



Water quality index forecast using artificial neural network techniques optimized with different metaheuristic algorithms

Hasanain Zamili¹ · Gulfem Bakan¹ · Salah L. Zubaidi^{2,3} · Mustafa A. Alawsi⁴

Received: 5 December 2022 / Accepted: 20 February 2023 / Published online: 17 March 2023
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2023

Abstract

An accurate water quality index (WQI) forecast is essential for freshwater resources management due to providing early warnings to prevent environmental disasters. This research provides a novel procedure to simulate monthly WQI considering water quality parameters and rainfall. The methodology includes data pre-processing and an artificial neural network (ANN) model integrated with the constraint coefficient-based particle swarm optimization and chaotic gravitational search algorithm (CPSOCGSA). The CPSOCGSA technique was compared with the marine predator's optimization algorithm (MPA) and particle swarm optimization (PSO) to increase the model's reliability. The Yesilirmak River data from 1995 to 2014 was considered to build and inspect the suggested strategy. The outcomes show the pre-processing data methods enhance the quality of the original dataset and identify the optimal predictors' scenario. The CPSOCGSA-ANN algorithm delivers the best performance compared with MPA-ANN and PSO-ANN considering multiple statistical indicators. Overall, the methodology shows good performance with $R^2 = 0.965$, $MAE = 0.01627$, and $RMSE = 0.0187$.

Keywords Data preprocessing · Metaheuristic algorithm · Water quality index prediction · Artificial neural network

Introduction

Surface water pollution is a significant concern for communities worldwide and requires more attention from environmental researchers (Hameed et al. 2016). In addition, the implications of contaminated surface water are deterioration

of water quality (WQ) (Panaskar et al. 2016), a direct threat to human health (Kadam et al. 2019), and disruption of the aquatic ecosystem's balance (Koranga et al. 2022). Accordingly, the United Nations (UN) announced that approximately 1.5 million people die annually from diseases driven by polluted water. Moreover, it is reported that 80% of health troubles in developing countries are brought on by polluted water (Aldhyani et al. 2020a; b). However, surface WQ is affected by various factors, including natural components such as temperature or precipitation (Gupta and Gupta 2021; Michalak 2016; Shanley 2017) and anthropogenic factors such as manufacturing practices, urbanization, and agriculture (Asadollah et al. 2021; Ustaoglu et al. 2020). In view of this, managing WQ in diverse climate conditions seems crucial to prevent the population from suffering from diseases and health troubles. Hence, various techniques have been created and developed to evaluate the WQ, such as the Water Quality Index (WQI) technique (Uddin et al. 2021). WQI has been commonly utilized "as a classification indicator" to evaluate and categorize the water bodies' quality based on measuring a wide range of physiochemical and organic variables and plays a vital role in water management (Das Kangabam et al. 2017; Judran and Kumar 2020).

✉ Hasanain Zamili
ha94.student@gmail.com

Gulfem Bakan
gbakan@omu.edu.tr

Salah L. Zubaidi
salahlafata@uowasit.edu.iq

Mustafa A. Alawsi
mustafa.abdulgani@mtu.edu.iq

¹ Department of Environment Engineering, Ondokuz Mayıs University, 55139 Samsun, Turkey

² College of Engineering, University of Warith Al-Anbiyaa, Karbala 56001, Iraq

³ Department of Civil Engineering, Wasit University, Wasit 52001, Iraq

⁴ Department of Building and Construction Techniques, Kut Technical Institute, Middle Technical University, Baghdad, Iraq

Prediction is crucial to WQ monitoring and is part of contemporary environmental management. Over the past few decades, many traditional WQ prediction methods have been used, e.g. auto-regressive integrated moving averages (ARIMA) (Araghinejad 2013) and multiple linear regression (MLR) (Rajaei and Boroumand 2015). However, with the expansion in the volume of data, traditional methods cannot effectively suit the requirements of researchers due to the increase in computing power and the inability to capture non-stationarity (Chang et al. 2016) and nonlinear (Huo et al. 2013) WQ owing to its sophisticated and complicated nature (Chen et al. 2020). Hence, artificial intelligence techniques such as artificial neural networks (ANNs) are becoming more popular in recent years because they can surmount the limitations or drawbacks of traditional models. In addition, due to their similarities with the brain's nervous system, ANNs are suitable for analyzing nonlinear and unpredictable problems and have become a hotspot in environmental quality research (Hajirahimi and Khashei 2022). The ANN's architecture comprises an input, hidden, and output layers (Kadam et al. 2019). artificial neural networks (ANNs) (Sakizadeh 2016). For instance, Vijay and Kamaraj (2021) employed ANN to predict WQI Drinking Water Distribution System. Also, Yilma et al. (2018) confirmed that ANN is useful for modelling the WQI.

Moreover, the imperative for raising data-driven techniques' reliability, accuracy, and capability has encouraged scientists to develop creative models. The fundamental goal of these new models is to maximise the potential of existing models by combining the benefits of several methodologies (Faruk 2010). These combined approaches generally integrate methods in a process where conventionally, one technology is considered the basic method, the others functioning as pre-processing or postprocessing techniques (Modaresi and Araghinejad 2014). In this context, various optimization techniques were utilized to integrate the machine learning models by finding their hyperparameters,

leading to increased accuracy and saving time (Ahmed et al. 2017). For example, particle swarm optimization (PSO) was employed in various fields, such as drought (Nabipour et al. 2020) and WQ (Aghel et al. 2018; Azad et al. 2019). Also, Faramarzi et al. (2020) proposed the marine predator's algorithm (MPA), which is applied in various applications, including friction stir welded (Abd Elaziz et al. 2020) and photovoltaic systems (Yousri et al. 2020). In addition, the slime mould algorithm (CPSOCGSA) has been devised by Rather and Bala (2019b) to solve optimization problems such as engineering design problems (Rather and Bala 2019a) and water drought forecasting (Alawsi et al. 2022).

Finally, this research aims to predict long-term WQI considering several WQ parameters and rainfall data. The following objectives will be carried out to accomplish this goal: (1) calculating the WQI for the Yesilirmak River. (2) Applying three stages of data pre-processing to increase data quality and choose the optimal independent factors. (3) To reduce uncertainty, combine the ANN model with the CPSOCGSA algorithm and compare the MPA and PSO methods.

Based on the authors' investigation, this is the first time utilizing update algorithms (i.e., CPSOCGSA and MPA) to predict the WQI.

