



# Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast)

Saber Kouadri<sup>1</sup> · Ahmed Elbeltagi<sup>2</sup> · Abu Reza Md. Towfiqul Islam<sup>3</sup> · Samir Kateb<sup>4</sup>

Received: 3 July 2021 / Accepted: 21 October 2021 / Published online: 6 November 2021  
© The Author(s) 2021

## Abstract

Groundwater quality appraisal is one of the most crucial tasks to ensure safe drinking water sources. Concurrently, a water quality index (WQI) requires some water quality parameters. Conventionally, WQI computation consumes time and is often found with various errors during subindex calculation. To this end, 8 artificial intelligence algorithms, e.g., multilinear regression (MLR), random forest (RF), M5P tree (M5P), random subspace (RSS), additive regression (AR), artificial neural network (ANN), support vector regression (SVR), and locally weighted linear regression (LWLR), were employed to generate WQI prediction in Illizi region, southeast Algeria. Using the best subset regression, 12 different input combinations were developed and the strategy of work was based on two scenarios. The first scenario aims to reduce the time consumption in WQI computation, where all parameters were used as inputs. The second scenario intends to show the water quality variation in the critical cases when the necessary analyses are unavailable, whereas all inputs were reduced based on sensitivity analysis. The models were appraised using several statistical metrics including correlation coefficient (R), mean absolute error (MAE), root mean square error (RMSE), relative absolute error (RAE), and root relative square error (RRSE). The results reveal that TDS and TH are the key drivers influencing WQI in the study area. The comparison of performance evaluation metric shows that the MLR model has the higher accuracy compared to other models in the first scenario in terms of 1,  $1.4572 \times 10^{-08}$ ,  $2.1418 \times 10^{-08}$ ,  $1.2573 \times 10^{-10}\%$ , and  $3.1708 \times 10^{-08}\%$  for R, MAE, RMSE, RAE, and RRSE, respectively. The second scenario was executed with less error rate by using the RF model with 0.9984, 1.9942, 3.2488, 4.693, and 5.9642 for R, MAE, RMSE, RAE, and RRSE, respectively. The outcomes of this paper would be of interest to water planners in terms of WQI for improving sustainable management plans of groundwater resources.

**Keywords** Artificial intelligence · Water quality index · Modelling · Sensitivity analysis · Random forest

✉ Saber Kouadri  
kouadrisaber8@gmail.com

Ahmed Elbeltagi  
ahmedelbeltagy81@mans.edu.eg

Abu Reza Md. Towfiqul Islam  
towfiq\_dm@brur.ac.bd

Samir Kateb  
katebsamir30@gmail.com

<sup>2</sup> Agricultural Engineering Department, Faculty of Agriculture, Mansoura University, Mansoura 35516, Egypt

<sup>3</sup> Department of Disaster Management, Begum Rokeya University, Rangpur 5400, Bangladesh

<sup>4</sup> Research Laboratory in Exploitation and Development of Natural Resources in Arid Zones, University of Kasdi Merbah-Ouargla, PB 147 RP, 30000 Ouargla, Algeria

<sup>1</sup> Laboratory of Water and Environment Engineering in Sahara Milieu (GEEMS), Department of Civil Engineering and Hydraulics, Faculty of Applied Sciences, Kasdi Merbah University Ouargla, Ouargla, Algeria

## Introduction

Groundwater quality assessment and monitoring is a crucial task for sustainable optimal management of groundwater resources (Egbueri 2020; Kawo and Karuppappan 2018; Li et al. 2018; Islam et al. 2020a). The continuous growth of the population is directly associated with the growth of clean water demand (Dos Santos et al. 2017; Islam et al. 2017; Rahman et al. 2020). This demand makes the researchers more encouraged to develop new models for the prediction of water quality (Uddin et al. 2021). As a key element of the water cycle and drinking water resource, groundwater becomes an issue under a huge pressure worldwide (Ahmed et al. 2019; Saha et al. 2020). Thus, appraising water quality is of an urgent interest in recent times. Horton (1965) developed the first water quality index (WQI) in order to transform the several parameters containing water into one single number to describe the overall water quality. After that, several indices have been developed (Hossain and Patra 2020; Mukate et al. 2019; Islam et al. 2020b). The parameters involved in the calculation of the WQI have to be chosen carefully in order to get expressive results (Abbasi & Abbasi 2012). Various WQIs have been adopted by many researchers to assess the drinking suitability of groundwater and the quality river water (Islam et al. 2017; 2019; Kabir et al. 2021). However, the deterioration of water quality could be caused by many factors, e.g., inadequate proper sanitation, pollutants derived from industries and excessive use of fertilizer in agricultural practices, climate change, and poor groundwater management plan (Loecke et al. 2017; Alam et al. 2007; Trevett et al. 2005; Islam et al. 2018). On the other hand, the water quality appraisal involves some issues like sample collection at an enormous scale, testing in the laboratory, and data manipulation, which are mostly time-consuming processes and more expensive in terms of equipment, chemical, reagent, and human capital (Tiyasha et al. 2020). Besides, the subindex calculation is a time-taking process. Ongley (2000) found that water quality appraisal using traditional methods triggers losses in the economic aspect which influences the policy-making ability for groundwater quality management plans. In addition to this circumstance, the recent Corona pandemic made laboratories suffer from the lack of chemical analysis reactors used for water analysis after the remarkable reduction of the quantities of imported goods in several countries. Thus, to overcome these circumstances, it is necessary to use a promising and cost-effect tool for rapid and precise water quality appraisal. In such a case, the artificial intelligence (AI) model is an alternative option to generate models during the pandemic period that would help predict the overall quality of groundwater based on

the results of analyses that do not need expensive reactors or very developed measurement instruments.

The AI technique is a potential and robust multifunctioning tool in water-science-related fields (Babbar and Babbar 2017; Kisi et al. 2018; Kim et al., 2019; Bui et al. 2020; Abba et al. 2020; Hayder et al. 2021; Singha et al. 2021; Bilali et al. 2021). Several research scholars have employed AI techniques worldwide including random forest (RF), support vector machine (SVM), and artificial neural network (ANN) in different water-related studies. The RF model was applied for the groundwater quality prediction (Singha et al. 2021), flood susceptibility study (Towfiqul Islam et al. 2021), river water quality prediction (Asadollah et al. 2021), and so on. Likewise, the SVM model was adopted for predicting marine water quality (Deng et al. 2021) and wastewater treatment plant monitoring (Nourani et al. 2018), with different precision levels. ANN-based prediction models have been extensively used in different fields including heavy metal pollution prediction (Singha et al., 2020), wetland vulnerability (Islam et al. 2021), and water level forecasting (Zhu et al. 2020).

Apart from these cited works, many studies have been performed for the prediction of WQI by appraising the performance of various AI models. For example, Gazzaz et al. (2012) adopted the ANN method to forecast river water quality and got a precision level of more than 90% (R<sup>2</sup>). Wang et al. (2017) applied a swarm optimization-based support vector regression model to predict WQI. A study performed by Ahmed et al. (2019) implemented 15 AI algorithms for the prediction of WQI, where the regression model and classification model outperformed the other models. Bui et al. (2020) found the better predictive performance of hybrid AI models over the conventional models for predicting WQI with 4 conventional and 12 hybrid AI techniques. Recently, Singha et al. (2021) applied deep learning for predicting WQI with 3 traditional models and found that the deep learning model is a more robust and accurate tool than the traditional model in the prediction of groundwater quality. Valentini et al. (2021) introduced a new WQI equation for Mirim Lagoon and evaluated its suitability based on 154 samples collected over three years at seven sampling points in Mirim Lagoon. For forecasting monthly WQI values at the Lam Tsuen River in Hong Kong, Asadollah et al. (2021) proposed a new ensemble machine learning algorithm called extra tree regression (ETR). The efficiency of the ETR model is comparable to that of traditional standalone models such as support vector regression (SVR) and decision tree regression (DTR) (Asadollah et al. 2021). Based on parameters such as pH, dissolved oxygen, conductivity, turbidity, fecal coliform, and temperature, Hu et al. (2021) investigated the classification of water quality using machine learning algorithms such as decision tree (DT), k-nearest neighbor (KNN), logistic regression (LogR),

multilayer perceptron (MLP), and Naive Bayes (NB) and found that the DT algorithm outperformed other models with a classification accuracy of 99%.

From the aforementioned literature review, it is obvious that different AI models have been performed under various hydro geological conditions with different accuracy levels. In this context, additive regression (AR), M5P tree (M5P), random subspace (RSS), multilinear regression (MLR), and locally weighted linear regression (LWLR) were applied in our research to improve the reliability of water quality appraisal; however, these AI models are scarcely used in the hydrology field in the prediction of groundwater quality.

