



# OPEN Development of soft computing-based models for forecasting water quality index of Lorestan Province, Iran

Balraj Singh<sup>1</sup>, Alireza Sepahvand<sup>2</sup>, Parveen Sihag<sup>3</sup>, Karan Singh<sup>4</sup>, Chander Prabha<sup>5</sup>, Anindya Nag<sup>6</sup>, Md. Mehedi Hassan<sup>6</sup>, S. Vimal<sup>7</sup> & Dongwann Kang<sup>8</sup>✉

The Water Quality Index (WQI) is widely used as a classification indicator and essential parameter for water resources management projects. WQI combines several physical and chemical parameters into a single metric to measure the status of Water Quality. This study explores the application of five soft computing techniques, including Gene Expression Programming, Gaussian Process, Reduced Error Pruning Tree (REPT), Artificial Neural Network with FireFly (ANN-FFA), and combinations of Reduced Error Pruning Tree with bagging. These models aim to predict the WQI of Khorramabad, Biranshahr, and Alashtar sub-watersheds in Lorestan province, Iran. The dataset consists of 124 observations, with input variables being sulfate ( $\text{SO}_4$ ), total dissolved solids (TDS), the potential of Hydrogen (pH), chloride (Cl), electrical conductivity (EC), Potassium (K), bicarbonate ( $\text{HCO}_3$ ), magnesium (Mg), sodium (Na), and calcium (Ca), and WQI as the output variable. For model creation (train subset) and model validation (test subset), the data were split into two subsets (train and test) in a ratio of 70:30. The performance evaluation parameters values of training and testing stages of various models indicate that the ANN-FFA based data-driven model performs better than the other modeling techniques applied with the values of coefficient of correlation 0.9990 & 0.9989; coefficient of determination 0.9612 & 0.9980; root mean square error 0.3036 & 0.3340; Nash–Sutcliffe error 0.9980 & 0.9979; and Mean average percentage error 0.7259% & 0.7969% for the train and test subsets, respectively. Taylor diagram results also suggest that ANN-FFA is the best-performing model, followed by the GEP model. This study introduces a novel model for predicting WQI using advanced soft computing models that have not been previously applied in this study area, highlighting its novelty and relevance. The proposed model significantly enhances predictive accuracy and efficiency, offering real-time, cost-effective WQI predictions that outperform traditional methods in handling complex, nonlinear environmental data.

**Keywords** Water quality index, Artificial neural network, FireFly algorithm, Gene expression programming, Reduced error pruning tree, Lorestan Province

Surface water is essential for ecology, social well-being, and economic growth<sup>1–3</sup>. Water quality (WQ) is influenced by various variables, including natural ones like rainfall and erosion and human ones like urban, agricultural, and industry operations<sup>4–6</sup>. Because surface water is the world's leading supplier of fresh water, its deterioration may have a considerable impact on the availability of drinking water and, more broadly, on economic growth and long-term plans<sup>7–9</sup>. Water pollution is caused by interactions with their surroundings and the subsequent interchange of toxins from urban, industrial, and agricultural sources along their course<sup>10–12</sup>,

<sup>1</sup>Panipat Institute of Engineering and Technology, Panipat 132102, India. <sup>2</sup>Lorestan University, Khorramabad 68151-44316, Iran. <sup>3</sup>Chandigarh University, Mohali 140413, India. <sup>4</sup>Department of Civil Engineering, Graphic Era (Deemed to be University), Dehradun 248002, India. <sup>5</sup>Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab, India. <sup>6</sup>Computer Science and Engineering Discipline, Khulna University, Khulna, Bangladesh. <sup>7</sup>Center of Excellence in Data Science, Department of Artificial Intelligence and Data Science, Sri Eshwar College of Engineering, Coimbatore 641202, India. <sup>8</sup>Department of Computer Science and Engineering, Seoul National University of Science and Technology, Seoul 01811, South Korea. ✉email: dongwann@seoultech.ac.kr

which results in polluting the freshwater ecosystems<sup>13,14</sup>, urban water systems<sup>15,16</sup>, and agriculture land<sup>17</sup>, due to present of microplastics and other polluted substances.

In order to assess and classify the quality of ground and surface waters, the WQI has been widely used as a classification indicator and is essential for managing water resources<sup>18–20</sup>. WQI combines several physical and chemical parameters into a single metric to measure the status of WQ<sup>21</sup>. This indicator's computation offers a practical method for evaluating the WQ. The WQI application was initially introduced by Horton<sup>22</sup> and Brown et al.<sup>23</sup>, and several practitioners later adopted and modified it<sup>24,25</sup>. WQI formulations often include extensive computations, which take time and effort. Additionally, the traditional methods for calculating the WQI need significant physical and chemical data, usually at daily intervals. Therefore, alternate methods for accurately and efficiently computing WQI are needed; environmental engineers may find this helpful innovation when monitoring and evaluating water quality.

In the form of machine learning models, soft computing models have been used increasingly in the last several decades to handle various environmental engineering challenges, such as river WQ modeling<sup>26–30</sup>. According to Yaseen et al.<sup>31</sup>, soft computing models significantly advance engineering process monitoring and control. Their methods may be used to make precise predictions without requiring intricate programming. Soft computing models are built on data mining and discovering patterns in data. For this, algorithms are built using a portion of the dataset (train), and the performance of predictions is tested using a different subset of the dataset (test)<sup>32–34</sup>. Our literature analysis shows that WQI simulation utilizing soft computing models has received much attention<sup>35</sup>. Tripathi and Singal<sup>36</sup> used the Principal Components Analysis (PCA) model to choose the ideal input variable combination and offer a novel way to compute the WQI in the Ganges River (India). By employing this technique, they could drastically cut the parameters from twenty-eight to just nine. Zali et al.<sup>37</sup> investigated the impacts of six primary input factors on the WQI using ANNs. They conducted a sensitivity analysis to determine the relative significance of each parameter in determining WQI, and they concluded that DO, SS, and NO<sub>3</sub> are the critical input factors. The ground WQI was calculated using a fuzzy-based model by Nigam and SM<sup>38</sup>, who also compared its prediction performance to other widely used calculation techniques. They discovered that the fuzzy-based model outperformed them. The Interactive Fuzzy model (IFWQI) was used by Srinivas and Singh<sup>39</sup> to construct a unique fuzzy decision-making technique for predicting WQI in rivers. Their findings show that the proposed model performs much better predicting WQI than the conventional fuzzy method. According to Yaseen et al.<sup>31</sup>, ANFIS-SC (Subtractive Clustering) was the best model for predicting WQI out of three hybrid methods based on the Adaptive Neuro-Fuzzy Inference System (ANFIS). These were ANFIS-FCM (Fuzzy C-Means data clustering), ANFIS-GP (Grid Partition), and ANFIS-SC (Subtractive Clustering).

Environmental scientists have been looking into other strong and reliable data-driven models, even though standard models based on ANN and ANFIS are well known for WQI modeling<sup>26,27,35</sup> to show how WQI affects different chemical factors in tropical environments. Another prominent strategy used effectively for different hydrological and environmental issues, such as rainfall forecasting, is tree-based models, such as Decision Trees (DTs)<sup>40</sup>. For predicting WQ, Granata et al.<sup>41</sup> made a Support Vector Regression (SVR) model, a Gaussian Process (GP) model, and a Regression Tree (RT) model. The SVR model worked the best for them. These relate to applying decision-tree and support vector regression models for WQ parameter prediction. Li et al.<sup>42</sup> suggested a hybrid SVR model with the FireFly Algorithm (FFA) to predict WQI using monthly data on the WQ parameter. This model was much better at making predictions than the standalone SVR model. Nitrate was discovered to be the most significant parameter for WQI prediction by Kamyab-Talesh et al.<sup>43</sup>. They explored the optimization of the SVM model to find the parameters that primarily impact the WQI. Wang et al.<sup>44</sup> investigated the performance of three machine learning models, SVR, SVR-GA (Genetic Algorithm), and SVR-PSO (Particle Swarm Optimization), to predict WQI using the spectral indicators Difference Index (DI), Normalized DI, and Ratio Index (RI) that were obtained from remote sensing, and found that the SVR-PSO was the best performing model.

Numerous studies have pointed out the uncertainty in soft computing models<sup>30,34,45–47</sup>. Enhancing the reliability and effectiveness of soft computing forecasts is crucial. Techniques such as Artificial Neural Networks (ANN), Fuzzy Logic, and Adaptive Neuro-Fuzzy Inference Systems (ANFIS) often operate within complex, nonlinear problem domains where inaccuracies in data, model parameters, and predictions are inevitable. It facilitates the quantification of model confidence, allowing for the provision of point forecasts and a range of probable outcomes accompanied by associated uncertainty levels<sup>48</sup>. Furthermore, it improves the reliability of soft computing models, making their predictions more resilient and comprehensible<sup>49</sup>. The hybrid model effectively addresses uncertainty since the FFA and bagging models enhance the robustness of the ANN and REPT models, respectively, and mitigate uncertainty.

Soft computing methods are advocated due to the infeasibility of performing consistent global monitoring of water quality in all rivers<sup>50</sup>. This research introduces several methodologies for predicting the WQI of three sub-watersheds in Iran, using soft computing models that have yet to be used in this region. The models included the Gaussian Process (GP), Gene Expression Programming (GEP), REP tree (REPt), Bagging REP tree (BREPt), and a hybrid Artificial Neural Network – FireFly Algorithm (ANN-FFA). The innovation of this research is in the development of hybrid models, namely combining Artificial Neural Networks (ANN) with Firefly Algorithm (FFA) and Bagging Random Ensemble Pruning Technique (REPt). These models have not been used before to predict the WQI in these sub-watersheds. Evaluating the Water Quality Index in a laboratory is costly and labor-intensive due to the processes of sample collection, transportation, and testing. This research introduces a real-time prediction system that employs soft computing models as an alternative method for predicting WQI. The objective is to rectify the deficiency in precise and reliable prediction of WQI by examining the efficacy of advanced soft computing models, namely Artificial Neural Network with Firefly Algorithm (ANN-FFA) and Bagging Random Enhanced Predictive Trees (REPt). These models are undergoing evaluation in comparison to conventional approaches like the Gaussian Process (GP), Gene Expression Programming (GEP), and

Randomized Exponential Perturbation Tree (REPt). This work introduces and validates ANN-FFA as a superior model for predicting WQI and enhancing accuracy and reliability in water quality management.