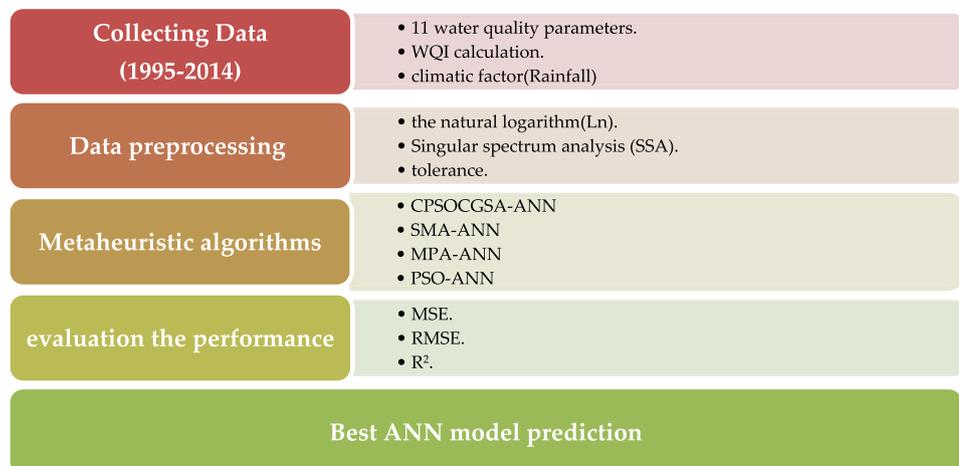
Methodology

The framework of the proposed methodology comprises five stages: study area and data collection, calculating the WQI, data pre-processing, CPSOCGSA algorithm, ANN model, and evaluation of the model's performance, as shown in (Fig. 1).

Area of study and data set

The Yesilirmak River is considered one of the longest rivers in Turkey and located between approximately 29' and

Fig. 1 Framework of the architecture of ANN to predict WQI



40' N. 31' 40' N and 44' 71' E. 44' 61' E, with 39,628 km² covering's area and discharges into the Black Sea in Samsun. The basin has transported partially and fully treated municipal wastewater and pollution from various diffuse sources, including farmyards and agricultural areas. Additionally, different delta parts have dried up in recent decades (Dinc et al. 2021). The station case study was located within the provincial borders of Çorum, Eastern Black Sea Basin. Therefore, the coastal sea climate is dominant in this basin zone, with annual precipitation of about 443 mm.

In several countries, the essential obstructive faced by researchers is the lack of data. Therefore, these stations were sufficiently selected for assessing and building prediction models regarding the number of data and availability of parameters. Additionally, the National Oceanic and Atmospheric Administration (NOAA) (NOAA 2021) was adopted for obtaining the climate variables. The dataset was collected for twenty years, from 1995 to 2014. It contained eleven WQ parameters with one influential climatic factor (rainfall), where the values are distributed regularly along the monitoring periods. Table 1 presents the WQ parameters used to calculate the WQI.

Water quality index calculation

The WQI is an evaluation method that reflects the influence of individual WQ variables on the overall quality of aquatic systems (Ramakrishnaiah et al. 2009). For calculating the WQI, 11 significant WQ parameters were specified regarding the World Health Organization standard for drinking WQ (Organization et al. 2004). The physicochemical parameters were: biological oxygen demand (BOD), pH, electrical conductivity (EC), dissolved oxygen (DO), chlorine (CL⁻¹), calcium(Ca⁺²), magnesium (Mg⁺²), nitrite (NO₂⁻¹), sodium (Na⁺¹), sulfate (SO₄), and

Table 1 The water standards, assigned weights, and relative weight for the WQI's equation (Sharma et al. 2014; Kulisz Monika et al. 2021; Kadam et al. 2019; Şener et al. 2017)

Parameters	WQ standard	AW	RW
BOD ₅ mg/l	5	4	0.102564
DO mg/l	5	5	0.128205
PH	6.5–8.5	4	0.102564
Ca ⁺² mg/L	300	2	0.051282
CL ⁻¹ mg/l	250	3	0.076923
EC µS/cm	250	5	0.128205
Mg ⁺² mg/L	30	2	0.051282
Na ⁺¹ mg/l	200	2	0.051282
NO ₂ ⁻¹ mg/l	3	4	0.102564
SO ₄ mg/l	250	4	0.102564
TDS mg/l	1000	4	0.102564

total dissolved solids (TDS) (Ewaid et al. 2018; Kulisz Monika et al. 2021). In this research, the WQI calculation can be categorized into three stages:

Firstly, weights (W_i) were assigned to all WQ parameters with a scale ranging from 1 to 5, reflecting their significance in affecting water bodies. Table 1 shows the WQ standards, assigned weights (AW.), and relative weights (RW.) of the WQI's equation.

Secondly, calculate the relative weights (W_i) by (Eq. 1):

$$W_i = \frac{w_i}{\sum_{i=1}^n w_i} \tag{1}$$

W_i Is the relative weight, w_i is the weight of every parameter, and n is the number of parameters.

$$q_i = (c_i/s_i) * 100 \tag{2}$$

$$q_i = ((c_i - v_i)/(s_i - v_i)) * 100 \tag{3}$$

Thirdly, from (Eq. 2), a quality rating scale (q_i) was assigned by dividing every parameter concentration in the water samples by its respective standard in WHO guidelines. In contrast, the quality rating for pH and DO was calculated by (Eq. 3).

Where C_i is the measured concentration of parameters in each simple, S_i is the drinking water standard for each parameter, V_i is the ideal value (for pH 7.0 and DO = 14.6), and S_{Ii} is the subindex of the parameter. Finally, WQI was derived from formulas (4) and (5).

$$S_{Ii} = q_i * W_i \tag{4}$$

$$WQI = \sum S_{Ii} \tag{5}$$

Table 2 shows the classification of WQI according to the range proposed by previous studies (Aldhyani et al. 2020a; b; Ewaid et al. 2018).

Data pre-processing

Three strategies were adopted in this research as pre-processing data methods, which are normalization, cleaning, and selecting the optimal model's input.

Table 2 water quality classification (WQC)

WQI range	0–25	26–50	51–75	76–100	>100
Classified	Excellent	Good	Poor	Very poor	Unsuitable for drinking

Normalization

Data normalization is a crucial step for mining datasets in soft computing. In various scientific research, techniques and assessment criteria mostly have multiple scale sizes and units, producing various analysis data results. Therefore, normalization will be a fundamental task to achieve comparability among the variables and realize the expectations of optimization data. Moreover, it minimizes the dimension's impact of different time series and eliminates the influence of outliers and multicollinearity of model parameters (Wu and Wang 2022). In this work, the SPSS 26 statistics package was employed to normalize the time series by applying the natural logarithm technique.

Cleaning

Data cleaning methods comprise locating outliers and noise and then treating outliers and eliminating noise to optimize the data analysis outcomes (Tabachnick et al. 2007). So, the box whisker technique was adopted in this study to identify the outliers were existed outside of the period in the formula below (Kossieris and Makropoulos 2018):

$$\pm 1.5 \times \text{CQR} (\text{CQR} = \text{third quartile}(Q3) - \text{first quartile}(Q1)) \quad (6)$$

Also, the technique of singular spectrum analysis (SSA) is used to detect and remove noise from time series. It is an efficient method for analyzing the initial time series into various components (Golyandina et al. 2018; Karami and Dariane 2022). This approach has shown to be effective in diverse speciality areas, such as rainfall prediction (Reddy et al. 2022), hydrology (Ouyang and Lu 2018), drought forecasting (Pham et al. 2021), and groundwater prediction (Polomčić et al. 2017).