Besides, after thoroughly reviewing earlier literature, to the best of the author's knowledge, no previous studies have tested and verified the performance of these above-mentioned AI models for the prediction of groundwater quality. Thus, to close this gap, the current study used 8 ML-based WQI prediction models in Illizi region of the southeast, Algeria. Groundwater acts as a vital source of human use and consumption in the study area, and groundwater quality is mainly affected by human-induced pollution; hence, a thorough systematic appraisal of groundwater quality is necessary for this region. Additionally, no such scientific investigation has been done in the current study region. The WQI prediction using 8 ML techniques is a more robust tool than appraising it with any standalone tool. Hence, to achieve this aim, this study has developed two scenarios. The first scenario is developed using 8 models to predict the WQI using all the analyzed parameters as inputs variables to reduce the time consumption of calculations. The second scenario is constructed to reduce the number of inputs based on sensitivity analysis and to select the main parameters controlling water quality to predict the WQI in the critical case.

## Materials and methods

### Study area

#### General setting

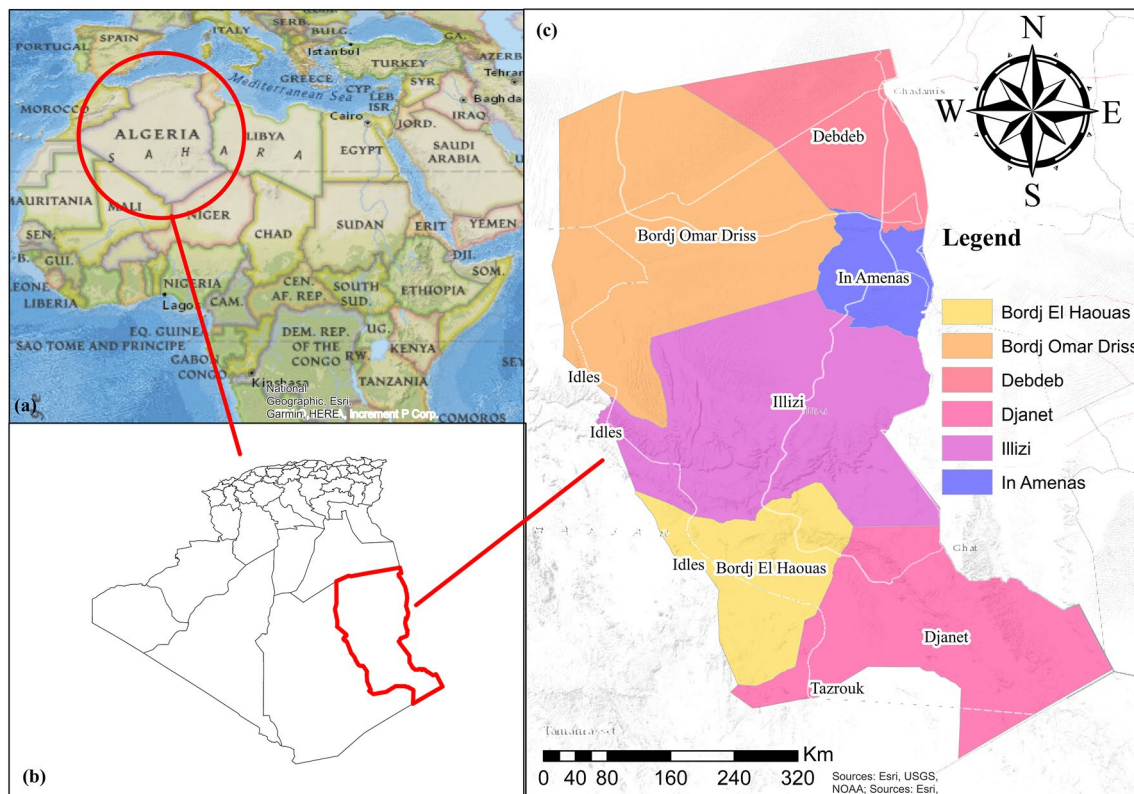
With 284,618 km<sup>2</sup> Illizi county is the third largest wilayah by area. It is located in the extreme southeast of Algeria, and it borders with three countries on a 1,233 km border with: Tunisia and Libya from the east and Niger from the south, where Ouargla county and Tamanrasset county border it from the north and the west, respectively (Kouadri and Samir 2021). Although the study area is very large, the climate has a homogenized distribution, with a very long hot summer and very short warm winter. The rains are extremely irregular. June is the hottest month of the year, while January is the coldest. Winds are generally light to moderate. Figure 1 presents the study area location.

### Hydrogeological settings

According to the authority of agricultural production in Saharian regions (CDARS), the hydrogeology of the Illizi area is distinguished by many aquifer deposits. The region has a large surface area, from which we can discern many aquifer horizons, such as Tassili's Cambro-Ordovician. Sandstone formations are traversed by a pattern of cracking and faults in addition to having a very low porosity. Tassili sandstones have a strong permeability due to these characteristics, which promote water circulation. The Devonian aquifer is located in Illizi and its surroundings, especially in the north, by exploitation from 250 to 1450 m in the Illizi and El Adeb Larach regions, respectively. The static level in regard to the land differs from one place to the next. In the high regions, it ranges from a few centimeters to a few meters; however, water is springing north and east of Illizi. The Carboniferous: This aquifer is extracted in the In Aménas area at various depths from 800 to 1100 m. The water drained by "lifting" is only used to keep the oil slicks under pressure and for irrigation; the static amount ranges between 200 and 300 m (Peterson 1985; Boudjema 1987; van de Weerd and Ware 1994; Kouadri and Kateb 2021).

The Continental Intercalaire (CI) aquifer system: It is found in the stratigraphic interval between the Triassic and the Albian summit. The Barremian and Albian, which are sandstone and sandy-clayey Lower Cretaceous continental deposits, form the majority of the aquifer layers. It drains the Triassic and Jurassic sandstone and clay-sandstone deposits in the Stah and In Aménas regions (where the CI is known as the Zaraitine and Taouratin Series), from Barremian and Albian to Deb Deb and Albian to BOD and Rhourd Nouss (Boudjema, 1987; Kouadri and Kateb 2021).

Medium-depth (400–500 m) drilling in, (T.F.T), Ohanet, and (B.O.D) capture the aquifer. Rhourd Nouss and the north of Deb Deb are comparatively wide (800–1200 m). The sheet's waters are gushing at Rhourd Nouss, Bordj Omar Driss, Tabankort, Maouar, Zemelet Mederba, and the north of Deb Deb; they are exploited by pumping at differing depths (from a few meters to 300 m) at Tinfouyé, Ohanet, the south of Deb Deb, and Stah; the useful tank's strength exceeds 250 m. Static pressure readings show that pressures can exceed 18 bars (e.g., Rhourd Nouss, gushing water). The Mio-Pliocene aquifer is made up of a rearrangement of sands and clays that stretches from the far northwest of the wilayah to the far northeast. Drilling 160–300 m deep is used to extract it in the Rhourd Nouss and El Hamra areas. The water is pumped out at different depths ranging from 80 to 100 m. Oued Djanet's Infero-Flux (Alluvial): The alluvial aquifer of Wadi Djanet was the region's largest and only water supply until the Cambro-Ordovician aquifer was discovered. It is a shallow aquifer spanning 17 km<sup>2</sup> of heterogeneous alluviums ranging from silty sand to small



**Fig. 1** Study area location map

pebbles resting on a twenty-kilometer stretch. Currently, 24 boreholes (including 1 well) have demanded it, with 9 boreholes and 1 well in operation. The water in this aquifer is of good quality, with dry residue ranging from 146 to 340 mg/l (Boudjema 1987; Montgomery 1993; Kouadri and Kateb 2021).

### Geological settings

According to the National Organization of Hydrographic Network (A.N.R.H), the city of Illizi is situated on a plateau land consisting of lower Devonian clay-sandstone and Emsian clay-sandstone deposits, as well as Quaternary. To the north, the middle to upper undifferentiated Devonian layers outcrop for around 12 km, before being surpassed much further north by Upper Devonian to Carboniferous layers created primarily by the Khenig sandstone, upper Famennian at Tournaisien, with average coastlines of 550–650 m and peaks exceeding 700 m. This disparity in elevation creates a landscape of canyons that favors river drainage and flow acceleration (Kouadri and Kateb 2021).

