Sigma J Eng Nat Sci, Vol. 44, No. 3, pp. 1783–1798, June, 2026

Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2026.2063

Research Article

Enhancing groundwater contaminant simulations: Predicting solute concentrations with GP and ANN models

Pappu KUMAR[1], Saurabh KUMAR[2,*], Uruya WEESAKUL[3], Apin WORAWIWAT[3], Anshuman SINGH[4], Dayanand SHARMA[5]

[1]*Department of Civil Engineering, GEC Palamu (under JUT Ranchi), Jharkhand, 822118, India*
[2]*Research Unit in Climate Change and Sustainability, Department of Civil Engineering, Faculty of Engineering, Thammasat School of Engineering, Thammasat University, Pathumthani 12120, Thailand*
[3]*Department of Civil Engineering, Faculty of Engineering, Thammasat School of Engineering, Thammasat University, Pathum Thani, 12120, Thailand*
[4]*Department of Civil Engineering, National Institute of Technology Patna, Bihar, 800005, India*
[5]*Department of Civil Engineering, Sharda University, Greater Noida, Uttar Pradesh, 201310, India*

ARTICLE INFO

Article history
Received: 30 August 2024
Revised: 18 November 2024
Accepted: 12 April 2025

Keywords:
Artificial Neural Networks; Contaminant Transport; Genetic Programming; Groundwater; Machine Learning

ABSTRACT

Groundwater contamination can affect people's lives seriously as well as have detrimental effects on ecosystems. Therefore, it is essential to use accurate predictive tools in order to predict and control groundwater contamination. This research aims at predicting solute concentrations in groundwater applying two advanced methods – genetic programming (GP) and artificial neural networks (ANN). These two methods are critical in dealing with the problems associated with risk assessment, contaminants remediation and water resources management. Advection-dispersion equation is used as an important tool in solute transport analysis and applied in combination with ANN and GP in order to optimize numerical solutions. In this paper 1,361 points (1,040 points for training and 321 points for testing) were used as data input. Analytical solute concentrations were considered in the development of ANN and GP models. The ANN models include five different learning algorithms which are Conjugate Gradient with Powell-Beale Restarts (CGB), Bayesian Regularization (BR), Conjugate Gradient with Feedback (CGF), Cartesian Genetic Programming (CGP) and Levenberg-Marquardt (LM). The analysis of the obtained results has proven that ANN-LM was the most successful model due to the highest correlation coefficient (~0.99) and the smallest root mean square error (RMSE) during both stages (training and testing). GP model correlated ~0.75 and provided 12% RMSE in the training process and 17% RMSE in the testing phase. The lowest values of AOC (area over the curve) were attained by ANN-BR (0.077 training/0.102 testing). ANN methods performed better when the amount of available hydrological data was limited. In conclusion, both ANN and GP prove themselves to be rather powerful tools for modelling groundwater contamination. It is noteworthy that the best results were obtained in the case of ANN-LM. Thus, it is reasonable to claim that ANN-based methods were more effective than others in this research. Novelty of this paper consists in combining artificial intelligence-based methods with advection-dispersion equation.

Cite this article as: Kumar P, Kumar S, Weesakul U, Worawiwat A, Singh A, Sharma D. Enhancing groundwater contaminant simulations: Predicting solute concentrations with GP and ANN models. Sigma J Eng Nat Sci 2026;44(3):1783–1798.

***Corresponding author.**
*E-mail address: arsaurabh12345@gmail.com

This paper was recommended for publication in revised form by Editor-in-Chief Ahmet Selim Dalkilic

INTRODUCTION

Groundwater aquifers store water for human and industrial use, and a large population depends upon groundwater as a source of water. According to [1], 40% and 25% of global groundwater resources are produced in the United States. Recently, groundwater has become the main source of drinking water. Several studies have shown that organic pollutants have become a major source of pollution. Contamination of the groundwater is a significant environmental problem which has serious consequences for both humans and the environment. It is important to be able to predict and understand how contaminants move through the groundwater system [2]. The transport of groundwater pollutants depends on various soil parameters, types of soil, grain sizes, and void ratios. The management of groundwater pollution may include and require analysis of the infiltration of water through porous soil. The increasing deposition of organic pollutants in groundwater has become a crucial issue. Groundwater aquifer studies have involved methods for groundwater pollution assessment, water level fluctuation analysis, and groundwater depth prediction using artificial intelligence. Studying aquifers both numerically and analytically can be complex. A significant threat to groundwater contamination arises from leachate produced from waste materials, which often contain toxic chemicals, particularly from industrial waste [3]. The study of groundwater contaminant transport is essential for the conservation, management and protection of aquifers. Contaminant transport is studied on the basis of field investigations and mathematical models.

Mathematical models are quintessentially used for the prediction of solution concentration [4]. Groundwater solute transport in saturated aquifers is understood to be transported through the processes of advection, dispersion, and diffusion. The interactions between the solid phase (soil) and various chemical reactions are critical components in modelling solute transport processes and are commonly represented by advection-dispersion-reaction equations (ADREs). There have been notable improvements in the formulation of analytical solutions for ADRE; but most times, these analytical solutions are limited to certain initial and boundary conditions. For instance, the majority of research work carried out focuses on the reaction kinetics of homogenous systems, which sufficiently account for practical situations like heterogeneous soil formations or varying reaction rates.

On the contrary, numerical techniques have been widely used in order to model ADRE with various initial and boundary conditions using finite difference, finite element, and finite volume techniques. These techniques are highly promising when it comes to modeling more complex situations, including heterogeneous reactions, multiphase transport problems, and even time-varying boundary conditions. However, there is one thing that needs improvement regarding these techniques and that is incorporating the effects of advanced chemical reactions along with transport problems.

The purpose of this study is to recognize the need for modeling frameworks that can capture these complex interactions within the ADRE model. It is essential to address this gap to enhance the accuracy of the models used to predict the transportation of the solutes. In the ADRE for porous media, the velocity is calculated using Darcy's law. The dispersion coefficient is dependent on the size of the domain [5]. The different geological formulations also impact molecular diffusion [6]. Traditional models, often based on physical principles such as the advection–dispersion equation (ADE), have proven useful but are limited by the need for detailed and sometimes unavailable data. Machine learning methods, for example, artificial neural networks (ANN) and genetic programming (GP), offer promising alternatives, offering improved techniques for simulating the transport of contaminants [7]. Groundwater transport of solutes has been widely studied using the conventional deterministic methods such as the advection-dispersion equation (ADE) [8]. Even though such deterministic models have proven to be useful in predicting the transport of contaminants through groundwater, the effectiveness of these models is limited by their inability to deal effectively with non-linearities, high-dimensionality of data and heterogeneity [9]. Also, deterministic models of contaminant transport in groundwater require extensive knowledge on hydrogeological properties that may not be available. Moreover, uncertainties in parameter estimation make it difficult to obtain accurate prediction [10]. ANN has been shown to be able to predict different attributes associated with groundwater contamination, for instance, solute concentration levels [11]. For example, the successful applications of ANNs in predicting solute concentration levels in groundwater include modelling nitrate levels in groundwater [12] and oil pollution spread [13]. Unlike ANN, GP develops explicit mathematical expressions explaining the interaction between dependent and independent variables in solving environmental problems. In particular, GP has been widely employed for contaminant transport in groundwater in recent years [14]. The application of GP to groundwater contaminant transport has been successfully applied in some of the recent studies carried out to model contaminant transport in groundwater. For instance, GP has been used in the modelling of pollutant behaviour in aquifers [15], where it has been found very useful in predicting contaminant concentration over time. Given its unique characteristics, the GP has become an increasingly popular approach for modelling contaminant behaviour in groundwater. However, GP still faces challenges in dealing with non-linearities and sparsity in the data. This motivates this study, whose goal is to apply a hybrid GP-ANN for predicting solute concentration levels in groundwater. This research introduces an innovative approach by integrating artificial intelligence methods

with the advection-dispersion equation, providing a more accurate and adaptable solution for modelling groundwater contamination.

Mathematical Model

The groundwater solute transport equation i.e., the ADRE with first-order decay for a single solute in one dimension, can be expressed as Eq. 1 [17], [18]:

$$\frac{\partial C}{\partial t} + v\frac{\partial C}{\partial x} = D\frac{\partial^2 C}{\partial x^2} - \frac{\rho_b}{\theta}\frac{\partial S}{\partial t} - \lambda C \quad (1)$$

where C is the solute concentration [ML^{-3}]; t is time [T]; v is Darcy's velocity [LT^{-1}]; D is the dispersion coefficient [L^2T^{-1}]; x indicates the direction [L]; S is the sorbed concentration of solute [MM^{-1}]; ρ_b is the bulk density of porous media [ML^{-3}]; θ is the porosity of porous media; and λ is the first-order decay constant [T^{-1}].

