Environment, Development and Sustainability
https://doi.org/10.1007/s10668-024-05075-6

Evaluation of water quality based on artificial intelligence: performance of multilayer perceptron neural networks and multiple linear regression versus water quality indexes

Selda Palabıyık[1] · Tamer Akkan[2]

Received: 26 June 2023 / Accepted: 22 May 2024

Abstract

A significant problem in the sustainable management of water resources is the lack of funding and long-term monitoring. Today, this problem has been greatly reduced by innovative, adaptive, and sustainable learning methods. Therefore, in this study, a sample river was selected and 14 variables observed at 5 different points for 12 months, traditionally reference values, were calculated by multivariate statistical analysis methods to obtain the water quality index (WQI). The WQI index was estimated using different algorithms including the innovatively used multiple linear regression (MLR), multilayer perceptron artificial neural networks (MLP-ANN) and various machine learning estimation algorithms including neural networks (NN), support vector machine (SVM), gaussian process regression (GPR), ensemble and decision tree approach. By comparing the results, the most appropriate method was selected. The determination of water quality was best estimated by the multiple linear regression (MLR) model. As a result of this MLR modeling, high prediction performance was obtained with accuracy values of $R^2 = 1.0$, RMSE = 0.0025, and MAPE = 0.0296. The root mean square error (RMSE), percent mean absolute error (MAE), and coefficient of determination (R^2) were used to determine the accuracy of the models. These results confirm that both MLR model can be used to predict WQI with very high accuracy. It seems that it can contribute to strengthening water quality management. As a result, as with the powerful results of the innovative approaches (MLR and MLP-ANN) and other assessments, it was found that the presence of intense anthropogenic pressure in the study area and the current situation needs immediate remediation.

Keywords Water quality index · Artificial neural networks · MLR · LM algorithm

✉ Tamer Akkan
biyoloji@yahoo.com

[1] Şebinkarahisar Health Services Vocational School, Giresun University, Şebinkarahisar, Giresun, Türkiye

[2] Department of Biology, Faculty of Arts and Sciences, Giresun University, Giresun, Türkiye

Published online: 01 June 2024

1 Introduction

Water is vital for all life forms. It is not only used for drinking, but also for industry, agriculture and global trade in maritime and interoceanic regions (Abuzir & Abuzir, 2022). Surface water resources, which directly affect the daily activities of humanity, are one of the most important types of water. Among these water supplies, freshwaters (river, lake etc.) are significant social resources that benefit society in a variety of ways, such as ecological habitat, fisheries, farming, and recreational assets (Mutlu et al., 2018; Nacar et al., 2020). However, nowadays freshwater ecosystems are suffering from a variety of threats, including over-exploitation, global warming, and man-made pollution (Brack et al., 2017).

The River Water Quality (RWQ) is a highly sensitive and essential topic in many countries. Similarly, a greater appreciation and definition of the implications of RWQ for daily use, cosystem, farming and industrial uses is badly needed (Gupta & Gupta, 2021). This is because rivers play an important role in creating habitat for many organisms and providing water for human activities. Anthropogenic activities are primarily responsible for the degradation and pollution of natural surface waters and surface sediments (Akkan & Mutlu, 2022; Mutlu et al., 2020; Withanachchi et al., 2018). In addition, rapid industrialization and population growth have increased water quality concerns (Bhattarai et al., 2021). It seems clear that quality monitoring in urban water bodies has become an increasingly important area of research for water scientists around the world over the past 20 years. The qualitative study of river water quality based on physical, chemical and biological parameters includes many water quality characteristics and analysis of a complex data matrix (Said & Khan, 2021).

The WQI is A mathematical tool and composite indicator. This index allows the water quality information to be converted from a single unit to smaller values, depending on the selected variables (Kadam et al., 2019). WQI is well-suited to the assessment of the suitability of RWQ in a range of applications, including agriculture, aquaculture and domestic use (Naubi et al., 2016). The use of water quality indices (WQIs) in the assessment of river water quality has been used since the 1960s. WQI has the potential to transform chosen WQ variables into a dimensionless number, which could enable for easy and clear visualisation of changes in river water character within a specific locality and period of time (Sutadian et al., 2016). Many different indices for WQ assessment based on different parameters are commonly utilized in research (Gad et al., 2022; Pan et al., 2022; Prasad & Sangita, 2008; Sobhan Ardakani et al., 2016). Determining water quality using conventional methods is often time consuming and expensive, especially in developing countries, as multiple parameters need to be determined (Dimple et al., 2022; Ragi et al., 2019). The increasing significance of water quality on a global scale is driving extensive research and the development of innovative and intelligent monitoring solutions. The traditional approach—known as the laboratory method—involves collecting samples from water resorces for analysis in a laboratory setting. While this approach has its merits, it is not without its limitations. It is costly, time-consuming, and a waste of human effort. Furthermore, it may not be the most cost-effective solution. Therefore, innovative methods are used to solve this problem, such as artificial neural network (ANN). This method eliminates the need for the chemical method to assess water quality parameters and is also cost-effective (Ragi et al., 2019).

ANN methodology is a configuration architecture that simulates the human brain and the biological organisation of the nervous system. The underlying technical components of this architecture need a high accuracy. For this reason, the raw data input is normalized and optimized in this simulated system. This is imperative for enhancing the computational speed and precision of ANN performance. They are automatically trained using the optimization algorithm of their designers and generate the outputs, such as WQ (Kadam et al., 2019). Successful estimation of RWQ has attracted the attention of various government agencies and environmental agencies worldwide because it is useful in determining watershed health, biodiversity, ecology, and suitability of drinking water needs of river basins (Satish et al., 2022). Some researchers have statistically correlated WQI results using regression analysis (Chauhan & Trivedi, 2022; Fernández del Castillo et al., 2022; Khan et al., 2022; Yıldız & Karakuş, 2020). Regression modelling is a valuable and powerful instrument for assessing time series. It allows the effect of influencing variables to be modelled, and it is particularly effective in dealing with outliers, lacking observations and disordered measurement models (Abaurrea et al., 2011). Recently, ANN modeling is often used to quantify the severity of WQ problems due to the fast training progress and its intuitive capability for handling both complex linear and nonlinear problems (Nathan et al., 2017). Moreover, novel leakage detection and water loss management methods are being used for the urban water management using very efficient methods (Geng et al., 2019; Hu et al., 2021). The utilisation of forecasting methods are beneficial in a multitude of disciplines, including economics, hydrology and meteorology. In the majority of cases, time series in the rational world offer decision-makers highly accurate forecasts (Cansu et al., 2024; Eğrioğlu & Bas, 2023; Egrioglu et al., 2024; Egrioglu et al., 2019). The importance of these innovative methods that are highly accurate, cost-effective, and adaptable to global changes is increasing in the sustainable management of water resources.

Over recent years, machine learning (ML) algorithms have become a valuable instrument for the effective resolution of numerous environmental issues, including the assessment of water quality (Akkan et al., 2022; Lap et al., 2023). However, because of the incompatibility of current WQI methods, many scientists started to use machine learning to minimize model fuzziness and estimate WQIs at an accurate level (Hassan et al., 2021; Kouadri et al., 2021).

One of the weaknesses of ANN models is that they are developed based on previous data and thus cannot be used if limited data is available (Nhantumbo et al., 2018). The Levenberg–Marquardt algorithm has some limitations: it is used for networks with only one source element, the algorithm requires a large amount of memory proportional to the square of the neural network, so it is not recommended for large networks (Kostiuk et al., 2022). However, the search for a structured method for selecting the suitable network structure to best predict water quality parameters has attracted attention (Ahmed et al., 2019). The single prediction models cannot cope with complex conditions in datasets, easily decrease to local optima and become liable to overfitting (Dong et al., 2023; Xu et al., 2017). The ANN faces some limitations due to the nonlinear and non-stationary properties of some time series (Dong et al., 2023; Yang et al., 2021). Deficiencies in ANN models are identified with the results obtained from different studies. Therefore, the most optimum models have to be developed through these studies. The obvious utilization of ANN to optimize the projection model for WQI forecasting for the currently surveyed regions is an application that has not yet been studied. The present paper is motivated by three objectives: (1) to conduct a preliminary assessment of RWQ for drinking and irrigation water by computing WQIs; (2) to apply ANN and MLR models for the prediction of WQI; and (3) to contrast ANN and MLR models to determine the exact values of WQIs for sustainable management of aquatic supplies.

2 Materials and method

2.1 Study area

The Aksu Creek flows into the Black Sea at the borders of Giresun Province, the central district of the eastern Black Sea region (Fig. 1). It rises in the Giresun Karagol region at an altitude of 3107 m, is fed by many streams in the Kızıltaş, Sarıyakup, Pınarlar and Gudul regions, and empties into the Black Sea after a distance of 60 km on the eastern border of the central district. Mount Kılıç (3107 m) in the south of the Aksu Basin is the highest area. In addition to the rather large altitude difference, the inclination values of the basin vary between 0° and 90°. The area of the basin that collects the water of the Aksu Creek is 731 km^2, its circumference is 129.4 km, the main waterway is 58.8 km long and has a slope of 4.5%. Moreover, the median value of the basin is 2102.3 m, the river grade is 4, the drainage density is 0.48 km^{-1}, and the channel frequency is 0.16 waterways/km^2. The main tributaries of Aksu Creek are Soğucaksu, Kargilimacun, Tehnelli, Karpuz, Kuçukaksu, Kırkgeçit, Bafadan, Tatlıçay, Çobanozu, Eğrioz, Hayıtlı, Karganlı, Asar, Naneli, Kuzgun creeks (Anli, 2003). The water area of Aksu Creek is 250 ha and its flow rate is 562.0 hm^3/year.

