

[16] Povstenko, Y.Z. and Klekot, J. The fundamental solutions to the central symmetric time-fractional heat conduction equation with heat absorption. *Journal of Applied Mathematics and Computational Mechanics,* 16(2), 101-112, (2017). [CrossRef]

[17] Povstenko Y.Z. and Kyrylych, T. Time-fractional heat conduction in an infinite plane containing an external crack under heat flux loading. *Computers and Mathematics with Applications,* 78(5), 1386-1395, (2019). [CrossRef]

[18] Povstenko, Y.Z. Theory of thermoelasticity based on the space-time-fractional heat conduction

equation. *Physica Scripta*, 2009(T136), 014017, (2009). [CrossRef]

[19] Sherief, H.H., El-Sayed, A.M.A. and Abd El-Latief, A.M. Fractional order theory of thermoelasticity. *International Journal of Solids and Structures,* 47(2), 269-275, (2010). [CrossRef]

[20] El-Karamany, A.S. and Ezzat, M.A. On fractional thermoelasticity. *Mathematics and Mechanics of Solids,* 16(3), 334-346, (2011). [CrossRef]

[21] Povstenko, Y.Z. Time-fractional radial heat conduction in a cylinder and associated thermal stresses. *Archive of Applied Mechanics,* 82, 345-362, (2012). [CrossRef]

[22] Povstenko, Y. Fractional heat conduction and related theories of thermoelasticity. In *Fractional Thermoelasticity* (pp. 13-33). Cham: Springer, (2015). [CrossRef]

[23] Özdemir, N., Povstenko, Y.Z., Avcı, D. and İskender, B.B. Optimal boundary control of thermal stresses in a plate based on time-fractional heat conduction equation. *Journal of Thermal Stresses,* 37(8), 969-980, (2014). [CrossRef]

[24] Povstenko, Y., Avcı, D., Eroğlu, B.B.İ. and Özdemir, N. Control of thermal stresses in axis-symmetric problems of fractional thermoelasticity for an infinite cylindrical domain. *Thermal Science,* 21(1), 19-28, (2017). [CrossRef]

[25] Hristov, J. Transient heat diffusion with a non-singular fading memory: from the Cattaneo constitutive equation with Jeffrey's kernel to the Caputo-Fabrizio time-fractional derivative. *Thermal Science,* 20(2), 757-762, (2016). [CrossRef]

[26] Hristov, J. Derivatives with non-singular kernels from the Caputo-Fabrizio definition and beyond: appraising analysis with emphasis on diffusion models. In *Frontiers in Fractional Calculus* (pp. 270-342). Bentham Science Publishers, (2017). [CrossRef]

[27] Hristov, J. Derivation of the fractional Dodson equation and beyond: transient diffusion with a non-singular memory and exponentially fading-out diffusivity. *Progress in Fractional Differentiation and Applications,* 3(4), 255-270, (2017). [CrossRef]

[28] Hristov, J. Linear viscoelastic responses and constitutive equations in terms of fractional operators with non-singular kernels-Pragmatic approach, memory kernel correspondence requirement and analyses. *The European Physical Journal Plus,* 134(6), 283, (2019). [CrossRef]

[29] Sene, N. Analytical solutions of Hristov diffusion equations with non-singular fractional derivatives. *Chaos: An Interdisciplinary Journal of Nonlinear Science,* 29(2), 023112, (2019). [CrossRef]

[30] Avci, D. and Eroglu, B.B.I. Semi-Analytical Solution of Hristov Diffusion Equation with Source. In *A Closer Look at the Diffusion Equation* (pp. 117-132). Nova Science Publishers (2020).