### Data-driven models

Recently, many researchers have used soft computing models in civil engineering and water resources<sup>51–62</sup>. The ANN-FFA, BREPt, REPt, GP, and GEP are used in this study, and the details of these models are as follows:

#### Gaussian process (GP)

GP regression is a state-of-the-art method straight over the function space based on the premise that neighboring studies must exchange information. Gaussian regression is the term for the extension of the Gaussian distribution. In GP regression, the covariance and mean are expressed as the matrix and vector of the Gaussian distribution. The validation for generalization is not necessarily due to prior knowledge of functional dependency and data. The GP Regression models can tell the difference between the forecast distribution and the input test data<sup>63</sup>. Any finite number of the random variables that make up a GP have a multivariate Gaussian distribution. Assuming  $p$  and  $q$  represent the input and target domains,  $x$  pairs  $(x_p, x_q)$  are selected randomly and equally. The mean function  $v_0$  represents a GP on  $p$  in regression, assuming that  $h \subseteq Re$ . The radial basis kernel (rbf) and Pearson VII kernel (puk) functions are used in this investigation. Kuss<sup>64</sup> is recommended to readers for further details. This investigation uses two kernels, rbf and puk, for model development in GP regression.

$$puk = \left( 1 / \left[ 1 + \left( 2 \sqrt{\|x_i - x_i\|^2} \sqrt{2 \left( \frac{1}{\omega} \right)^{-1} / \sigma} \right)^2 \right]^{\omega} \right) \quad (1)$$

$$rbf = \left( e^{-\gamma \|x_i - x_i\|^2} \right) \quad (2)$$

Here,  $\omega$ ,  $\gamma$ , and  $\sigma$  are the parameters of kernels.

GP offers probabilistic predictions, interpolation, versatility, and compatibility with small data sets. It can handle privacy constraints by adding synthetic noise, model complex phenomena, and provide uncertainty estimates by learning noise and smoothness parameters from training data. However, it has several disadvantages, including being not sparse, inefficient in high dimensions, unsuitable for outliers, positive-only variables, computationally expensive, difficult to choose a kernel function, and potentially requiring careful hyperparameter tuning, which can be computationally expensive and require careful optimization techniques<sup>65</sup>.

#### GEP (gene expression programming)

GEP is a strategy suggested by Ferreira<sup>66</sup> that uses software programs. It is an advanced technique built on the GA foundation frequently used in recent research. The linear chromosomes that comprise the GEP's software programs are articulated or translated into ETs. The initial step in this program's problem-solving process is to create the initial population, accomplished through the arbitrary birth of chromosomes. Later, the chromosomes transform into extracellular organisms, which are then evaluated according to performance criteria to represent the solubility of the produced ETs. If the results satisfy the performance standards, population generating ceases. If the outcomes are unsatisfactory, the system regenerates with some improvements to create a new generation with improved value. The benefit of GEP is the ability to contrast chromosomes in a symbolic and linear string of a predetermined length. The step involved in the GEP is depicted in Fig. 1. For this study, the mutation rate is 0.044, inversion rate, incessant and root scale transport rate is 0.1. one -point and two-point crossover rate is 0.3, gene recombination and transportation rate is 0.1, no. of chromosomes is 30 with three head size, no. of gene per chromosomes is 3. Researchers are directed to Ferreira<sup>67</sup> and Ebtehaj et al.<sup>68</sup> for further information on GEP.

GEP offers flexibility, efficiency, simplicity, and power in evolutionary computation methods. It allows for varying solutions and linear representation of chromosomes and can solve complex problems by evolving intricate models representing nonlinear relationships. Despite its strengths, it faces challenges like complex solutions, computational resources, and parameter tuning, which can be challenging to interpret and optimize.

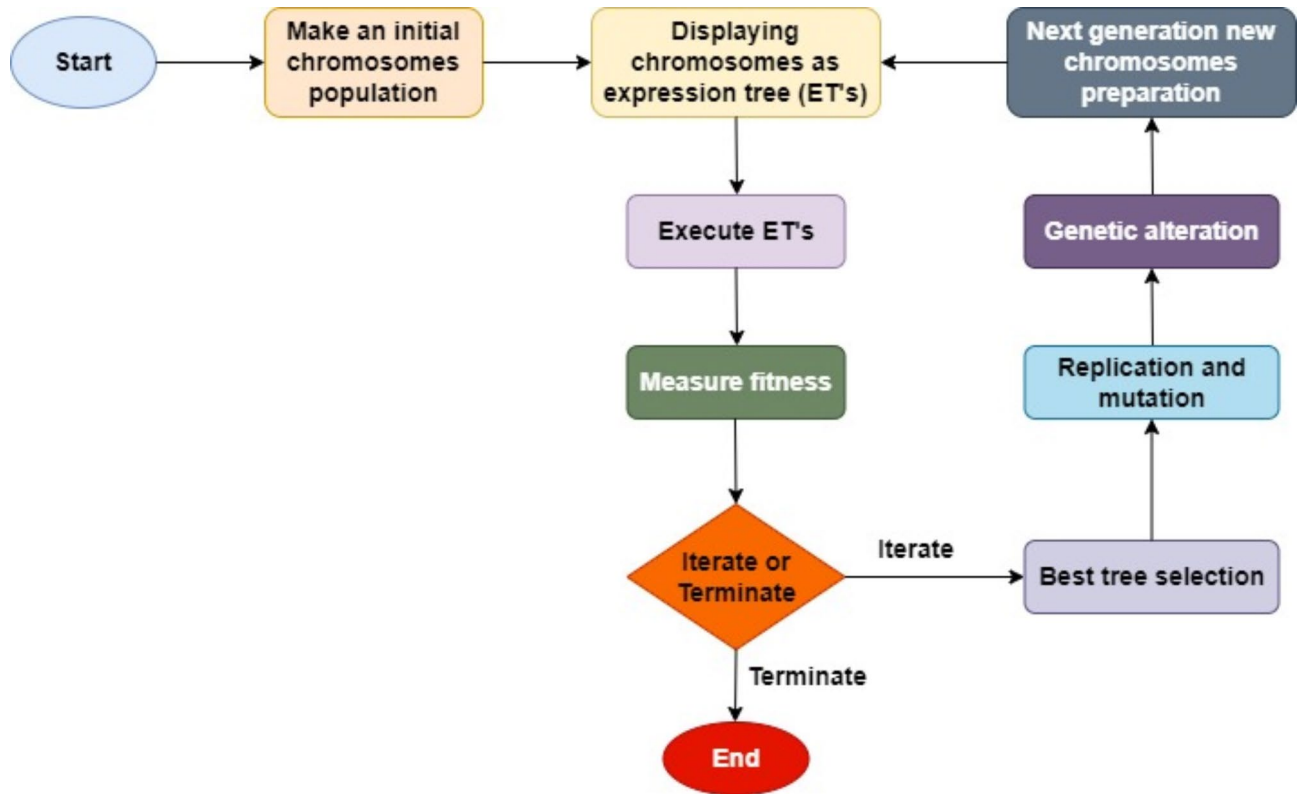
#### Reduced error pruning tree (REPt)

The REPt model employs the idea of randomly chosen characteristics determined by computer technology to speed up classification tree logic techniques and reduce variance inaccuracy<sup>69</sup>. The REPt employs the logistic regression technique and creates many trees through various computation processes; the most straightforward tree was selected from all the created trees. When the conclusion is significant, and the complexity of the tree's internal structure is minimized, the REPt has provided a flexible and straightforward modeling technique by monitoring training data sets. The pruning algorithm considers the backward over-fitting complexity of this technique. It uses the post-pruning algorithm to push for the most miniature possible representation of the best precision tree logic<sup>70</sup>. It only chooses values once for numeric characteristics<sup>71</sup>.

REPt is a simple, intuitive, fast, and easy method for improving model generalization and reducing complexity and tree size. It requires a validation set and accuracy measure, and unnecessary nodes are removed for better prediction. However, a greedy algorithm makes the best decision at each step without considering future consequences, leading to suboptimal solutions. It is sensitive to the choice of validation set, affected by data noise or randomness, and biased towards simpler trees.

#### Bagging

Bagging is the technique that enhances the results of weak, soft computing models<sup>72</sup>. The amount of the original database that will be merged depends on the bagging factor<sup>73</sup>. Bootstrap-resampled observed data are



**Fig. 1.** Flowchart for GEP (created using diagrams.net).

used in each model. Algorithms for bagging include three steps: Bootstrapping produces a fresh training set using replacements. The outputs of the classifier are associated with distributed voting. This strategy enhances classification variance and generalization. The fundamental classifier must be unbalanced for this model to work; otherwise, no classification will result. In this study, bagging is used to investigate the reliability of the Reduced Error Pruning Tree in predicting WQL.

Bagging allows weak learners to outperform strong ones, reducing variance and eliminating overfitting. However, it can introduce bias and be computationally expensive, potentially discouraging its use in certain situations.

### Artificial neural network (ANN)

A computational model that replicates how nerve cells in the human brain function is known as an ANN. The multi-layer perceptron (MLP) kind of ANN employed in this study is trained using the backpropagation learning technique. An input layer, a hidden layer, and an output layer comprise the three layers of the MLP. The data are accepted by the input layer, processed by the hidden layer, and then shown by the output layer as the model's outputs. During the learning process, each layer's neurons are connected to the layers below it through a weight. The external world provides input to the neurons in the input layer. The calculations in this layer are nonexistent. The input layer sends data to the hidden layer, which performs calculations and sends the results to the output layer. The system's output is transmitted via neurons in the output layer. The Neural Network is a two-way process. The first process is training the model, which seeks a suitable nonlinear relationship by generating appropriate weights between the various variables. The second process involves processing the sum using a nonlinear transfer function to produce a prediction. Then, an ANN learns by running a backward process to update the weights until the error has been minimized and done in response to the errors between the actual output values and the intended output values.

In the subsequent testing phase, the neural network is fed a different data set. The learned weights-based neural network predictions are tested against the desired output values. It is done to determine if the ANN over- or under-fits a particular quantity of data. The ANNs have performed satisfactorily while dealing with a variety of engineering simulations.

ANNs offer fault tolerance, self-learning, parallel processing, nonlinear modeling, and complete data. They can tolerate long training times, perform quickly, and predict output values for specific input values. They can also detect complex relationships and nonlinear problems. ANNs have drawbacks such as their black-box nature, computational expense, long development time, overfitting, high data requirements, and reliance on numerical input, making them difficult to understand<sup>74</sup>. For ANN models, a low bias and low variance are crucial. Haykin<sup>74</sup> is recommended to readers for further details. One hidden layer of the ANN model is utilized in the current investigation.

### FireFly algorithm (FFA)

Yang<sup>75</sup> introduced the firefly method at Cambridge University. This method is a swarm intelligence optimization model based on firefly movement and was created using the firefly's natural behavior and radiance pattern<sup>76</sup>. Due to its population-based search, it efficiently handles multi-modal functions, allowing candidate solutions to benefit from building blocks from diverse solutions. Fireflies constantly go toward sources of light when they have less of them<sup>76</sup>. Additionally, fireflies can create a specific pattern. The patterns that fireflies follow are special. The rhythmic light (Flash), light rate (Rate of Flash), and the separation between the light signals are three of the most crucial elements in the absorption of fireflies. The brightness varies in attractiveness with absorption and is proportionate to the distance from its source<sup>75</sup>.

$$P(l) = \frac{P_s}{l^2} \quad (3)$$

For an environment with a fixed light gain coefficient, the intensity of light  $P$  fluctuates with  $l$  in the above relationship, where  $P_s$  is the intensity of the light source<sup>75</sup>.

$$P = P_0 e^{-\gamma l} \quad (4)$$

Where  $P_0$  is the initial light intensity. Given that a firefly's attractiveness is inversely correlated with the amount of light it receives from its neighbors, the following definition of  $\beta$  is used:

$$\beta = \beta_0^{-\gamma l^2} \quad (5)$$

The hybridization of the artificial neural network-based Firefly algorithm is depicted in Fig. 2. The appeal of  $l=0$  is that  $\beta_0$ . The charm function  $\beta(l)$  in actual implementation can be any uniform descending function, as in the generic form below<sup>75</sup>.

$$\beta(l) = \beta_0^{-\gamma l^m} \quad (6)$$

However, it also has drawbacks, such as high computational time complexity and slow convergence speed due to its full-attracted model, which focuses on firefly oscillation during movement.