2.2 Collection of surface water samples

To evaluate the physicochemical variables, the sample containers used for the study were washed in a bath of weak acid or distilled water 1 day before being used in the field. Then, the sample vessels rinsed with distilled water were dried in an oven and made ready for use. The water sample was taken with a Nansen bottle according to the relevant guideline of TS EN ISO 5667 and brought to the laboratory without losing time in the cold chainn. In this paper, data were collected from five different stations for this dataset for 1 year. In other studies,

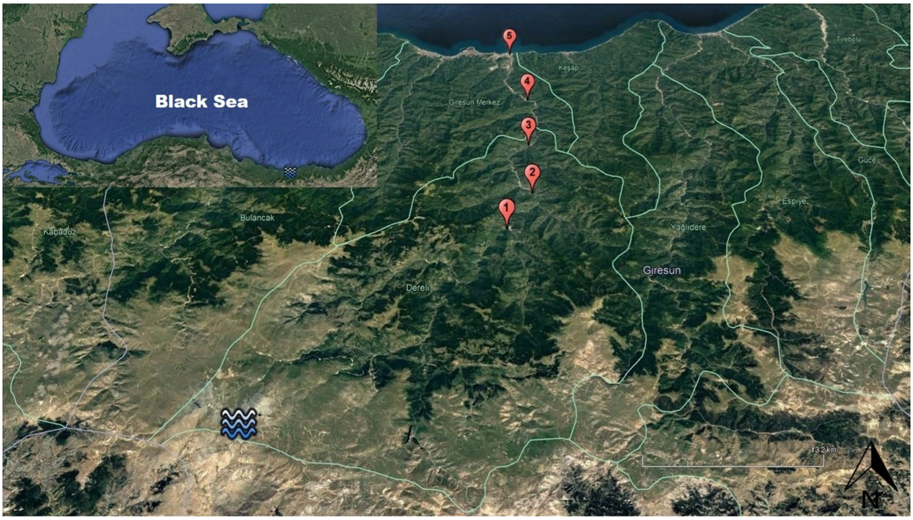


Fig. 1 Sampling area (Google Earth)

similar to our study, water samples were taken for a year as a data set to determine the quality of aquatic resources (Huang & Yang, 2019; Krtolica et al., 2021; Najah Ahmed et al., 2019; Ucun Ozel et al., 2020).

2.3 Analysis of water samples

Analyzes of surface water samples from Aksu Creek were conducted in two phases, under field and laboratory conditions. During field studies, water temperature, pH, dissolved oxygen, salinity, electrical conductivity, total dissolved solids, and oxidation potential of water samples were measured using YSI 556 MPS and turbidimeters WTW-355 IR. Nutrients, which must be measured immediately under field conditions, were also analyzed using the YSI 9300 photometer and appropriate commercial kits. During the field studies, measurement calibrations of all variables were performed each month using standard calibration solutions and the instruments were made ready for use.

2.4 Water quality indexes assessment

A first basic step in the calculation of water quality indices is the selection of the variables, the determination of the value of the partial index, the creation of the weights, and the use of the aggregation processes of the partial indices to obtain the value of the final index that can be used to comment on water quality. In this study, the basic water quality variables and their weights were obtained from literature studies (Gupta & Gupta, 2021; Khalid and others 2019; Pan et al., 2022; Qi et al., 2022). In accordance with the expert opinion, the weights calculated with different equations were used to eliminate the possible differences that may result from the conventional weighting processes. WHO (2011; 2017), CCME (2007), and FAO (1994) were used for the default values used in the WQI assessment.

WQI: the weighted arithmetic water quality index method, uses the most frequently measured water quality variables to classify water quality according to quality levels. For this purpose, it was calculated and evaluated using the following equation to evaluate the WQ adequacy of Aksu Creek (Brown et al., 1972, Table 1).

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

$$Qi = \frac{E_{m-}E_{id}}{E_{s-}E_{id}} \times 100 \tag{2}$$

Table 1 WQI rating scale

WQI	Quality	Class
0–25	Perfect	A
26–50	Good	B
51–75	Weak	C
76–100	Too weak	D
> 100	Inappropriate	E

$$\mathrm{SI} = \mathrm{W_{İ}} \times \mathrm{Q_{İ}} \tag{3}$$

$$\mathrm{WQI} = \sum_{i=1}^{n} \mathrm{SI} \tag{4}$$

In equality

W_i: weight of each variable, w_i: relative weight of each variable, E_m: element measured value, E_{id}: element ideal value, E_s: element standard value, Q_i: rating value, SI: the sub-index value represents.

Evaluation

The nutrient pollution index (NPI) was determined and evaulated by Isiuku and Enyoh, (2020):

$$NPI = {}^{C_{NO3}}\!/_{MAC_{NO3}} + {}^{C_{TP}}\!/_{MAC_{TP}} \tag{5}$$

In equality

$C_{x:}$ NO_3-N and TP (average concentration), MAC_x: Maximum permitted level (Turkish Surface Water Quality Regulation, 2016).

2.5 Statistical calculations

The normality test, analysis of variance, multiple comparison tests, correlation analysis, cluster analysis, factor analysis, and principal component analysis used to specifically test the data obtained in this work were analyzed using the statistical program SPSS 17.0. MATLAB deep statistics and machine learning toolbox was used to estimate the WOI.

2.6 Machine learning (ML) models

Machine learning (ML) is a key element of artificial intelligence (AI)—it allows a system automatically to learn and evolve from its experience, without the need for a specific programme (Sun & Scanlon, 2019). The term machine learning describes an artificial intelligence (AI) and computer science technique that uses data and algorithms to model and progressively improve the accuracy of the human intelligence process. Additionally, in water management applications, it is used to evaluate real-time data, enhance water quality monitoring, as well as assess and forecast present and predicted water quality caused by various factors like acidity, turbidity, salts, nutrients and pollutants (Jafar et al., 2023).

In order to ascertain the most accurate prediction model for water quality, five machine learning (ML) algorithms were employed: support vector machine (SVM), neural network/multilayer perceptron (MLP), ensemble, Gaussian process regression (GPR), and decision tree. The efficacy of each algorithm was evaluated through comparison of their respective accuracy, with the aim of identifying the most suitable prediction model. The performance evaluation of the ML models in question revealed that Gaussian Process Regression (GPR) exhibited the lowest training error and provided the most accurate prediction of the WQI input dataset.

2.6.1 Gaussian process regression

Machine learning is a crucial aspect of the programming domain. The primary benefit of GPR is its capacity to discern that the learning sample adheres to the prior probabilities of a Gaussian process regression (Elbeltagi et al., 2021). Gaussian process regression (GPR) is kernel machine learning methodology that does not require the specification of a parametric model. This method has gained considerable attention in the literature over the past few years (Sharifzadeh et al., 2019). The Gaussian process regression (GPR) technique is concerned with the conditioning and bounding via the use a priori knowledge of a Gaussian fit in regression-based fields (Ali et al., 2022). The application of GPR in the fields of aquatic sciences encompasses a number of diverse areas, including the forecasting of water flow, the estimation of pipe bursting rates in water distribution systems and the monitoring of groundwater quality (Zare Farjoudi & Alizadeh, 2021).

2.7 Multiple linear regression (MLR) based WQI model

MLR is widely used for water quality estimation in different parts of the world (Egbueri & Agbasi, 2022). Two water quality indices were estimated using MLR. In the context of (MLR, the estimators are unknown variables that are estimated from two or more known variables. In other words, multiple regression analysis helps to estimate the value Y for given values X_1, X_2, …, X_k. The commonly known multiple regression equation of X_1, X_2, …, X_k and Y (the dependent variable) is Said and Khan, (2021):

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k \tag{6}$$

The assumption of independence of variable components in MLR, as a parametric statistical model, may deviate from the real situation (Golfinopoulos & Arhonditsis, 2002). Multiple linear regression (MLR) analysis was performed using the stepwise method. This methodology entails the utilisation of variables that exert a considerable influence on the dependent variable. In this technique, the WQI was calculated as the independent variable and the factors affecting this factor were accepted as independent variables.

2.8 Artificial neural networks

The ANN model is based on the human brain (Choden et al., 2022). ANNs represents a form of artificial intelligence that attempts to emulate—in both structure and function—the biological architecture of the human neurosystem and nervous system (Malinova & Guo, 2004). ANNs are suitable for use in nonlinear mapping and exhibit advanced fault tolerance, self-adaptation, self-regulation, and self-learning capabilities, in addition to other beneficial attributes. The suitability of ANNs for high-dimensional and non-linear system problems has been idenitified (Bas et al., 2022; Che & Wan, 2022). The ANNs activates each neuron's inputs (WQ variables) to produce an output signal through (WQI) the application of an activation function (Jain et al., 2022).

2.9 Application of artificial neural network (ANN) based WQI model (MLP-ANN)

ANN adapted to MATLAB software was used as a computational tool to check the correlation between input and output variables and to make a prediction regression of the obtained data (Adeogun et al., 2021; Igwegbe et al., 2019). The computational tool ANN, adapted to MATLAB software, was employed to assess the correlation between input and output variables Furthermore, a regression analysis was conducted to predict the outcomes of the obtained data (Adeogun et al., 2021; Igwegbe et al., 2019). The predicted results by the ANN were comparing with the WQI outcomes by using nntool to train the dataset.