A plain landscape stretches from the northeast to the side of Tin-Tourha, east to the field of Halloufa, and south to the side of Gara Souf Mellene, passing through Adjnadjane to the Gara Tan Harab. This plain, which has an 8-km radius, is

mostly made up of post-Mesozoic (Quaternary) formations with an altimetry of 560–570 m. The lower Devonian formations, known as the Oued Samène formations, are located in the south and beyond 8 km (Siegenien). Their elevations are in excess of 700 m. With frank deformations and large fractures, these formations form a tectonic domain. Less significant faults run east–west as well.

In a strict sense, the geology of the state of Illizi is divided into two broad units: the crystalline basement and the sedimentary cover, which are lithostratigraphically distinct.

### Data collection

In order to prepare this work, the results of water analysis provided by the Directorate of Water Resources (DRE) of the State of Illizi were relied on. The presented data set consists of the results of analysis of 114 samples taken from 57 exploited wells of 6 different layers. The samples were taken between 1999 and 2020. The analyses of each sample consisted of physical elements represented by TDS, CE, and T°C and chemical elements represented in pH, Ca<sup>2+</sup>, Mg<sup>2+</sup>, Na<sup>+</sup>, K<sup>+</sup>, anions as Cl<sup>-</sup>, HCO<sub>3</sub><sup>-</sup>, SO<sub>4</sub><sup>2-</sup>, and pollution indicators as NO<sub>3</sub><sup>-</sup>. The different used models in this work to deal with this type of data considered a new challenge, where the efficiency



and performance of the models will be tested with an irregular data set.

### Calculation of water quality index (WQI)

WQI is one of the most widely used tools for determining the quality of water and its suitability for human use (El Baba et al. 2020; Reyes-Toscano et al. 2020; Zhang et al. 2020; Maskooni et al. 2020; Bahir et al. 2020). The following are the measures for estimating WQI: In the beginning, a weight must assign on to each factor ranging from 1 to 5, based on its significance and impact on drinking water and human health. Mineralization,  $SO_4^{2-}$ ,  $Cl^-$ , and  $NO_3^-$  are awarded the highest rating of “5” due to their direct impact on water quality and human health (Seifi, A. et al. 2020). The bicarbonates  $HCO_3^-$ , on the other hand, have a minimum value of “1”. Assigned weights, relative weights, and the limits required by WHO are shown in Table 1.

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

where “ $W_i$ ” is the relative weight. “ $w_i$ ” is the weight/parameter. “ $n$ ” is the number of parameters.

Then, a quality rating scale ( $q_i$ ) for each parameter is calculated based on Eq. (2).

$$q_i = (C_i/S_i) \times 100 \tag{2}$$

where “ $q_i$ ” is the quality rating. “ $C_i$ ” is the chemical concentration/water sample (mg/L). “ $S_i$ ” is the WHO drinking water quality standard (mg/L).

Furthermore, a subindex of the  $i$ th parameter is calculated using Eq. (3).

$$S_{Li} = q_i \times W_i \tag{3}$$

where

**Table 1** Assigned weights and relative weights of physicochemical parameters

Parameter	WHOs	Weight ( $w_i$ )	Relative weight ( $W_i$ )
pH	6.5–8.5	4	0.086957
TDS	1200 mg/l	5	0.108696
$Ca^{2+}$	200 mg/l	3	0.065217
$Mg^{2+}$	150 mg/l	3	0.065217
$K^+$	30 mg/l	2	0.043478
$Na^+$	200 mg/l	2	0.043478
$Cl^-$	250 mg/l	5	0.108696
$HCO_3^-$	380 mg/l	1	0.021739
$SO_4^{2-}$	400 mg/l	5	0.108696
TH	10 meq/l	3	0.065217
$NO_3^-$	50 mg/l	5	0.108696
Total		38	1

“ $S_{Li}$ ” is the subindex rating.

“ $q_i$ ” is the quality rating.

“ $W_i$ ” is the relative weight.

Finally, the water quality index calculated as follows:

$$WQI = \sum S_{Li} \tag{4}$$

### Artificial intelligence models

In this study, ANN, MLR, SVM, M5P tree, RF, LWLR, RS, and AR models were proposed for the estimation of WQI of ILLIZI groundwater. Data set was partitioned into two parts. 70% of the data were employed for calibration phase and the 30% of the data for verification purposes. Selection of dominant inputs parameters is one of the important parts in any AI-based modeling. MATLAB (R2018b) was used for the analysis of ANN and MLR, while the rest of models were developed using Waikato Environment for Knowledge Analysis (WEKA-version 3.8.4).

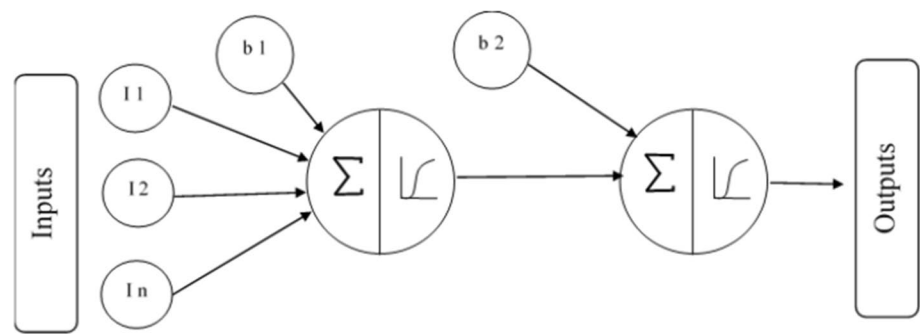
#### Artificial neural network (ANN)

Artificial neural network (ANN) is a system that inspired its dynamic functionality from the simulation of human nervous system. It was used for the first time by McCulloch and Pitts (1943), where the method works to create a relationship between inputs and outputs through assigned weights which plays the role of a mathematical memory (Elbeltagi et al. 2020c).

As seen in Fig. 2, the ANN is made up of three groups of layers: The hiding layers are intermediate layers between the independent input and dependent output layers where all the computations are performed, and the output layer outputs the result for the given inputs (Babae et al. 2021). The input layers' circles are denoted by the vector “ $i$ .” The secret neuron layers are represented by the middle circles. The “activation” nodes are represented by these circles, which are often referred to as the weights ( $W_s$ ). The final circle reflects the output sheet, which displays the water quality index's expected value (Elbeltagi et al., 2020a, b, c, d).

In order to optimize the performance of the network, training algorithm was founded; such as feed-forward back propagation algorithm. This algorithm works to minimize the error rate by calculating the difference between calculated and predicted values. Based on the error amount, new weights will be assigned in order to have better predicted results. Depending on the main factors affecting the performance of an ANN system, we can find the number of the hidden neurons and the activation function (Kouadri et al 2021; Elbeltagi et al. 2021a, b). In an attempt to select the optimal number of hidden neurons, an iterative algorithm had been used in order to plot the performance of the ANN

**Fig. 2** Architecture of ANN model



model in function, of MSE in training and validation phase versus the number of hidden neurons number. The optimal number of hidden neurons is the one that give the lowest rate error in both training and validation phases.

### Multi linear regression (MLR)

Multilinear regression analysis is considered as one of the simplest mathematical models. It is based on the linear relationships between inputs and outputs. In other words, it extracts the linear relationships between dependent and independent variables by involving a regression that is constant in the formula (Sihag et al., 2020). MLR work is based on the equation below:

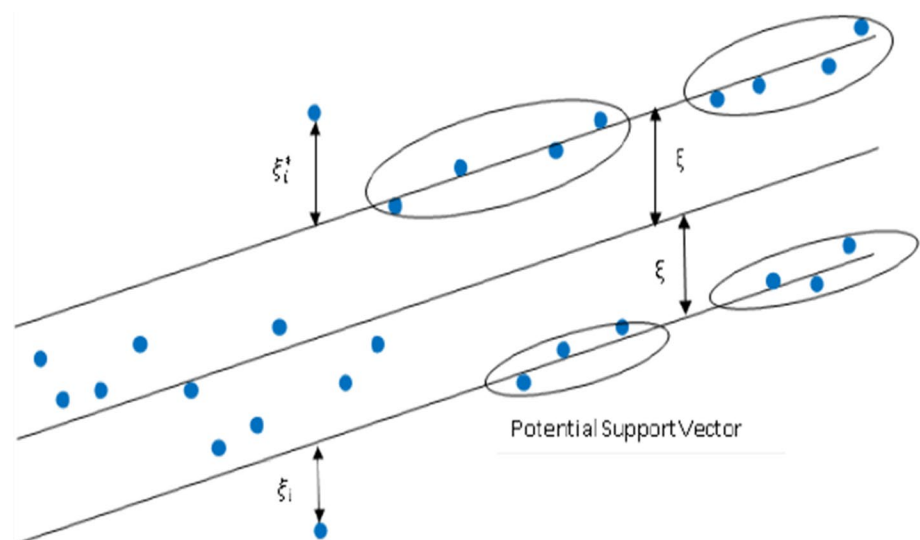
$$y = b_0 + b_1x_1 + b_2x_2 + \dots b_ix_i$$

where Y: the independent variable. B: the regression constant. X: the  $i_{th}$  predictor.