### Materials and methodology

#### Study area

The Khorramabad, Biranshahr, and Alashtar sub-watersheds, situated in the Iranian province of Lorestan between 48°03'0 10''E and 48°59'0 07''E and between 33°11'0 47''N and 34°03'0 27''N with an area of 3,562.1

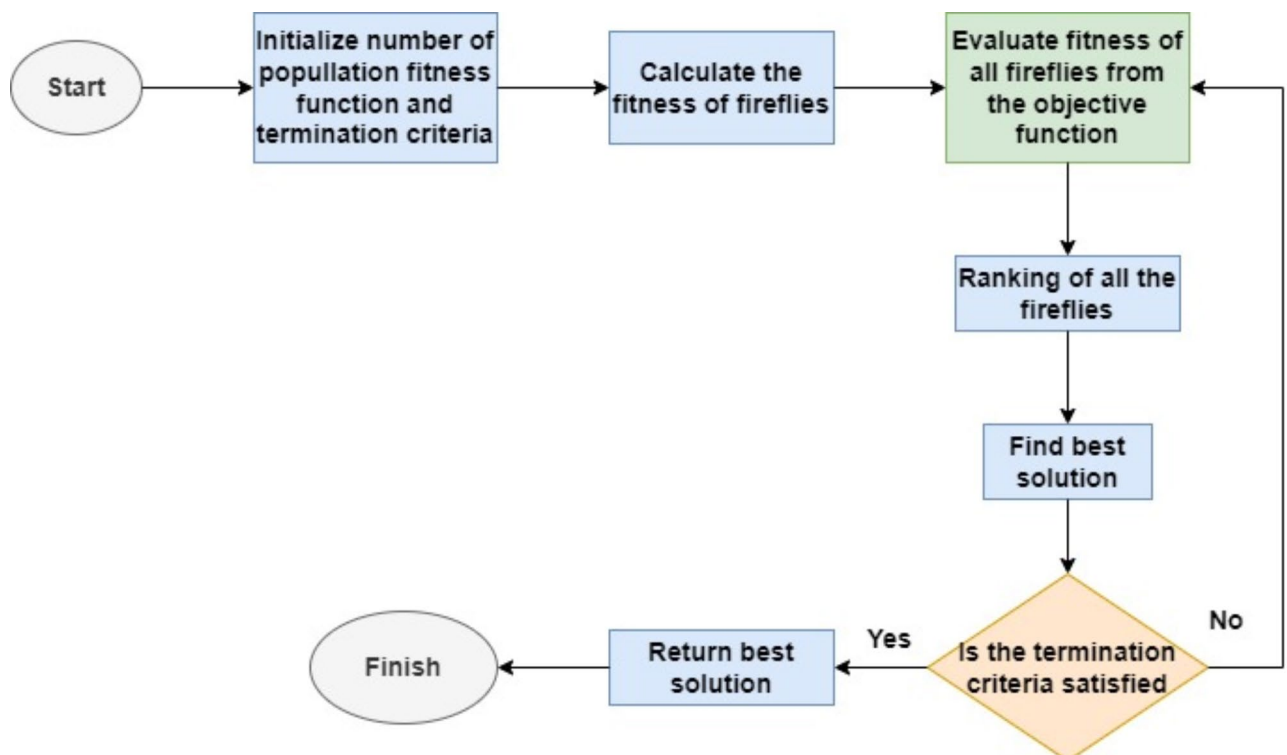


Fig. 2. Flowchart for FFA (created using diagrams.net).

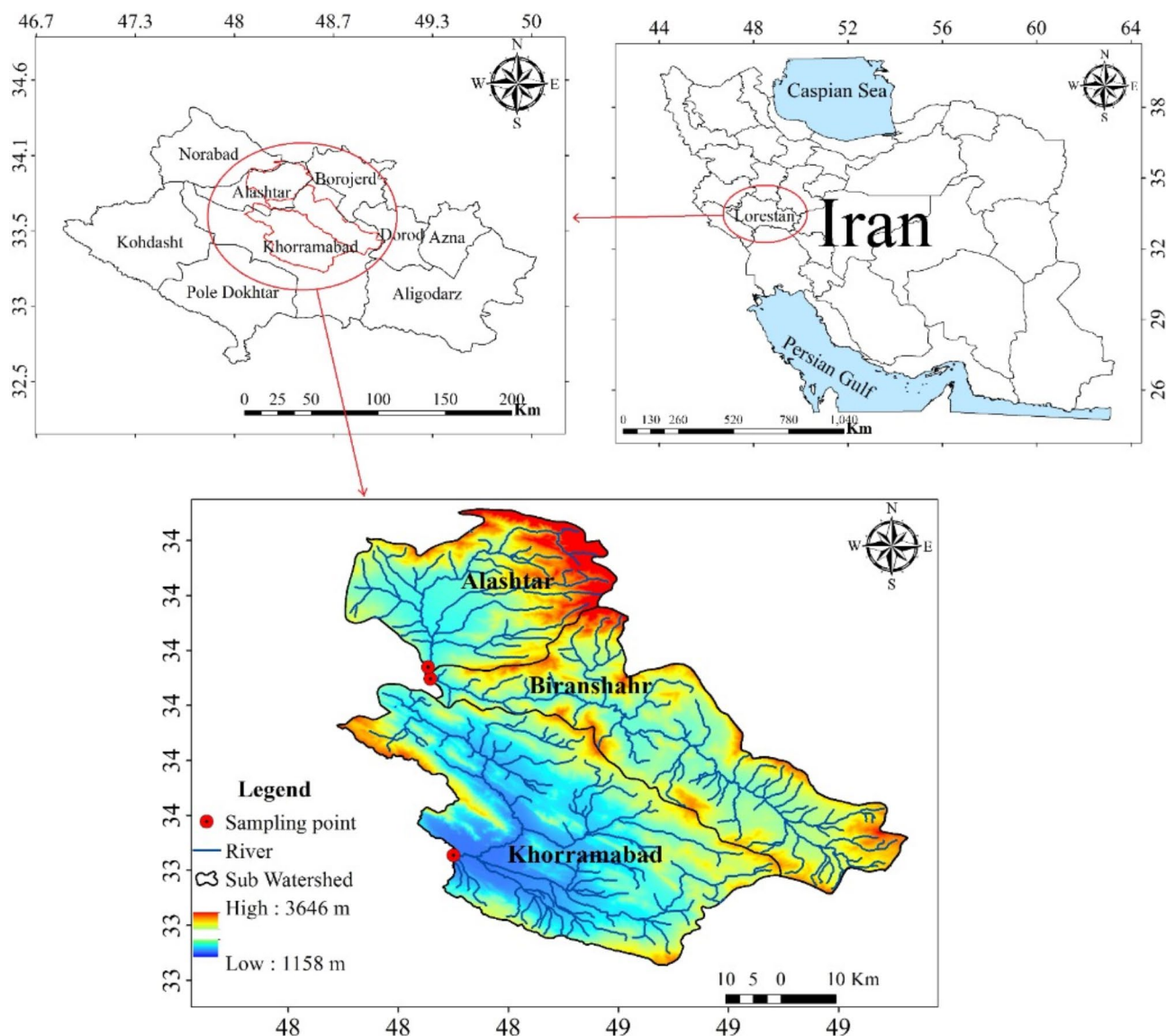


km<sup>2</sup>, were used as the source of flow and water quality data for the study. The catchment area's elevation ranges from 1,158 to 3,646 m MSL. The data measurements were made between September 2014 and August 2017. The average rainfall for the Khorrarnabad, Biranshahr, and Alashtar sub-watersheds is 442 millimeters, 484 millimeters, and 556 millimeters, respectively. The study area is depicted in detail in Fig. 3; the red dots in the figure indicate the location from where the samples were collected.

### WQI and data preparations

The WQI consistently summarizes WQ data for reporting to the public. It provides a straightforward assessment of drinking water quality from a source comparable to the UV or air quality index. The water quality data are compared to “BIS and WHO” to determine the WQI<sup>77</sup>. The WQI calculates a single score by combining three measurements: the scope, frequency, and amplitude of water quality exceedances. This computation yields a score that ranges from 0 to 100. The better the water quality, the lower the score. The results are then categorized into one of the five groups. If the value of WQI is less than 50, the water quality will be ‘Excellent’. If it comes in 50–100, 100–200, and 200–300, then the quality is ‘Good,’ ‘Poor,’ and ‘Very Poor.’ If the value exceeds 300, the water is “Not Suitable for Drinking”.

The dataset used for the study is collected from Singh et al.<sup>54</sup>. It comprises water quality measurements from three sub-watersheds in Iran, collected from Sept. 2014 to Aug. 2017. The dataset includes vital water quality parameters such as sulfate (SO<sub>4</sub>), total dissolved solids (TDS), the potential of Hydrogen (pH), chloride (Cl), electrical conductivity (EC), Potassium (K), bicarbonate (HCO), magnesium (Mg), sodium (Na), and calcium (Ca). The WQI is calculated using the formula given in Singh et al.<sup>54</sup>. Further, the whole dataset is divided



**Fig. 3.** Study area (generated using ArcGIS, v10.8, <https://www.esri.com/en-us/arcgis/products/arcgis-desktop/overview>).

into two subsets: train and test, on the ratio of 70:30. The input variables of the study are  $\text{SO}_4$ , TDS, pH, Cl, EC, K, HCO, Mg, Na, and Ca, while WQI is the output variable. Statistical analysis, including mean, standard deviation, minimum, and maximum values for each parameter, has been performed to understand the dataset characteristics better. This analysis helps understand the data variability and its impact on model predictions. Table 1 gives the values of min., max., ranges, mean, standard error, standard deviation, kurtosis, and skewness of various variables used in this study. Using the kurtosis values, it is observed that EC, pH, HCO, and WQI give the negative values (Platykurtic) and TDS, Cl,  $\text{SO}_4$ , Ca, Na, and k give the positive values (Leptokurtic) for complete data; EC, Mg, and WQI give negative values (Platykurtic) and TDS, Ph, HCO, Cl,  $\text{SO}_4$ , Ca, NA, and K give the positive values (Leptokurtic) for train subset. WQI, pH, and HCO give negative values (Platykurtic), and TDS, EC, Cl,  $\text{SO}_4$ , CA, Mg, Na, and K give negative values (Leptokurtic). The mean, max., min., standard error, and deviation values are approximately the same for full, train, and test data subsets. Finally, the correlation of the various input and output variables is calculated and plotted as a heat map in Fig. 4. The figure suggests that the correlation of WQI with pH is negative, while it is positive for all other variables. The EC and pH have the highest correlation with WQI, while pH gives the lowest correlation.