Identifying the best model input

Determining the appropriate predictors is fundamental in establishing a forecast model's structure and enhancing the performance of model. Therefore, the tolerance approach was applied to identify the optimal scenario for the predictors to avoid multicollinearity (Cali et al. 2016; Pallant 2020). It recommended a tolerance coefficient value equal to or higher than 0.2 for selecting the model's predictors.

Constriction coefficient-based particle swarm optimization and chaotic gravitational search algorithm (CCPSOCGSA)

CPSOGSA is a hybrid heuristic optimization that utilizes the intensification potential of the CPSO algorithm with the diversification capability of GSA's algorithm. The components of this hybridization technique will be clarified in the subsections below.

A. Constriction Coefficient based Particle Swarm Optimization (CCPSO.)

The PSO algorithm is a popular optimization approach inspired by natural swarm behaviour for birds or fish. The PSO architecture consists of three principal parameters: *abest*, *bbest* and inertia weight. Where the *abest* and *bbest* assist the finding of the search-space region. The inertia weight has a significant impact on the process of global exploration. In (Eqs. 7,8), the Particle Swarm mathematical formulation explains the updating of the process for the particle's location and velocity during the alteration of the particle values.

$$v_x^d(t+1) = w(t)v_x^d(t) + c_1 r_{x1} (abest - x_x^d(t)) + c_2 r_{x2} (bbest - x_x^d(t)) \quad (7)$$

$$x_x^d(t+1) = x_x^d(t) + v_x^d(t+1) \quad (8)$$

where the c_1 , c_2 are the learning constants, while r_{x1} and r_{x2} are the numbers range from 0 to 1.

Constriction coefficients were employed to improve Particle Swarm Optimisation (PSO) exploitation stage by minimizing the impact of particle movements outside the solution space and hastening convergence during the optimization stage. The coefficients are described as follows:

$$\varphi_1 = 2.05, \varphi_2 = 2.05, \varphi = \varphi_1 + \varphi_2 \quad (9)$$

$$K = 2 / \left(\varphi - 2 + \sqrt{(\varphi^2 - 4)} \right) \quad (10)$$

Substituting the inertia weight by the notation K , (Eq. 7) can be rewritten as in below:

$$v_x^d(t+1) = K v_x^d(t) + K \varphi_1 r_{x1} (abest_x(t) - x_x^d(t)) + K \varphi_2 r_{x2} (bbest - x_x^d(t)) \quad (11)$$

where $K\varphi_1 = c_1$, $K\varphi_2 = c_2$.

B. Chaotic Gravitational Search algorithm CGSA

Gravity and motion Newton's law is considered fundamental to the configuration of the GSA algorithm. Newton's law states that "the gravitational force between two masses is directly proportional to the product of their masses and inversely proportional to the square of the distance between them". Accordingly, the gravitational force (F_{ij}) between masses (i.e., searching agents) (x) and (y) at time (t) can be represented as in the following (Eq. 12):

$$F_{xy} = G(t) \frac{m_{px}(t)m_{ay}(t)}{R_{xy}(t)+ \epsilon} (x_x^d(t) + x_y^d(t)) \tag{12}$$

where m_{px} is the attractive mass, $and m_{ay}$ is the passive masses. While $R_{xy}(t)$ is the Euclidian distance between the two masses at the time (t), while ϵ is a small coefficient, and G is the constant help for controlling the solution space to secure a feasible region.

The (G) constant can be represented by (Eq. 13):

$$G(t) = G(t_o) e^{(-\alpha \frac{CI}{MI})} \tag{13}$$

where $G(t_o)$ & $G(t)$ is the initial and final values of G , respectively. CI is the current iteration, α is the small constant, and MI is the maximum value of iterations. The behaviour of G over time is proposed using a chaotic normalization (Rather and Bala 2019a, b). Hence, (Eq. 14) describe the final formula of the gravitational constant's:

$$G^c(t) = C_i^{norm}(t) + G(t_o) e^{(-\alpha \frac{CI}{MI})} \tag{14}$$

The aggregate force exerted by the masses (i.e., searching agents) could be found as shown in (Eq. 15):

$$F_x^d(t) = \sum_{y=1, y \neq x}^m \gamma_y F_{xy} \tag{15}$$

where γ value ranges from 0 to 1, it is essential to calculate the position and velocity to find the global optimum, which could be represented according to (Eqs. 16, 17):

$$v_x^d(t+1) = \gamma_y v_x^d(t) + a_x^d(t) \tag{16}$$

$$x_x^d(t+1) = x_x^d(t) + v_x^d(t+1) \tag{17}$$

where $a_x^d(t)$ is the mass acceleration.

C. The Combination algorithm (CPSOCGSA)

Combining the two heuristic techniques (CPSO and CGSA) could assist to exceeded each technique's

imperfections. The hybridization equation formula can be described in (Eq. 18):

$$v_x^d(t+1) = (2 / (\varphi - 2 + \sqrt{\varphi^2 - 4})) v_x^d(t) + K\varphi_1 r_{x1} (a_x^d(t) - x_x^d(t)) + K\varphi_2 r_{x2} (bbest - x_x^d(t)) \tag{18}$$

The particle's location is indicated by (Eq. 19):

$$x_x^d(t+1) = x_x^d(t) + v_x^d(t+1) \tag{19}$$

Artificial neural network (ANN)

ANNs are an information-processing approach designed to simulate the functioning of the human brain by attempting the same connectivity and processes as biological neurons (Kouadri et al. 2022). In this research, the ANN structure was composed of four layers of neurons: an input layer which has WQ parameters and rainfall data, two hidden layers to address the nonlinearity relationship, and an output layer which has the target (WQI) (see Fig. 2). The multilayer perceptron (MLP) is performed with feed-forward backpropagation (FFBB). The efficiency with a low error rate and speed of the learning algorithm (Levenberg–Marquardt, LM) was an underlying factor employed with FFBB for training the ANN model (Payal et al. 2015). Matlab Toolbox was running to implement the ANN technique.

The data were separated into three groups: training, testing, and validation, with 70%, 15%, and 15% of values utilised for each set, respectively, such as (Kulisz Monika et al. 2021; Zubaidi et al. 2018).