### Support vector regression (SVM)

Initially, support vector machine (SVM) was developed in order to help identify the distribution pattern of data samples in order to classify them into categories and help in making good decisions. The main idea of this method depends on using a set of studied sample points as supports to draw vectors separating the various classes in the studied data. When SVM was used to solve discontinuous issues, support vector machine regressor (SVMR) was created to deal with continuous issues. This system is characterized by many features that make it a permanent target for use in solving linear and nonlinear correlation problems (Elbeltagi et al., 2021a, b). Among its advantages is the dependence on structural risk minimization (SRM) principle which showed greater effectiveness than traditional empirical risk minimization (ERM). SRM is characterized by its great ability to reduce error, unlike some other methods, such as artificial neural networks that reduce error only in the results of training phase; this has given the SVM method a greater effectiveness in treating prediction Issues. Using the one-dimensional example in Fig. 3, SVR problem formulation is often best obtained from a geometrical perspective. The equation below represents

**Fig. 3** One-dimensional linear SVR



the continuous-valued equation that is being approximated (Awad& Khanna, 2015).

$$y = f(x) = \langle w, x \rangle + b = \sum_{j=1}^M w_j x_j + b, y, b \in \mathbb{R}, x, w \in \mathbb{R}^M$$

To simplify the mathematical notation for multidimensional data, multiply  $x$  by one and add  $b$  in the  $w$  vector to obtain the multivariate regression in equation below:

$$f(x) = \begin{bmatrix} w \\ b \end{bmatrix}^T \begin{bmatrix} x \\ 1 \end{bmatrix} = w^T x + bx, w \in \mathbb{R}^{M+1}$$

### M5P tree

M5P tree model has been presented by Quinlan (Quinlan 1992). It is a model that is a learner tree that deals with regression situations. The basis of this algorithm is based on dividing the overall problem into smaller problems by dividing the data, so that a multivariate model is constructed for each small problem and assigning linear regression functions into the final nodes. This method is characterized by its ability to deal with complex problems with many variables, with the condition that they are continuous class problems instead of discrete classes (Adnan et al. 2021; Sihag et al. 2020; Singh et al. 2017).

Figure 4 presents an M5P tree architecture. Depending on the amount of error calculated in each node, the M5P tree determines information about the criteria for dividing it. After studying the error, based on the standard deviation at the entrance to the node, the correction characteristic of this error is determined by testing all the characteristics of the studied node. The reduction of standard deviation is calculated by the following equation:

$$SDR = sd(K) - \sum \frac{|K_i|}{|K|} sd(K_i)$$

where  $K$ : a set of instances that attain the node.  $K_i$ : the subset of illustrations that have the  $i$  th product of the possible set.  $sd$ : the standard deviation.

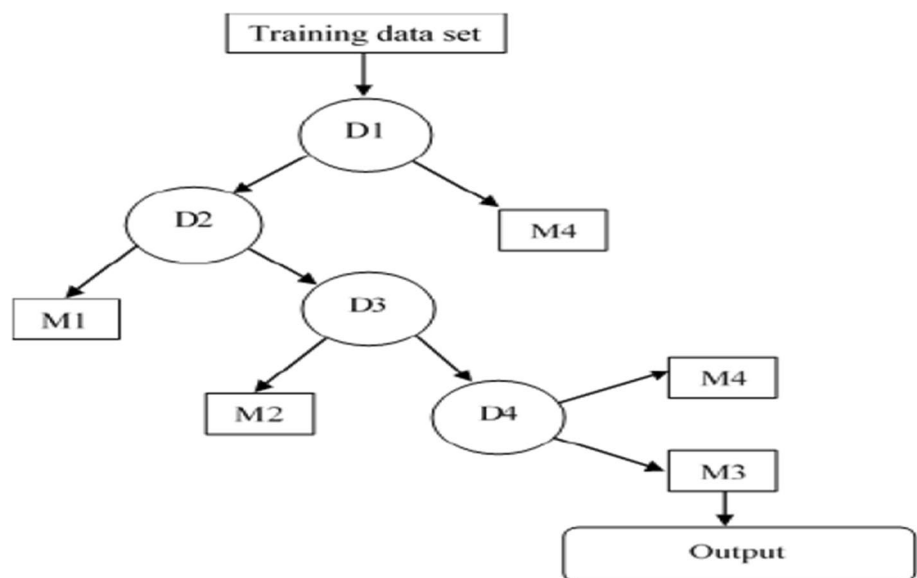
### Random forest(RF)

The random forest method was first introduced by Breiman (Breiman 2001). This method is considered as one of the machine learning systems that depend mainly on a group of decision trees targeting the middle separation of the target groups using individual trees. The construction of this method depends on two factors in the random regression of forests, namely, first the number of trees to be planted in the forest, and it is symbolized by the symbol  $(k)$ , second the number of variables specified at each node for the growth of the tree which is symbolized by  $(m)$  (Bournas et al. 2003; Pham et al. 2017; Sihag et al. 2019). The architecture of random forest model is presented in Fig. 5.

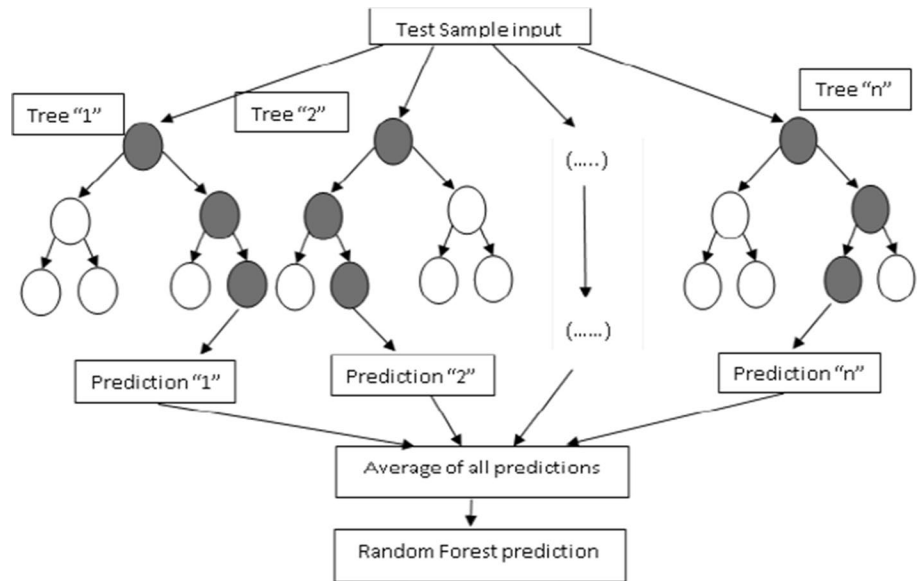
### Locally weighted linear regression (LWLR)

LWLR is a multivariate smoothing technique for fitting a regression surface to data. In a moving fashion, the dependent variable is smoothed as a function of the independent variables, similar to how a moving average for a time series is calculated. The fundamental structure is as follow, let  $x_i, \dots, x_{ip}, i = 1, \dots, n$ , be  $\langle n \rangle$  measurements of  $p$  independent variables, and let  $y_i, i = 1, \dots, n$  be measurements of the dependent variable. Assume that  $y_i = g(x_i) + \xi_i$  generates the results. We assume that the  $\xi_i$  are independent normal variables with mean 0 and variance  $\sigma^2$ , as in the most commonly

Fig. 4 M5P tree architecture model



**Fig. 5** Random forest architecture model



used regression framework. If  $g$  is a member of a parametric class of functions, such as polynomials, in the ordinary setting, we will assume that  $g$  is a smooth function of the independent variables, but in this case, we will only assume that  $g$  is a smooth function of the independent variables. We can approximate a large class of smooth functions with local fitting, well more than we might possibly predict from any one parametric class of functions (Cleveland and Devlin 1988; Kisi and Ozkan 2017).