### Statistical parameters

Statistical parameters are a formal and productive procedure to measure and validate the results of data-driven models. Four statistical parameters are used in this investigation, and these parameters are coefficient of correlation (COC), coefficient of determination (COD), root mean square error (RMSE), Nash–Sutcliffe error (NSE), and mean absolute percentage error (MAPE). The range of the COC and COD lies between  $-1$  and  $1$ ; the range of RMSE lies between  $0$  to  $\infty$ , and the output of the MAPE and RRSE is in percentage<sup>78–80</sup>.

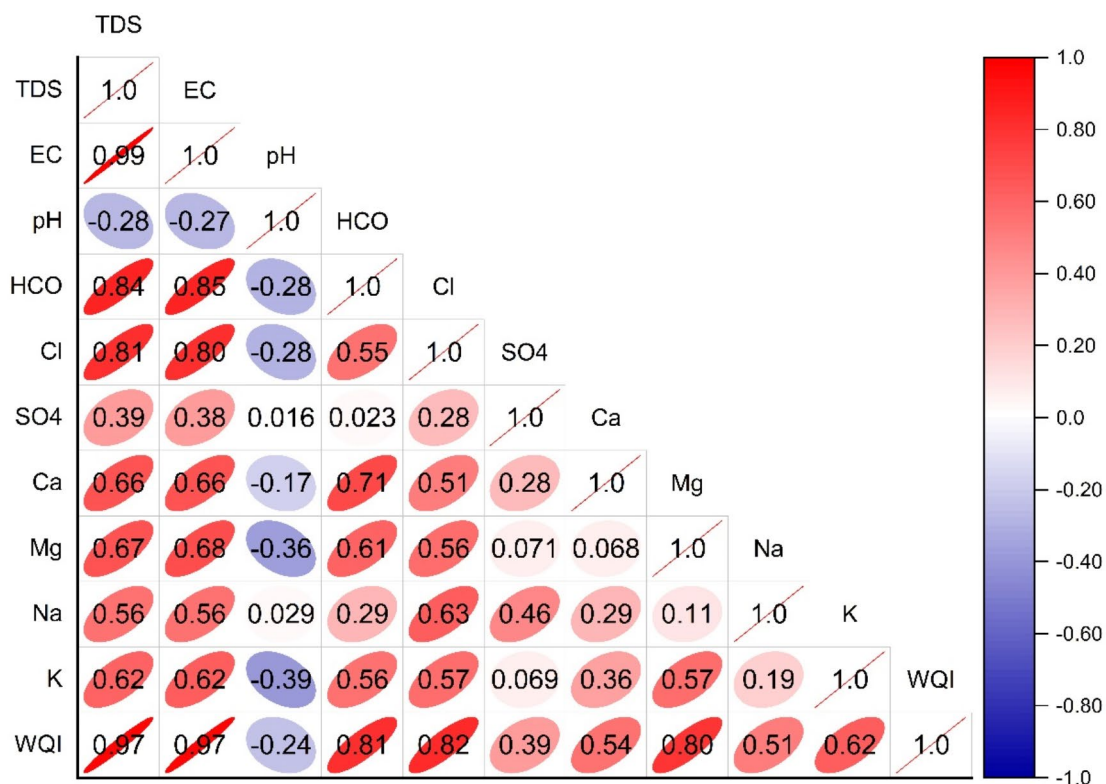
### Proposed work flow

The workflow of the proposed data-driven models, described in Fig. 5, is as follows:

- Dataset: The dataset used in the study is collected from Singh et al.<sup>54</sup>. It contains ten physio-chemical WQ parameters, including  $\text{SO}_4$ , TDS, pH, Cl, EC, K, HCO, Mg, Na, and Ca.
- Data Splitting: The whole dataset is divided into two subsets, train and test, with a ratio of 70:30. In the training subset, there are 86 observations, while in the test subset, 38 observations are there.
- Data-Driven Models: Five soft computing models ANN-FFA, BREPt, REPt, GP, and GEP) are used to predict the WQI of three watersheds in Iran.
- Statistical Parameters: The potential of the soft computing models is assessed using statistical parameters. Five statistics, such as COD, COC, RMSE, NSE, and MAPE, are used.

	TDS	EC	pH	HCO	Cl	$\text{SO}_4$	Ca	Mg	Na	K	WQI
Full Data											
Min.	150.00	241.00	6.47	56.44	3.55	0.00	24.05	2.43	0.69	0.00	23.65
Max.	573.00	882.00	8.39	180.01	92.17	74.45	110.22	48.62	45.98	11.73	56.64
Mean	316.17	492.00	7.70	115.68	26.79	17.49	59.84	18.58	8.73	2.60	35.61
Standard Error	8.03	12.34	0.04	2.37	1.72	1.33	1.40	0.80	0.82	0.21	0.65
Standard Deviation	89.38	137.43	0.50	26.44	19.18	14.80	15.54	8.94	9.17	2.37	7.23
Kurtosis	0.07	-0.04	-0.03	-0.09	0.78	3.83	0.35	0.51	4.42	1.17	-0.09
Skewness	0.69	0.62	-0.99	0.38	1.25	1.83	-0.23	0.85	2.19	1.23	0.76
Training											
Min.	160.00	241.00	6.47	56.44	3.55	0.00	24.05	3.65	0.69	0.00	24.82
Max.	573.00	882.00	8.39	180.01	92.17	74.45	110.22	38.90	45.98	8.60	56.64
Mean	317.63	492.49	7.71	114.27	27.23	17.45	59.01	18.41	9.40	2.57	35.61
Standard Error	10.08	15.59	0.06	2.83	2.19	1.62	1.75	0.94	1.08	0.24	0.81
Standard Deviation	93.52	144.60	0.51	26.22	20.35	15.07	16.25	8.74	10.00	2.26	7.49
Kurtosis	0.01	-0.11	0.21	0.02	0.62	5.22	0.37	-0.02	3.37	0.33	-0.03
Skewness	0.73	0.68	-1.11	0.39	1.23	2.18	0.01	0.73	2.00	1.10	0.84
Testing											
Minimum	150.00	249.00	6.80	64.07	3.55	0.48	26.05	2.43	0.69	0.00	23.65
Maximum	498.00	767.00	8.36	180.01	70.90	55.71	82.16	48.62	31.96	11.73	49.57
Mean	312.87	490.89	7.68	118.88	25.79	17.58	61.72	18.96	7.22	2.65	35.63
Standard Error	13.03	19.71	0.07	4.38	2.67	2.33	2.24	1.54	1.11	0.43	1.09
Standard Deviation	80.33	121.48	0.46	26.99	16.44	14.36	13.80	9.48	6.81	2.64	6.69
Kurtosis	0.13	0.06	-0.83	-0.19	0.89	0.36	0.80	1.58	8.11	2.31	-0.30
Skewness	0.50	0.37	-0.65	0.35	1.25	0.94	-0.99	1.07	2.69	1.41	0.51

**Table 1.** Statistical descriptions of variables.



**Fig. 4.** Heatmap of input and output variables (generated using Origin Pro, v2024b).

## Results

River water quality affects groundwater quality due to the direct percolation of water<sup>81</sup>. Additionally, using river water for irrigation might impact groundwater resources. Hence, areas where aquifers are protected should be used for river water irrigation<sup>82</sup>. Iran experiences an average annual precipitation of 730 mm due to its humid environment. Its rivers' daily streamflow varies and occasionally experiences spikes of discharge. Seasonal variations are seen in the measured physio-chemical parameters, with rainy seasons showing the most incredible values. A total of five soft computing models (ANN-FFA, BREPt, REPt, GP, and GEP) are used in this study to predict the WQI of three watersheds (Khorramabad, Biranshahr, and Alashtar) in Lorestan province, Iran. The performance of the soft computing models strictly depends upon the user-defined parameters (UDFs). The UDFs are calculated using trial and error methods. Several sets of UDFs are used; on these sets, the performance of the soft computing models is checked. The set that gives the best results of soft computing models is chosen. These chosen values of UDFs for different soft computing models used to predict WQI in this study are summarized in Table 2.

Statistical parameters are one measure that checks the performance of the data-driven models. This study uses four statistical parameters (Eqs. 7–10). The outcomes of the statistical parameters for various soft computing models are tabulated in Table 3 for the train and test subsets. According to Table 3, ANN-FFA is the model that has the edge over other models in the prediction of the WQI in training and test and test subsets. It gives the most efficient values of statistical parameters i.e. COC = 0.9990 & 0.9989; COD = 0.9612 & 0.9980; RMSE = 0.3036 & 0.3340; NSE = 0.9980 & 0.9979; and MAPE = 0.7259% & 0.7969% for train and test subsets, respectively. Regarding preciseness, the ANN-FFA is followed by GEP, GP\_puk, BREPt, REPt, and GP\_rbf. In between the kernel function, the puk kernel with GP gives better results than the rbf kernel with GP. The GP-rbf kernel gives the worst result in the prediction of WQI. Thus, the ANN-FFA data-driven model gives the best result of WQI and is supreme among all data-driven models.

Scattered plots, variation plots, box plots, and Taylor diagrams are also used to check the potential of the soft computing models. The scattered plots of WQI with all soft computing models are plotted in Fig. 6 (a to f). It is plotted between the actual values of WQI ( $WQI_{Actual}$ ) vs. predicted values of WQI ( $WQI_{Predicted}$ ), and the diagonal line is the best fit. The model in which all the points lie on best-fit lines is the best. Figure 6 (a) shows