The architecture structure of ANN includes three layers: input, hidden and output, and these layers consist of one or more simple artificial neural cells, called neurons or processing elements. For this study, 14 parameters correspond to neurons in the input layer, which is the number of parameters examined. A trial-and-error approach was used to estimate neurons in the hidden layer. For that purpose, initially a network is trained and tested using a minimum set of networks (2 nodes of the hidden layer). Afterwards, the number of hidden layer nodes was gradually increased (up to 5) to measure the overall performance during training and testing. To build an ANN model, all datasets were divided into training (70%), validation (15%) and testing (15%), respectively. The training input set is employed to compute gradients and update the weights at layers of the network. In contrast, the validation dataset is utilised to make decisions, complete training, and avoid overfitting the network. Levenberg Marquardt (trainlm) technique was used to test the function of the network. The proposed structure of the ANN model is shown in Fig. 2.

The MLP characterized by its adaptability to any application of the learning procedure applying the enormously widespread back-propagation algorithm technique. Nonetheless, convergence is not without limitations. These include a tendency to be slow, unstable, and to remain at local minima. Thus, the Levenberg–Marquardt algorithm, which has been enhanced, provides a solution to these shortcomings (Toha & Tokhi, 2008). The Levenberg–Marquardt algorithm is sufficiently adequate for use with hundreds of weights in models. The problems of approximating functions that demand precision in formation often profit by utilizing it (Bekas et al., 2021).

2.10 Determining the accuracy of the model

The models were evaluated by means of coefficient of determination (R^2), root mean squared error (RMSE) and mean absolute percentage error (MAPE). RMSE, MAPE, and R^2 were calculated using Eqs. (7), (8), and (9), respectively. The coefficient of determination (R^2) range is 0 to 1, and it idenifies the degree of correlation between the observed and predicted values. The 1 represents an excellent correlation within the observed values and

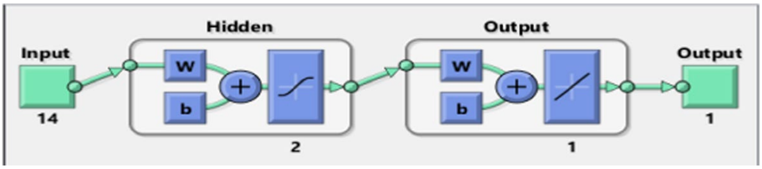


Fig. 2 General scheme for the Levenberg Marquardt (trainlm) technique

the line drawn through them, and 0 represents there is no statistical correlation within the observed and forecasted data (Barzegar et al., 2016).

The majority of the aforementioned tasks were completed using MATLAB, which included model training, statistical analysis of parameters, calculation of correlation coefficients, error analysis, and so forth.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(output_t - target_t\right)^2} \tag{7}$$

$$MAPE = \frac{1}{n}\sum\nolimits_{t=1}^{n}\left|\frac{output_t - target_t}{target_t}\right| \tag{8}$$

$$R^2 = 1 - (\sum_{t=1}^{n}(output_t - target_t)^2/(output_t - target_t)^2 \tag{9}$$

3 Discussion and conclusion

3.1 Statistical analysis results

Multivariate statistical methods, such as correlation analysis, principal component analysis, factor analysis, and cluster analysis, are frequently employed to identify the key variables that influence water and sediment quality, as well as to ascertain the dominant factors that affect water and sediment quality. These methods are also used to investigate the sources of these factors and to determine the long-distance relationship between them (Al-Ani et al., 2019; Basatnia et al., 2018).

According to the results of Pearson correlation of the variables observed in the surface water samples of Aksu Stream, the significant correlation pairs are respectively: pH/Turb, pH/DO, Alk/EC, Alk/TDS, Hard/EC, Hard/TDS, Hard/Alk, TAN/DO, TAN/Alk, TAN/Hard, NO_3/EC, NO_3/TDS, NO_3/DO, SO_4/EC, SO_4/EC, SO4/Hard, Na/NO_3 were observed (Fig. 3).

According to the results of factor analysis, 5 factors explained 78.585 of the total variance in this study (Table 2). In the first factor, the variance explained is 29.629. A strong positive weight and a positive weight of alkalinity and hardness were found in the variables EC and TDS. The variance value of the second factor explained 14.648. NO_3 and DO were strongly positive and TAN was determined with positive weight. Based on these factors, we can say that there are climatic, agricultural input and erosion factors on water quality of Aksu Creek. The variance value of the third factor explains 13.807. Na and K have strong positive weights. The variance value of the fourth factor explains 10.537. TP has a negative weight and NO_2 has a medium positive weight. The variance value of the fifth factor explains 9.965. pH has a medium positive weight and turbidity is weighted toward a positive weight. Using these factors, we can show the impact of anthropogenic influences on the water quality of Aksu Creek, especially heavy metals released by mining activities and inputs from erosion.

In order to determine the variable factors affecting water quality in the Aksu Stream, a total of 14 variables were studied from the physicochemical parameters determined

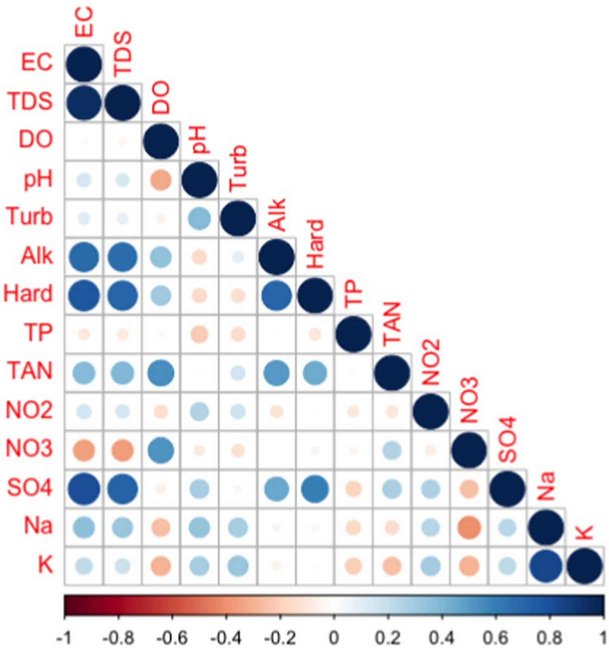


Fig. 3 Pearson correlation graph of water variables

Table 2 Varimax rotated factor matrix of water variables

Eigenvalues	4.148	2.051	1.933	1.475	1.395
% of Variance	29.629	14.648	13.807	10.537	9.965
Cumulative %	29.629	44.276	58.083	68.620	78.585
	Component (Varimax with Kaiser Normalization)				
	F1	**F2**	**F3**	**F4**	**F5**
EC	.952	− .139	.157	.090	.072
TDS	.925	− .158	.120	.058	.082
DO	.143	.852	− .106	− .185	− .046
pH	.009	− .236	.033	.553	.631
Turb	− .002	.060	.295	.012	**.844**
Alk	.801	.279	.033	− .256	.047
Hard	.853	.260	.008	.028	− .238
TP	− .037	− .309	− .308	− .616	.075
TAN	.525	.527	− .258	− .144	.371
Na	.166	− .220	.865	.118	.184
K	.040	− .175	.887	.207	.137
NO_2	.068	− .157	.086	**.652**	.107
NO_3	− .241	.757	− .298	.143	− .049
SO_4	.792	− .162	.031	.392	.041

The format has been modified to a normal style

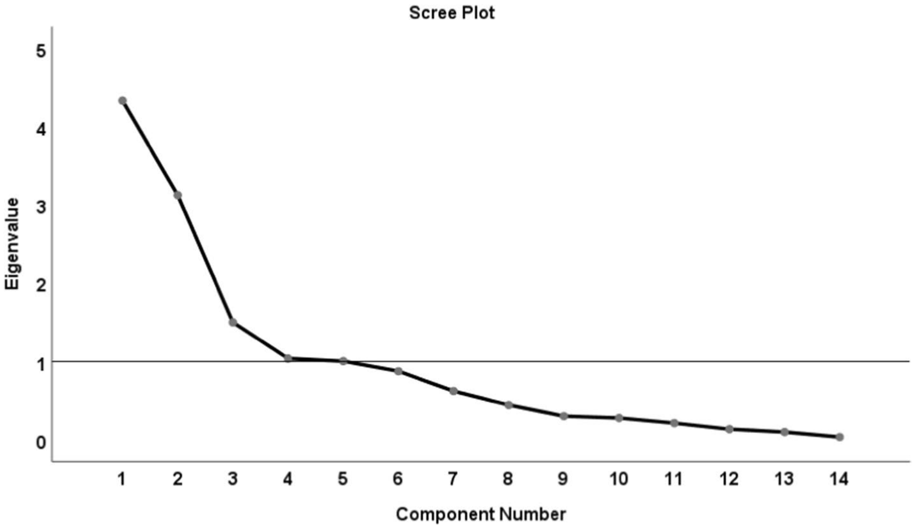
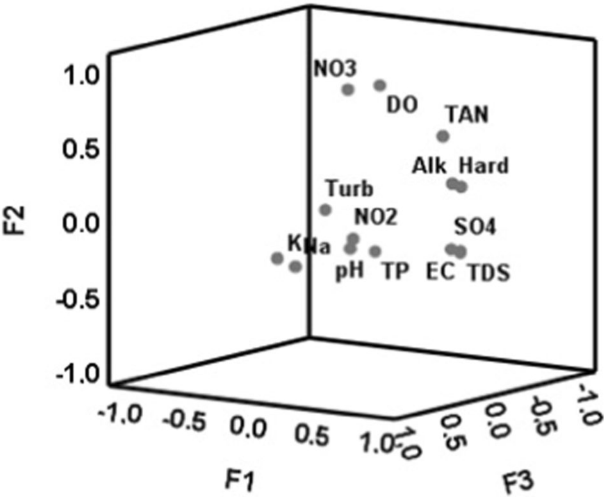


Fig. 4 Scree Plot diagram of water variables

Fig. 5 Component plot, R-mode factor analysis plot of the physicochemical parameters in the studied



in the water. As criteria for evaluating the principal components, values with eigenvalues greater than one were determined as sources of variance to be explained from the data used. The diagram expressing the eigenvalues of the principal components is shown in Fig. 4.