The equilibrium linear sorption relationship between S and C can be expressed as:

$$S = K_d C \quad (2)$$

where k_d is the linear sorption coefficient [L^3M]. Dispersion, sorption, and degradation can be "turned off" by setting the dispersion coefficient (D), the sorption distribution coefficient (k_d), and the first-order decay rate constant (λ) to zero, respectively, in Equation number (1) [19].

Additionally, for rate-limited sorption, Eq. (3) can be expressed as follows:

$$f\frac{\partial C}{\partial t} = \alpha(K_d C - S) \quad (3)$$

where α [T^{-1}] is a constant. A higher value of α leads to equilibrium sorption.

The study in this paper is based on the solution of the above ADRE with the following initial and boundary conditions.

$$\text{Initial Condition: } C = S = 0, t = 0, 0 \le x \le L \quad (4)$$

$$\text{Boundary Condition: } C = C_0 H(t_s - t), x = 0,\ t > 0 \quad (5)$$

$$\frac{\partial C}{\partial x} = 0, x = 0, t > 0$$

The above conditions correspond to the step input and zero gradient conditions at the exit for the domain of length L.

Analytical Solution

The analytical solution to equation (1), considering the initial and boundary conditions specified in Eqs. (4) and (5), is provided below [7]:

$$\bar{C}(x,s) = \frac{1}{s}C_o[1 - \exp(-t_s s)]\frac{r_2\exp(r_2L + r_1x) - r_1\exp(r_1L + r_2x)}{r_2\exp(r_2L) - r_1\exp(r_1L)} \quad (6)$$

were,

$$r_{1,2} = \frac{1}{2D_x}\left[v \pm (v^2 + 4D_x\Theta^{1/2}\right]$$
$$\Theta = s + \lambda + \frac{\rho_b\alpha k_d s}{\theta(s+\alpha)} \quad (2)$$

The above solution is used for generating data for applying the prediction modes. The different input parameters use equation (3) to calculate the solute concentration. The generated output solute concentration is used as output data for machine learning models. The mathematical governing ADE equation (2) given parameters such as bulk density ρ_b is taken as a constant of 1.5, C (mg/m) is the solute concentration, the velocity (v) (m/min) lies between the minimum value of 0.35 and the maximum value of 1.5, the dissolving mass (R) lies between the minimum value of 1.06 and the maximum value of 7.22, the dispersion coefficient (Dx) (m^2/min) is the minimum value of 0 and the maximum value of 0.32, the hydraulics conductivity (Kd) (mg/kg)/(mg/m) is the minimum value of 0.0 and the maximum value of 0.10, and the time (t) (min) is the minimum value of 0.1 and the maximum value of 14.26.

Groundwater contamination has traditionally been predicted the aforementioned equations, either through numerical or analytical solutions implemented in various existing software tools. Numerous soft computing techniques have been employed to forecast hydrological events, with artificial intelligence (AI), particularly artificial neural networks (ANNs), also being utilized to predict contaminants in groundwater [20].

The size parameter (x), the relative complex refractive index of the nanosuspension (m), can be estimated using the following relation:

MATERIALS AND METHODS

The groundwater solute transport model relies on an incomplete differential condition, as illustrated in Figure 1, which presents a streamlined process diagram for deriving a solution. The iterative process continues until the specified termination criterion is met, with the entire methodology detailed in the accompanying flowchart. Key parameters, including solute concentration, soil properties, hydraulic conductivity, porosity, pH, and electrical conductivity, significantly influence the behaviour of solute particles. The results are calculated and analysed using MATLAB software, ensuring accurate computational outcomes.

Artificial intelligence approaches were applied in predictive modeling, especially concerning the prediction of groundwater contamination. In this respect, ANN was chosen since it can effectively display the non-linear relations between the model parameters, inputs, and outputs. At the same time, GP could act as an auxiliary tool aimed at increasing the level of accuracy and interpretation of the models being developed. This work presents a comprehensive review of the application of ANN and GP in predicting

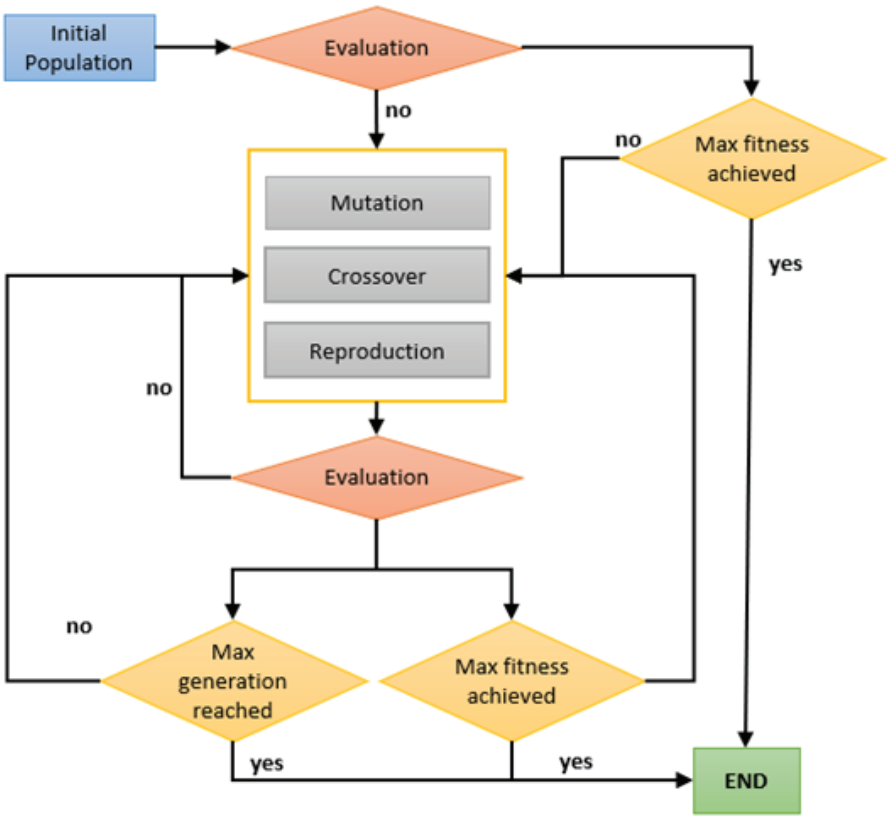


Figure 1. Flowchart of the proposed work.

the level of contamination of groundwater. Hence, artificial intelligence and ANN prove to be able to provide accurate estimations of the levels of contamination and increase the effectiveness of groundwater pollution modeling. Many studies concerning ANN and genetics prove that there exist computer programs which can be used in predicting the pollution of groundwater [21]. Genetic programming (GP) concept [22] can be viewed as an approach to creating and optimizing computer programs based on evolutionary algorithms. Genetic programming aims to develop computer programs based on Darwinian theory of 'natural selection', which means that it is a part of evolutionary computations. Genetic programming (GP) mechanism is associated with creating a random initial population of computer programs which further evolve into another population of programs. This transformation involves such actions as recombination, mutation, and reproduction. With each following generation, genetic programming (GP) transforms initial populations of computer programs into new populations. Throughout this process, GP optimizes the created computer programs through its genetic operators in order to produce the required computer programs. Many individuals (computer programs) must be created throughout this optimization process. They are designed according to a trial-and-error approach involving combinations of numerous functions. The type of the function could vary. It may consist of such actions as applying Boolean logic operation (AND, OR, NOT), conducting arithmetic operation (addition, division, multiplication, and subtraction) and implementing geometry operation (sine, cosine, tangent) along with others. These functions must be applied within the framework of genetic operators which involve reproduction, recombination (or crossover), and mutation. Reproduction is related to duplication of an individual without any change in its characteristics, recombination – to exchange of genes between two individuals and mutation – to modification of genes randomly selected from an individual. Thus, genetic programming (GP) helps to optimize problems through its genetic operations. Figure 2 and 3 illustrate a graphical depiction of the tree-like genetic operations of GP in the course of crossover.

An initial population is generated randomly and subsequently analysed for solutions within a universe of possible outcomes. Initially, a model is developed, after which the genetic programming (GP) algorithm is used to evaluate individuals and compare their fitness to that of the developed model. All the genetic algorithms like mutation, crossover, and reproduction happen in parallel with the fitness function. Genetic programming uses all these genetic functions to make changes to the individuals based on their

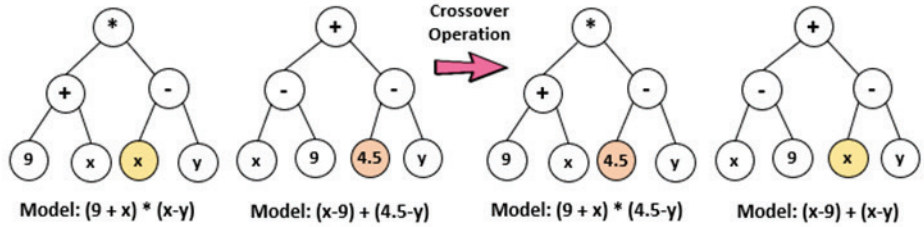


Figure 2. A graphical representation of a tree-like structure showing the genetic operations involved in the crossover process is presented.