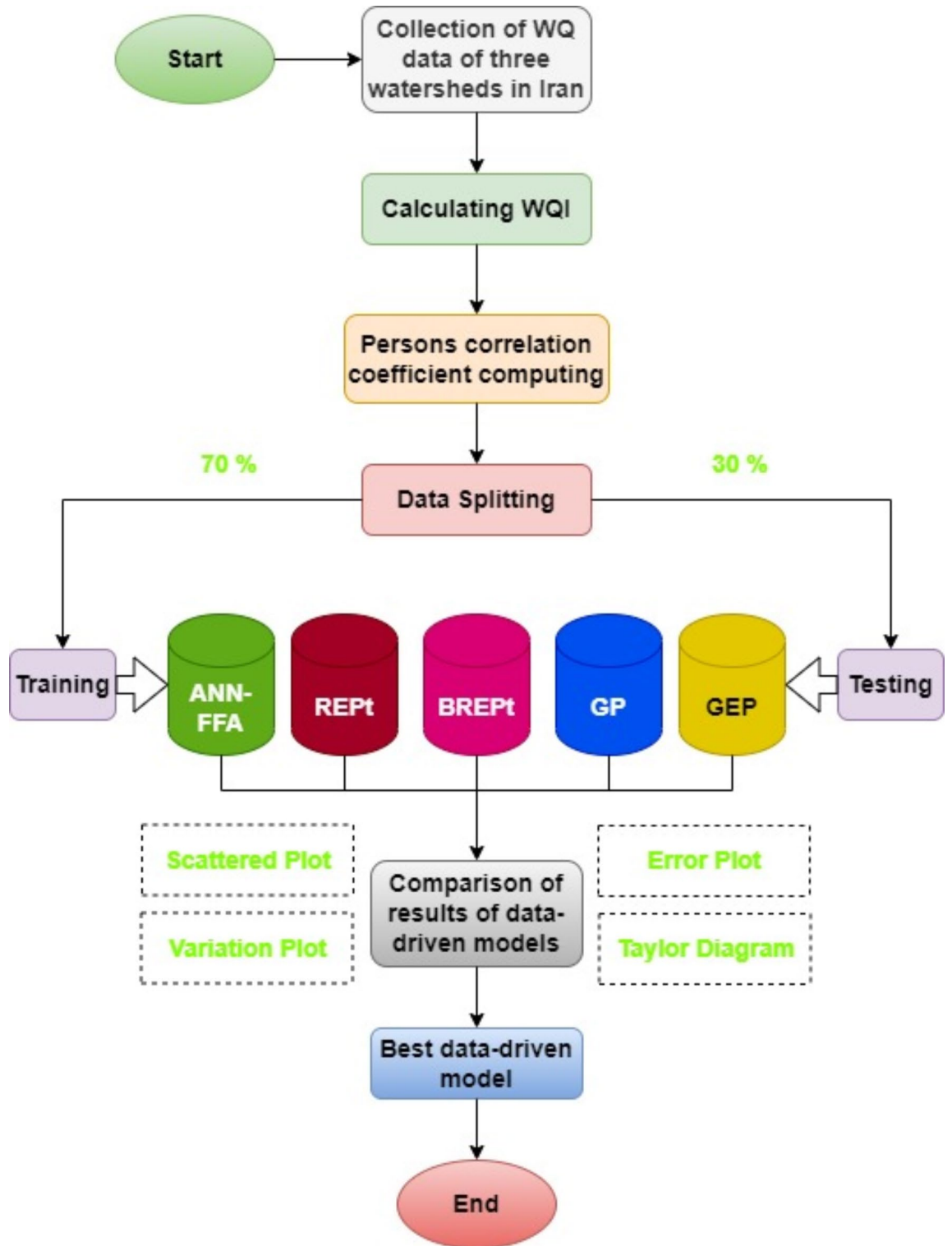


Fig. 5. Proposed workflow of the study (created using diagrams.net).

that all the points of  $WQI_{Actual}$  vs.  $WQI_{Predicted}$  lie on the best-fit line compared to other data-driven models. Figure 6 (f) shows the points of GEP, which has the second position; Fig. 6 (d) shows points of GP\_puk, which got the third position; Fig. 6 (c) shows points of BREPt, which got the fourth position; Fig. 6 (b) shows points of GP\_puk which got the fifth position and Fig. 6 (e) shows points of GP\_rbf which got the last position in the prediction of WQI as per the scattered diagram. The scattered diagram (Fig. 6) suggests the same trend, which is suggested in Table 3. Hence, the ANN-FFA model has the highest accuracy in predicting WQI.

Data-driven Models		UDFs
GP	puk kernel	Gaussian noise (0.01), $\sigma(1)$ , $\omega(0.1)$
	rbf kernel	Gaussian noise $\gamma(0.01)$ , (1)
GEP		Mutation rate (0.044), inversion rate (0.1), incessant rate (0.1), root scale transport rate (0.1), one-point crossover rate (0.13), two-point crossover rate (0.3), gene recombination rate (0.1), transportation rate (0.1), number of chromosomes (30), head size (3), and no. of gene per chromosomes (3).
REPt		Maximum tree depth: -1; minimum total instance weight in the leaf: 2; minimum likelihood of variance: 0.001
BREPt		Batch size-80, bag Size percent = 100, Classifier = REPTree, numbers of executions slots = 1, number of iterations = 100
ANN-FFA		Iteration (1000), population (150), $\alpha$ (0.05), $\beta$ (0.5), $\gamma(0.8)$ and neuron (12)

**Table 2.** UDFs of soft computing models.

Statistical Parameters	ANN-FFA	REPt	BREPt	GP_puk	GP_rbf	GEP
Training						
COC	0.9990	0.9804	0.9821	0.9911	0.9786	0.9935
COD	0.9980	0.9612	0.9645	0.9823	0.9577	0.9871
RMSE	0.3036	1.4668	1.4230	1.5925	5.2067	0.8899
NSE	0.9980	0.9612	0.9635	0.9543	0.5110	0.9857
MAPE (%)	0.7259	2.5430	2.5919	2.6894	11.6093	2.0344
Testing						
COC	0.9989	0.9659	0.9695	0.9710	0.9619	0.9842
COD	0.9980	0.9330	0.9399	0.9428	0.9252	0.9686
RMSE	0.3340	1.7215	1.6266	1.9545	4.6343	1.4299
NSE	0.9979	0.9320	0.9393	0.9124	0.5073	0.9531
MAPE (%)	0.7969	4.3074	3.5236	3.5735	12.5262	2.9659

**Table 3.** Results of statistical parameters for various data-driven models.

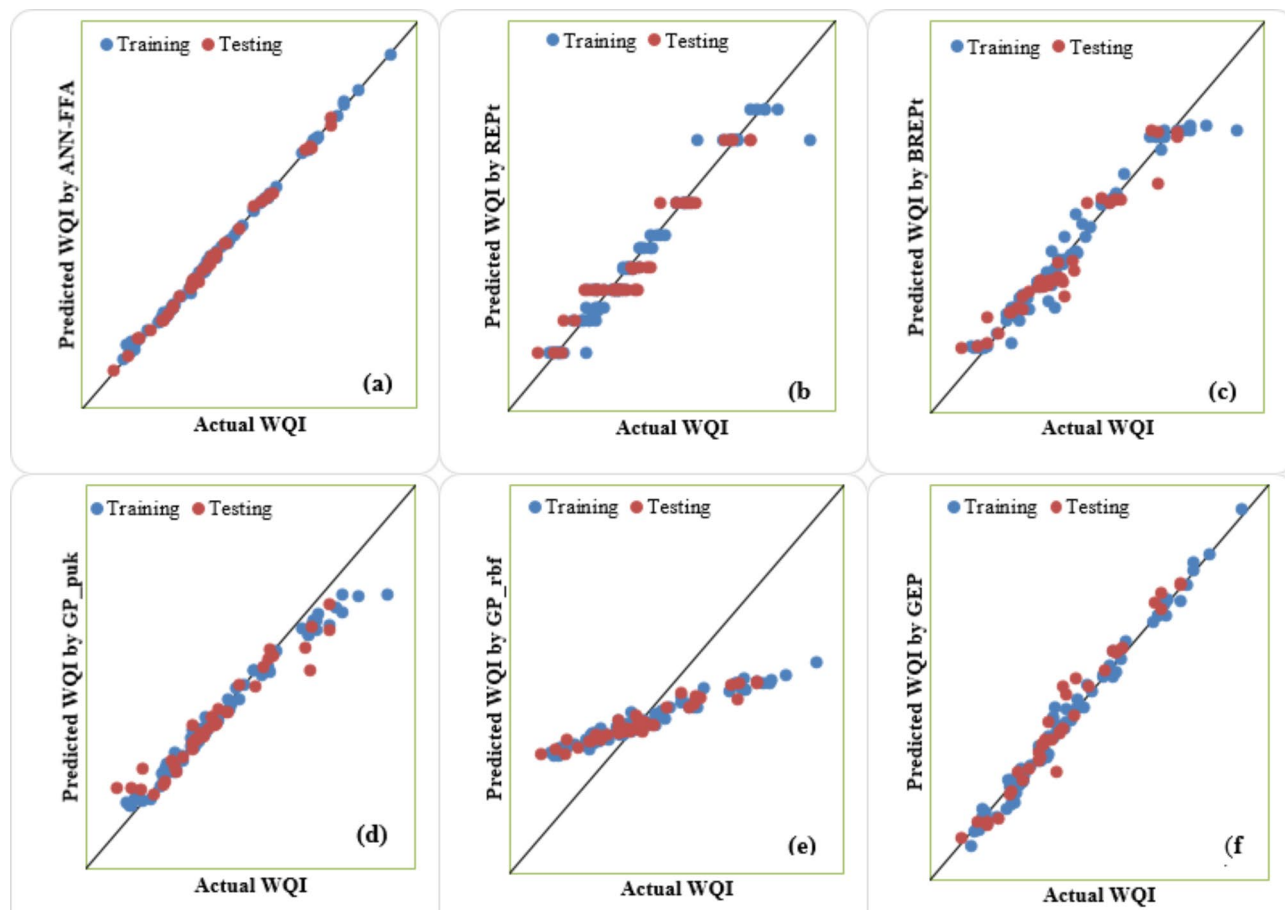
The variation plots of various soft computing models are depicted in Fig. 7 (a to f). As the name indicates, the variations plot shows the visual interpretation of the variation among the  $WQI_{Actual}$  and  $WQI_{Predicted}$ . Figure 7 (a) shows the variation plot of ANN-FFA; Fig. 7 (b) shows the variation plot of REPt; Fig. 7 (c) shows the variation plot of BREPt; Fig. 7 (d) shows the variation plot of GP\_puk; Fig. 7 (e) shows the variation plot of GP\_rbf; Fig. 7 (f) shows the variation plot of GEP. Figure 7 suggests a minimum difference between  $WQI_{Actual}$  and  $WQI_{Predicted}$  for ANN-FFA data-driven models, followed by GEP, GP\_puk, BREPt, REPt, and GP\_rbf. Hence, the variations plots (Fig. 7a to f) also suggest that the ANN-FFA is the model that can predict the WQI accurately and precisely.

The distributions of relative errors (%) in the form of an open box plot for all models are plotted in Fig. 8 to illustrate the efficacy of the soft computing models. This figure shows that the ANN-FFA model had the slightest errors compared to the other soft computing models for the training subset. Also, the ANN-FFA models performed well through the test subset. ANN-FFA points are not distributed and present near zero, while the points of other models are distributed from +10 to -15. Hence, Fig. 8 concludes that the ANN-FFA model is the best data-driven model for predicting WQI.

The Taylor diagrams, a graphical method for assessing the performance of a data-driven model, are displayed in Fig. 9. This figure shows that for the testing subset, the red solid circle point from the ANN-FFA model is closer to the actual (black hollow) point than those from the other models based on the distance between the points acquired by the soft computing models and the actual point. The Taylor diagram (Fig. 9) also concluded that ANN-FFA is the best-performing model, followed by the GEP model for predicting WQI. The performance of the GP\_rbf (solid orange circle) model is the lowest among all applied models for predicting WQI.

### Comparison of obtained results with previous literature

The result of the best model, i.e., ANN-FFA, is compared with the previously published literature. The previously published literature selected for this study are Hu et al.<sup>83</sup>, Hussein et al.<sup>84</sup>, Mohseni et al.<sup>85</sup>, and Kim et al.<sup>86</sup>. Table 4 shows the comparison of these published models with the best-selected model of the study, ANN-FFA. The comparison is based on four statistical parameters: COD, RMSE, NSE, and MAPE. The results suggest that the obtained model (ANN-FFA) is superior to the models published in the literature based on statistical parameters. Thus, the ANN-FFA model outperforms the comparative soft computing models and has superior results to the model published in the literature. It implies that the ANN-FFA model is a robust and reliable tool for predicting WQI, with potential applications in various fields such as environmental science, water resource management, and public health.