However, the traditional method (trial and error) was time-consuming and did not always introduce the optimal solution. As a result, combining metaheuristic algorithms with the ANN model is regarded as a superior technique for

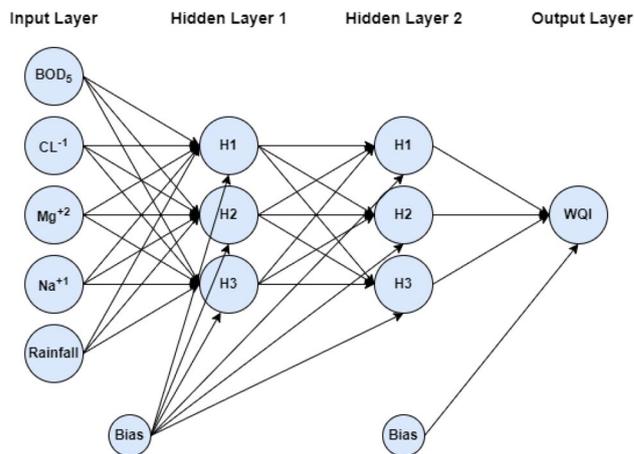


Fig. 2 ANN structure for CPSOCGSA-ANN

determining the optimal neurons' number in the hidden layers (N_1 , N_2) and selecting the best learning rate coefficient (Lr). These hyperparameters are responsible for choosing the best predictors and target mapping to avoid underfitting or overfitting the model.

Model performance assessment

Three statistical criteria are used in this study to evaluate the WQI prediction model's performance. The criteria are the root means squared error (RMSE, Eq. 20), mean absolute error (MAE, Eq. 21), and determination coefficient (R^2 , Eq. 22) (Mohammadi et al. 2020; Mohammadi and Mehdi-zadeh 2020):

$$MAE = \frac{\sum_{i=1}^N |O_i - F_i|}{N} \quad (20)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (O_i - F_i)^2}{N}} \quad (21)$$

$$R^2 = \left[\frac{\sum_{i=1}^N (O_i - \bar{O}_i)(F_i - \bar{F}_i)}{\sqrt{\sum_{i=1}^N (O_i - \bar{O}_i)^2 \sum_{i=1}^N (F_i - \bar{F}_i)^2}} \right]^2 \quad (22)$$

where O_i represents observed WQI, F_i is the forecast WQI, N is the sample size, \bar{F}_i is the mean of forecast WQI, and \bar{O}_i is the mean of observed WQI.

Additionally, the Taylor diagram was utilised to evaluate the outcomes of prediction models by creating a visual comprehension of performance by showing diverse points on a polar plot for multisets of modelling outcomes (Taylor 2001).

Results and discussion

Water quality index assessment

The assigned weight index approach was juxtaposed with the WHO standards for drinking WQ to calculate the WQI for the Yesilrmak river. Table 3 presents the eleven parameters that were combined for the WQI prediction and the descriptive statistics values for them. As can be observed, the WQI varies from 37.41 to 181.06, and the mean was 51.36; this variation reflects the unsuitability for drinking, as shown in Table 2. Calcium (Ca^{+2}), Nitrite (NO_2^{-1}), chloride (CL^{-1}), and sodium (Na^{+1}) contents did not exceed the acceptable range in any tested values. While pH, biological oxygen demand (BOD_5), electrical conductivity (EC), magnesium

Table 3 Descriptive statistics for the Yesilrmak river

Variables	Mean	Min	Max	Std. error
WQI	51.36	37.41	181.06	1.51
BOD_5 (mg/l)	4.06	0.0	40.00	0.46
DO (mg/l)	7.60	0.04	12.70	0.18
PH	8.25	6.90	9.00	0.03
Ca^{+2} (mg/L)	55.59	19.00	129.00	1.55
CL^{-1} (mg/l)	22.78	1.28	102.90	1.04
EC ($\mu S/cm$)	710.03	399.00	1903.00	14.66
Mg^{+2} (mg/L)	39.81	17.60	105.20	0.94
Na^{+1} (mg/l)	37.83	10.16	73.80	1.02
NO_2^{-1} (mg/l)	0.08	0.0	0.80	0.01
SO_4 (mg/l)	54.50	16.80	352.00	3.20
TDS (mg/l)	461.16	241.00	1217.90	9.49

(Mg^{+2}), sulfate (SO_4), and total dissolved solids (TDS) were higher than the specified values in the WHO guidelines. In the same context, the dissolved oxygen exceeded the minimum acceptable value in six samples during the monitoring periods.

Preparation of predictors and target variables

The input data [i.e., WQ parameters and rainfall) and target data (WQI)] were normalized by applying a natural logarithm (Ln) to minimize the adverse impact of extreme values and achieve the normal time series distribution. Then, if any outliers remained after the normalization, it was adjusted. The box plot technique displayed in (Fig. 3) decreased the variation scale of normalization data compared to the raw data. It also showed how the data had been cleaned from outliers.

After that, the SSA was applied to decompose the normalized and cleaned time series into three different components to eliminate the noise. Also, the data pre-processing technique enhanced the correlation coefficient (R) between the dependent and independent factors, such as improving the correlation between WQI and chlorine from 0.609 to 0.7106.

Then, a tolerance approach was applied to locate the optimal predictors scenario to simulate the WQI with high accuracy and prevent multicollinearity by clearing superfluous parameters. According to (Pallant 2020), the tolerance coefficient for the nominated input variables must be higher than 0.2. As shown in Table 4 below, five variables were chosen; biological oxygen demand, Chlorine, magnesium, potassium and Rainfall as the optimal input data model.

In the last step, according to (Gharghan et al. 2016; Kuljis and Kujawska 2021), the data were categorized into 70% training, 15% testing, and 15% validation to create and evaluate the forecasting model.

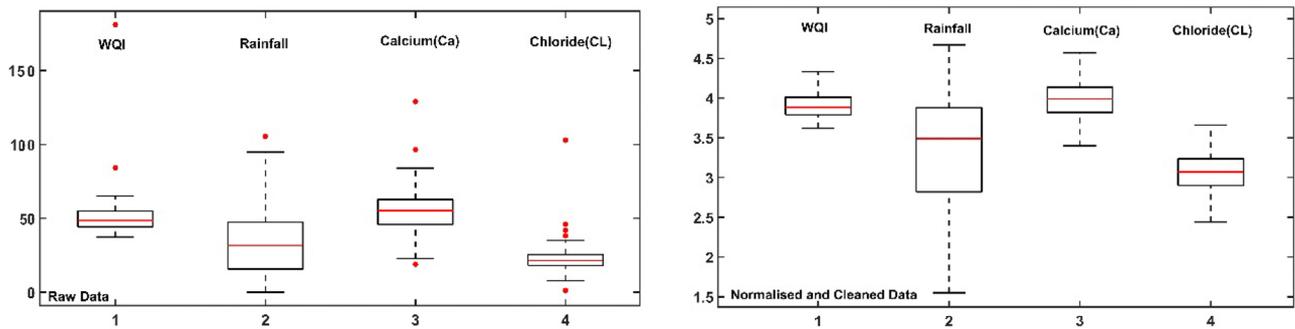


Fig. 3 Box plot drawings of WQI with some of the selected parameters before and after normalized and cleaned data

Table 4 Collinearity statistics for the specified predictors

Predictors	Tolerance value
BOD ₅	0.645
CL ⁻¹	0.352
Mg ⁺²	0.543
Na ⁺¹	0.520
Rainfall	0.729

Configuring the hybrid model

The artificial neural network was hybridized with a metaheuristic approach to locate the optimum LR, N₁, and N₂. The Matlab program was used for running the hybrid models (i.e., MPA-ANN, CPSOCSGA-ANN, and PSO-ANN) to identify the optimum ANN's hyperparameters (LR, N₁, and N₂). This research utilized swarm sizes (10, 20, 30, 40, and 50 pop) for every swarm repeated five times with (200 iterations) to achieve the minimum fitness function (MSE). For example, Fig. 4 depicts the CPSOCSGA-ANN performance and exposes the optimum fitness function for all swarms of WQI.