**Random subspace (RSS)**

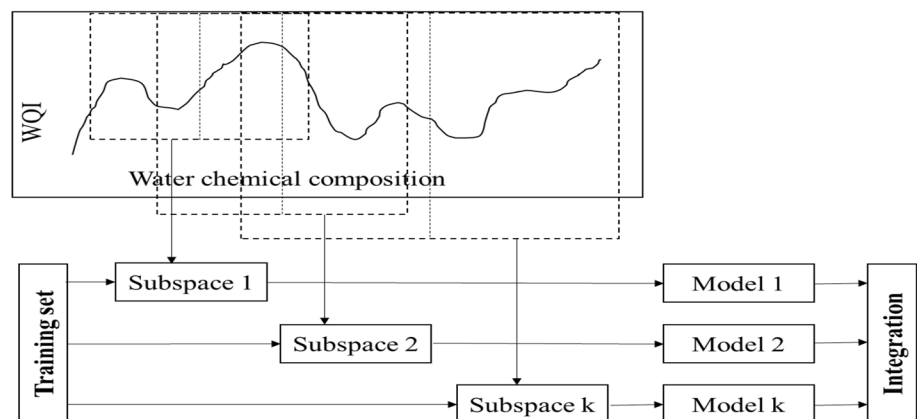
Ho (Ho 1998) was the first who implemented the RS model as a novel coupled algorithms for resolving nature issues based on artificial intelligence. This model uses combination and training of multiple classifier on altered feature space. The training basis of this model are the generated multiple training subsets for the classifiers (Ho 1998). The training set ( $x$ ), the base-classifier ( $w$ ), and the number of subspaces ( $L$ )

are the RS inputs (Kuncheva and Plumpton 2010; Luo et al. 2019; Garca-Pedrajas and Ortiz-Boyer 2008; Lai et al. 2006; Wang et al. 2018, 2015). This technique is highly advocated by (Pham et al. 2017) to avoid over-fitting problems and to deal with the most unnecessary data sets. Figure 6 presents the architecture of an RSS model.

**Additive regression (AR)**

Hastie and Tibshirani (1986) have introduced the generalized additive model (GAM). The GAM, an extension of the generalized linear model (GLM) (McCullagh and Nelder 1989), has several benefits over the latter model. The GAM assumes no form of dependence, unlike the GLM, which is based on the clear assumption of linearity of the parameters, and the relationship is not generally linear. Its theory is based on the use of a sum of nonlinear functions to model the response, which helps one to model the effect of each explanatory variable more specifically. In modeling the

**Fig. 6** Random subspace architecture model





effects of environmental variables, this precision makes it a common technique since these effects are often nonlinear and are difficult to specify parametrically (Peng and Dominici 2008; Bruneau and Grégoire 2011). The Jbilou and El Adlouni (2012) literature review described the capacity of the GAM in environmental health studies as a powerful technique to detect nonlinear associations between an environmental explanatory variable and a variable dependent on health. The equation used for this algorithm is written as:

$$g(E(y)) = \beta_0 + f_1x_1 + f_2x_2 + \dots + f_px_p + \epsilon \tag{5}$$

The nonlinear smooth functions are used in the estimation of this model's application.

$f_i(x_i)$ ,  $i = 1, \dots, p$ , for any single explanatory vector  $x_i$ .

Several data set split features are selected using the standard deviation error (SDR) as a parameter for the best characteristics to segment the data set into each node. The selected attribute is meant to reduce errors.

$$SD(Tree) - \sum \frac{Tree\_i}{Tree} * SD(Tree\_i) \tag{6}$$

where Tree (i) denotes the subset of examples with the product of the possible evaluations, SD() denotes the standard deviation of the statement. The stop criteria are the number of instances needed to reach a certain number or a small form value shift. All models' parameters used for modeling the WQI are clarified in Table 2.

### Sensitivity analysis

When there are several input variables, feature selection is one of the most important steps in developing a soft computing model to forecast and simulate engineering phenomena. There are many methods for determining the best possible combinations, including the best subset regression, shared knowledge, forward stepwise filtering, and so on. The best subset regression analysis was used in this research to find the best input combinations for the WQI model. Six statistical parameters were computed for this reason, including MSE, decision

coefficients ( $R^2$ ), adjusted  $R^2$ , Mallows' Cp (Gilmour 1996), Akaike's AIC, and Amemiya's PC (Claeskens and Hjort 2008).

### Model's performance criteria

Throughout the course of the analysis, actual WQI data and modeled values were compared. The following statistical metrics were chosen to determine the accuracy of models: root mean square error, coefficient of determination, and mean absolute error (Malone et al. 2017; Elbeltagi et al. 2020a, b, d).

All parameters are defined as follows:

$WQI_A^i$  is the calculated or actual value.

$WQI_P^i$  is predicted or foreseen value.

$WQI^-$  is the mean value of reference samples, and N is the total number of data points.

### Root mean square error

The sample standard deviation of the variations between expected and real values is known as the RMSE. It is given by:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (WQI_A^i - WQI_P^i)^2} \tag{7}$$

### Mean absolute error

The mean absolute error assesses the extent of errors in a series of predictions without taking their sign into account. It's an estimation of the absolute differences between expected and observed values over the test sample. It is defined as follows:

$$MAE = \frac{1}{N} \sum_{i=1}^N |WQI_P^i - WQI_A^i| \tag{8}$$

### Relative absolute error

The total absolute error is normalized by dividing it by the total absolute error of the basic indicator in the relative absolute error.

**Table 2** Parameters of machine learning algorithms used for WQI modeling

Model name	Description of parameters
Locally weighted linear regression (LWLR)	Batch size-100, KNN=0, Nearest neighbor search algorithm = linear NN
Random subspace (RSS)	Batch size-100, Classifier=REPTree, random seed-1, subspace size=0.5, numbers of executions slots = 1, number of iterations = 10
Support vector machine (SVM)	Batch size-100, C=0.1, kernel used = poly_kernel
Random forest (RF)	Batch size-100, bag Size percent = 100, max depth=0, numbers of executions slots = 1, number of iterations = 100, random seed = 1
M5P tree	Batch size-100, Minimum number of instances = 4
Additive regression (AR)	Batch size-100, Classifier = Decision-stump, shrinkage = 1, number of iterations = 10

$$RAE = \frac{WQI_A^i - WQI_P^i}{WQI_P^i} \times 100 \tag{9}$$

**Root relative squared error**

The total squared error is normalized by dividing it by the total squared error of the basic indicator in the relative squared error. The error is reduced to the same dimensions as the quantity being predicted by taking the square root of the relative squared error.

$$RRSE = \frac{\sqrt{\sum_{i=1}^N (WQI_P^i - WQI_A^i)^2}}{\sqrt{\sum_{i=1}^N (WQI_A^i - WQI^-)^2}} \tag{10}$$

**Results and discussion**

**Statistical analysis**

Table 3 presents the descriptive statistics for 114 ground-water samples. The correlation matrix is useful since it illustrates the importance of each parameter independently and their effect on the hydrochemistry mechanism (Helena et al. 2000; Khan 2011; Patil et al. 2020; Islam et al. 2017; 2020b). If the values of (r) are +1 or -1 in the Pearson’s correlation matrix (Table 4), they are treated as strong correlation coefficients values and signify total correlation, i.e., functional dependency, between two variables. If the values are closer to zero, it means there is no meaningful interaction between two variables at the p < 0.05 level (Singh et al. 2011; Patil et al. 2020). If r is bigger than 0.7, the parameters are highly.

**Table 3** Descriptive statistics for all input and output variables

Variable	Mean	SE Mean	StDev	Min	Q1	Median	Q3	Max
EC	2499	478	4250	63	530	2334	2887	28,570
TH	32.85	3.54	31.46	6.54	15.52	20.00	34.00	147.00
pH	7.8205	0.0581	0.5165	6.1600	7.6000	7.8500	8.1000	9.5000
TDS	1327.7	83.7	743.9	154.0	540.0	1509.0	1926.0	2668.0
HCO <sub>3</sub> <sup>-</sup>	123.53	6.09	54.12	3.33	98.00	116.69	142.00	270.00
Ca <sup>2+</sup>	141.76	9.78	86.94	10.00	60.00	132.26	212.00	350.00
Mg <sup>2+</sup>	73.31	5.75	51.08	4.80	29.16	75.15	100.93	235.00
Na <sup>+</sup>	163.6	11.9	105.7	4.0	44.0	214.2	249.0	394.0
K <sup>+</sup>	19.40	1.52	13.48	0.80	5.00	23.12	27.67	51.77
SO <sub>4</sub> <sup>2-</sup>	490.1	36.1	320.7	6.0	185.0	521.0	723.0	1330.3
Cl <sup>-</sup>	293.5	23.7	210.6	0.0	67.0	337.0	489.3	663.0
NO <sub>3</sub> <sup>-</sup>	22.32	3.24	28.82	0.00	0.36	11.00	41.00	164.83
WQI	105.64	6.17	54.82	23.45	61.21	107.73	134.40	278.10