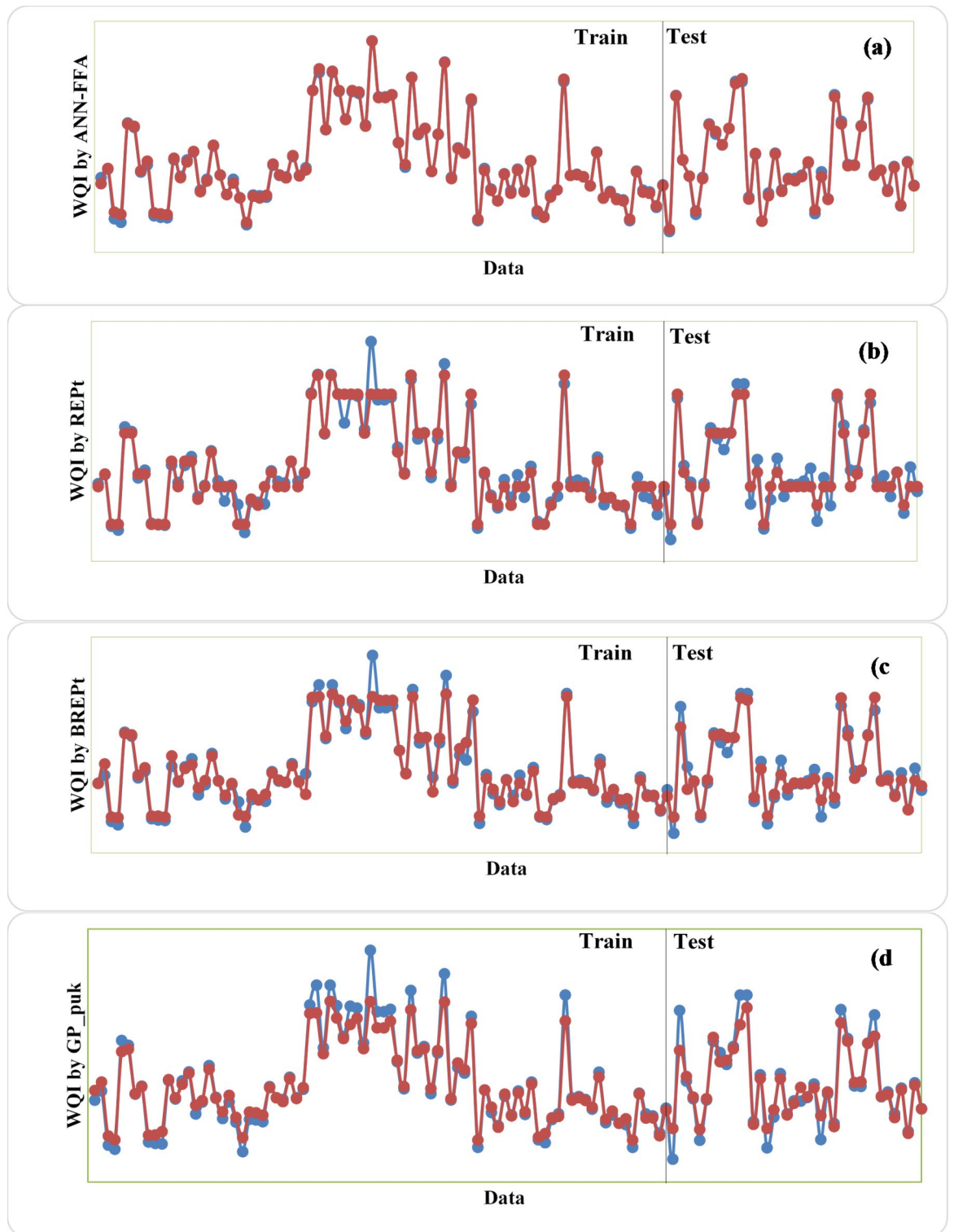


**Fig. 6.** (a–f) Scattered plots of various soft computing models in the prediction of WQI (generated using MS Office, v2019).

## Discussions

This research aimed to examine the efficacy of several soft computing methods in predicting WQI for three sub-watersheds located in Iran. The investigated models included Gene Expression Programming (GEP), Gaussian Process (GP), Reduced Error Pruning Tree (REPt), Bagging REPt, and Artificial Neural Network optimized with the FireFly Algorithm (ANN-FFA). The ANN-FFA model exhibited exceptional performance due to its strong correlation coefficients and fewer errors. The accuracy of ANN-FFA may be ascribed to its robust and efficient optimization capabilities. The GEP model demonstrated commendable performance. Nevertheless, it attained a different level of accuracy than shown by ANN-FFA. The approach is evolutionary, progressively enhancing solutions over generations, efficiently capturing the data's intrinsic relationships. Nonetheless, the system's performance may be influenced by the complexity of the problem and the choice of parameters. The GP models, using radial basis function (RBF) and Pearson VII kernel (PUK) functions, demonstrated robust albeit somewhat worse performance than ANN-FFA and GEP. The efficacy of water quality data may have been constrained by the challenges posed by its noisy and complex nature. The decreased prediction accuracy may also be ascribed to the selection of kernel functions since the RBF and PUK kernels may not have been the most appropriate for this dataset. The REPt and Bagging REPt models had the lowest accuracy relative to the other models analyzed. However, acknowledging that these models may exhibit constrained simplicity and interpretability when used for highly complex and nonlinear data, such as water quality measurements, is essential.

This research's findings have substantial implications for water resource management, particularly in regions like Lorestan province, where accurate and timely measurement of the WQI is crucial for sustainable water management. The ANN-FFA algorithm's ability to provide precise and accurate WQI forecasts with few mistakes makes it an excellent option for integration into decision support systems used by water resource managers. The study shows that soft computing models, especially those that are enhanced with optimization algorithms like FireFly, can be used in addition to or instead of traditional laboratory methods for testing WQI. The shift to model-based predictions offers a cost-efficient, time-saving, and scalable method for water quality monitoring, which is particularly beneficial in resource-constrained scenarios.



**Fig. 7.** (a–f) Variation plots of various soft computing models in the prediction of WQI (generated using MS Office, v2019).

## Conclusion

This research introduces advanced soft computing models for predicting the WQI of three sub-watersheds in Lorestan Province, Iran. This work's primary contribution is the invention and implementation of hybrid models, namely the Artificial Neural Network optimized by the FireFly Algorithm (ANN-FFA), which has not been used before in this research domain. The paper illustrates the higher prediction accuracy of the hybrid