The three-dimensional representation of the rotated factor matrix, which shows which factor the variables are in relationship with, is given in Fig. 5.

3.2 Water quality index results

WQI comprehensively represents the quality of groundwater and surface aquaticresources as a combination of various WQ parameters (Acharya et al., 2018; Deshmukh & Aher, 2016; Gupta & Gupta, 2021; Khalid and others 2019). In the Yellow River (China), the highest and lowest WQI values were calculated as 92.1 & 52.6 and 95.3 & 57, respectively, where water quality was "good" and "moderate" (Pan et al., 2022), WQ assessment in the Yihe River (China) were reported to vary from upstream to downstream, with average WQI values ranging from 78.54 to 83.67, with the highest WQI values (82.43) (Qi et al., 2022). In our study, it was found that the WQI values of Aksu Creek ranged from 103 to 141, with an average value of 113.6, which can be expressed as "not suitable for drinking water use" (Fig. 6). The highest value was obtained at the discharge point of Aksu Creek, which is expressed as the 5th station. This can be explained by the fact that downstream stations receive more pollutants from upstream stations due to runoff after rainfall and runoff from high altitude is a possible additional source of pollutants in these regions (Singh et al., 2015). Similarly, NPI is frequently applied to evaluate nutrient contamination effects at surface water bodies. In a practical application, it is calculated according to the NO_3 and TP concentrations and shows the quality of the water (Isiuku & Enyoh, 2020). The NPI values were higher at Station 3. The NPI results showed that station S4 was “considerable polluted”, and other stations were “very high polluted” in Aksu Creek. This undesired condition may be reflection from anthropogenic effects.

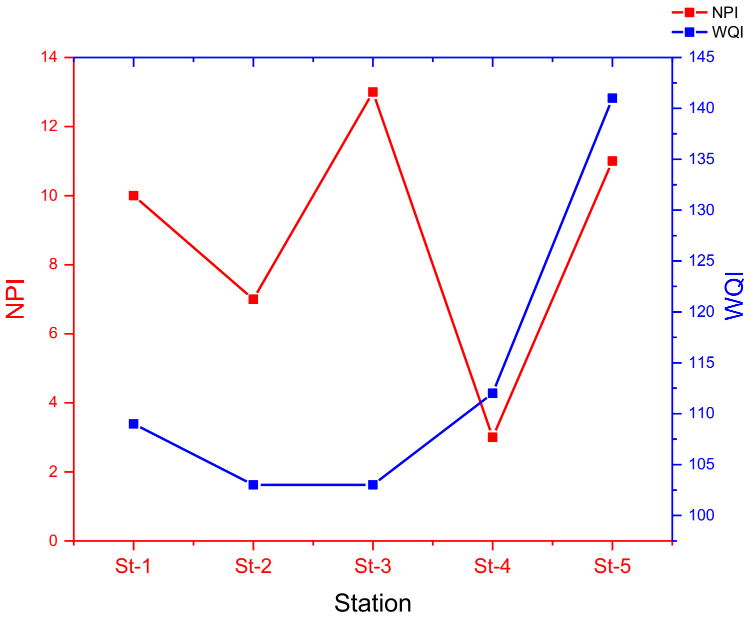


Fig. 6 WQI and NPI results in water samples

Table 3 Various performance criteria were used to determine the best model after the evaluation of the machine learning algorithms

Model type	RMSE (Validation)	MSE (Validation)	R^2 (Validation)	MAE (Validation)
Neural network/multilayer perceptron (MLP)	4.88949	23.90717	0.98205	2.84736
Gaussian process regression	0.00362	0.00001	0.99999	0.00256
SVM	3.00524	9.03152	0.99322	2.51747
Decision Tree	8.61947	74.29530	0.94423	6.33886
Ensemble	8.75033	76.56831	0.94252	6.51294

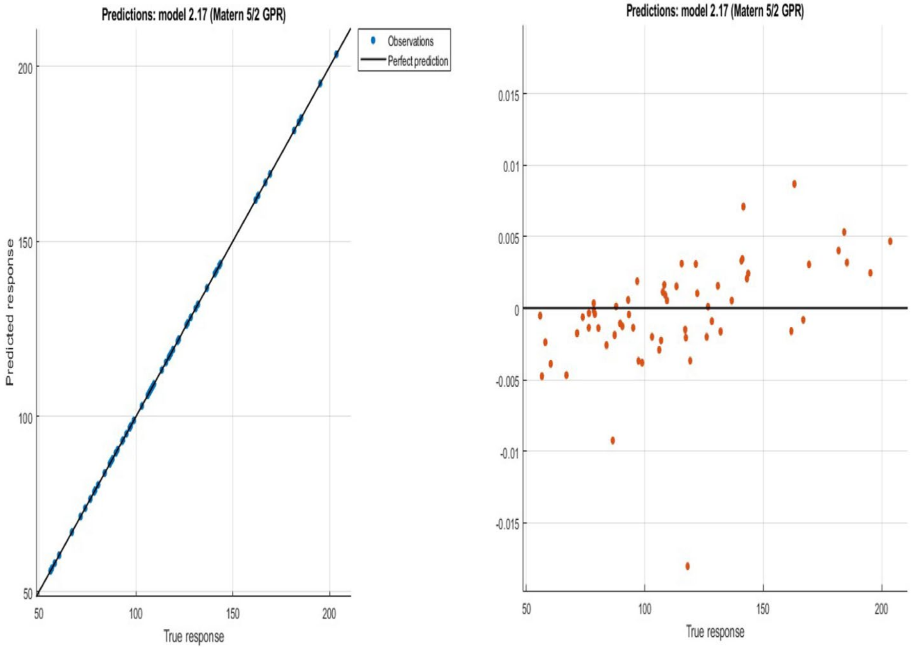


Fig. 7 Gaussian process regression model of the points representing the calculated WQI values

3.3 Machine learning (ML) algorithms results for WQI prediction

Various performance criteria were used to compare different algorithms to determine the best model. After the evaluation of the machine learning algorithms, as shown in Table 3, the results of linear gaussian process regression; RMSE of 0.00362, MSE of 0.00001, R^2 of 0.99999, MAE of 0.00256 were found to be the best algorithm.

Figure 7 shows that the Gaussian Process Regression model of the points representing the calculated WQI values and the prediction points with the Gaussian Process Regression model showed an overall 1:1 perfect line match. It is evident that this model is the most suitable for forecasting the water quality parameter values.

3.4 MLR and MLP-ANN application results for WQI prediction

In this study, water quality variables were measured from surface water stations and used to predict water quality using MLR methods. This approach is essentially a basic least squares technique. Due to its realistic return, the MLR technique is relatively straightforward and requires less time (Sahoo & Jha, 2013). The MLR is a method modelled to reveal linear relationships between two random vectors, X and Y. Among the reasons why multivariate regression is generally preferred are (1) predicting Y with respect to X, (2) testing assumptions about the relationship between X and Y, and (3) the adaptability of Y to forecasted time series or spatial patterns (DelSole & Tippett, 2022). The R^2 (R-squared) metric is the ratio of the regression model to the independent variables and means how much of the variance of the dependent variable it measures. The R^2 value takes a range from 0 to 1. The high R^2 shows that the model explains a large amount of the variance in the dependent variable and a good model is achieved. However, the R^2 metric can lead to the problem of over-fitting; in other terms, the R^2 value can be high in overly complex models, but this weakens the generalisability of the model (Akdağ, 2023). A non-linear model may not be preferred unless a linear relationship is expected to be present in the data set or the model is required to be complex. Linear models are simpler to interpret and are quite adequate for predicting complex biological systems (Heil et al., 2023). Furthermore, linear models can be combined with artificial neural networks (ANNs) by constructing a linear model of the data and then using an ANN to model the residuals. In this way, a good model performances can be achieved while preserving the interpretations obtained from the linear part of the model (Laarne et al., 2022). The high R^2 value of the water parameters used in our study, which may be due to a linear relationship, may indicate an overfitting problem. This result can be interpreted as the model is too specific to the data. It provides an advantage in the sustainable policy-making of a specific area in water resources management, such as a basin, wetland, and water reservoir etc. In order to overcome this situation, analyses can be performed using various adjustment techniques or different model options. In addition, it would be more advantageous to use MLR when there is a linear relationship between the variables in the data set. The reasons why linear regression is a better method for small sample sizes and low-dimensional data have been discussed and data have been presented that it is a good method (Kipruto & Sauerbrei, 2022; Santana et al., 2021; Wang & Yao, 2020; Yan & Wang, 2022).