Figure 5A shows that the swarm size (30₅) was superior in the CPSOCSGA-ANN algorithm and produced the minimum fitness function (MSE=0.001283, with five Iterations). While, swarm size (40₅) for the MPA-ANN algorithm offers the best solution within the fitness function (MSE=0.01341, with nine Iterations), as shown in Fig. 5B. Finally, the PSO-ANN algorithm within swarm size (40₄) was superior to other PSO-ANN swarms and delivers less error (MSE=0.01479, with 94 Iterations), as shown in Fig. 5C.

Based on the best swarms for each metaheuristic method, Table 5 shows the hyperparameters for ANN models.

Performance evaluation

After determining the optimal hyperparameters for the ANN approach, each ANN method was run several times

to identify the optimal network that delivers an accurate result. Three statistical metrics (RMSE, MAE, and R²) were employed to assess the model's effectiveness in predicting WQI data. The results of the performance criteria indices for the CPSOCSGA-ANN, PSO-ANN, and MPA-ANN strategy in the validation section are depicted in Table 6 (Dawson et al. 2007). The CPSOCSGA-ANN was superior to other models because it realized the minimum values of MAE, RMSE criteria and the highest value of R².

The Taylor diagram (Fig. 6) was utilized to evaluate the hybrid models' performance in the validation phase. The measured WQI time series is constituted by the red character (Reference) on the Taylor diagram's X-axis. Taylor's graphical diagram compares three statistics, including correlation coefficient (R), standard deviation (SD), and root mean square error difference (RMSD). It thus delivers a dependable evaluation of the relative performance of various strategy (Ghorbani et al. 2018; Tao et al. 2021). Regarding the Taylor diagram, the CPSOCSGA-ANN model was indicated to be the best forecast model compared with the PSO-ANN and MPA-ANN models because it is the nearest to the reference point.

Moreover, to increase the reliability of the CPSOCSGA-ANN strategy, the Kolmogorov_Smirnov (K_S) and Shapiro_Wilk (S_W) tests were adopted to test the normality of the error data. The results showed that the (p-values) of K_S and S_W tests were more than 0.05, indicating that the errors have a normal distribution according to Valentini et al. (2021) (see Table 7).

Based on the above results, it is possible to conclude that:

- The assessment result of Yesilrmak river quality varied from good to poor, except for one point (June 2000) that was unsuitable for drinking.
- The SSA and tolerance approach has high benefits for enhancing data quality and picking the optimized scenario for input parameters.

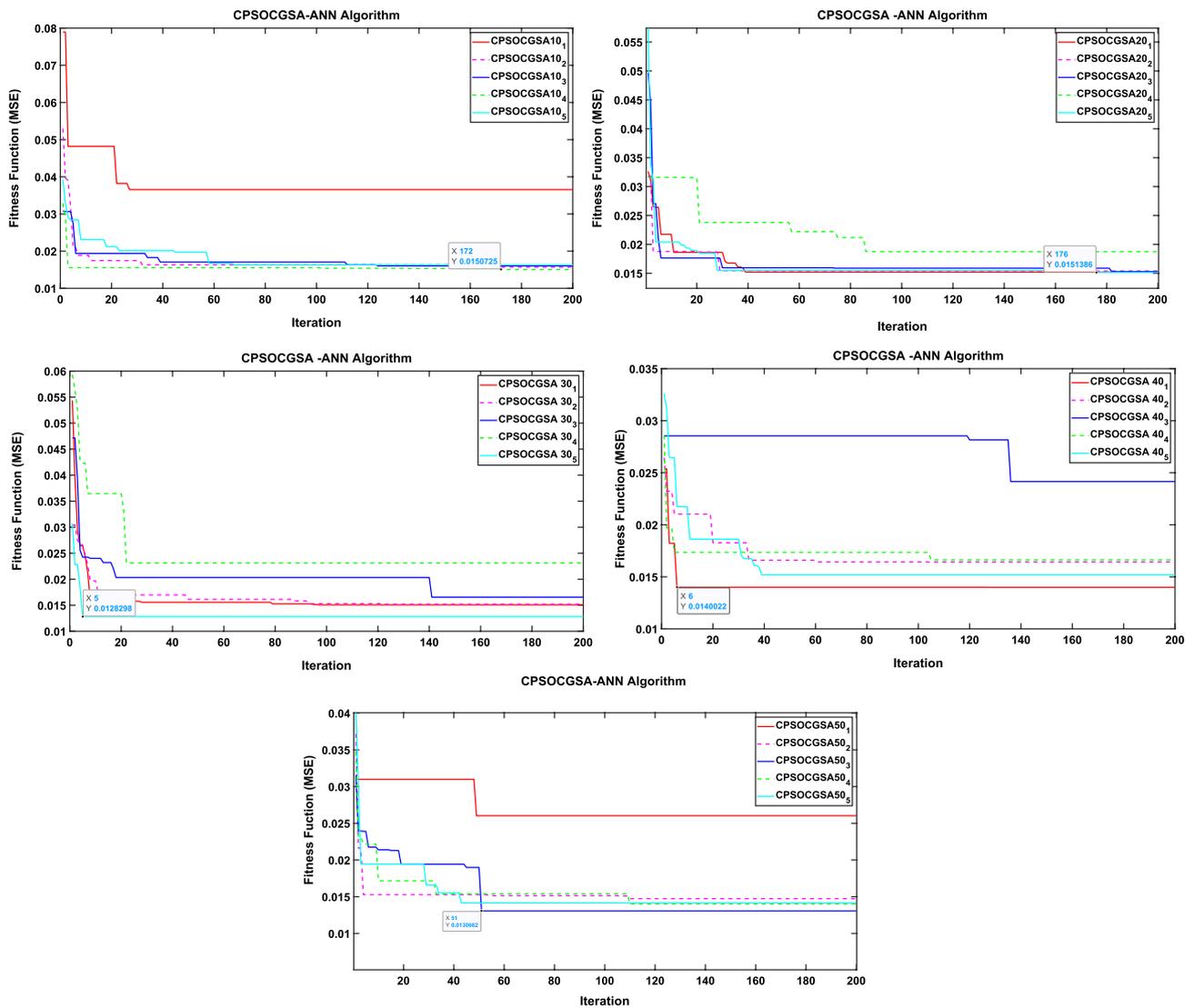


Fig. 4 CPSOCGSA-ANN algorithm performance for simulating WQI

- Comparing the performance accuracy of the hybrid models, including CPSOCGSA-ANN, PSO-ANN, and MPA-ANN, showed that the CPSOCGSA-ANN was a reliable and superior model to predict WQI data.