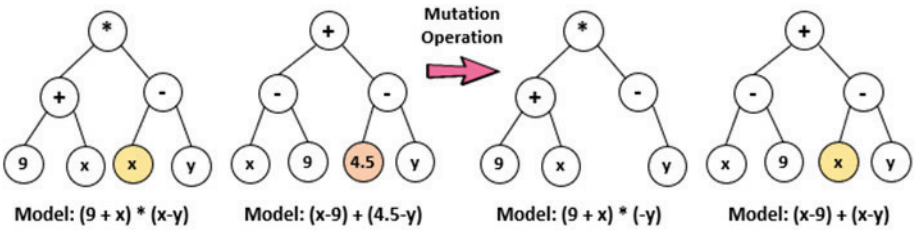


Figure 3. The mutation process is shown using a graphical tree-like structure that illustrates the related genetic operations.

fitness level. The fitness function determines the proximity of the individuals towards the goal. This process continues until the predetermined termination criteria are met. The entire procedure is illustrated in the flowchart provided below (Fig. 4).

Multigene Symbolic Regression

In a few intellectual [23] and modern applications [24], representative relapses have utilized and tracked extraordinary accomplishments. In this specific situation, a remarkable kind of emblematic relapse called multigene representative relapse (MGSR) is created. It is ordinarily brought out through GP to propel a populace of trees, which encodes a numerical articulation. Such numerical articulation predicts a (N × 1) yield vector (reaction variable, y) utilizing a related (N × M) network of data sources (autonomous factors, x1, x2, etc.), where N and M are the quantity of perceptions of the resulting variable (y) and the quantity of information factors (x1, x2, etc.).

Unlike typical genetic programming methods, the MGSR method involves the formation of symbolic models, each of which represents a weighted linear combination of the evolution of GP trees and/or a blend of GP trees. In such a scenario, each GP tree can be considered a gene. The figure below shows an example of a multigene model that uses the independent variables (x1, x2, x3, x4) to predict the output variable y.

In the construction of MGSR model, there exist several nonlinear terms like sin and tan; however, the model remains linear in terms of coefficient estimates of d0, d1 and d2. While designing the model, the user is free to set the maximum number of genes and maximum tree depth of a gene denoted as Gmx and Dmx, respectively. These two settings provide the flexibility to design complex models easily. Restricting Dmx to about 5-6 nodes gives rise to models that remain compact as linear functions of nonlinear transformations of independent variables. Ordinary

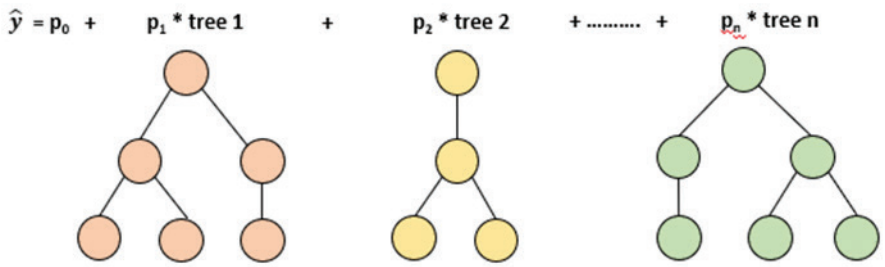


Figure 4. Example of multigene archetypal.

least squares OLS is used to estimate the linear coefficients of all models from training data sets. In the literature, it has been found out that the efficiency of MGSR in symbolic regression is higher than genetic programming [25]. Moreover, the multigene technique can be easily incorporated into the nonlinear partial least squares algorithm [26].

The creation of the initial population in the case of the genetic programming (GP) method involves the creation of individuals that have genetically programmed trees with a depth of between 1 and Gmax. In the GP method, the acquisition and elimination of traits occur via an operator termed two-point high-level crossover, allowing the sharing of traits among the individuals. The two-point high-level crossover works together with the recombination methods in the genetic programming method. Figure 5 illustrates two-point high-level crossover.

Here, (I) has the qualities (Ga, Gb, Gc, Gd) and (II) has the qualities (Gp, Gq, Gr) with the maximum qualities being five (Gmx=5) respectively. The hybrid focuses are generated randomly for each individual in the group. In the way illustrated in Figure 4e, the qualities (Gb, Gc, Gd) in Individual-I and the quality (Gp) in Individual-II can be swapped leading to two new individuals. In the case of two-point hybridization, this will provide the opportunity to develop qualities for both individuals and also allow separation of qualities. Depending on the likelihood that swapping of qualities results in a new individual with more qualities than the maximum qualities allowed, then a number of qualities are randomly selected and discarded until the total reaches the maximum qualities (Gmx).

Artificial Neural Network (ANN)

Artificial intelligence (ANN) uses the data analysis method, and the algorithm is based on the nervous system of humans. ANNs are used for various fields of research and science technology and perform nonlinear mapping with inputs and outputs. Nasseri and Mohanty [27] formulated a feed-forward network with an ANN and a genetic algorithm for river training. In every field of engineering and related fields, artificial neural networks (ANNs) are computational models or mathematical models that are stimulated by the nervous system (like the brain) and are used for regression analysis and pattern recognition [28]. The neural network is used for clustering primitive artificial networks to solve real problems. The general construction of the ANN is shown in Fig. 6.

Artificial Neural Network (ANN) mimics the way that human brains process information using their nervous systems. In an ANN, nodes are known as neurons, while the processing units are referred to as units, and their interconnection forms a network. A neuron comprises many processing units connected with each other through weighted

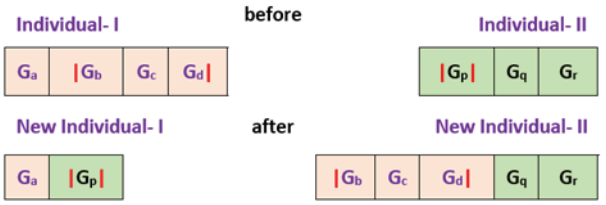


Figure 5. Genes (Gb, Gc, Gd) from Individual -I and genes (Gp) from Individual -II are then swapped, resulting in the two new individuals.

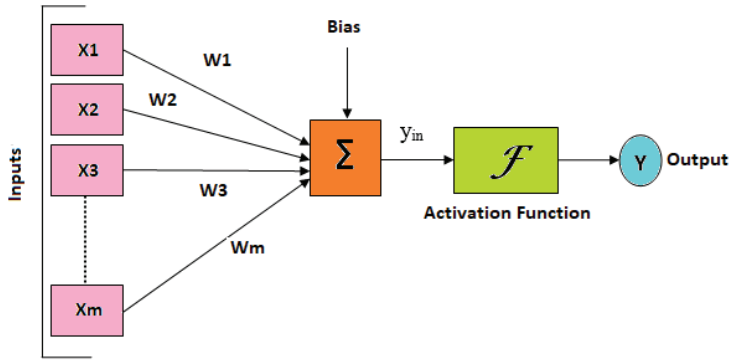


Figure 6. General construction of the ANN.

links. Feed-forward architecture is widely used in the field of ANN and is quite effective due to the error minimization resulting from the backpropagation technique. A feedforward network has an input layer, one or more hidden layers, and an output layer.

a. A neural network functions as a mathematical model with a defined input (X) and output (Y) and is often associated with a specific learning algorithm. ANNs are applied in forecasting and modelling complex relationships between inputs and outputs. They are considered nonlinear data modelling tools capable of establishing complex connections between inputs and outputs. Transfer potential: This term totals the sources of information and its loads.
b. Activation function: This function applies a nonstraight exchange capacity or "initiation" to the exchange potential.
c. Threshold function: Depending on the actuation work, the limit capacity to all things considered "initiate" the neuron or not.

Sigmoid activation functions help to map the data in a nonlinear fashion, allowing the network to learn more sophisticated relationships. Its goal is to model the probability that a given input belongs to a certain class. The logistic model outputs a value between 0 and 1, representing the probability of an instance belonging to class 1, making the sigmoid a natural choice [7]. Typically, the activation function used is a logistic sigmoid as follows:

$$f(\theta) = \frac{1}{1 + e^{-\theta}} \tag{8}$$

Transfer potential function $(\theta) = \sum w_i * x_i$, where w_i is the weight function.