R^2 was employed to assess the robustness of the MLR models presented in this paper. Water variables of Aksu Creek such as pH, EC, TDS, DO, turbidity, alkalinity, hardness, TP, TAN, NO_2, NO_3, SO_4, Na and K were used in the study. Furthermore, an artificial neural network (ANN) structure is employed, comprising a number of neurons in the input layer, which correlate with the aforementioned water variables, and a single output variable. By this process, MLR modeling demonstrated high prediction performance with $R^2 = 1.0$, RMSE = 0.0025 and MAPE = 0.0296 accuracy values. A summary The MLR estimates and performances are presented in Table 4. It is sufficient to mention an MLR as a very useful tool for estimating the WQI.

ANN modeling utilized for the present paper also showed high prediction performance. Figure 8 shows the parity plots and regression models related to MLP-ANN estimation of WOIs. To ascertain the degree of error within the MLP-ANNs, the sum of squared errors was determined. Generally, there is obtained lower modeling inaccuracies obtained in all MLP-ANN methods. The present investigation shows that MLP-ANNs provide precise and reliable estimates for the WQ output variables of Aksu Creek.

Table 4 The performance results for modelling

Confidence intervals for beta		Performace for the test set		R^2	F	*P*-Value For F Test	Std. error Est.
Lower bound	Upper bound	Test values for WQ	Predicted values	1.0000	2.73E + 10	0.0000	1.22E−07
−43.147150	−43.136728	161.85	161.852100				
0.000441	0.000446	166.90	166.899700				
−0.003502	0.002367	185.33	185.326400				
−0.575238	−0.574990	203.56	203.563800				
7.362009	7.363194	181.77	181.765800				
1.104354	1.104367	117.32	117.319900				
0.000602	0.000830	143.67	143.675200				
−0.000011	0.000122	141.24	141.241900				
1.104175	1.104402	143.13	143.131500				
12.270570	12.270787	122.18	122.177700				
110.428562	110.444106	106.11	106.108500				
0.275835	0.276525	109.37	109.368300				
0.000406	0.000448	126.72	126.724200				
0.000484	0.000666	141.65	141.654400				
0.276122	0.278907	**RMSE**	0.0025112				
		MAPE	0.0296950				

Root mean square error (RMSE) and mean absolute percentage error (MAPE) are used in this research to interpret the model performance, so these criteria have been bolded in the table.

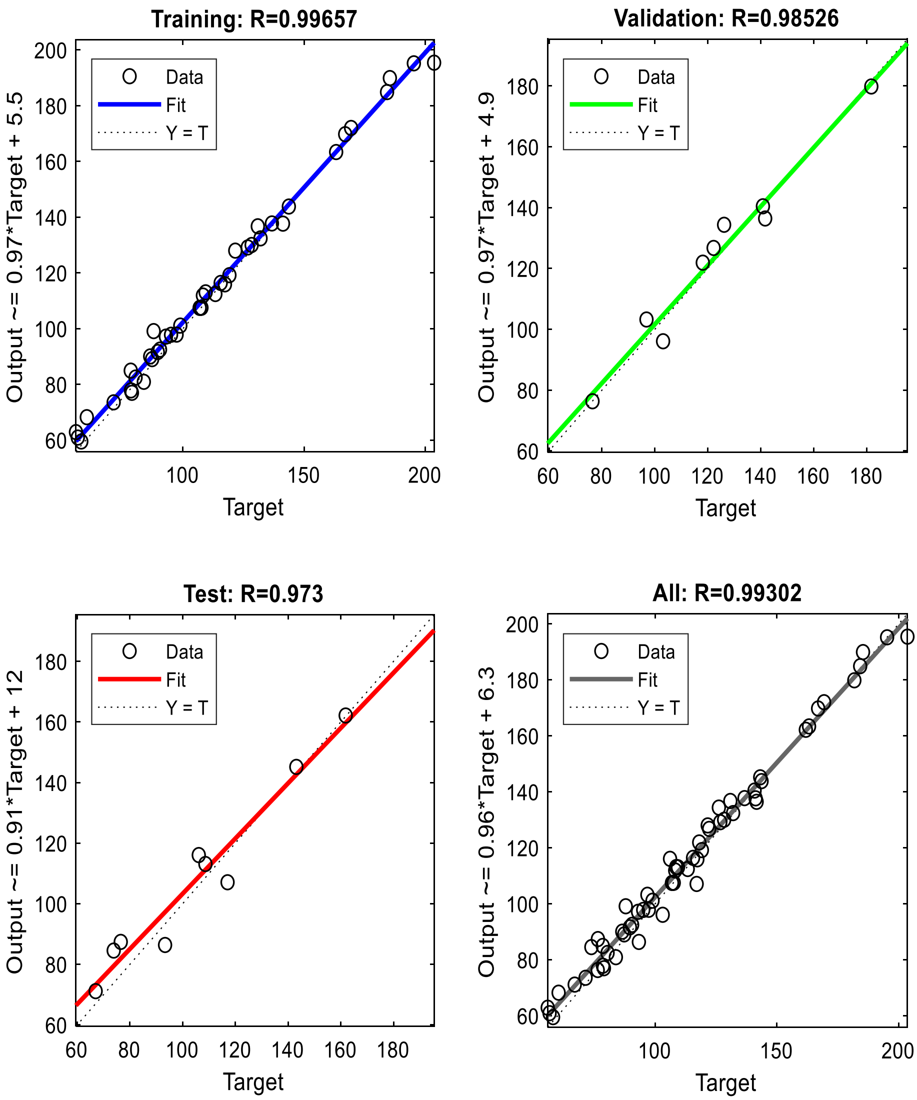


Fig. 8 The regression plot of the ANN model of **a** Training, **b** Validation, **c** Testing, and **d** All

Figure 8 also shows the data regression both individually and as a whole. The dotted line function in these plots is the objective function, i.e., the best mode determined by the neural network. At that, the correlation coefficient is equal to 1 (R = 1), or else it is less than 1. The function represents the function along the vertical axis whose line fits the data points using a neural network. This figure shows all three different datasets, including training, validation, and testing. In addition, the regression of each data category was found to be above 0.97 in all figures. A plot of the calculated WQI values compared to those predicted by ANN can be found in Fig. 8.

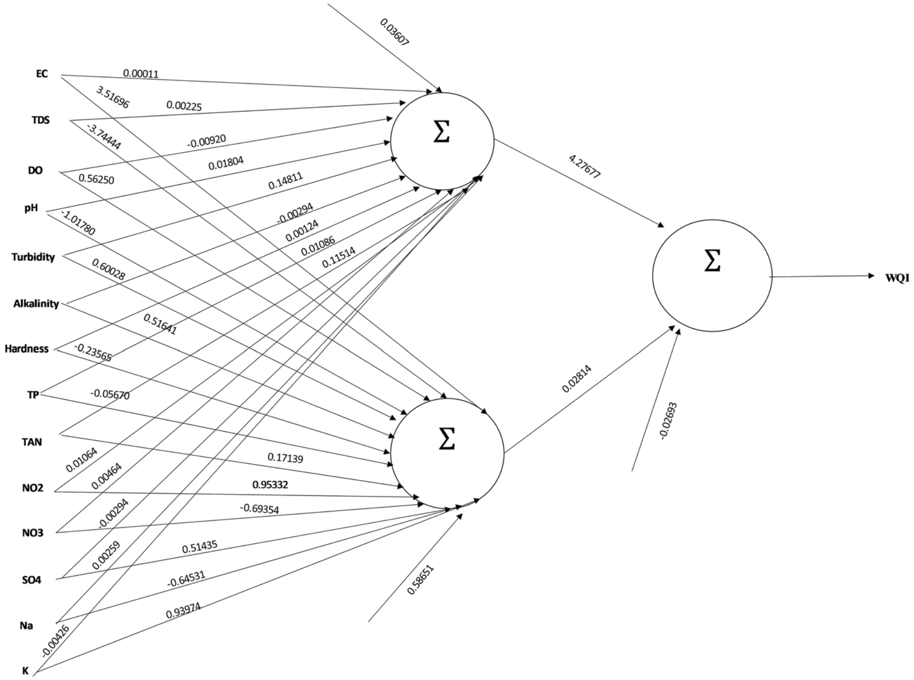


Fig. 9 Optimal multilayer perceptron neural network for WQI

Table 5 Performance statistics of the model in training, validation and testing for MLP-ANN (Levenberg-marudt)

Number of hidden node	MLP-ANN (Levenberg–Marudt)			
	Dataset	RMSE	MAPE	R^2
14–2–1–1	Validation	0.3187	0.3301	0.9852
	Training	0.1136	0.0904	0.9965
	Testing	0.3014	0.2141	0.9703
14–3–1–1	Validation	15.8757	13.5745	0.9601
	Training	30.5511	21.2610	0.6670
	Testing	23.9380	15.3488	0.6736
14–4–1–1	Validation	19.5297	13.8006	0.6700
	Training	19.3978	14.3178	0.8548
	Testing	27.9296	19.4390	0.8210
14–5–1–1	Validation	22.1110	20.0086	0.6678
	Training	30.9996	23.9498	0.6190
	Testing	29.2329	16.6684	0.6190

The format has been modified to a normal style

From Fig. 8, the predicted outcomes of the proposed ANN model is in good accordance with the computed results. ANN was observed as a powerful technique for WQI modeling with strong [R^2 = training (0.99)], testing (0.97) and validation (0.98).