Conclusion

The present research has assessed the WQ during the monitoring period (1995–2014) for the Yesilrmak river within the provincial borders of Çorum, Turkey. To this

end, eleven physicochemical parameters were assigned to calculate the WQI. The result varied from good to poor, except for one point (June 2000) that was unsuitable for drinking, as shown in Table 3. Additionally, this study employed novel hybridization models combining pre-processing data and ANN approach optimized by multiple metaheuristic methods (i.e., CPSOCGSA, MPA, and PSO) to predict the WQI. Also, to raise the accuracy of models, the implications of climate factors cannot be neglected; hence, adopting rainfall beside the WQ parameters as predictors. The results for the models presented two crucial

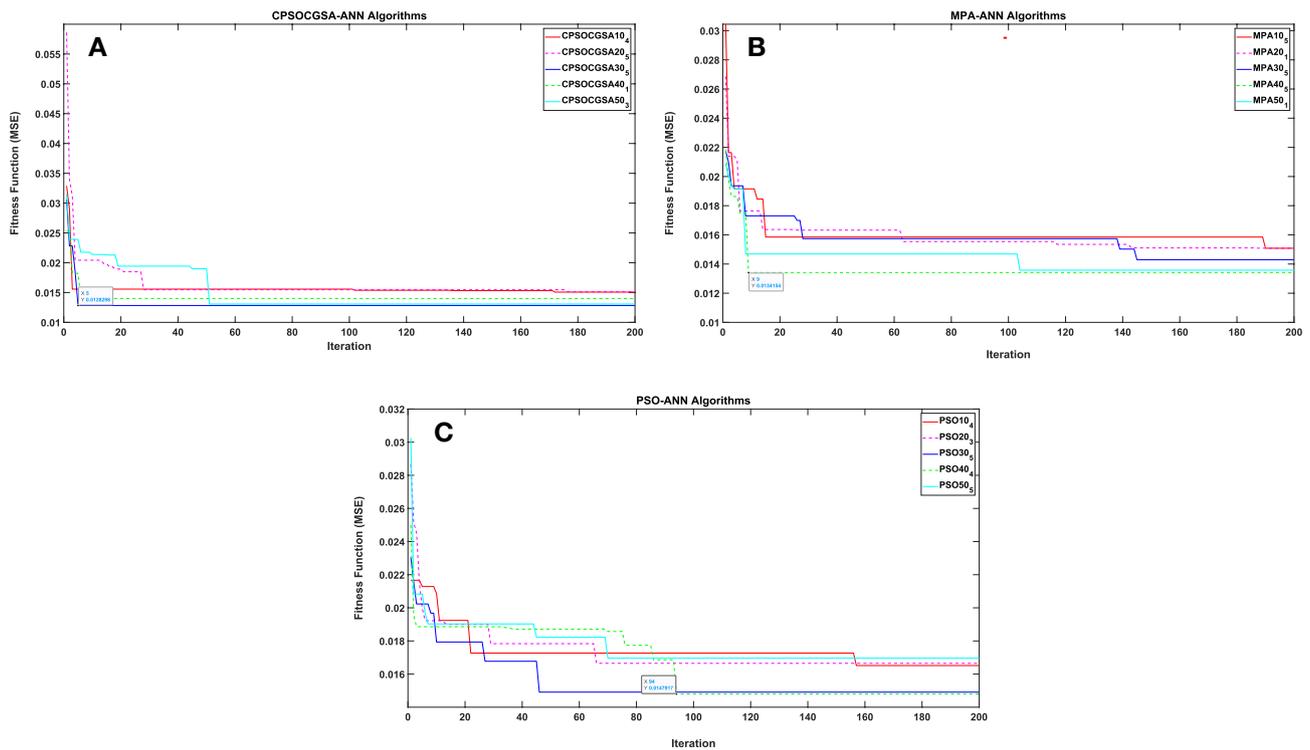


Fig. 5 Best performed of CPSOCGSA-ANN, PSO-ANN and MPA-ANN algorithms for simulating WQI

Table 5 ANN hyperparameters of all models

Model	Lr	N1	N2
CPSOCGSA-ANN	0.6622	3	3
MPA-ANN	0.013	4	1
PSO-ANN	0.2321	1	4

Table 6 Performance evaluation for validation data phase

Model	MAE	RMSE	R ²
CPSOCGSA-ANN	0.01627	0.0187	0.965
MPA-ANN	0.0308	0.0391	0.937
PSO-ANN	0.0324	0.0399	0.924

aspects. The first aspect confirms that pre-processing methods (i.e., SSA and tolerance) enhanced the dataset quality and yielded appropriate scenario of predictors. The second aspect, the performance of CPSOCGSA-ANN, was superior to other models (MPA-ANN and PSO-ANN) that yielded R² = 0.965, MAE = 0.01627, and RMSE = 0.0187.

For future research, it is recommended to use these metaheuristic algorithms to integrate other machine learning models. Also, assess the impact of additional predictors such as an expansion of urbanization, the volume of wastewater discharges, and water consumption.

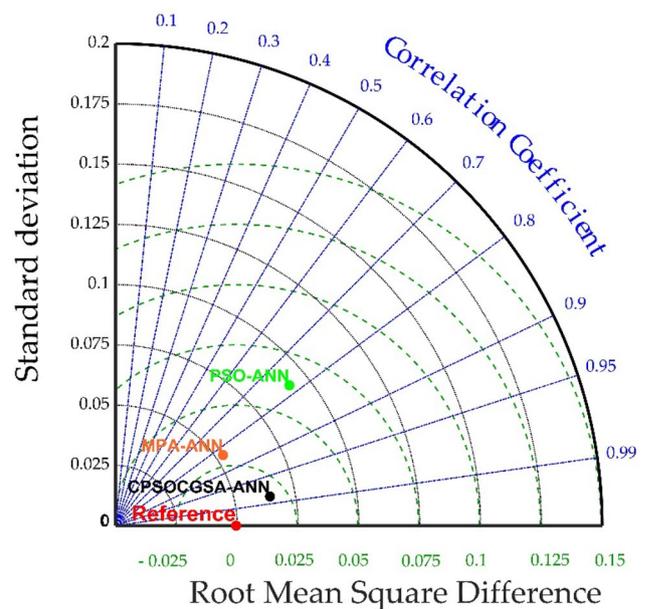


Fig. 6 Taylor diagrams for different forecasting models

Table 7 Tests of normality

Target	Kolmogorov–Smirnova (K–S)	Shapiro–Wilk (S–W)
WQI	0.200	0.224

Declarations

Conflict of interest The authors declare that they have no personal or financial interest, which could influence the work presented in this paper.