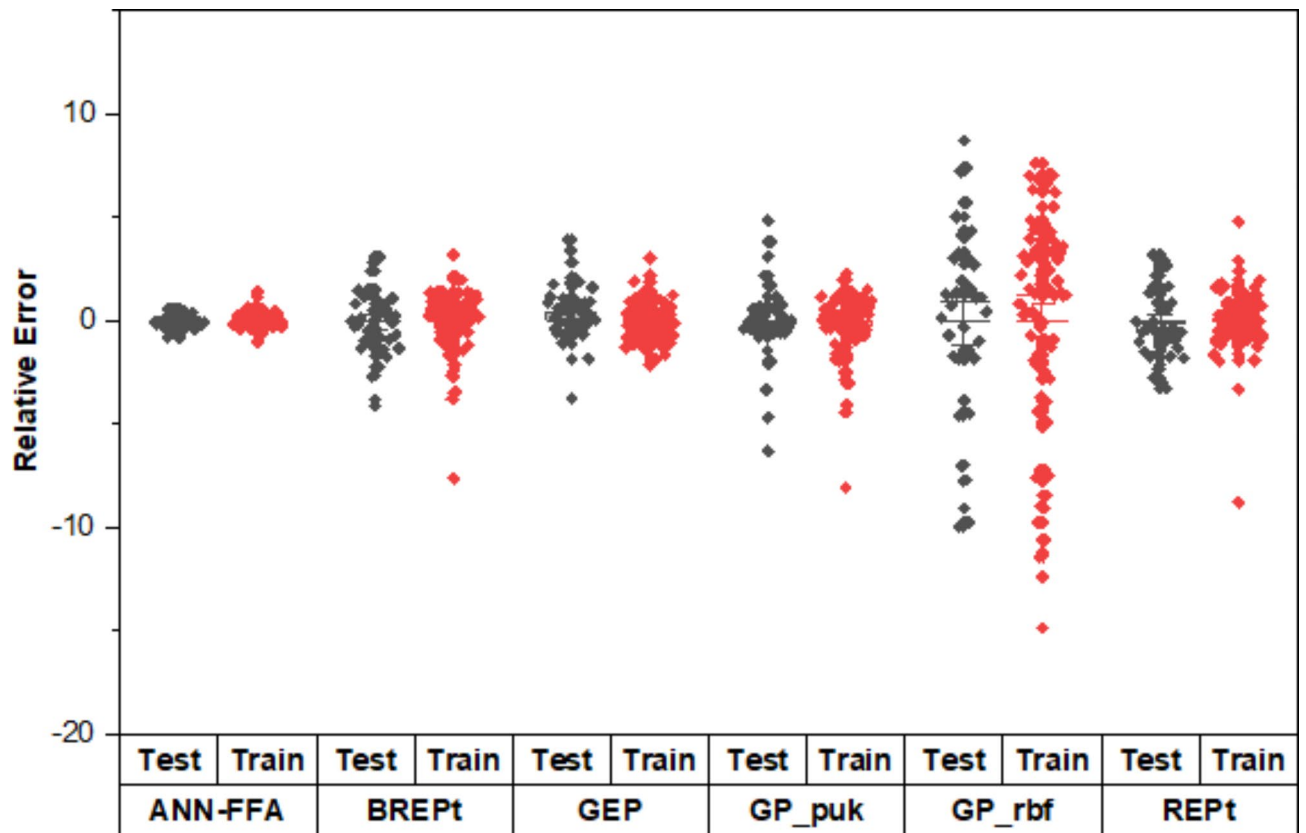


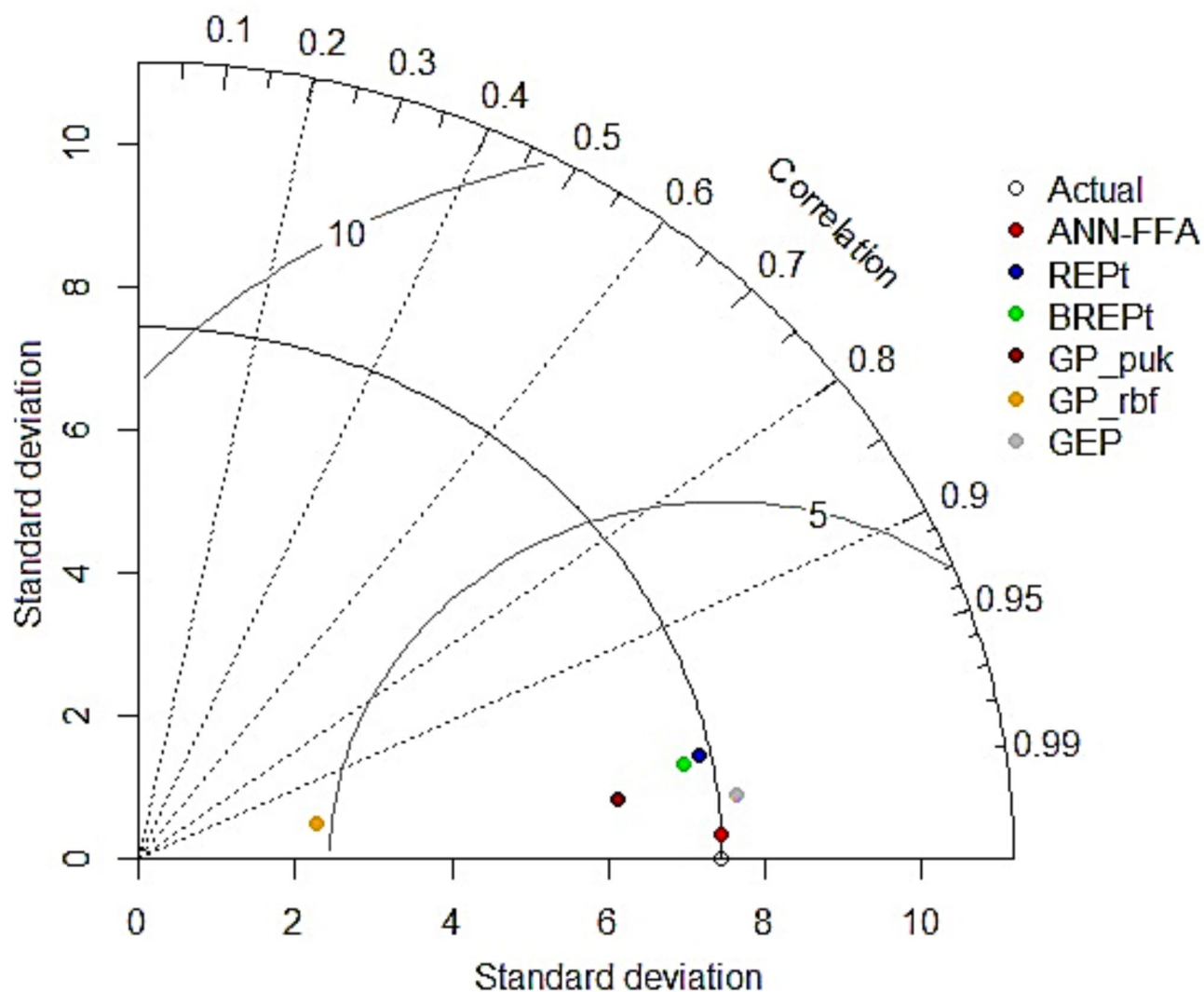
Fig. 8. Relative error for various data-driven models (generated using Origin Pro, v2024b).

technique by comparing the performance of ANN-FFA with other models, such as the Gaussian Process (GP), Gene Expression Programming (GEP), and REP Tree (REPt).

The main results indicate that the ANN-FFA model surpassed all other models, achieving a correlation coefficient (COC) of 0.9989, a coefficient of determination (COD) of 0.9980, a root mean square error (RMSE) of 0.3340, a Nash–Sutcliffe error of 0.9979, and a mean absolute percentage error (MAPE) of 0.7969%. The research indicates that the Gaussian Process model using the Puk kernel function outperforms the model employing the RBF kernel function for WQI prediction. These results are crucial for the literature on water quality modeling since they provide a new standard for using hybrid models in environmental monitoring. The ANN-FFA model, by providing a cost-efficient, real-time prediction system, can significantly contribute to water resource management and environmental conservation. This method offers a dependable resource for policymakers and environmental managers to make educated choices, particularly in areas with inadequate water quality monitoring equipment.

Notwithstanding the encouraging outcomes, the research had several limitations. The dataset was confined to three sub-watersheds, perhaps failing to capture the heterogeneity in water quality over larger areas or diverse climatic zones. The models depend significantly on historical data, and their accuracy may diminish in regions where such data is less abundant or inconsistent. Subsequent research needs to broaden the investigation to include a more comprehensive array of datasets from other places and environmental situations. The possibility of integrating machine learning with conventional physics-based models to enhance forecast accuracy warrants investigation.





**Fig. 9.** Taylor diagram for various data-driven models (generated using R, v4.4.1).

Literature Studies	Model	Statistical Parameters			
		COD	RMSE	NSE	MAPE (%)
This study	ANN-FFA	0.9980	0.3340	0.9979	0.7969
Hu et al. <sup>83</sup>	IMSSA-SVR	0.9880	0.9630	-	5.0620
Hussein et al. <sup>84</sup>	SVR	-	2.6925	0.9611	-
Mohseni et al. <sup>85</sup>	XGBOOST	0.9900	3.2730	-	-
Kim et al. <sup>86</sup>	LSTM	0.9100	0.1100	-	-

**Table 4.** Comparison of this study’s results with previously published results based on statistical parameters.

### Data availability

The data that support the findings of this study are available from the authors upon reasonable request. For further inquiries, please contact Parveen Sihag at parveen12sihag@gmail.com.

Received: 11 June 2024; Accepted: 17 October 2024

Published online: 29 October 2024

## References

- Pandhiani, S. M., Sihag, P., Shabri, A. B., Singh, B. & Pham, Q. B. Time-series prediction of streamflows of Malaysian rivers using data-driven techniques. *J. Irrig. Drain. Eng.* **146** (7), 04020013 (2020).
- Gupta, A. D. Implication of environmental flows in river basin management. *Phys. Chem. Earth Parts A/B/C.* **33** (5), 298–303 (2008).
- Grabowski, R. C. & Gurnell, A. M. Hydrogeomorphology—Ecology interactions in river systems. *River Res. Appl.* **32** (2), 139–141 (2016).
- Singh, A. P., Dhadse, K. & Ahalawat, J. Managing water quality of a river using an integrated geographically weighted regression technique with fuzzy decision-making model. *Environ. Monit. Assess.* **191** (6), 1–17 (2019).
- Cordier, C. et al. Culture of microalgae with ultrafiltered seawater: A feasibility study. *SciMedicine J.* **2** (2), 56–62 (2020).
- Bhatti, N. B., Siyal, A. A., Qureshi, A. L. & Bhatti, I. A. Socio-economic impact assessment of small dams based on t-paired sample test using SPSS software. *Civil Eng. J.* **5** (1), 153–164 (2019).
- Sihag, P., Jain, P. & Kumar, M. Modelling of impact of water quality on recharging rate of storm water filter system using various kernel function based regression. *Model. Earth Syst. Environ.* **4** (1), 61–68 (2018).
- Shahzad, G., Rehan, R. & Fahim, M. Rapid performance evaluation of water supply services for strategic planning. *Civil Eng. J.* **5** (5), 1197–1204 (2019).
- Cheng, H., Hu, Y. & Zhao, J. Meeting China's water shortage crisis: Current practices and challenges. *Environ. Sci. Technol.* **43** (2), 240–244 (2009).
- Vörösmarty, C. J. et al. Global threats to human water security and river biodiversity. *Nature.* **467** (7315), 555–561 (2010).
- Solangi, G. S., Siyal, A. A. & Siyal, P. Analysis of Indus Delta groundwater and surface water suitability for domestic and irrigation purposes. *Civil Eng. J.* **5** (7), 1599–1608 (2019).
- Katyal, D. Water quality indices used for surface water vulnerability assessment. *Int. J. Environ. Sci.*, **2**(1). (2011).
- Mohebbi, M. R. et al. Assessment of water quality in groundwater resources of Iran using a modified drinking water quality index (DWQI). *Ecol. Ind.* **30**, 28–34 (2013).
- Guo, M., Noori, R. & Abolfathi, S. Microplastics in freshwater systems: dynamic behaviour and transport processes. *Resour. Conserv. Recycl.* **205**, 107578 (2024).
- Cook, S., Abolfathi, S. & Gilbert, N. I. Goals and approaches in the use of citizen science for exploring plastic pollution in freshwater ecosystems: A review. *Freshw. Sci.* **40** (4), 567–579 (2021).
- Stride, B. et al. Microplastic transport dynamics in surcharging and overflowing manholes. *Sci. Total Environ.* **899**, 165683 (2023).
- Mahdian, M. et al. Anzali Wetland crisis: unraveling the decline of Iran's ecological gem. *J. Geophys. Research: Atmos.* **129** (4), e2023JD039538 (2024).
- Tian, H. et al. Biodegradation of microplastics derived from controlled release fertilizer coating: selective microbial colonization and metabolism in plastisphere. *Sci. Total Environ.* **920**, 170978 (2024).
- Lumb, A., Sharma, T. C. & Bibault, J. F. A review of genesis and evolution of water quality index (WQI) and some future directions. *Water Qual. Exposure Health.* **3** (1), 11–24 (2011).
- Debels, P., Figueroa, R., Urrutia, R., Barra, R. & Niell, X. Evaluation of water quality in the Chillán River (Central Chile) using physicochemical parameters and a modified water quality index. *Environ. Monit. Assess.* **110** (1), 301–322 (2005).
- Bordalo, A. A., Teixeira, R. & Wiebe, W. J. A water quality index applied to an international shared river basin: the case of the Douro River. *Environ. Manage.* **38** (6), 910–920 (2006).
- Horton, R. K. An index number system for rating water quality. *J. Water Pollut Control Fed.* **37** (3), 300–306 (1965).
- Brown, R. M., McClelland, N. L., Deininger, R. A. & Tozer, R. G. A water quality index—do we dare. *Water Sew. Works*, **117**(10). (1970).
- Pesce, S. F. & Wunderlin, D. A. Use of water quality indices to verify the impact of Córdoba City (Argentina) on Suquía River. *Water Res.* **34** (11), 2915–2926 (2000).
- Cude, C. G. Oregon water quality index a tool for evaluating water quality management effectiveness 1. *JAWRA J. Am. Water Resour. Association.* **37** (1), 125–137 (2001).
- Kargar, K. et al. Estimating longitudinal dispersion coefficient in natural streams using empirical models and machine learning algorithms. *Eng. Appl. Comput. Fluid Mech.* **14** (1), 311–322 (2020).
- Alizadeh, M. J. et al. Effect of river flow on the quality of estuarine and coastal waters using machine learning models. *Eng. Appl. Comput. Fluid Mech.* **12** (1), 810–823 (2018).
- Singh, B., Sihag, P. & Deswal, S. Modelling of the impact of water quality on the infiltration rate of the soil. *Appl. Water Sci.* **9** (1), 1–9 (2019).
- Mandal, S., Mahapatra, S. S., Adhikari, S. & Patel, R. K. Modeling of arsenic (III) removal by evolutionary genetic programming and least square support vector machine models. *Environ. Processes.* **2** (1), 145–172 (2015).
- Dehghani, M., Saghafian, B., Nasiri Saleh, F., Farokhnia, A. & Noori, R. Uncertainty analysis of streamflow drought forecast using artificial neural networks and Monte-Carlo simulation. *Int. J. Climatol.* **34** (4), 1169–1180 (2014).
- Yaseen, Z. M. et al. Hybrid adaptive neuro-fuzzy models for water quality index estimation. *Water Resour. Manage.* **32** (7), 2227–2245 (2018).
- Singh, B., Sihag, P. & Singh, K. Modelling of impact of water quality on infiltration rate of soil by random forest regression. *Model. Earth Syst. Environ.* **3**, 999–1004 (2017).
- Najafzadeh, M., Rezaie-Balf, M. & Tafarajnoruz, A. Prediction of riprap stone size under overtopping flow using data-driven models. *Int. J. River Basin Manage.* **16** (4), 505–512 (2018).
- Singh, B., Ebtehaj, I., Sihag, P. & Bonakdari, H. An expert system for predicting the infiltration characteristics. *Water Supply.* **22** (3), 2847–2862 (2022a).
- Tung, T. M. & Yaseen, Z. M. A survey on river water quality modelling using artificial intelligence models: 2000–2020. *J. Hydrol.* **585**, 124670 (2020).
- Tripathi, M. & Singal, S. K. Use of principal component analysis for parameter selection for development of a novel water quality index: A case study of river Ganga India. *Ecol. Ind.* **96**, 430–436 (2019).
- Zali, M. A. et al. Sensitivity analysis for water quality index (WQI) prediction for Kinta River. *Malaysia World Appl. Sci. J.* **14**, 60–65 (2011).
- Nigam, U. & SM, Y. Development of computational assessment model of fuzzy rule based evaluation of groundwater quality index: comparison and analysis with conventional index. In Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM), Amity University Rajasthan, Jaipur-India. (2019), February.
- Srinivas, R. & Singh, A. P. Application of fuzzy multi-criteria model to assess the water quality of river Ganges. In Soft Computing: Theories and Applications (513–522). Springer, Singapore. (2018).
- Xiang, B., Zeng, C., Dong, X. & Wang, J. The application of a decision tree and stochastic forest model in summer precipitation prediction in Chongqing. *Atmosphere.* **11** (5), 508 (2020).
- Granata, F., Papirio, S., Esposito, G., Gargano, R. & De Marinis, G. Machine learning algorithms for the forecasting of wastewater quality indicators. *Water.* **9** (2), 105 (2017).
- Li, J. et al. Hybrid soft computing model for determining water quality indicator: Euphrates River. *Neural Comput. Appl.* **31** (3), 827–837 (2019).