The ANN implementation was performed utilizing an algorithm introduced in the MATLAB platform to select the minimum number of principal components to be utilized as input and the number of neurons in the hidden layer, resulting in the most suitable ANN model. Figure 9 is illustrates the proposed structure of the ANN.

An ANN-based validation evaluation (RMSE, MAPE and R^2) of the WQI from surface water in Aksu Stream (RMSE, MAPE, and R^2) is also given (Table 5). The error difference with the simulated outcome and the observed data set is utilized to assess the efficiency of algorithm. The number of hidden neurons was determined to be 2, 3, 4, and 5, but the model ANN with 2 hidden neurons showed the best performance. Therefore, 2 hidden neurons were selected in the study. The selected model 14-2–1-1 proved to be the most reliable in terms of R^2.

3.5 Comparative performance of water quality prediction models

The results of the study show that both MLR and MLP-ANN modeling are accurate and reliable options for calculating the WQI values of the RWQ. The R^2 values of both models ranged from 0.970 to 1.000. Moreover, the 70/15/15 partitioning of the dataset, comprising training, validation, and testing subsets, has been demonstrated to be an effective approach for estimating the WQI with the Levenberg–Marquardt algorithm (LMA). The optimum hidden neurons number was determined to be 2, 3, 4, or 5. The model ANN with 2 hidden neurons demonstrated the best performance. Therefore, 2 hidden neurons were chosen in the study. However, the MLR model outperformed the ANN used to estimate the WQI of the RWQ.

It is high that the WQI values estimated by MLR model are consistent with the observed values and therefore provide successful results in estimating WQI values. The results of present paper indicate that the MLR model is suitable according to the MLP-ANN model. With this process, MLR modeling showed high prediction performance with $R^2 = 1.0$, $\text{RMSE} = 0.0025$ and $\text{MAPE} = 0.0296$ (Table 6).

The MLR technique, which has a significant realistic success, functions as much simpler and less time-consuming (Sahoo & Jha, 2013), and our findings have been supported in the literature (Egbueri & Agbasi, 2022). In addition, a low RMSE value indicates that the model is performing well. The MAPE value is the set aside value used for a model to execute the forecast. The value of the smallest MAPE shows the good performance of the model (Olyaie et al., 2015).

Table 6 Comparative performance of water quality prediction models

	For WQI Methods				
	Levenberg–Marudt %70–%15-%15			Multiple linear regression (MLR)	The best machine learning and deep learning method
	Validation	Training	Testing		(Gaussian Process Regression (GPR))
RMSE	0.3187	0.1136	0.3014	0.0025	0.00362
MAPE	0.3301	0.0904	0.2141	0.0296	MAE = 0.00256 MSE = 0.00001
R^2	0.9852	0.9965	0.9703	1	0.99999

Root mean square error (RMSE), mean square error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) are used in this research to interpret the model performance because these criteria have been considered in most studies (Asadollah et al., 2021; Uddin et al., 2023; Zaghloul & Achari, 2022). The application of multiple regression algorithms has significantly demonstrated its efficacy and reliability in accurately predicting the WQI (Ahmed et al., 2019).

4 Conclusion

The main objective of present paper is identify the WQI to assess the availability of river water for sustainable management for drinking and irrigation purposes. The WQI results demonstrate that 80% of the analyzed river water samples exhibited satisfactory quality for public use, whereas 20% exhibited poor quality and were unsuitable for public consumption. In a similar viewpoint, the NPI classified the aforementioned option as being highly unsuitable. Nevertheless, the river water was limitedly suitable for irrigation purposes. Therefore, the majority of the river water resources within the basin is suitable for both human consumption and household usage. Pearson correlation analysis and principal component analysis were used to effectively rule out possible sources of contamination. It was determined that the water chemistry and quality of Aksu Creek were affected by a confluence of geogenic and anthropogenic factors.

As a further objective of present paper, ANN, ML and MLR models were comparisons to identify the precision of WQI for WQ prediction in the future. WQI values estimation was verified using ANN and MLR models. However, the MLR model outperformed the ANN and ML used to estimate the WQI of the surface water of Aksu Creek. Based on presented findings, multiple regression and ANN methods for estimating parameters used in determining RWQ and WQI, which is an important quality index, can be used. In this way, errors caused by factors such as expert opinions in WQIs will be eliminated, and purer results will be achieved.

The results of this study will contribute to the sustainable monitoring, assessment, and management of surface water resources, as this is the first estimation in this region. In addition, the knowledge gained here will make an important contribution to the diversification and growth of the global literature on WQ prediction. This fundamental study will shape similar studies. Finally, the information contained herein will provide important information to water managers, policy makers, and water researchers, particularly locally and globally, and will promote relative evaluation of models.

Acknowledgements This study was produced from the doctoral thesis prepared by Selda PALABIYIK. The authors are grateful to Prof. Dr. Erol EGRIOGLU, Prof. Dr. Eren BAS and PhD. Hakan IŞIK for their insightful comments and suggestions.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Selda Palabıyık and Tamer Akkan. The first draft of the manuscript was written by Selda Palabıyık and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK). The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability The datasets used and/or analyzed during the current study are available from the corresponding author on reasonablenrequest.

Declarations

Confict of interest The authors declare that they have no known competing fnancial interests or personal relationships that could appear to have infuenced the work reported in this paper.

References

Abaurrea, J., Asín, J., Cebrián, A. C., & García-Vera, M. A. (2011). Trend analysis of water quality series based on regression models with correlated errors. *Journal of Hydrology, 400*(3–4), 341–352.

Abuzir, S. Y., & Abuzir, Y. S. (2022). Machine learning for water quality classification. *Water Quality Research Journal, 57*(3), 152–164.

Acharya, S., Sharma, S., & Khandegar, V. (2018). Assessment of groundwater quality by water quality indices for irrigation and drinking in South West Delhi, India. *Data in Brief, 18*, 2019–2028.

Adeogun, A. I., Bhagawati, P., & Shivayogimath, C. (2021). Pollutants removals and energy consumption in electrochemical cell for pulping processes wastewater treatment: Artificial neural network, response surface methodology and kinetic studies. *Journal of Environmental Management, 281*, 111897.

Najah Ahmed, A., Othman, F. B., Afan, H. A., Ibrahim, R. K., Fai, C. M., Hossain, M. S., Ehteram, M., & Elshafie, A. (2019). Machine learning methods for better water quality prediction. *Journal of Hydrology, 578*, 124084. https://doi.org/10.1016/j.jhydrol.2019.124084

Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J. (2019). Efficient water quality prediction using supervised machine learning. *Water, 11*(11), 2210. https://doi.org/10.3390/w11112210

Akkan, T., & Mutlu, T. (2022). Assessment of Heavy Metal Pollution of Çoruh River (Turkey). *The Black Sea Journal of Sciences*, 12(1), 355-367. https://doi.org/10.31466/kfbd.1073227

Akkan, T., Mutlu, T., & Baş, E. (2022). Forecasting sea surface temperature with feed-forward artificial networks in combating the global climate change: The sample of Rize Türkiye. Ege Journal of Fisheries and Aquatic Sciences, 39(4), 311-315. https://doi.org/10.12714/egejfas.39.4.06

Akdağ, S. (2023). *Analysis of y-balance data used in sports sciences with machine learning algorithms* (Thesis). Karabük University Institute of Graduate Programs. Retrieved from. http://acikerisim.karabuk.edu.tr:8080/xmlui/handle/123456789/2960

Al-Ani, R., Al Obaidy, A., & Hassan, F. (2019). Multivariate analysis for evaluation the water quality of Tigris River within Baghdad City in Iraq. *The Iraqi Journal of Agricultural Science, 50*(1), 331–342.

Ali, O., Ishak, M. K., Ahmed, A. B., Salleh, M. F. M., Ooi, C. A., Khan, M. F. A. J., & Khan, I. (2022). On-line WSN SoC estimation using Gaussian Process regression: An adaptive machine learning approach. *Alexandria Engineering Journal, 61*(12), 9831–9848. https://doi.org/10.1016/j.aej.2022.02.067

Anli, A. S. (2003). *A study on the rainfall and flow characteristics of Aksu Creek watershed in Giresun province* (Master's Thesis). Graduate School of Natural and Applied Sciences, Ankara, 173p.

Asadollah, S. B. H. S., Sharafati, A., Motta, D., & Yaseen, Z. M. (2021). River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. *Journal of Environmental Chemical Engineering, 9*(1), 104599. https://doi.org/10.1016/j.jece.2020.104599

Barzegar, R., Adamowski, J., & Moghaddam, A. A. (2016). Application of wavelet-artificial intelligence hybrid models for water quality prediction: A case study in Aji-Chay River. *Iran. Stochastic Environmental Research and Risk Assessment, 30*(7), 1797–1819. https://doi.org/10.1007/s00477-016-1213-y

Bas, E., Egrioglu, E., & Karahasan, O. (2022). A Pi-Sigma artificial neural network based on sine cosine optimization algorithm. *Granular Computing, 7*(4), 813–820. https://doi.org/10.1007/s41066-021-00297-9

Bas, E., & Eğrioğlu, E. (2023). A new recurrent pi-sigma artificial neural network inspired by exponential smoothing feedback mechanism. *Journal of Forecasting*, 42(4), 802-812. https://doi.org/10.1007/s41066-024-00474-6

Basatnia, N., Hossein, S. A., Rodrigo-Comino, J., Khaledian, Y., Brevik, E. C., Aitkenhead-Peterson, J., & Natesan, U. (2018). Assessment of temporal and spatial water quality in international Gomishan Lagoon, Iran, using multivariate analysis. *Environmental Monitoring and Assessment, 190*, 1–17.