[31] Alkahtani, B.S.T. and Atangana, A. A note on Cattaneo-Hristov model with non-singular fading memory. *Thermal Science,* 21(1), 1-7, (2017). [CrossRef]

[32] Koca, I. and Atangana, A. Solutions of Cattaneo-Hristov model of elastic heat diffusion with Caputo-Fabrizio and Atangana-Baleanu fractional derivatives. *Thermal Science,* 21(6), 2299-2305, (2017). [CrossRef]

[33] Sene, N. Solutions of fractional diffusion equations and Cattaneo-Hristov diffusion model. *International Journal of Analysis and Applications,* 17(2), 191-207, (2019). [CrossRef]

[34] Singh, Y., Kumar, D., Modi, K. and Gill, V. A new approach to solve Cattaneo-Hristov diffusion model and fractional diffusion equations with Hilfer-Prabhakar derivative. *AIMS Mathematics,* 5(2), 843-855, (2020). [CrossRef]

[35] Eroglu, B.B.I. and Avci, D. Separable solutions of Cattaneo-Hristov heat diffusion equation in

a line segment: Cauchy and source problems. *Alexandria Engineering Journal,* 60(2), 2347–2353, (2021). [CrossRef]

[36] Avci, D. and Eroglu, B.B.I. Optimal control of the Cattaneo–Hristov heat diffusion model. *Acta Mechanica*, 232, 3529–3538, (2021). [CrossRef]

[37] Avci, D. and Eroglu, B.B.I. Oscillatory heat transfer due to the Cattaneo-Hristov Model on the real line. In *Fractional Calculus: New Applications in Understanding Nonlinear Phenomena* (pp. 108-123). Singapore: Bentham Science Publishers, (2022). [CrossRef]

[38] Avci, D. Temperature profiles and thermal stresses due to heat conduction under fading memory effect. *The European Physical Journal Plus,* 136, 356, (2021). [CrossRef]

[39] Ångström, A.J. Neue Methode, das Wärmeleitungsvermöogen der Köorper zu bestimmen. *Annalen der Physik und Chemie,* 190(12), 513–530, (1861). [CrossRef]

[40] Nowacki, W. State of stress in an elastic space due to a source of heat varying harmonically as function of time. *Bulletin of the Polish Academy of Sciences Technical Sciences,* 5(3), 145–154, (1957).

[41] Nowacki, W. *Thermoelasticity*. Pergamon Press: Oxford, UK, (1986).

[42] Baehr, H.D. and Stephan, K. *Heat and Mass Transfer*. Springer: Berlin/Heidelberg, Germany, (2006).

[43] Povstenko, Y.Z. and Ostoja-Starzewski, M. Doppler effect described by the solutions of the Cattaneo telegraph equation. *Acta Mechanica,* 232, 725-740, (2021). [CrossRef]

[44] Datsko, B., Podlubny, I. and Povstenko, Y. Time-Fractional diffusion-wave equation with mass absorption in a sphere under harmonic impact. *Mathematics,* 7(5), 433, (2019). [CrossRef]

[45] Povstenko, Y.Z. Fractional thermoelasticity problem for an infinite solid with a cylindrical hole under harmonic heat flux boundary condition. *Acta Mechanica,,* 230, 2137-2144, (2019). [CrossRef]

[46] Povstenko, Y.Z. and Kyrylych, T. Time-fractional diffusion with mass absorption in a half-line domain due to boundary value of concentration varying harmonically in time. *Entropy,* 20(5), 346, (2018). [CrossRef]

[47] Povstenko, Y.Z. and Kyrylych, T. Time-fractional diffusion with mass absorption under harmonic impact. *Fractional Calculus and Applied Analysis,* 21(1), 118-133, (2018). [CrossRef]

[48] Podlubny, I., Magin, R.L. and Trymorush, I. Niels Henrik Abel and the birth of fractional calculus. *Fractional Calculus and Applied Analysis,* 20(5), 1068-1075, (2017). [CrossRef]

[49] Caputo, M. and Fabrizio, M. Applications of new time and spatial fractional derivatives with exponential kernels. *Progress in Fractional Differentiation and Applications,* 2(1), 1-11, (2016). [CrossRef]

[50] Povstenko, Y.Z. *Linear fractional diffusion-wave equation for scientists and engineers*. Switzerland: Springer International Publishing, (2015). [CrossRef]

[51] Prudnikov, A.P., Brychkov, Y.A., Maricheva, O.I., Romer, R.H. *Integrals and series*, (1988).

[52] Povstenko, Y.Z. Fundamental solutions to time-fractional advection diffusion equation in a case of two space variables. *Mathematical Problems in Engineering,* 2014(3), 1-7, (2014). [CrossRef]

Mathematical Modelling and Numerical Simulation with Applications (MMNSA)
(https://dergipark.org.tr/en/pub/mmnsa)