Bayesian Regularized Artificial Neural Network

The term Bayesian derives from the 18th-century mathematician and provides the mathematical analysis of static data using Bayesian interface. Bayesian regularization improves ANN performance by reducing overfitting through adaptive weight optimization. Bayesian regularization is changed from nonlinear relapse to unbending relapse. The Bayesian regularization ANN provides power, and the Bayesian Regularization (BR) model boundaries of foreign substance transport are also effectively used to predict groundwater pollution in this review. This cycle is equivalent to a punished log probability.

$$\mathrm{F} = \beta \mathrm{E_D}(\mathrm{D/W, M}) + \alpha \mathrm{E_W}(\mathrm{W/M}) \tag{9}$$

In this context, (EW (w/M)) represents the sum of the squares of the architecture weights, where (M) denotes the ANN architecture or model in statistical terms. The parameters (α) and (β) are objective function parameters, also known as regularization parameters or hyperparameters, which must be adaptively estimated and take positive values [3].

Levenberg–Marquardt (LM)

In the backpropagation algorithm, the minimum value is directly proportional to the local gradient. The descent gradient becomes slow and takes many small steps for learning. To overcome this, second-order Levenberg–Marquardt algorithm multilayer FNN training is used. The Levenberg–Marquardt method is designed such that the value is approximate and the optimal solution is computed [29].

$$x_{i+1} = x_i - [J^T J + \mu I]^{-1} J^T e \tag{10}$$

where x = loads of the brain organization, J = Jacobian lattice of the presentation standards to be limited, μ = a scalar that controls the learning system, and e = remaining blunder vector. The Blunder vector might appear in an equation such as: e = r + b, where r is the residual vector and b is the blunder vector. When the scalar (mu) is zero, the method corresponds to Newton's strategy using an approximate Hessian matrix. As (mu) increases, the method resembles gradient descent with a small step size. Newton's method is faster and more precise near an error minimum, so the goal is to transition to Newton's method as quickly as possible. The Levenberg–Marquardt algorithm is one of the fastest methods for training feedforward neural networks. Owing to its high memory requirements, it is typically used for small networks [30]. Therefore, many researchers have successfully utilized it [31].

Data Preparation and Statistical Visualization

The analytical solution of the groundwater solute transport model was developed by [32] via MATLAB software. In this work, we used a mathematical solution with various input conditions. The input parameters, such as the diffusion coefficient and velocity, change, and we obtain the solute concentration as an output. A total of 1361 data points are generated randomly, including the dissolve mass (R), dispersion coefficient (D_x), velocity (v), hydraulic conductivity (K_d), time (t), and concentration (c). Then, the mathematical model is implemented, and travel-timewise solute concentrations are obtained. The results obtained from the mathematical model are divided into training (TRn) and testing (TSt) datasets. The statistical details of the datasets are provided in Table 1.

The correlation matrix and Pearson correlation between the input and output variables are illustrated in Figure 7, and pair plots of the input and output parameters are shown in Figure 8.

Data Analysis

Two machine learning models, genetic programming and artificial neural networks (ANNs), are used to predict solute concentrations. For the ANN, five different training algorithms are employed. ANN models make use of different learning algorithms such as Conjugate Gradient with Powell Beale Restarts (CGB), Bayesian Regularization (BR), Conjugate Gradient with Feedback (CGF), Cartesian

Table 1. Statistical details of the input and response variables

Statistical Parameters	R -	Dx (m^2/min)	v (m/min)	k_0 (mg/kg)/(mg/m)	t (min)	c (mg/m)
Minimum	1.06	0.00	0.35	0.00	0.01	0.00
1st Quartile	4.83	0.04	0.69	0.03	2.26	0.23
Median	5.54	0.05	0.85	0.05	4.51	0.75
Mean	5.33	0.06	0.91	0.05	5.37	0.61
3rd Quartile	6.14	0.08	0.99	0.07	8.26	0.97
Maximum	7.22	0.32	1.95	0.10	14.26	1.25
Trimmed	5.44	0.05	0.87	0.05	5.26	0.76
MAD	1.03	0.03	0.21	0.03	4.45	0.35
Standard error	0.03	0.00	0.01	0.00	0.11	0.01
Standard deviation	1.19	0.05	0.31	0.03	3.85	0.37
Variance	1.42	0.00	0.09	0.00	14.80	0.14
Skewness	-1.35	2.67	1.11	2.78	0.40	-0.43
Kurtosis	2.53	11.07	1.08	20.86	-0.86	-1.45
Geometric mean	5.12	0.04	0.86	0.04	2.83	0.38
Harmonic mean	4.77	0.02	0.82	0.02	0.10	0.03

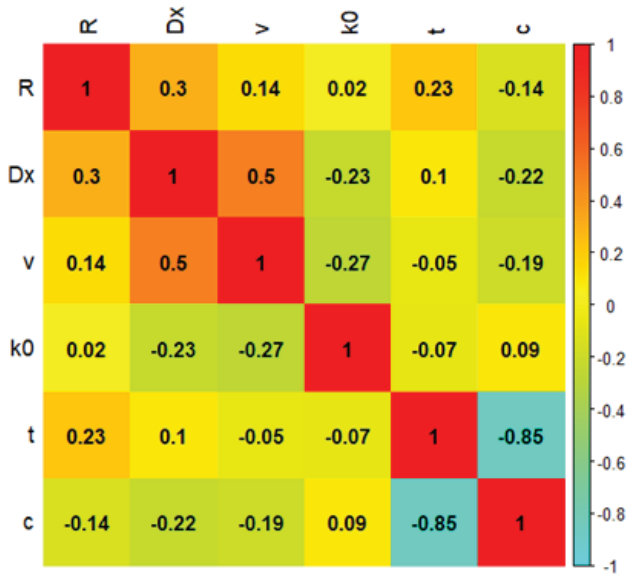


Figure 7. Correlation matrix of the input and output parameters.

Genetic Programming (CGP), and Levenberg-Marquardt (LM). Statistical values are computed in order to compare the accuracy and predictability of the models. The training dataset makes up 70% of the total data set, while the rest serves as test data. The whole procedure is represented in the flowchart shown in Figure 9.

Various statistical parameters are computed to evaluate the performance of the models. These include the adjusted coefficient of determination Adj. (R^2, bias factor, mean absolute error (MAE), mean absolute percentage error (MAPE), performance index (PI), coefficient of determination (R2), root mean square error (RMSE), variance accounted for (VAF), welfare improvement (WI), and weighted mean absolute percentage error (WMAPE). The mathematical equations for these parameters are provided below [33].

$$Adj.R^2 = 1 - \frac{(n-1)}{(n-p-1)}(1-R^2) \tag{11}$$

$$\text{Bias Factor} = \frac{1}{n}\sum_{i=1}^{n}\frac{a_i}{p_i} \tag{12}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|(\hat{y}_i - a_i)| \tag{13}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{a_i - \hat{y}_i}{a_i}\right| x\ 100 \tag{14}$$

$$PI = adj.R^2 + 0.01VAF - RMSE \tag{15}$$

$$R^2 = \frac{\sum_{i=1}^{n}(a_i - a_{mean})^2 - \sum_{i=1}^{n}(a_i - p_i)^2}{\sum_{i=1}^{n}(a_i - a_{mean})^2} \tag{16}$$