**Table 4** The Pearson correlation coefficient between input and output variables

	EC	TH	pH	TDS	HCO <sub>3</sub> <sup>-</sup>	Ca <sup>2+</sup>	Mg <sup>2+</sup>	Na <sup>+</sup>	K <sup>+</sup>	SO <sub>4</sub> <sup>2-</sup>	Cl <sup>-</sup>	NO <sub>3</sub> <sup>-</sup>	WQI
EC	<b>1</b>												
TH	0.494	<b>1</b>											
pH	-0.065	0.099	<b>1</b>										
TDS	0.492	0.294	0.071	<b>1</b>									
HCO <sub>3</sub> <sup>-</sup>	0.370	0.480	-0.005	0.204	<b>1</b>								
Ca <sup>2+</sup>	0.340	0.124	0.060	0.883	0.010	<b>1</b>							
Mg <sup>2+</sup>	0.529	0.635	0.128	0.817	0.387	0.589	<b>1</b>						
Na <sup>+</sup>	0.443	0.089	0.034	0.930	0.116	0.780	0.639	<b>1</b>					
K <sup>+</sup>	0.442	-0.038	0.075	0.768	0.138	0.702	0.490	0.752	<b>1</b>				
SO <sub>4</sub> <sup>2-</sup>	0.541	0.388	0.070	0.947	0.172	0.818	0.784	0.849	0.723	<b>1</b>			
Cl <sup>-</sup>	0.195	0.045	0.094	0.869	0.002	0.805	0.688	0.848	0.635	0.689	<b>1</b>		
NO <sub>3</sub> <sup>-</sup>	0.767	0.231	-0.189	0.652	0.272	0.538	0.494	0.635	0.668	0.655	0.397	<b>1</b>	
WQI	0.637	0.659	0.089	0.908	0.357	0.744	0.911	0.766	0.608	0.905	0.695	0.665	<b>1</b>

correlated, and if  $r$  is between 0.4 and 0.7, the parameters are moderately correlated. A correlation matrix is used to consider the correlation between chemical parameters and WQI values in this study. The WQI which is the parameter focus on in this study has very weak correlations with pH and  $\text{HCO}_3^-$ , moderate correlations with EC, TH,  $\text{K}^+$ ,  $\text{Cl}^-$ , and  $\text{NO}_3^-$ , and strong correlations with TDS,  $\text{Ca}^{2+}$ ,  $\text{Mg}^{2+}$ ,  $\text{Na}^+$ , and  $\text{SO}_4^{2-}$ .

The Electrical conductivity of water (EC) has a negative correlation with the pH, and positive correlation of  $r < 0.4$  with  $\text{HCO}_3^-$ ,  $\text{Ca}^{2+}$  and  $\text{Cl}^-$ ,  $0.4 > r > 0.7$  with TH, TDS,  $\text{Mg}^{2+}$ ,  $\text{Na}^+$ ,  $\text{K}^+$ ,  $\text{SO}_4^{2-}$ , and WQI,  $r > 0.7$  with  $\text{NO}_3^-$  which has a strong correlation. The total hardness (TH) moderately correlated with  $\text{HCO}_3^-$ ,  $\text{Mg}^{2+}$ , and WQI, where no correlation exists with the rest of parameters. pH is observed to have no correlation with other parameters with an  $r$  coefficient ranged between  $-0.189$  and  $0.128$ . The correlation of TDS with  $\text{HCO}_3^-$  and  $\text{NO}_3^-$  is found to be weak and moderate, respectively, where all of  $\text{Ca}^{2+}$ ,  $\text{Mg}^{2+}$ ,  $\text{Na}^+$ ,  $\text{K}^+$ ,  $\text{SO}_4^{2-}$ ,  $\text{Cl}^-$ , and WQI have a strong correlation with it.  $\text{HCO}_3^-$  have no existing relationship with  $\text{Ca}^{2+}$ ,  $\text{Mg}^{2+}$ ,  $\text{Na}^+$ ,  $\text{K}^+$ ,  $\text{SO}_4^{2-}$ ,  $\text{Cl}^-$ ,  $\text{NO}_3^-$ , and WQI in the other hand the  $\text{Ca}^{2+}$ ,  $\text{Mg}^{2+}$ ,  $\text{Na}^+$ ,  $\text{K}^+$ ,  $\text{SO}_4^{2-}$ , and  $\text{Cl}^-$  are characterized with strong and moderate correlation with each other.

### Sensitivity analysis

In this section, a sensitivity analysis is performed to determine the most sensitive parameters in the considered combination set in predicting WQI. The selection of 2 best input combinations is mainly based on the nonlinear subset regression and sensitivity analysis. The advantage of using

the nonlinear sensitivity input variables selection approach to carefully determine the most relevant factors has been reported in several studies (Bui et al. 2020; Kisi et al. 2018; Liu et al. 2019). The best subset regression analysis for determining the best input combinations is presented in Table 5. We found that the best combination was TH / pH / TDS / Ca / Mg / Na / K / SO4 / Cl / NO3 and achieved high correlation and less statistical errors. Besides, all founded combinations generated good results.

Figure 7 presents the standardized coefficients of inputs variables for sensitivity analysis. We conclude that TH is identified as the most sensitive parameter. It has the highest standardized coefficient (0.453) among the considered parameters. After TH, the TDS earn the second place in the list of the most sensitive variables with standardized coefficient equal to 0.243. On the other hand,  $\text{SO}_4^{2-}$ ,  $\text{Cl}^-$ , and  $\text{NO}_3^-$  have 0.152, 0.176, and 0.135 as standardized coefficient, respectively, where the rest of parameters are considered as non-influential variables in predicting the WQI (Table 6).

Based on the results obtained from Tables 5, 6 and Fig. 7, and in order to achieve the objective targeted in this paper, two inputs combinations have been chosen: the first combination encloses all the parameters, where the second contains only the two strong influential inputs in predicting WQI which are TH and TDS.

### Evaluation of several ML models in WQI prediction

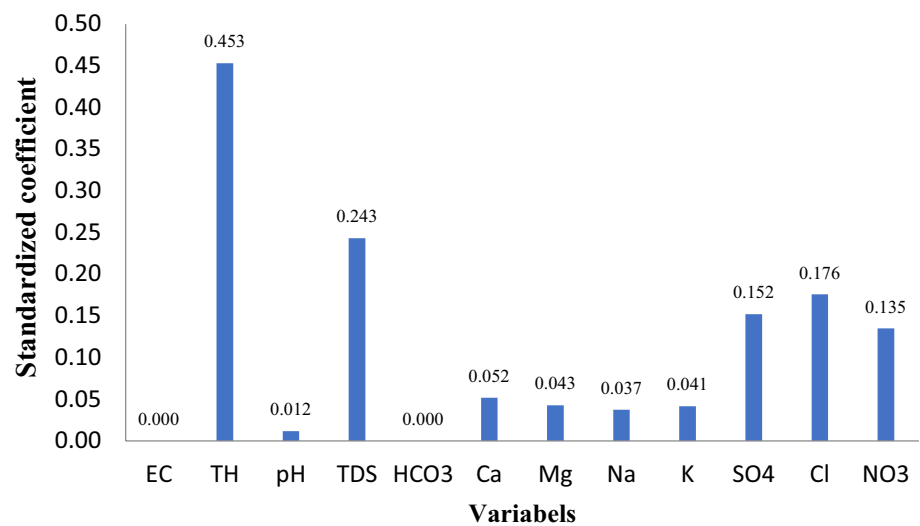
This study included the results of performing eight different methods of predicting the water quality parameter (WQI). The eight models used were as follows: MLR, ANN, M5P