Bekas, G. K., Alexakis, D. E., & Gamvroula, D. E. (2021). Forecasting discharge rate and chloride content of karstic spring water by applying the Levenberg–Marquardt algorithm. *Environmental Earth Sciences, 80*(11), 404.

Bhattarai, A., Dhakal, S., Gautam, Y., & Bhattarai, R. (2021). Prediction of nitrate and phosphorus concentrations using machine learning algorithms in watersheds with different landuse. *Water, 13*(21), 3096.

Brack, W., Dulio, V., Ågerstrand, M., Allan, I., Altenburger, R., Brinkmann, M., Bunke, D., Burgess, R. M., Cousins, I., Escher, B. I., & Hernández, F. J. (2017). Towards the review of the European Union Water Framework Directive: Recommendations for more efficient assessment and management of chemical contamination in European surface water resources. *Science of the Total Environment, 576*, 720–737.

Brown, R. M., McClelland, N. I., Deininger, R. A., & O'Connor, M. F. (1972). A water quality index—crashing the psychological barrier. In *Indicators of Environmental Quality: Proceedings of a symposium held during the AAAS meeting in Philadelphia, Pennsylvania, December 26–31, 1971* (pp. 173–182). Springer.

Cansu, T., Bas, E., Egrioglu, E., & Akkan, T. (2024). Intuitionistic fuzzy time series forecasting method based on dendrite neuron model and exponential smoothing. *Granular Computing,* 9, 49. https://doi.org/10.1007/s41066-024-00474-6

Chauhan, S. S., & Trivedi, M. K. (2023). Artificial neural network-based assessment of water quality index (WQI) of surface water in Gwalior-Chambal region. *International Journal of Energy and Environmental Engineering, 14*(1), 47–61.

Che, L., & Wan, L. (2022). Water quality analysis and evaluation of eutrophication in a swamp wetland in the permafrost region of the lesser khingan mountains, China. *Bulletin of Environmental Contamination and Toxicology, 108*(2), 234–242.

Choden, Y., Chokden, S., Rabten, T., Chhetri, N., Aryan, K. R., & Al Abdouli, K. M. (2022). Performance assessment of data driven water models using water quality parameters of Wangchu river. *Bhutan. SN Applied Sciences, 4*(11), 290.

DelSole, T., & Tippett, M. (2022). *Statistical methods for climate scientists*. Cambridge University Press.

Deshmukh, K. K., & Aher, S. P. (2016). Assessment of the impact of municipal solid waste on groundwater quality near the Sangamner City using GIS approach. *Water Resources Management, 30*, 2425–2443.

Dimple, D., Rajput, J., Al-Ansari, N., & Elbeltagi, A. (2022). Predicting irrigation water quality indices based on data-driven algorithms: Case study in semiarid environment. *Journal of Chemistry, 2022*, 1–17. https://doi.org/10.1155/2022/4488446

Dong, Y., Wang, J., Niu, X., & Zeng, B. (2023). Combined water quality forecasting system based on multiobjective optimization and improved data decomposition integration strategy. *Journal of Forecasting, 42*(2), 260–287. https://doi.org/10.1002/for.2905

Egbueri, J. C., & Agbasi, J. C. (2022). Combining data-intelligent algorithms for the assessment and predictive modeling of groundwater resources quality in parts of southeastern Nigeria. *Environmental Science and Pollution Research, 29*(38), 57147–57171.

Egrioglu, E., Bas, E., & Chen, M. Y. (2024). A fuzzy Gaussian process regression function approach for forecasting problem. *Granular Computing*, 9(2), 47. https://doi.org/10.1007/s41066-024-00475-5

Egrioglu, E., Yolcu, U., & Bas, E. (2019). Intuitionistic high-order fuzzy time series forecasting method based on pi-sigma artificial neural networks trained by artificial bee colony. *Granular Computing, 4*, 639–654.

Elbeltagi, A., Azad, N., Arshad, A., Mohammed, S., Mokhtar, A., Pande, C., Etedali, H. R., Bhat, S. A., Islam, A. R. M. T., & Deng, J. (2021). Applications of Gaussian process regression for predicting blue water footprint: Case study in Ad Daqahliyah. *Egypt. Agricultural Water Management, 255*, 107052. https://doi.org/10.1016/j.agwat.2021.107052

Fernández del Castillo, A., Yebra-Montes, C., Verduzco Garibay, M., de Anda, J., Garcia-Gonzalez, A., & Gradilla-Hernández, M. S. (2022). Simple prediction of an ecosystem-specific water quality index and the water quality classification of a highly polluted river through supervised machine learning. *Water, 14*(8), 1235.

Gad, M., Saleh, A. H., Hussein, H., Farouk, M., & Elsayed, S. (2022). Appraisal of surface water quality of nile river using water quality indices, spectral signature and multivariate modeling. *Water, 14*(7), 1131.

Geng, Z., Hu, X., Han, Y., & Zhong, Y. (2019). A novel leakage-detection method based on sensitivity matrix of pipe flow: Case study of water distribution systems. *Journal of Water Resources Planning and Management, 145*(2), 04018094. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001025

Golfinopoulos, S. K., & Arhonditsis, G. B. (2002). Multiple regression models: A methodology for evaluating trihalomethane concentrations in drinking water from raw water characteristics. *Chemosphere, 47*(9), 1007–1018.

Gupta, S., & Gupta, S. K. (2021). Development and evaluation of an innovative enhanced river pollution index model for holistic monitoring and management of river water quality. *Environmental Science and Pollution Research, 28*, 27033–27046.

Hassan, M., Mehedi, H., Mahedi, M., Akter, L., Rahman, M. M., Zaman, S., et al. (2021). Efficient prediction of water quality index (WQI) using machine learning algorithms. *Human-Centric Intell. Syst., 1*, 86.

Heil, B. J., Crawford, J., & Greene, C. S. (2023). The effect of non-linear signal in classification problems using gene expression. *PLOS Computational Biology, 19*(3), e1010984. https://doi.org/10.1371/journal.pcbi.1010984

Hu, X., Han, Y., Yu, B., Geng, Z., & Fan, J. (2021). Novel leakage detection and water loss management of urban water supply network using multiscale neural networks. *Journal of Cleaner Production, 278*, 123611. https://doi.org/10.1016/j.jclepro.2020.123611

Huang, W., & Yang, Y. (2019). Water quality sensor model based on an optimization method of RBF neural network. *Computational Water, Energy, and Environmental Engineering, 9*(1), 1–11. https://doi.org/10.4236/cweee.2020.91001

Igwegbe, C. A., Mohmmadi, L., Ahmadi, S., Rahdar, A., Khadkhodaiy, D., Dehghani, R., & Rahdar, S. (2019). Modeling of adsorption of methylene blue dye on Ho-CaWO4 nanoparticles using response surface methodology (RSM) and artificial neural network (ANN) techniques. *MethodsX, 6*, 1779–1797.

Isiuku, B. O., & Enyoh, C. E. (2020). Pollution and health risks assessment of nitrate and phosphate concentrations in water bodies in South Eastern. *Nigeria. Environmental Advances, 2*, 100018.

Jafar, R., Awad, A., Hatem, I., Jafar, K., Awad, E., & Shahrour, I. (2023). Multiple linear regression and machine learning for predicting the drinking water quality index in Al-Seine lake. *Smart Cities, 6*(5), 2807–2827. https://doi.org/10.3390/smartcities6050126

Jain, A., Rallapalli, S., & Kumar, D. (2022). Cloud-based neuro-fuzzy hydro-climatic model for water quality assessment under uncertainty and sensitivity. *Environmental Science and Pollution Research, 29*(43), 65259–65275.

Kadam, A. K., Wagh, V. M., Muley, A. A., Umrikar, B. N., & Sankhua, R. N. (2019). Prediction of water quality index using artificial neural network and multiple linear regression modelling approach in Shivganga River basin, *India. Modeling Earth Systems and Environment, 5*(3), 951–962. https://doi.org/10.1007/s40808-019-00581-3

Khalid, S. (2019). An assessment of groundwater quality for irrigation and drinking purposes around brick kilns in three districts of Balochistan province, Pakistan, through water quality index and multivariate statistical approaches. *Journal of Geochemical Exploration, 197*, 14–26.

Khan, M. S. I., Islam, N., Uddin, J., Islam, S., & Nasir, M. K. (2022). Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. *Journal of King Saud University-Computer and Information Sciences, 34*(8), 4773–4781.

Kipruto, E., & Sauerbrei, W. (2022). Comparison of variable selection procedures and investigation of the role of shrinkage in linear regression-protocol of a simulation study in low-dimensional data. *PLoS ONE, 17*(10), e0271240.