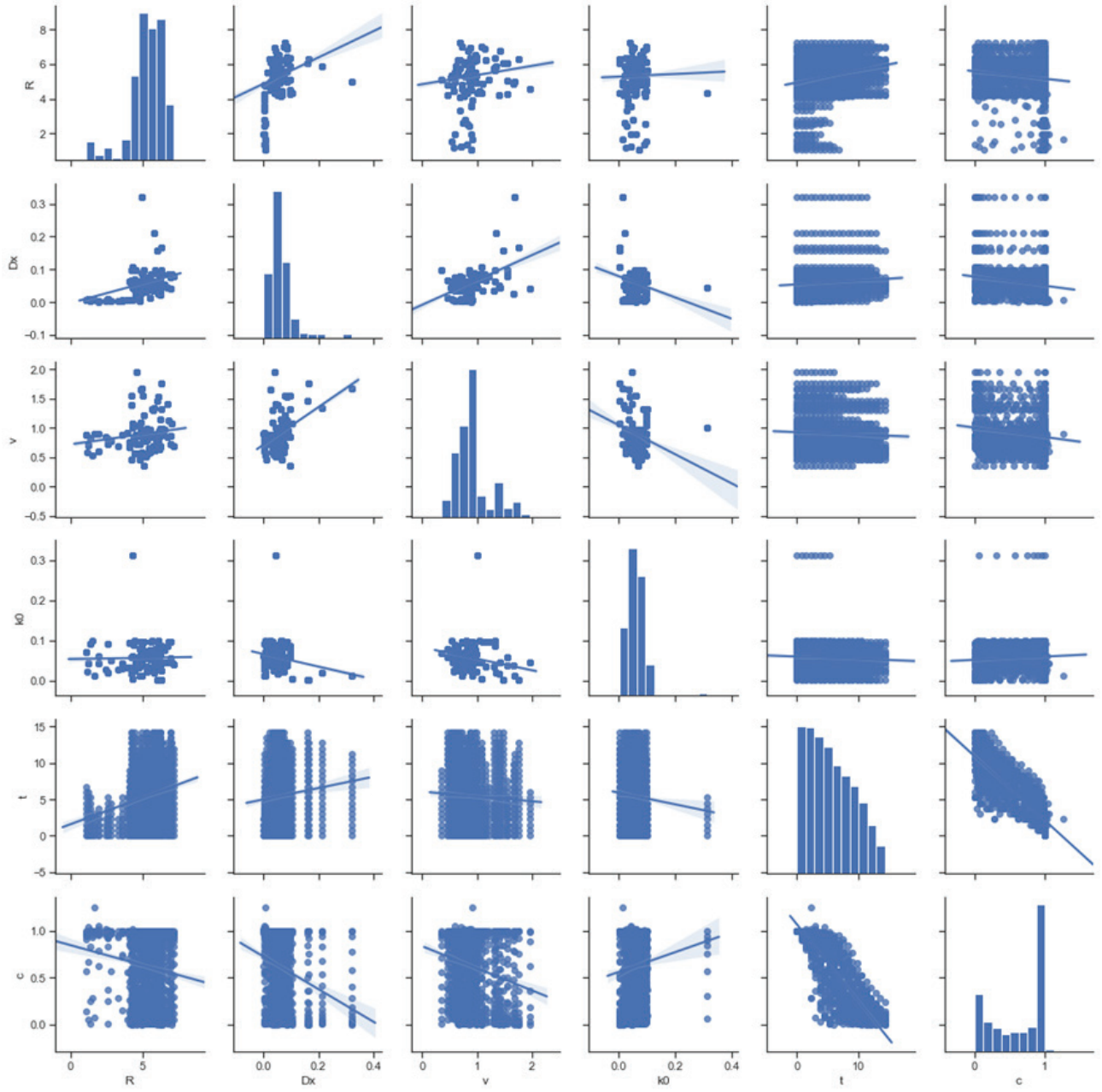


Figure 8. Pair plot of the input and output parameters.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_i - p_i)^2} \quad (17)$$

$$VAF\ (\%) = (1 - \frac{var(a_i - p_i)}{var(a_i)}) \times 100 \quad (18)$$

$$WI = U_{ai} - U_{pi} \quad (19)$$

$$WMAPE = \frac{\sum_{i=1}^{n}\left|\frac{a_i - p_i}{y_i}\right| \times a_i}{\sum_{i=1}^{n} a_i} \quad (20)$$

In this context, (a_i) and (p_i) represent the observed and predicted values for the ith sample, respectively. The total number of samples in the dataset is denoted by n, where a_{mean} refers to the mean of the observed values. Ideally, these statistical parameters should match their theoretical values for a perfect model. However, the model's reliability can be assessed by comparing these values with the upper

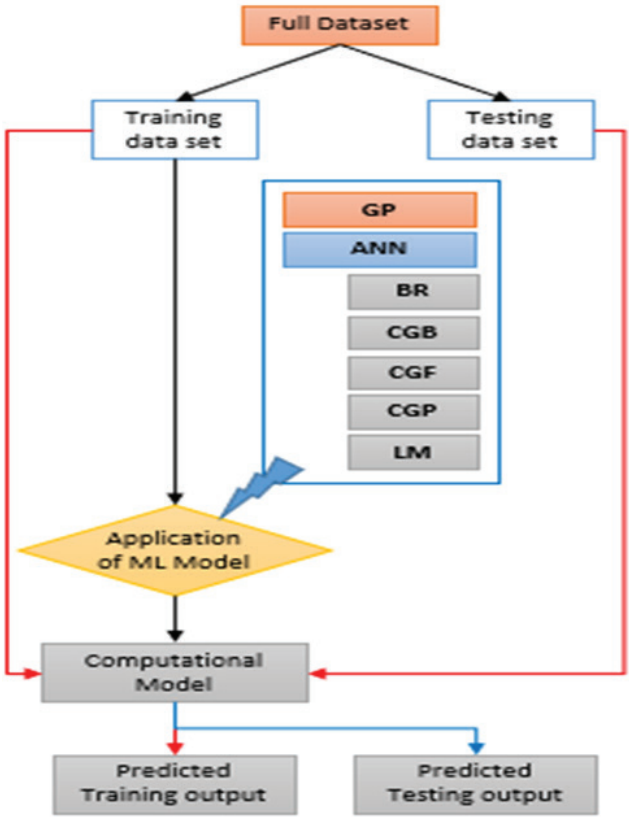


Figure 9. A flowchart depicting the data analysis process.

Table 2. Optimal values for various statistical parameters

Parameters	Ideal Value	Parameters	Ideal Value
AdjR²	1	R²	1
Bias Factor	1	RMSE	0
MAE	0	VAF	100
MAPE	0	WI	0
PI	Greater than 1	WMAPE	0

Table 3. Range of important statistical parameters

Parameters	Range	Parameters	Range
AdjR²	-1 to 1	R²	0 to 1
MAE	0 to ∞	RMSE	0 to ∞
MAPE	0 to ∞	WMAPE	0 to ∞

and lower ranges. The ideal values and ranges of these statistical parameters are presented in Table 2 and Table 3.

RESULTS AND DISCUSSION

Performance Analysis of the ANN Model

It has been analysed whether the use of an ANN approach for obtaining accurate predictions of contamination concentrations in groundwater systems is feasible. The results obtained in terms of accuracy of prediction, as well as information regarding the value of RMSE, are shown in Table 4.

Performance Analysis of the Genetic Programming (GP) Model

The feasibility of employing the GP model to achieve accurate predictions of contaminant concentrations in groundwater is assessed. The accuracies achieved during the training and testing phases are 90% and 79%, respectively, while the recorded RMSE values are 12% and 17%, respectively. The actual and predicted RMS values are shown in Figure 10.

Statistical Details of the Results

Various performance metrics are calculated to assess each model during both the training and validation phases. The results are shown in Table 5.

Visual Construal of Results

A comparison of the coefficients of determination and adjusted coefficients of determination is shown in Figures 11(a) and 11(b).

Taylor Diagram

It is a two-dimensional representation [34] that shows the relationship between actual and predicted values using metrics such as the correlation coefficient, root mean square error, and ratio of standard deviations. This diagram is useful for evaluating various aspects of complex models or comparing the performance of different models. Figure 12(a) and Figure 12(b) show the Taylor diagrams of training and testing stages for all AI-based models. The distance of each point from the origin represents the standard deviation of the normalized model or the standard deviation of two reference datasets. The radial distance of 1 (indicated by the solid arc) corresponds to perfect consistency

Table 4. Comparison of performance for various ANN models

Parameters	Training phase					Testing phase				
	ANN-BR	ANN-CGB	ANN-CGF	ANN-CGP	ANN-LM	ANN-BR	ANN-CGB	ANN-CGF	ANN-CGP	ANN-LM
MAE	0.08	0.08	0.08	0.08	0.08	0.11	0.11	0.11	0.10	0.11
R²	0.88	0.88	0.89	0.88	0.89	0.83	0.84	0.83	0.84	0.85
RMSE	0.13	0.13	0.13	0.13	0.13	0.16	0.16	0.16	0.15	0.15