**Table 5** The best subset regression analysis for determining the best input combinations to model WQI

Variables	MSE	R <sup>2</sup>	Adjusted R <sup>2</sup>	Mallows' Cp	Akaike's AIC	Schwarz's SBC	Amemiya's PC
Mg	516.623	0.830	0.828	9,529,247.202	495.512	500.251	0.174
TH / TDS	22.745	0.993	0.992	414,025.294	249.766	256.874	0.008
TH / TDS / NO3	6.137	0.998	0.998	110,180.065	147.222	156.700	0.002
TH / TDS / Cl / NO3	3.326	0.999	0.999	58,896.876	99.784	111.631	0.001
TH / TDS / HCO3 / Cl / NO3	2.121	0.999	0.999	37,030.353	65.174	79.391	0.001
TH / TDS / HCO3 / K / Cl / NO3	0.670	1.000	1.000	11,489.095	-24.980	-8.394	0.000
TH / pH / TDS / HCO3 / K / Cl / NO3	0.296	1.000	1.000	4977.243	-88.517	-69.562	0.000
TH / pH / TDS / HCO3 / Na / K / Cl / NO3	0.054	1.000	1.000	843.444	-222.230	-200.905	0.000
TH / pH / TDS / HCO3 / Ca / Na / K / SO4 / NO3	0.000	1.000	1.000	-56.278	-678.898	-655.203	0.000
TH / pH / TDS / Ca / Mg / Na / K / SO4 / Cl / NO3	0.000	1.000	1.000	-56.953	-997.271	-971.208	0.000
EC / TH / pH / TDS / Ca / Mg / Na / K / SO4 / Cl / NO3	0.000	1.000	1.000	-54.953	-995.851	-967.418	0.000
EC / TH / pH / TDS / HCO3 / Ca / Mg / Na / K / SO4 / Cl	0.004	1.000	1.000	12.000	-421.838	-393.405	0.000

The best model for the selected selection criterion is displayed in blue

**Fig. 7** The standardized coefficients of input variable for sensitivity analysis



**Table 6** The regression analysis for identifying the most effective parameters

Source	Value	Standard error	t	Pr >  t	Lower bound (95%)	Upper bound (95%)
EC	0.000	0.000	–	–	–	–
TH	0.453	0.000	62,811.849	< 0.0001	0.453	0.453
pH	0.012	0.000	3014.997	< 0.0001	0.012	0.012
TDS	0.243	0.000	3537.901	< 0.0001	0.243	0.243
HCO3	0.000	0.000	–	–	–	–
Ca	0.052	0.000	3362.499	< 0.0001	0.052	0.052
Mg	0.043	0.000	2496.162	< 0.0001	0.043	0.043
Na	0.037	0.000	1729.869	< 0.0001	0.037	0.037
K	0.041	0.000	5516.609	< 0.0001	0.041	0.041
SO4	0.152	0.000	5891.180	< 0.0001	0.152	0.152
Cl	0.176	0.000	9144.842	< 0.0001	0.176	0.176
NO3	0.135	0.000	20,334.708	< 0.0001	0.135	0.135

tree, SVM, RF, AR, RSS, and LWLR. Two combinations of variables were relied upon. The first configuration contained all the chemical elements used in the calculation of the water quality factor (WQI), while the second configuration was limited to only two components, namely the sum of dissolved salts (TDS) and water hardness (TH). These two elements were identified as the most controlling water quality index (WQI) based on sensitivity analysis results. It is worth mentioning that the Continental Intercalaire (CI) aquifer system received non-point sewage from different industries and agricultural inputs which highly attributed in deteriorating WQI. Generally, in groundwater studies, some factors affect the predictive precision of the models. However, there are some possible factors affecting the precision in this work could definitely be the low correlation values between pH, and WQI, TDS, and TH. It could also be caused by the enhanced pollution that is triggered by human inputs

on the side of the industry, which drastically decrease the precision of the models. This result is in good agreement with the studies done by Zhu and Heddam (2019).

Five statistical parameters were selected in order to determine the performance of the different models and compare them. Table 7 represents the results of the models depending on the first combination of inputs in the training and testing phases. As shown in Table 7, the MLR model was performed perfectly in the prediction process for the training phase, as it obtained a correlation coefficient of  $R = 1$  and the performance indicators were the smallest value by  $MAE = 1.4 \times 10^{-8}$ ,  $RMSE = 2.14 \times 10^{-8}$ ,  $RAE = 1.25 \times 10^{-10}\%$ , and  $RRSE = 3.17 \times 10^{-10}$ . It was followed directly by the ANN model which had a correlation coefficient of  $R = 0.9996$ ,  $MAE = 0.925$ ,  $RMSE = 1.4013$ ,  $RAE = 1.89\%$ , and  $RRSE = 0.024$ , whereas the lowest performing model in the training phase was the LWLR model with correlation

**Table 7** First input combination model result

Method	Phase	R	MAE	RMSE	RAE (%)	RRSE (%)
MLR	Training	1	1.4572E-08	2.1418E-08	1.2573E-10	3.1708E-08
	Testing	1	4.8359E-09	7.7286E-09	7.7919E-11	2.5293E-08
SVM	Training	0.9824	5.9522	10.4416	14.0075	19.1688
	Testing	0.9412	5.1607	11.3869	22.6331	37.2655
ANN	Training	0.9996	0.92517	1.4013	1.8914	2.44
	Testing	0.9987	1.40755	2.7351	1.6879	4.45
RSS	Training	0.9872	6.0524	9.2816	14.2432	17.0393
	Testing	0.9592	6.0023	9.9209	26.3241	32.4677
RF	Training	0.9974	2.7173	4.9228	6.3947	9.0373
	Testing	0.9931	2.287	3.9022	10.0301	12.7705
AR	Training	0.9893	6.3745	8.3915	15.0014	15.4051
	Testing	0.9947	2.4634	3.1467	10.8038	10.2979
LWLR	Training	0.9423	15.5235	18.3922	36.5317	33.7646
	Testing	0.9595	7.0762	8.6779	31.0341	28.3999
M5P tree	Training	0.9993	1.4857	2.029	3.4963	3.7249
	Testing	0.9923	2.4137	3.7963	10.5855	12.4239

coefficient  $R = 0.9423$ ,  $MAE = 15.52$ ,  $RMSE = 18.39$ ,  $RAE = 36\%$ , and  $RRSE = 33.76$ . Through the values of the performance of indicators, we note generally acceptable performance for the eight models. Yaseen et al. (2019) reported that RMSE is the most significant predictive numerical index for measuring the performance of the model in any data-mining modeling and time series forecasting. Our finding is in line with that of Yaseen et al. (2018), where the performance accuracy increases as the input variables are increased for the prediction of WQI.

For the test phase, the MLR model had the highest correlation value of  $R = 1$  and the smallest error indicators that closely approximated zero.  $MAE = 4.8 * 10^{-9}$ ,  $RMSE = 7.7 * 10^{-9}$ ,  $RAE = 7.7 * 10^{-11}\%$ , and  $RRSE = 2.5 * 10^{-10}$ . It was followed by the ANN model which obtained a correlation coefficient of  $R = 0.9987$  and MAE performance indicators = 1.4,  $RMSE = 2.7$ ,  $RAE = 1.68\%$ , and  $RRSE = 0.044$ , whereas the weakest performance was recorded in the testing phase when the SVM model consists of the correlation coefficient of  $R = 0.9412$  and MAE performance indicators = 5.16,  $RMSE = 11.386$ ,  $RAE = 22.6\%$ , and  $RRSE = 37.265$ . The predictive capability of the MLR model is definitely not surprising, because it is an evolving nonlinear system identification tool and has shown better predictive ability in many studies (Abba et al. 2020; El Bilali et al. 2021).

Table 8 represents the performance results of the eight models depending on the second configuration of inputs, which includes the elements TH and TDS. Through Table 8, we note that during the training phase, the best results were recorded on the RF model with a correlation coefficient of  $R = 0.9984$  and MAE performance indicators = 1.99,  $RMSE = 3.248$ ,  $RAE = 4.6\%$ , and  $RRSE = 5.96\%$ . The

ANN model came in second place with a correlation coefficient of  $R = 0.9969$ , MAE performance indicators = 2.46,  $RMSE = 3.88$ ,  $RAE = 3.3\%$ , and  $RRSE = 7.01\%$ . For the ANN model that provided the best performance based on the first combination of inputs, it regressed to the fifth place when using the second combination of inputs with correlation coefficient of  $R = 0.9958$  and performance indicators of  $MAE = 3.48$ ,  $RMSE = 4.98$ ,  $RAE = 4.23\%$ , and  $RRSE = 7.37$ . The weakest performance was recorded when using the LWLR model with a correlation coefficient of  $R = 0.9406$  and MAE performance indicators = 15.33,  $RMSE = 18.74$ ,  $RAE = 36.08\%$ , and  $RRSE = 34.42\%$ . For the test phase, the ANN model outperformed the rest of the models with a correlation coefficient of  $R = 0.9957$ , MAE performance indicators = 3.85,  $RMSE = 6.19$ ,  $RAE = 3.96\%$ , and  $RRSE = 9.35\%$ . Followed by the RF model which obtained a correlation coefficient of  $R = 0.9926$  and performance indicators of  $MAE = 2.15$ ,  $RMSE = 3.82$ ,  $RAE = 9.45\%$ , and  $RRSE = 12.51$ . The weakest performance was recorded on the MLR model with a correlation coefficient of  $R = 0.9325$ , MAE performance indicators = 7.94,  $RMSE = 11.04$ ,  $RAE = 12.51\%$ , and  $RRSE = 36.15\%$ . The main reason for the poor performance of the other models in both input combinations can be related to the inverse association, which was identified by the negative correlation between the observed pH concentration and the  $NO_3^-$  and  $HCO_3^{2-}$  parameters except for the TH and TDS values. This observation was analogous to the results reported by Zhu and Heddam (2019).