Kostiuk, Y., Kryvoruchko, O., Tsiutsiura, M., Yerukaiev, A., & Rusan, N. (2022). Research of methods of control and management of the quality of butter on the basis of the neural network. In *2022 International conference on smart ınformation systems and technologies (SIST)*. Presented at the 2022 ınternational conference on smart information systems and Technologies (SIST). (pp. 1–6). https://doi.org/10.1109/SIST54437.2022.9945764

Kouadri, S., Elbeltagi, A., Islam, A. RMd. T., & Kateb, S. (2021). Performance of machine learning methods in predicting water quality index based on irregular data set: Application on Illizi region (Algerian southeast). *Applied Water Science, 11*(12), 190. https://doi.org/10.1007/s13201-021-01528-9

Krtolica, I., Cvijanović, D., Obradović, Đ, Novković, M., Milošević, D., Savić, D., Vojinović-Miloradov, M., & Radulović, S. (2021). Water quality and macrophytes in the Danube River: Artificial neural network modelling. *Ecological Indicators, 121*, 107076. https://doi.org/10.1016/j.ecolind.2020.107076

Laarne, P., Amnell, E., Zaidan, M. A., Mikkonen, S., & Nieminen, T. (2022). Exploring non-linear dependencies in atmospheric data with mutual information. *Atmosphere, 13*(7), 1046. https://doi.org/10.3390/atmos13071046

Lap, B. Q., Nguyen, H. D., Hang, P. T., Phi, N. Q., Hoang, V. T., Linh, P. G., & Hang, B. T. T. (2023). Predicting water quality index (WQI) by feature selection and machine learning: A case study of An Kim Hai irrigation system. *Ecological Informatics, 74*, 101991. https://doi.org/10.1016/j.ecoinf.2023.101991

Malinova, T., & Guo, Z. (2004). Artificial neural network modelling of hydrogen storage properties of Mg-based alloys. *Materials Science and Engineering: A, 365*(1–2), 219–227.

Mutlu, C., Eraslan Akkan, B., & Verep, B. (2018). The Heavy Metal Assessment Of Harsit Stream (Giresun, Turkey) Using Multivariate Statistical Techniques. *Fresenius Environmental Bulletin*, 27(12B), 9851-9858.

Mutlu, C., Bayraktar, F., & Verep, B. (2020). An Example of Sediment Quality Assessment Studies Boğacık Creek (Giresun). *Journal of Anatolian Environmental and Animal Sciences,* 5(3), 433-438. https://doi.org/10.35229/jaes.793295

Nacar, S., Mete, B., & Bayram, A. (2020). Estimation of daily dissolved oxygen concentration for river water quality using conventional regression analysis, multivariate adaptive regression splines, and TreeNet techniques. *Environmental Monitoring and Assessment, 192*, 1–21.

Nathan, N. S., Saravanane, R., & Sundararajan, T. (2017). Application of ANN and MLR models on groundwater quality using CWQI at Lawspet, Puducherry in India. *Journal of Geoscience and Environment Protection, 5*(03), 99.

Naubi, I., Zardari, N. H., Shirazi, S. M., Ibrahim, N. F. B., & Baloo, L. (2016). Effectiveness of water quality index for monitoring Malaysian river water quality. *Polish Journal of Environmental Studies, 25*(1), 231–239.

Nhantumbo, C., Carvalho, F., Uvo, C., Larsson, R., & Larson, M. (2018). Applicability of a processes-based model and artificial neural networks to estimate the concentration of major ions in rivers. *Journal of Geochemical Exploration, 193*, 32–40. https://doi.org/10.1016/j.gexplo.2018.07.003

Olyaie, E., Banejad, H., Chau, K. W., & Melesse, A. M. (2015). A comparison of various artificial intelligence approaches performance for estimating suspended sediment load of river systems: a case study in United States. *Environmental Monitoring and Assessment*, 187, 1–22. https://doi.org/10.1007/s10661-015-4381-1

Pan, B., Han, X., Chen, Y., Wang, L., & Zheng, X. (2022). Determination of key parameters in water quality monitoring of the most sediment-laden Yellow River based on water quality index. *Process Safety and Environmental Protection, 164*, 249–259.

Prasad, B., & Sangita, K. (2008). Heavy metal pollution index of ground water of an abandoned open cast mine filled with fly ash: A case study. *Mine Water and the Environment, 27*(4), 265–267.

Qi, J., Yang, L., & Liu, E. (2022). A holistic framework of water quality evaluation using water quality index (WQI) in the Yihe River (China). *Environmental Science and Pollution Research, 29*(53), 80937–80951.

Ragi, N. M., Holla, R., & Manju, G. (2019). Predicting water quality parameters using machine learning. In *2019 4th ınternational conference on recent trends on electronics, ınformation, communication & technology (RTEICT)* (pp. 1109–1112). IEEE.

Sahoo, S., & Jha, M. K. (2013). Groundwater-level prediction using multiple linear regression and artificial neural network techniques: A comparative assessment. *Hydrogeology Journal, 21*(8), 1865.

Said, S., & Khan, S. A. (2021). Remote sensing-based water quality index estimation using data-driven approaches: a case study of the Kali River in Uttar Pradesh, India. *Environment, Development and Sustainability, 23*, 18252–18277.

Santana, A. C., Barbosa, A. V., Yehia, H. C., & Laboissière, R. (2021). A dimension reduction technique applied to regression on high dimension, low sample size neurophysiological data sets. *BMC Neuroscience, 22*(1), 1. https://doi.org/10.1186/s12868-020-00605-0

Satish, N., Anmala, J., & Varma, M. R. R. (2022). Prediction of stream water quality in Godavari River Basin, India using statistical and artificial neural network models. *H2Open Journal, 5*(4), 621–641.

Sharifzadeh, M., Sikinioti-Lock, A., & Shah, N. (2019). Machine-learning methods for integrated renewable power generation: A comparative study of artificial neural networks, support vector regression, and Gaussian Process Regression. *Renewable and Sustainable Energy Reviews, 108*, 513–538. https://doi.org/10.1016/j.rser.2019.03.040

Singh, S. K., Srivastava, P. K., Singh, D., Han, D., Gautam, S. K., & Pandey, A. (2015). Modeling groundwater quality over a humid subtropical region using numerical indices, earth observation datasets, and X-ray diffraction technique: A case study of Allahabad district, India. *Environmental Geochemistry and Health, 37*, 157–180.

Sobhan Ardakani, S., Yari, A. R., Taghavi, L., & Tayebi, L. (2016). Water quality pollution indices to assess the heavy metal contamination, case study: Groundwater resources of Asadabad Plain in 2012. *Archives of Hygiene Sciences, 5*(4), 221–228.

Sun, A. Y., & Scanlon, B. R. (2019). How can Big Data and machine learning benefit environment and water management: A survey of methods, applications, and future directions. *Environmental Research Letters, 14*(7), 073001. https://doi.org/10.1088/1748-9326/ab1b7d

Sutadian, A. D., Muttil, N., Yilmaz, A. G., & Perera, B. (2016). Development of river water quality indices—a review. *Environmental Monitoring and Assessment, 188*, 1–29.

Toha, S. F., & Tokhi, M. O. (2008). MLP and Elman recurrent neural network modelling for the TRMS. In *2008 7th IEEE international conference on cybernetic intelligent systems* (pp. 1–6). IEEE.

Ucun Ozel, H., Gemici, B. T., Gemici, E., Ozel, H. B., Cetin, M., & Sevik, H. (2020). Application of artificial neural networks to predict the heavy metal contamination in the Bartin River. *Environmental Science and Pollution Research, 27*(34), 42495–42512. https://doi.org/10.1007/s11356-020-10156-w

Uddin, M. G., Nash, S., Rahman, A., & Olbert, A. I. (2023). Assessing optimization techniques for improving water quality model. *Journal of Cleaner Production, 385*, 135671. https://doi.org/10.1016/j.jclepro.2022.135671

Wang, X., & Yao, J. (2020). Linear regression estimation methods for inferring standard values of snow load in small sample situations. *Mathematical Problems in Engineering, 2020*, 1–10.

Withanachchi, S. S., Ghambashidze, G., Kunchulia, I., Urushadze, T., & Ploeger, A. (2018). Water quality in surface water: A preliminary assessment of heavy metal contamination of the Mashavera River, Georgia. *International Journal of Environmental Research and Public Health, 15*(4), 621.

Xu, Y., Yang, W., & Wang, J. (2017). Air quality early-warning system for cities in China. *Atmospheric Environment, 148*, 239–257. https://doi.org/10.1016/j.atmosenv.2016.10.046

Yan, R., & Wang, S. (2022). *Applications of machine learning and data analytics models in maritime transportation*. Institution of Engineering and Technology. Retrieved 13 January 2024, from http://www.scopus.com/inward/record.url?scp=85147391909&partnerID=8YFLogx

Yang, S., Deng, Z., Li, X., Zheng, C., Xi, L., Zhuang, J., Zhang, Z., & Zhang, Z. (2021). A novel hybrid model based on STL decomposition and one-dimensional convolutional neural networks with positional encoding for significant wave height forecast. *Renewable Energy, 173*, 531–543. https://doi.org/10.1016/j.renene.2021.04.010

Yıldız, S., & Karakuş, C. B. (2020). Estimation of irrigation water quality index with development of an optimum model: A case study. *Environment, Development and Sustainability, 22*, 4771–4786.

Zaghloul, M. S., & Achari, G. (2022). Application of machine learning techniques to model a full-scale wastewater treatment plant with biological nutrient removal. *Journal of Environmental Chemical Engineering, 10*(3), 107430. https://doi.org/10.1016/j.jece.2022.107430

Zare Farjoudi, S., & Alizadeh, Z. (2021). A comparative study of total dissolved solids in water estimation models using Gaussian process regression with different kernel functions. *Environmental Earth Sciences, 80*(17), 557. https://doi.org/10.1007/s12665-021-09798-x

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