References

- Abd Elaziz M, Shehabeldeen TA, Elsheikh AH, Zhou J, Ewees AA, Al-qaness MAA (2020) Utilization of random vector functional link integrated with marine predators algorithm for tensile behavior prediction of dissimilar friction stir welded aluminum alloy joints. *J Market Res* 9(5):11370–11381. <https://doi.org/10.1016/j.jmrt.2020.08.022>
- Aghel B, Rezaei A, Mohadesi M (2018) Modeling and prediction of water quality parameters using a hybrid particle swarm optimization-neural fuzzy approach. *Int J Environ Sci Technol* 16(8):4823–4832. <https://doi.org/10.1007/s13762-018-1896-3>
- Ahmed MS, Mohamed A, Khatib T, Shareef H, Homod RZ, Ali JA (2017) Real time optimal schedule controller for home energy management system using new binary backtracking search algorithm. *Energy Build* 138:215–227. <https://doi.org/10.1016/j.enbuild.2016.12.052>
- Alawsi MA, Zubaidi SL, Al-Ansari N, Al-Bugharbee H, Ridha HM (2022) Tuning ann hyperparameters by CPSOCSA, MPA, and SMA for short-term spi drought forecasting. *Atmosphere* 13(9):1436
- Aldhyani TH, Al-Yaari M, Alkahtani H, Maashi M (2020a) Water quality prediction using artificial intelligence algorithms. *Appl Bionics Biomech*
- Aldhyani THH, Al-Yaari M, Alkahtani H, Maashi M (2020b) Water quality prediction using artificial intelligence algorithms. *Appl Bionics Biomech* 2020:6659314. <https://doi.org/10.1155/2020/6659314>
- Araghinejad S (2013) Data-driven modeling: using Matlab® in water resources and environmental engineering, vol 67. Springer
- Asadollah SBHS, Sharafati A, Motta D, Yaseen ZM (2021) River water quality index prediction and uncertainty analysis: a comparative study of machine learning models. *J Environ Chem Eng*. <https://doi.org/10.1016/j.jece.2020.104599>
- Azad A, Karami H, Farzin S, Mousavi S-F, Kisi O (2019) Modeling river water quality parameters using modified adaptive neuro fuzzy inference system. *Water Sci Eng* 12(1):45–54. <https://doi.org/10.1016/j.wse.2018.11.001>
- Cali D, Osterhage T, Streblov R, Müller D (2016) Energy performance gap in refurbished german dwellings: lesson learned from a field test. *Energy Build* 127:1146–1158
- Chang F-J, Chen P-A, Chang L-C, Tsai Y-H (2016) Estimating spatio-temporal dynamics of stream total phosphate concentration by soft computing techniques. *Sci Total Environ* 562:228–236
- Chen Y, Song L, Liu Y, Yang L, Li D (2020) A review of the artificial neural network models for water quality prediction. *Appl Sci* 10(17):5776
- Das Kangabam R, Bhoominathan SD, Kanagaraj S, Govindaraju M (2017) Development of a water quality index (WQI) for the lohtak lake in india. *Appl Water Sci* 7(6):2907–2918
- Dawson CW, Abrahart RJ, See LM (2007) Hydrotest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. *Environ Model Softw* 22(7):1034–1052
- Dinc B, Çelebi A, Avaz G, Canlı O, Güzel B, Eren B, Yetis U (2021) Spatial distribution and source identification of persistent organic pollutants in the sediments of the yeşilirmak river and coastal area in the black sea. *Mar Pollut Bull* 172:112884
- Ewaid SH, Abed SA, Kadhum SA (2018) Predicting the tigris river water quality within Baghdad, Iraq by using water quality index and regression analysis. *Environ Technol Innov* 11:390–398
- Faramarzi A, Heidarinejad M, Mirjalili S, Gandomi AH (2020) Marine predators algorithm: a nature-inspired metaheuristic. *Expert Syst Appl* 152:113377
- Faruk DÖ (2010) A hybrid neural network and arima model for water quality time series prediction. *Eng Appl Artif Intell* 23(4):586–594
- Gharghan SK, Nordin R, Ismail M (2016) A wireless sensor network with soft computing localization techniques for track cycling applications. *Sensors* 16(8):1043
- Ghorbani MA, Deo RC, Karimi V, Yaseen ZM, Terzi O (2018) Implementation of a hybrid MLP-FFA model for water level prediction of lake Egirdir, Turkey. *Stoch Env Res Risk Assess* 32(6):1683–1697
- Golyandina N, Korobeynikov A, Zhigljavsky A (2018) Singular spectrum analysis with r. Springer, New York
- Gupta S, Gupta SK (2021) A critical review on water quality index tool: genesis, evolution and future directions. *Ecol Inform*. <https://doi.org/10.1016/j.ecoinf.2021.101299>
- Hajirahimi Z, Khashei M (2022) Hybridization of hybrid structures for time series forecasting: a review. *Artif Intell Rev*. <https://doi.org/10.1007/s10462-022-10199-0>
- Hameed M, Sharqi SS, Yaseen ZM, Afan HA, Hussain A, Elshafie A (2016) Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, malaysia. *Neural Comput Appl* 28(S1):893–905. <https://doi.org/10.1007/s00521-016-2404-7>
- Huo S, He Z, Su J, Xi B, Zhu C (2013) Using artificial neural network models for eutrophication prediction. *Proc Environ Sci* 18:310–316
- Judran NH, Kumar A (2020) Evaluation of water quality of Al-Gharraf river using the water quality index (WQI). *Model Earth Syst Environ* 6(3):1581–1588. <https://doi.org/10.1007/s40808-020-00775-0>
- Kadam A, Wagh V, Muley A, Umrikar B, Sankhua R (2019) Prediction of water quality index using artificial neural network and multiple linear regression modelling approach in Shivganga river basin, India. *Model Earth Syst Environ* 5(3):951–962
- Karami F, Dariane AB (2022) Melody search algorithm using online evolving artificial neural network coupled with singular spectrum analysis for multireservoir system management. *Iran J Sci Technol Trans Civ Eng* 46(2):1445–1457
- Koranga M, Pant P, Kumar T, Pant D, Bhatt AK, Pant R (2022) Efficient water quality prediction models based on machine learning algorithms for Nainital Lake, Uttarakhand. In: *Materials today: proceedings*
- Kossieris P, Makropoulos C (2018) Exploring the statistical and distributional properties of residential water demand at fine time scales. *Water* 10(10):1481
- Kouadri S, Pande CB, Panneerselvam B, Moharir KN, Elbeltagi A (2022) Prediction of irrigation groundwater quality parameters using ANN, LSTM, and MLR models. *Environ Sci Pollut Res* 29(14):21067–21091
- Kulisz M, Kujawska J, Przysucha B, Cel W (2021) Forecasting water quality index in groundwater using artificial neural network. *Energies* 14(18):5875
- Kulisz M, Kujawska J (2021) Application of artificial neural network (ANN) for water quality index (WQI) prediction for the River Warta, Poland. In: Paper presented at the *Journal of Physics: conference series*
- Michalak AM (2016) Study role of climate change in extreme threats to water quality. *Nature* 535(7612):349–350. <https://doi.org/10.1038/535349a>