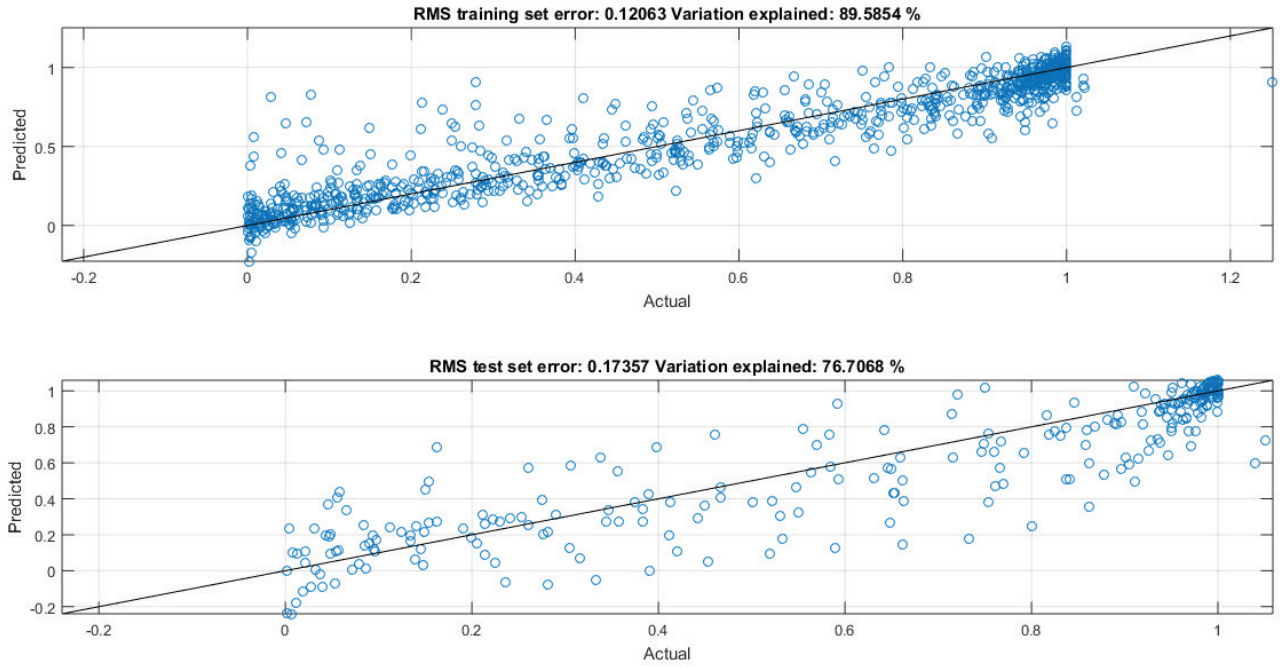


Figure 10. Actual and predicted RMS value.

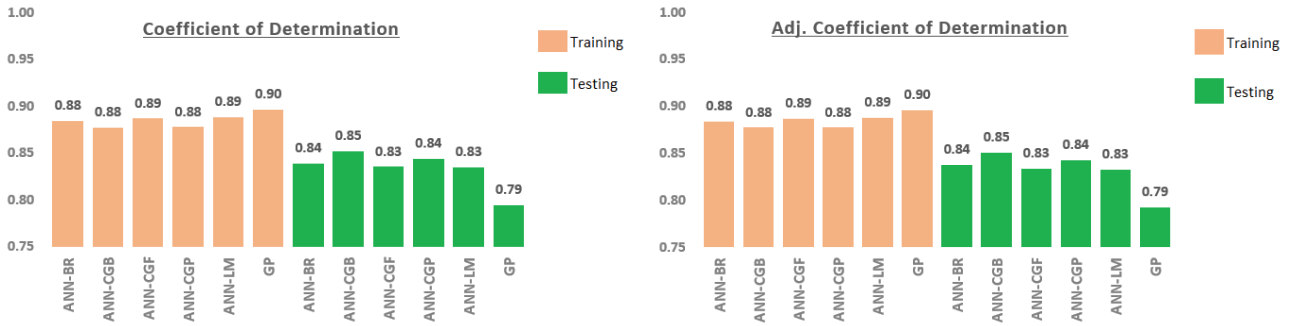
Table 5. Performance parameters of the training and validation phases

Parameters	ANN BR (TRn)	ANN BR (TSt)	ANN CGB (TRn)	ANN CGB (TSt)	ANN CGF (TRn)	ANN CGF (TSt)	ANN CGP (TRn)	ANN CGP (TSt)	ANN LM (TRn)	ANN LM (TSt)	GP (TRn)	GP (TSt)
Adj.R^2	0.88	0.84	0.88	0.85	0.89	0.83	0.88	0.84	0.89	0.83	0.90	0.79
Bias Factor	4.30	2.20	3.22	1.31	3.64	1.44	2.96	0.83	3.46	1.68	1.50	0.65
MAE	0.08	0.10	0.08	0.11	0.08	0.11	0.08	0.11	0.08	0.11	0.08	0.12
MAPE	3.52	1.50	2.63	0.83	2.78	0.75	2.80	1.34	3.88	1.01	3.28	1.48
PI	1.64	1.52	1.62	1.55	1.65	1.50	1.62	1.52	1.65	1.50	1.67	1.40
R^2	0.88	0.84	0.88	0.85	0.89	0.83	0.88	0.84	0.89	0.83	0.90	0.79
RMSE	0.13	0.15	0.13	0.15	0.13	0.16	0.13	0.16	0.13	0.16	0.12	0.17
VAF	88.38	83.41	87.73	84.95	88.65	82.89	87.74	83.83	88.79	82.81	89.59	78.14
WI	0.98	0.95	0.97	0.95	0.98	0.95	0.97	0.95	0.98	0.95	0.98	0.94
WMAPE	0.13	0.16	0.14	0.17	0.13	0.17	0.13	0.16	0.13	0.17	0.13	0.19

in variance with the reference dataset. The angle of each point relative to the horizontal axis represents the correlation coefficient. The points closest to the straight edge (with a correlation of 1.0) exhibit the highest correlation with the reference dataset. The distance between a model point and the reference point (points at standard deviation = 1.0, correlation = 1.0) indicates the unified RMSE.

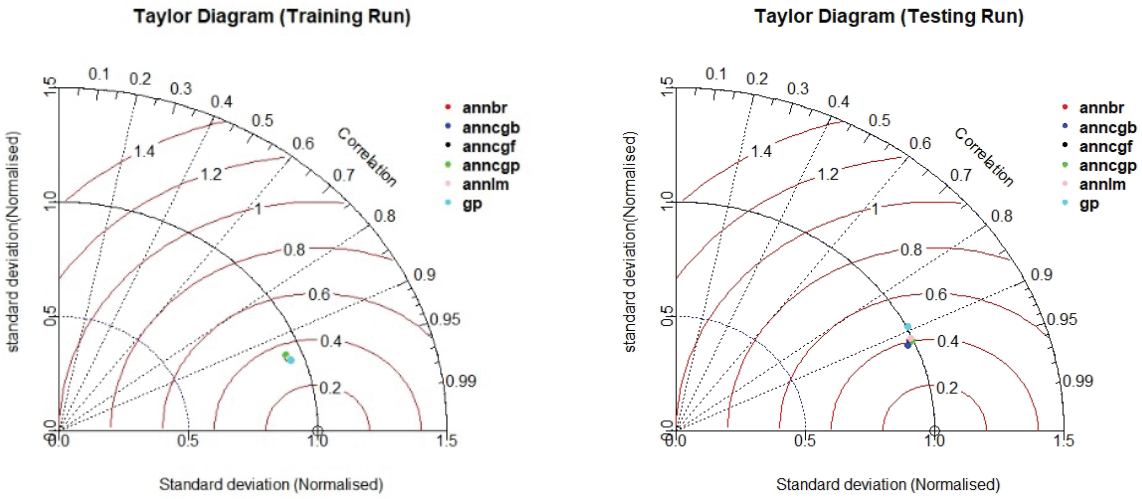
Figure 12(a) illustrates a moderate correlation (~0.75) for ANN-BR (yellow dots). Compared with the other models, the model results in a slightly greater deviation (~1.2), indicating an overestimation of variance and a larger RMSE, positioning it further from the reference point. ANN-CGB (blue points) shows a correlation coefficient of about 0.95, as can be seen from Figure 12(a). The deviations of its pattern are very close to 1.0, showing a consistent variance, while its low RMSE means that it performs well. On the other hand, ANN-CGF (green points) shows a correlation coefficient of about 0.85, as depicted in Figure 12(a).

The correlation coefficient of ANN-CGP (blue dots) is ~0.9 (Fig. 12a), where model deviation is closely correlated with that of the benchmark (~1.0). The low RMSE implies high consistency for this model. ANN-LM has the highest correlation (~0.99) out of all models analyzed (Fig. 12a). Model deviation is close to the benchmark (~1.0) and also



(a) (b)

Figure 11. (a) Coefficient of determination and (b) adjusted coefficient of determination comparison.



(a) (b)

Figure 12. (a) Taylor diagram in training phase and **(b)** Taylor diagram in testing phase

has the minimum RMSE, hence the best model accuracy. The correlation coefficient for the GP model (light blue dots) is ~0.85 (Fig. 12a), and it slightly underestimates the variance (~0.9 standard deviation).

Generally, "ANN-LM" (represented in cyan color) shows optimal results due to having the highest correlation coefficient (around 0.99), best fit to the deviation of the reference model, and the lowest RMSE value. There are other models like "ANN-CGB" and "ANN-CGP" that also show excellent performance but are slightly inferior compared to "ANN-LM". On the contrary, "ANN-BR" model shows slightly inferior results.

The ANN-BR model demonstrates a fairly high correlation (≈0.65) as seen in Figure 12(b). The deviation of the ANN-BR model is substantially higher than 1.0 (≈1.4), which means that the ANN-BR model overestimates the variance, having a high RMSE value. This puts ANN-BR away from the reference point in comparison to the other two models. On the contrary, the ANN-CGB model displays a very high correlation coefficient (≈0.9) as illustrated in Figure 12(b). The deviation is slightly higher than 1.0 (≈1.05).