43. Kamyab-Talesh, F., Mousavi, S. F., Khaledian, M., Yousefi-Falakdehi, O. & Norouzi-Masir, M. Prediction of water quality index by support vector machine: A case study in the Sefidrud Basin, Northern Iran. *Water Resour.* **46** (1), 112–116 (2019).
44. Wang, X. et al. Genome-wide and gene-based association mapping for rice eating and cooking characteristics and protein content. *Sci. Rep.* **7** (1), 1–10 (2017).
45. Ghiasi, B. et al. Uncertainty quantification of granular computing-neural network model for prediction of pollutant longitudinal dispersion coefficient in aquatic streams. *Sci. Rep.* **12** (1), 4610 (2022).
46. Singh, B. & Minocha, V. K. Comparative Study of Machine Learning Techniques for Prediction of Scour Depth around Spur Dikes. In *World Environmental and Water Resources Congress 2024* (pp. 635–651).
47. Moosavi, A., Rao, V. & Sandu, A. Machine learning based algorithms for uncertainty quantification in numerical weather prediction models. *J. Comput. Sci.* **50**, 101295 (2021).
48. National Research Council, Division on Engineering, Physical Sciences. Board on Mathematical Sciences, Their Applications, Committee on Mathematical Foundations of Verification, & Uncertainty Quantification. (2012). *Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification*. National Academies.
49. Heo, J. et al. Uncertainty-aware attention for reliable interpretation and prediction. *Advances in neural information processing systems*, **31**. (2018).
50. Wagener, A. D. L. et al. Distribution and source apportionment of hydrocarbons in sediments of oil-producing continental margin: a fuzzy logic model. *Environ. Sci. Pollut. Res.* **26** (17), 17032–17044 (2019).
51. Vand, A. S., Sihag, P., Singh, B. & Zand, M. Comparative evaluation of infiltration models. *KSCE J. Civ. Eng.* **22** (10), 4173–4184 (2018).
52. Sihag, P., Singh, B., Gautam, S. & Debnath, S. Evaluation of the impact of fly ash on infiltration characteristics using different soft computing techniques. *Appl. Water Sci.* **8** (6), 1–10 (2018).
53. Singh, B. Prediction of the sodium absorption ratio using data-driven models: a case study in Iran. *Geol. Ecol. Landscapes.* **4** (1), 1–10 (2020).
54. Singh, B., Sihag, P., Singh, V. P., Sepahvand, A. & Singh, K. Soft computing technique-based prediction of water quality index. *Water Supply.* **21** (8), 4015–4029 (2021).
55. Sepahvand, A., Singh, B., Ghobadi, M. & Sihag, P. Estimation of infiltration rate using data-driven models. *Arab. J. Geosci.* **14** (1), 1–11 (2021a).
56. Sepahvand, A. et al. Assessment of the various soft computing techniques to predict sodium absorption ratio (SAR). *ISH J. Hydraulic Eng.* **27** (sup1), 124–135 (2021b).
57. Singh, B. & Singh, T. Soft Computing-based prediction of compressive strength of high strength concrete. In *Applications of Computational Intelligence in Concrete Technology* (207–218). CRC. (2022).
58. Singh, B., Sihag, P., Parsaie, A. & Angelaki, A. Comparative analysis of artificial intelligence techniques for the prediction of infiltration process. *Geol. Ecol. Landscapes.* **5** (2), 109–118 (2021).
59. Singh, T., Singh, B., Bansal, S. & Saggiu, K. Prediction of Ultrasonic pulse velocity of concrete. In *Applications of Computational Intelligence in Concrete Technology* (235–251). CRC. (2022b).
60. Ghiasi, B., Sheikhan, H., Zeynolabedin, A. & Niksokhan, M. H. Granular computing-neural network model for prediction of longitudinal dispersion coefficients in rivers. *Water Sci. Technol.* **80** (10), 1880–1892 (2019).
61. Noori, R., Ghiasi, B., Sheikhan, H. & Adamowski, J. F. Estimation of the dispersion coefficient in natural rivers using a granular computing model. *J. Hydraul. Eng.* **143** (5), 04017001 (2017).
62. Najafzadeh, M. et al. A comprehensive uncertainty analysis of model-estimated longitudinal and lateral dispersion coefficients in open channels. *J. Hydrol.* **603**, 126850 (2021).
63. Williams, C. K. & Rasmussen, C. E. *Gaussian Processes for Machine Learning* (Vol2p. 4 (MIT Press, 2006). No. 3.
64. Kuss, M. Gaussian process models for robust regression, classification, and reinforcement learning (Doctoral dissertation, echnische Universität Darmstadt Darmstadt, Germany). (2006).
65. Donnelly, J., Daneshkhah, A. & Abolfathi, S. Forecasting global climate drivers using gaussian processes and convolutional autoencoders. *Eng. Appl. Artif. Intell.* **128**, 107536 (2024).
66. Ferreira, C. Gene expression programming in problem solving. In *Soft Computing and Industry* (635–653). Springer, London. (2002).
67. Ferreira, C. *Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence* Vol. 21 (Springer, 2006).
68. Ebtehaj, I., Bonakdari, H., Zaji, A. H., Azimi, H. & Sharifi, A. Gene expression programming to predict the discharge coefficient in rectangular side weirs. *Appl. Soft Comput.* **35**, 618–628 (2015).
69. Quinlan, J. R. Simplifying decision trees. *Int. J. Man. Mach. Stud.* **27** (3), 221–234 (1987).
70. Kalmegh, S. Analysis of weka data mining algorithm reptime, simple cart and randomtree for classification of Indian news. *Int. J. Innovative Sci. Eng. Technol.* **2** (2), 438–446 (2015).
71. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
72. Moretti, F., Pizzuti, S., Panziera, S. & Annunziato, M. Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling. *Neurocomputing.* **167**, 3–7 (2015).
73. Donnelly, J., Daneshkhah, A. & Abolfathi, S. Physics-informed neural networks as surrogate models of hydrodynamic simulators. *Sci. Total Environ.* **912**, 168814 (2024).
74. Haykin, S. *Neural networks. A comprehensive foundation.* (1994).
75. Yang, X. S. *Nature-inspired Metaheuristic Algorithms* (Luniver, 2010).
76. Gazi, V. & Passino, K. M. Stability analysis of social foraging swarms. *IEEE Trans. Syst. Man. Cybernetics Part. B (Cybernetics).* **34** (1), 539–557 (2004).
77. Raheja, H., Goel, A. & Pal, M. Prediction of groundwater quality indices using machine learning algorithms. *Water Pract. Technol.* **17** (1), 336–351 (2022).
78. Azamathulla, H. M., Rathnayake, U. & Shatnawi, A. Gene expression programming and artificial neural network to estimate atmospheric temperature in Tabuk, Saudi Arabia. *Appl. Water Sci.* **8**, 1–7 (2018).
79. Amaratunga, V., Wickramasinghe, L., Perera, A., Jayasinghe, J. & Rathnayake, U. Artificial neural network to estimate the paddy yield prediction using climatic data. *Math. Probl. Eng.* **2020**, 1–11 (2020).
80. Anushka, P., Md, A. H. & Upaka, R. Comparison of different artificial neural network (ANN) training algorithms to predict the atmospheric temperature in Tabuk. *Saudi Arabia Mausam.* **71** (2), 233–244 (2020).
81. Kapetas, L., Kazakis, N., Voudouris, K. & McNicholl, D. Water allocation and governance in multi-stakeholder environments: insight from Axios Delta, Greece. *Sci. Total Environ.* **695**, 133831 (2019).
82. Busico, G. et al. A novel hybrid method of specific vulnerability to anthropogenic pollution using multivariate statistical and regression analyses. *Water Res.* **171**, 115386 (2020).
83. Hu, Y., Lyu, L., Wang, N., Zhou, X. & Fang, M. Application of machine learning model optimized by improved sparrow search algorithm in water quality index time series prediction. *Multimedia Tools Appl.* **83** (6), 16097–16120 (2024).
84. Hussein, E. E., Derdour, A., Zerouali, B., Almaliki, A., Wong, Y. J., Ballesta-de los Santos, M., ... Elbeltagi, A. (2024). Groundwater Quality Assessment and Irrigation Water Quality Index Prediction Using Machine Learning Algorithms. *Water*, **16**(2), 264.
85. Mohseni, U., Pande, C. B., Pal, S. C. & Alshehri, F. Prediction of weighted arithmetic water quality index for urban water quality using ensemble machine learning model. *Chemosphere.* **352**, 141393 (2024).

86. Kim, H. I. et al. Incorporation of Water Quality Index models with Machine Learning-based techniques for Real-Time Assessment of aquatic ecosystems. *Environ. Pollut.*, 124242. (2024).

### Acknowledgements

This study was supported by the Research Program funded by the Seoul National University of Science and Technology (SeoulTech).

### Author contributions

Balraj Singh, Alireza Sepahvand, Parveen Sihag, Karan Singh, and Dongwann Kang wrote the main manuscript text. Balraj Singh and Alireza Sepahvand primarily prepared the figures and tables. All authors, including Chander Prabha, Anindya Nag, Md. Mehedi Hasan, and S. Vimal, contributed to designing the experiments and analyzing the results. All authors reviewed and revised the manuscript.

### Declarations

### Competing interests

The authors declare no competing interests.

### Ethical approval

The manuscript is conducted ethically.

### Consent to publish

The research is scientifically consented to be published.

### Additional information

**Correspondence** and requests for materials should be addressed to D.K.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024