- Modaresi F, Araghinejad S (2014) A comparative assessment of support vector machines, probabilistic neural networks, and k-nearest neighbor algorithms for water quality classification. *Water Resour Manag* 28(12):4095–4111
- Mohammadi B, Mehdizadeh S (2020) Modeling daily reference evapotranspiration via a novel approach based on support vector regression coupled with whale optimization algorithm. *Agricult Water Manag* 237:106145. <https://doi.org/10.1016/j.agwat.2020.106145>
- Mohammadi B, Linh NTT, Pham QB, Ahmed AN, Vojteková J, Guan Y, Abba SI, El-Shafie A (2020) Adaptive neuro-fuzzy inference system coupled with shuffled frog leaping algorithm for predicting river streamflow time series. *Hydrol Sci J* 65(10):1738–1751. <https://doi.org/10.1080/02626667.2020.1758703>
- Nabipour N, Dehghani M, Mosavi A, Shamshirband S (2020) Short-term hydrological drought forecasting based on different nature-inspired optimization algorithms hybridized with artificial neural networks. *IEEE Access* 8:15210–15222. <https://doi.org/10.1109/access.2020.2964584>
- NOAA (2021) National oceanic and atmospheric administration. Data tools: find a station. <https://www.ncdc.noaa.gov/cdo-web/datatools/findstation>
- Ouyang Q, Lu W (2018) Monthly rainfall forecasting using echo state networks coupled with data preprocessing methods. *Water Resour Manag* 32(2):659–674
- Pallant J (2020) SPSS survival manual: a step by step guide to data analysis using ibm. SPSS, Routledge
- Panaskar D, Wagh V, Muley A, Mukate S, Pawar R, Aamalawar M (2016) Evaluating groundwater suitability for the domestic, irrigation, and industrial purposes in Nanded Tehsil, Maharashtra, India, using GIS and statistics. *Arab J Geosci* 9(13):1–16
- Payal A, Rai CS, Reddy BR (2015) Analysis of some feedforward artificial neural network training algorithms for developing localization framework in wireless sensor networks. *Wirel Pers Commun* 82(4):2519–2536
- Pham QB, Yang T-C, Kuo C-M, Tseng H-W, Yu P-S (2021) Coupling singular spectrum analysis with least square support vector machine to improve accuracy of SPI drought forecasting. *Water Resour Manag* 35(3):847–868
- Polomčić D, Gligorić Z, Bajić D, Cvijović Č (2017) A hybrid model for forecasting groundwater levels based on fuzzy c-mean clustering and singular spectrum analysis. *Water* 9(7):541
- Rajaei T, Boroumand A (2015) Forecasting of chlorophyll-*a* concentrations in south san francisco bay using five different models. *Appl Ocean Res* 53:208–217
- Ramakrishnaiah C, Sadashivaiah C, Ranganna G (2009) Assessment of water quality index for the groundwater in Tumkur Taluk, Karnataka State, India. *E-J. Chem.* 6(2):523–530
- Rather SA, Bala PS (2019a) Hybridization of constriction coefficient-based particle swarm optimization and chaotic gravitational search algorithm for solving engineering design problems. In: Paper presented at the international conference on advanced communication and networking
- Rather SA, Bala PS (2019b) Hybridization of constriction coefficient based particle swarm optimization and gravitational search algorithm for function optimization. In: Paper presented at the proceedings of the international conference on advances in electronics, electrical & computational intelligence (ICAEEC)
- Reddy PCS, Yadala S, Goddumarrri SN (2022) Development of rainfall forecasting model using machine learning with singular spectrum analysis. *IJUM Eng J* 23(1):172–186
- Sakizadeh M (2016) Artificial intelligence for the prediction of water quality index in groundwater systems. *Model Earth Syst Environ* 2(1):1–9
- Şener Ş, Şener E, Davraz A (2017) Evaluation of water quality using water quality index (WQI) method and GIS in Aksu River (SW-Turkey). *Sci Total Environ* 584:131–144
- Shanley K (2017) Climate change and water quality: keeping a finger on the pulse. *Am J Public Health* 107(1):e10. <https://doi.org/10.2105/ajph.2016.303504>
- Sharma P, Meher PK, Kumar A, Gautam YP, Mishra KP (2014) Changes in water quality index of ganges river at different locations in allahabad. *Sustain Water Qual Ecol* 3:67–76
- Tabachnick BG, Fidell LS, Ullman JB (2007) Using multivariate statistics, vol 5. Pearson, Boston
- Tao H, Al-Bedyry NK, Khedher KM, Shahid S, Yaseen ZM (2021) River water level prediction in coastal catchment using hybridized relevance vector machine model with improved grasshopper optimization. *J Hydrol* 598:126477
- Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. *J Geophys Res Atmos* 106(D7):7183–7192. <https://doi.org/10.1029/2000jd900719>
- Uddin MG, Nash S, Olbert AI (2021) A review of water quality index models and their use for assessing surface water quality. *Ecol Ind* 122:107218
- Ustaoglu F, Tepe Y, Taş B (2020) Assessment of stream quality and health risk in a subtropical turkey river system: a combined approach using statistical analysis and water quality index. *Ecol Indic.* <https://doi.org/10.1016/j.ecolind.2019.105815>
- Valentini M, dos Santos GB, Muller Vieira B (2021) Multiple linear regression analysis (MLR) applied for modeling a new WQI equation for monitoring the water quality of Mirim Lagoon, in the State of Rio Grande do Sul—Brazil. *SN Appl Sci* 3(1):70. <https://doi.org/10.1007/s42452-020-04005-1>
- Vijay S, Kamaraj K (2021) Prediction of water quality index in drinking water distribution system using activation functions based ann. *Water Resour Manag* 35(2):535–553
- WH Organization, WHO, Staff WHO (2004) Guidelines for drinking-water quality, vol 1. World Health Organization
- Wu J, Wang Z (2022) A hybrid model for water quality prediction based on an artificial neural network, wavelet transform, and long short-term memory. *Water* 14(4):610
- Yilma M, Kiflie Z, Windsperger A, Gessese N (2018) Application of artificial neural network in water quality index prediction: a case study in little akaki river, addis ababa, ethiopia. *Model Earth Syst Environ* 4(1):175–187
- Yousri D, Babu TS, Beshr E, Eteiba MB, Allam D (2020) A robust strategy based on marine predators algorithm for large scale photovoltaic array reconfiguration to mitigate the partial shading effect on the performance of pv system. *IEEE Access* 8:112407–112426. <https://doi.org/10.1109/access.2020.3000420>
- Zubaidi SL, Gharghan SK, Dooley J, Alkhaddar RM, Abdellatif M (2018) Short-term urban water demand prediction considering weather factors. *Water Resour Manag* 32(14):4527–4542. <https://doi.org/10.1007/s11269-018-2061-y>

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.