For instance, ANN-CGF shows a strong correlation (about 0.95) (Fig. 12(b)), while the value for its standard deviation is almost equal to 1.0 (about 1.02). This shows that this model has good consistency with the reference data. The second model, ANN-LM, exhibits the highest correlation (about 0.99) and perfect consistency (model deviation equals 1.0) with the variance of the reference data. It is the most accurate with the lowest RMSE among all other models.

In the GP test, the performance of the proposed model exhibits a correlation of about 0.75 (Fig. 12(b)) with a standard deviation marginally lower than 1.0 (0.9), indicating a modest tendency to underestimate the variance. Nevertheless, the proposed model has a higher RMSE compared to the ANN-LM and ANN-CGP models. In general,

the ANN-LM (blue) demonstrates a superior correlation coefficient (~0.99), a perfect variance coefficient (model deviation=1.0), and minimal RMSE. Similarly, the ANN-CGP (red) and ANN-CGB (blue) have relatively high correlations, consistent variances, and minimum RMSE. ANN-BR (red) and GP (light blue) display the poorest correlation and highest RMSE.

Regression Error Characteristic Curve

The receiver operating characteristic (ROC) curve is used to compare the classification results and generalize the regression curve [35]. The REC is the generalized term of the ROC curve for regression. REC curves plot the error tolerance on the x-axis against the percentage of points predicted within that tolerance on the y-axis. The resulting curve estimates the cumulative distribution function of the error. The REC curve visually presents commonly used statistics, allowing users to assess the relative merits of various regression functions by comparing their REC curves. The curve shape also provides additional insights to guide modelling. Figures 13(a) and 13(b) display the REC curves for the training and testing phases.

The area over the curve (AOC) values quantifies the expected error of regression models, with lower values indicating superior performance. The AOC values reveal that the ANN-BR model outperforms all the other models in both the training and testing phases. Specifically, the ANN-BR model has the lowest AOC value of 0.077 during training and 0.102 during testing. A graphical representation of the AOC values is shown in Figure 14.

Flow contamination through groundwater is calculated via the advection-dispersion equation. However, solving

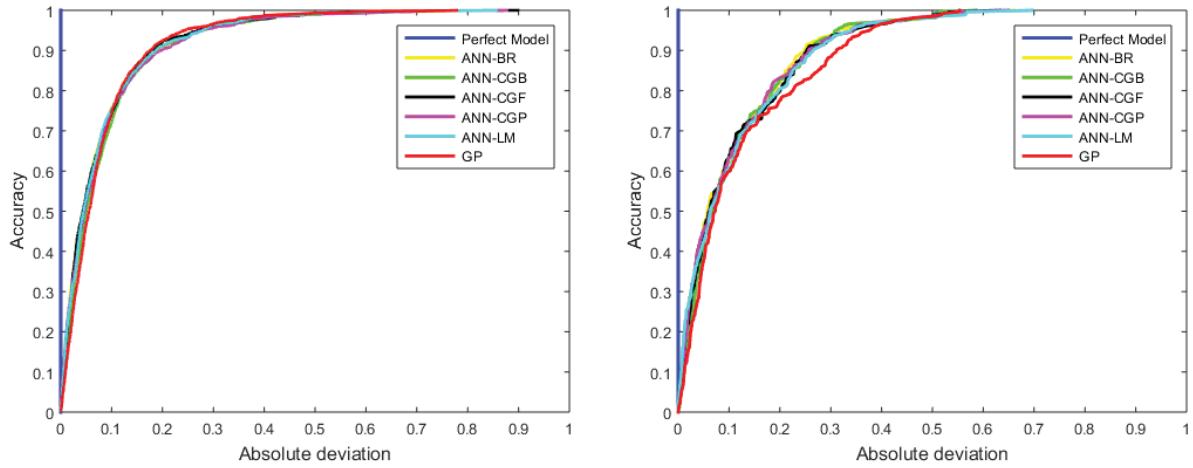


Figure 13. (a) REC curve for training phase and **(b)** REC curve for testing phase.

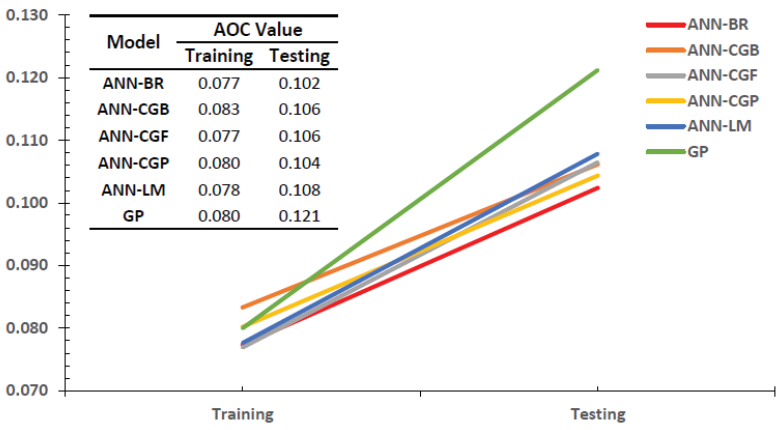


Figure 14. Graphical representations of AOC.

Table 6. Best model in the training and testing phases (statistical parameters)

Phase	AdjR2	MAE	R2	RMSE	VAF
	Highest accuracy achieved				
Training phase	GP	All Model	GP	GP	GP
Testing Phase	ANN-CGB	ANN-BR	ANN-CGB	ANN-BR ANN-CGB	ANN-CGB

this mathematical formulation can be quite complex, necessitating the application of artificial intelligence to model contaminant concentrations in groundwater. Artificial intelligence has been employed to develop groundwater models, with artificial neural networks (ANNs) and genetic algorithms being widely used for predicting solute concentrations and various groundwater contamination applications. Genetic algorithms and ANNs are used to train data generated from the analytical solution of the advection–dispersion equation, enabling the forecasting of contaminant concentrations. ANN-BR, ANN-CGB, ANN-CGP, and ANN-LM are some of the ANN models used for predicting the concentration of solutes in groundwater. Their results are provided in Table 6, from which it is possible to determine the most appropriate regression model. The use of hydrological data in these models enables efficient simulation and pollution prediction. They can be used as effective decision-making tools in the process of rehabilitation. The speed and precision of their predictions are useful when dealing with limited hydrological data.

The performance efficiency and prediction accuracy of the developed models are evaluated using a dataset of 1,361 data points. Of these, 1,040 data points are used for training, and the remaining 321 are used for testing the ANN model. The RMSE and R^2 values for the various models are provided in Table 5. The ANN-BR model delivers the best performance, with operating frequencies computed using MATLAB software. The training and testing results show that the GP model slightly outperforms the ANN models in terms of the RMSE and R^2, as shown in Table 6.

The errors obtained in the ANN and GP during training and testing stages have been evaluated in terms of Absolute Fraction of Variance (R^2), Mean Absolute Percentage Error (MAPE), and Root Mean Square Error (RMSE) [36]. For validation purposes, the RMSE, MAE, and R^2 measures have been employed. Based on the results obtained, it has been found that the proposed ANN-BR model, having frequencies calculated with MATLAB software, has provided the best outcome with a maximum value of R^2 as 0.87. The ANN model has used the Conjugate Gradient with Powell-Beale Restarts (CGB) learning algorithm, which possesses a fast-learning capability and ensures good convergence. Based on the above observations, it can be inferred that the feed-forward neural network utilizing the CGB algorithm is efficient in predicting groundwater contamination. It has been noticed that ANN based algorithms prove to be more efficient in predicting groundwater contaminants with minimal hydrological parameters.

Differently from the previous studies that applied an ANN model optimized with genetic algorithm for the prediction of groundwater table depths, for instance Pandey et al. [37], this study presents a new model of hybrid GP-ANN designed for prediction of solute concentrations. The work differs in terms of issues being addressed and a novel approach that uses combination of genetic programming and ANN to improve forecast accuracy. Also, unlike most of previous studies that used the advection-dispersion equation to simulate solute transport, this study distinguishes itself through incorporation of AI-based methods for better forecasting. In general, the results show that the hybrid GP-ANN model, especially its ANN-LM form, provides more accurate forecasting of solute concentrations.

Some basic uncertainties can arise during the research, affecting the results that would be obtained through it. First, random input data, generated for the study within a specified time, might not reflect real-life conditions, leading to inconsistencies between the simulation and the occurrence of pollution. Secondly, the use of convection-diffusion equations in the research makes some assumptions about the aquifer such as its homogeneity and isotropy.

CONCLUSION

This research study successfully showcases how state-of-the-art computational models such as Genetic Programming (GP) and Artificial Neural Networks (ANN) can be applied to estimate the concentration of a particular solute in groundwater, an essential parameter involved in groundwater pollution mitigation techniques. The model of advection-dispersion equation was trained using the data obtained through the analytical model. According to the findings, the GP and ANN algorithms were both able to give accurate estimation of groundwater pollution levels. Performance assessment was conducted using different parameters like R^2, RMSE, and AOC. Amongst all RNA-based models, the ANN-LM proved to be the most efficient. It gave a correlation value close to 0.99. While the GP model exhibited a slightly poor performance compared to the ANN model when it comes to the aspects of correlation (0.75) and RMSE, its capability to predict accurately right

from the beginning phase of the study could be considered as impressive. It underscores the capabilities of GP in the prediction of water pollution at an early stage. In addition to that, the performance of ANN-BR model, despite being less efficient compared to ANN-LM, can be considered impressive since it achieved the least AOC value.

The findings from this research emphasize the possibility of combining AI-based approaches such as ANN and GP with the existing methods of mathematical modeling to enhance the precision of predictions concerning groundwater pollution. It also emphasizes the applicability of these two machine learning algorithms, especially ANN, in dealing with complicated environmental systems using little information about hydrology. The knowledge generated from this research will be instrumental in the field of environmental monitoring and water resource management.

AUTHORSHIP CONTRIBUTIONS

Authors equally contributed to this work.

DATA AVAILABILITY STATEMENT

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

CONFLICT OF INTEREST

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICS

There are no ethical issues with the publication of this manuscript.

STATEMENT ON THE USE OF ARTIFICIAL INTELLIGENCE

Artificial intelligence was not used in the preparation of the article.

REFERENCES

[1] Ford SR, Dreger DS, Walter WR. Source characterization of the 6 August 2007 Crandall Canyon Mine seismic event in central Utah. Seismol Res Lett 2008;79:637–644. [CrossRef]

[2] Tecirli ES, Akgün K, Çaglak A, Sari Erkan H, Onkal Engin G. Treatment of textile wastewater in combined granular activated carbon-membrane bioreactor (GAC-MBR). Sigma J Eng Nat Sci 2024;42:1490–1499. [CrossRef]

[3] O'Donnell DE, Aaron S, Bourbeau J, Hernandez P, Marciniuk DD, Balter M, et al. Canadian Thoracic Society recommendations for management of chronic obstructive pulmonary disease-2007 update. Can Respir J 2007;14(Suppl B):5B–32B. [CrossRef]

[4] Bettenbrock K, Fischer S, Kremling A, Jahreis K, Sauter T, Gilles ED. A quantitative approach to catabolite repression in Escherichia coli. J Biol Chem 2006;281:2578–2584. [CrossRef]

[5] Taylor GI. Dispersion of soluble matter in solvent flowing slowly through a tube. Proc R Soc Lond A Math Phys Sci 1953;219:186–203. [CrossRef]

[6] Bryan A, Pepper G, Wener MH, Fink SL, Morishima C, Chaudhary A, et al. Performance characteristics of the Abbott Architect SARS-CoV-2 IgG assay and seroprevalence in Boise, Idaho. J Clin Microbiol 2020;58:e00941-20. [CrossRef]

[7] Parker JC, van Genuchten MT. Flux-averaged and volume-averaged concentrations in continuum approaches to solute transport. Water Resour Res 1984;20:866–872. [CrossRef]

[8] Bear J. Hydraulics of groundwater. New York, USA: Dover Publications; 2012.

[9] Delleur JW, (editor). Handbook of Groundwater Engineering. 1st ed. Berlin: Springer; 2014.

[10] Bierkens MFPA. Modelling groundwater flow and transport with uncertainty quantification: A review. J Hydrol 1999;220:1–13.

[11] Kisi O, Zounemat-Kermani M. A novel artificial intelligence model for predicting groundwater contamination. Water Resour Manag 2020;34:65–83.

[12] Zhou X. Application of artificial neural networks in predicting groundwater quality parameters. Environ Earth Sci 2020;79:152.

[13] Shahmoradi B. Modeling solute concentration in contaminated groundwater using ANN and statistical methods. J Contam Hydrol 2021;239:103802.

[14] Pujol J, Obregón-Neira N. Genetic programming for environmental modeling: Case study on groundwater contaminant transport. J Environ Manage 2017;91:905–911.

[15] Ghosh S, Ghosh S, Datta B. Genetic programming approach to simulate groundwater contaminant transport. J Hydrol 2015;522:385–397.

[16] Fukuda M, Datta B, Oishi S. Application of genetic programming for groundwater contamination modeling. Groundw Monit Remediat 2012;32:68–76.

[17] Kumar P, Kumar S, Kumar M, Chaubey A, Noyak J. Groundwater flow and solute transport modelling: A mathematical analysis. IJEP 2024;44:257–264.

[18] Yalman Kosunalp H, Gulsu M. Operational matrix for multi-order fractional differential equations with Hermite polynomials. Sigma J Eng Nat Sci 2024;42:1050–1057. [CrossRef]

[19] Huang J, Goltz MN. Analytical solutions for a soil vapour extraction model that incorporates gas phase dispersion and molecular diffusion. J Hydrol 2017;549:452–460. [CrossRef]

[20] Abrahart R, Kneale PE, See LM. Neural networks for hydrological modelling. Boca Raton, Florida: CRC Press; 2004. [CrossRef]
[21] Coulibaly P, Anctil F, Aravena R, Bobée B. Artificial neural network modeling of water table depth fluctuations. Water Resour Res 2001;37:885–896. [CrossRef]
[22] Khatau SB, Hale CM, Stewart-Hutchinson PJ, Patel MS, Stewart CL, Searson PC, et al. A perinuclear actin cap regulates nuclear shape. Proc Natl Acad Sci U S A 2009;106:19017–19022. [CrossRef]
[23] Angeline PJ, Saunders GM, Pollack JB. An evolutionary algorithm that constructs recurrent neural networks. IEEE Trans Neural Netw 1994;5:54–65. [CrossRef]
[24] Searson DP. GPTIPS 2: An open-source software platform for symbolic data mining. arXiv 2010;1:77–80.
[25] Koza JR. Evolution of subsumption using genetic programming. Cambridge: MIT Press; 1992. p. 110–119.
[26] Alfaro-Cid E, Esparcia-Alcázar AI, Moya P, Femenia-Ferrer B, Sharman K, Merelo JJ. Modeling pheromone dispensers using genetic programming. In: Giacobini M, Brabazon A, Cagnoni S, Caro GDA, Ekart A, Esparcia-Alcazar A, et al. (editors). Applications of Evolutionary Computing. New York: Springer; 2009. p. 635–644. [CrossRef]
[27] Nasseri MHB, Mohanty B. Fracture toughness anisotropy in granitic rocks. Int J Rock Mech Min Sci 2008;45:167–193. [CrossRef]
[28] Premack D. Human and animal cognition: Continuity and discontinuity. Proc Natl Acad Sci U S A 2007;104:13861–13867. [CrossRef]
[29] Mohanty S, Jha MK, Kumar A, Sudheer KP. Artificial neural network modeling for groundwater level forecasting in a river island of eastern India. Water Resour Manag 2010;24:1845–1865. [CrossRef]
[30] Maier HR, Dandy GC, Burch MD. Use of artificial neural networks for modelling cyanobacteria Anabaena spp. in the River Murray, South Australia. Ecol Modell 1998;105:257–272. [CrossRef]
[31] Daliakopoulos IN, Coulibaly P, Tsanis IK. Groundwater level forecasting using artificial neural networks. J Hydrol 2005;309:229–240. [CrossRef]
[32] Goltz JH. Analytical modeling of solute transport in groundwater: Using models to understand the effect of natural processes on contaminant fate and transport. Vol. 1. John Wiley & Sons; 2017. [CrossRef]
[33] Kumar S, Kumar DR, Wipulanusat W, Keawsawasvong S. Development of ANN-based metaheuristic models for the study of the durability characteristics of high-volume fly ash self-compacting concrete with silica fume. J Build Eng 2024;94:109844. [CrossRef]
[34] Taylor KE. Summarizing multiple aspects of model performance in a single diagram. J Geophys Res Atmos 2001;106:7183–7192. [CrossRef]
[35] Bennett KP, Bi J, Edu B. Regression error characteristic curves. In: Proceedings of the 20th International Conference on Machine Learning (ICML-03). 2003:43–50.
[36] Topçu İB, Sarıdemir M. Prediction of compressive strength of concrete containing fly ash using artificial neural networks and fuzzy logic. Comput Mater Sci 2008;41:305–311. [CrossRef]
[37] Pandey K, Kumar S, Malik A, Kuriqi A. Artificial neural network optimized with a genetic algorithm for seasonal groundwater table depth prediction in Uttar Pradesh, India. Sustainability 2020;12:1–24. [CrossRef]