It is noted that the ensemble tree-based model such as RF outperformed all the other models with considerable accuracy in second input combination model due to its robustness deal with complicated pathways which can perform



**Table 8** Second input combination model result

Method	Phase	R	MAE	RMSE	RAE (%)	RRSE (%)
MLR	Training	0.9958	3.4879	4.9802	4.23	7.37
	Testing	0.9325	7.9427	11.0452	12.51	36.15
SVM	Training	0.9961	3.1387	5.1847	7.3863	9.5181
	Testing	0.9837	2.4438	5.5898	10.7176	18.2935
ANN	Training	0.9969	2.4618	3.8889	3.3028	7.01
	Testing	0.9957	3.8551	6.1968	3.9680	9.35
RSS	Training	0.9853	8.7124	11.346	20.503	20.8291
	Testing	0.9505	6.425	9.8266	28.1781	32.1591
RF	Training	0.9984	1.9942	3.2488	4.693	5.9642
	Testing	0.9926	2.1563	3.8228	9.4568	12.5107
AR	Training	0.9797	8.7182	11.1211	20.5167	20.4163
	Testing	0.9901	2.9985	4.4495	13.1505	14.5617
LWLR	Training	0.9406	15.3324	18.7494	36.082	34.4204
	Testing	0.9463	8.1329	9.9722	35.668	32.6355
M5P tree	Training	0.9963	3.3738	4.6778	7.9398	8.5875
	Testing	0.9856	2.8955	5.1745	12.6987	16.9343

predictions without requiring regular large datasets. Our results showed that the RF model is superior to other models in terms of precision. The key reason is that RF model can accommodate high-dimensional factors to improve water quality prediction accuracy, e.g., the inclusion of a monthly physicochemical variable in this study. Besides, according to the RF model, Castrillo and García (2020) reported a high prediction precision of the RF model compared to the MLR model. In addition, there is in line with earlier published works in classification problem (Salamand Islam 2020; Chen et al. 2020).

Figure 8 describes the dispersion of points representing the calculated WQI values against the predicted WQI values based on each model separately using the first set of inputs. Through Document 1, it appears that the MLR model is the most suitable for predicting the values of the water quality parameter due to the total match of the points with the perfect line 1:1. Fig. 9 describes the dispersion of points representing the calculated WQI values against the predicted WQI values based on each model separately using the second combination of inputs. The document shows a large dispersion of the MLR model points, while the RF model points are more ideally positioned compared to the rest of the models. The largest dispersion of points was in the case of using both the LWLR and RSS model, which indicates the poor performance of the two models in the case of using the second set of inputs.

The best model in each scenario is presented in Fig. 10 using scatter plot with smooth lines, blue for calculated WQI and purple for predicted WQI values, and markers present samples. Part (a) presents results of MLR model from the first scenario, where an optimal fitness is shown between calculated and predicted WQI values. In part (b), we notice

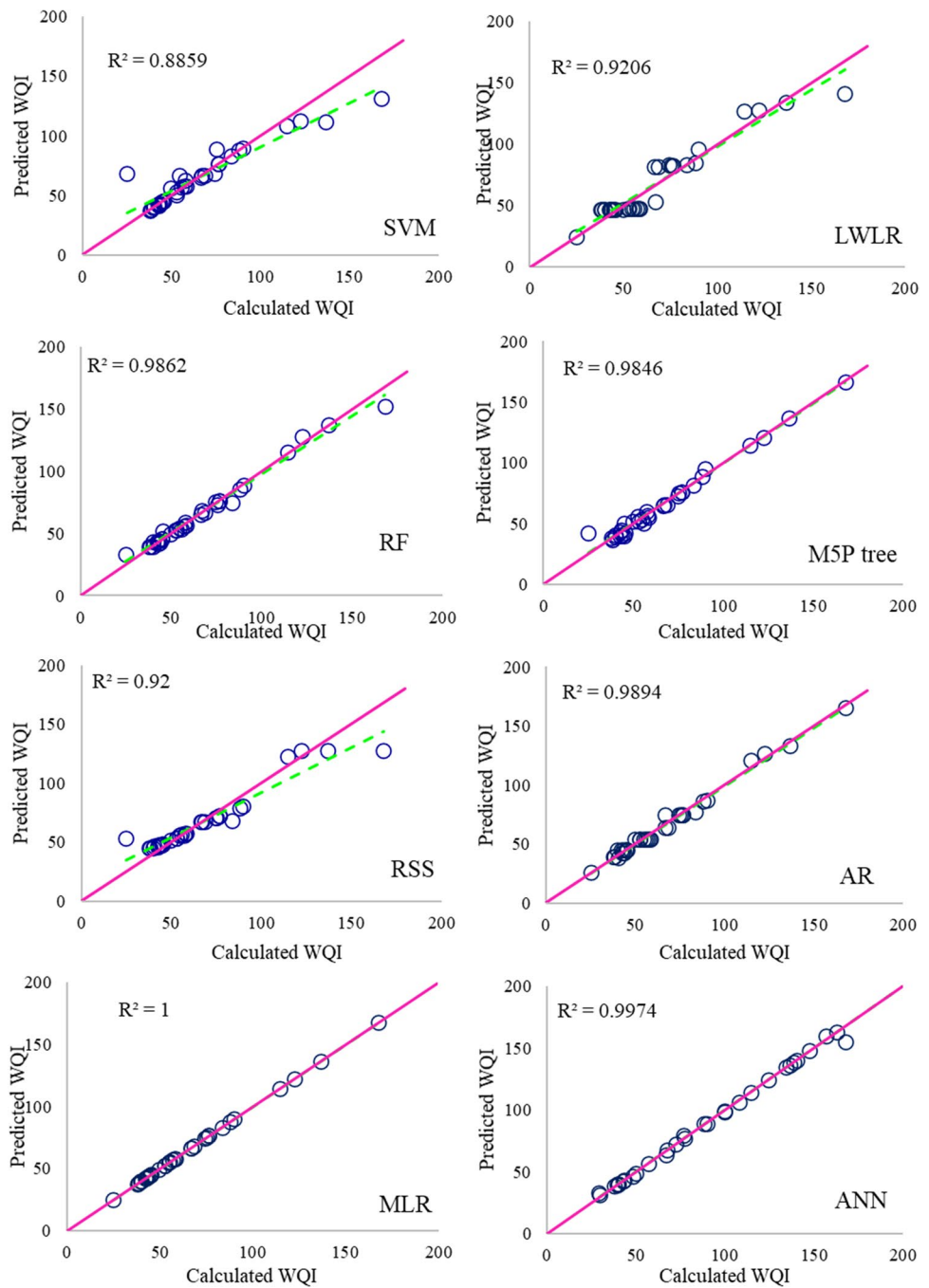
a presentation of RF model from the second scenario. The fitness in second scenario is not as in the first one, because a reduction in inputs had been made; this is why some predicted points does not fit with their calculated versus.

In addition to the aforementioned, the Wilcoxon rank-sum test was also relied upon in order to confirm the results mentioned in the previous paragraphs. This test is a non-parametric statistical news, used to compare two groups. The test calculates the difference between the pairs and the results are used to determine whether the two groups are statistically different from each other or not. In this work, this method was used to test the null hypothesis, which states that every two identical groups have the same continuous distribution. Some conditions must be met to apply this test, which is that the data should be from the same community and be associated. With random and independent data selection, Table 9 represents the *P* values for each model based on the first and second input configurations. In the case of using the first input combination, the highest probability was recorded when using the MLR and AR models with a value of  $P=0.9951$  for both models, whereas the lowest probability was recorded when using the RSS model with a value of  $P=0.4730$ .

The use of the second group of inputs witnessed noticeable changes in the performance of the models. The highest probability of match was recorded when using the RF model with a value of  $P=0.9951$ . Both the MLR and AR models reported significant decreases in performance with values of  $P=0.8588$  and  $P=0.7585$ , respectively. The weakest performance was recorded again when using the RSS model with a value of  $P=0.5519$ .

The physicochemical parameters chosen in the current study may also pose a drawback due to possible

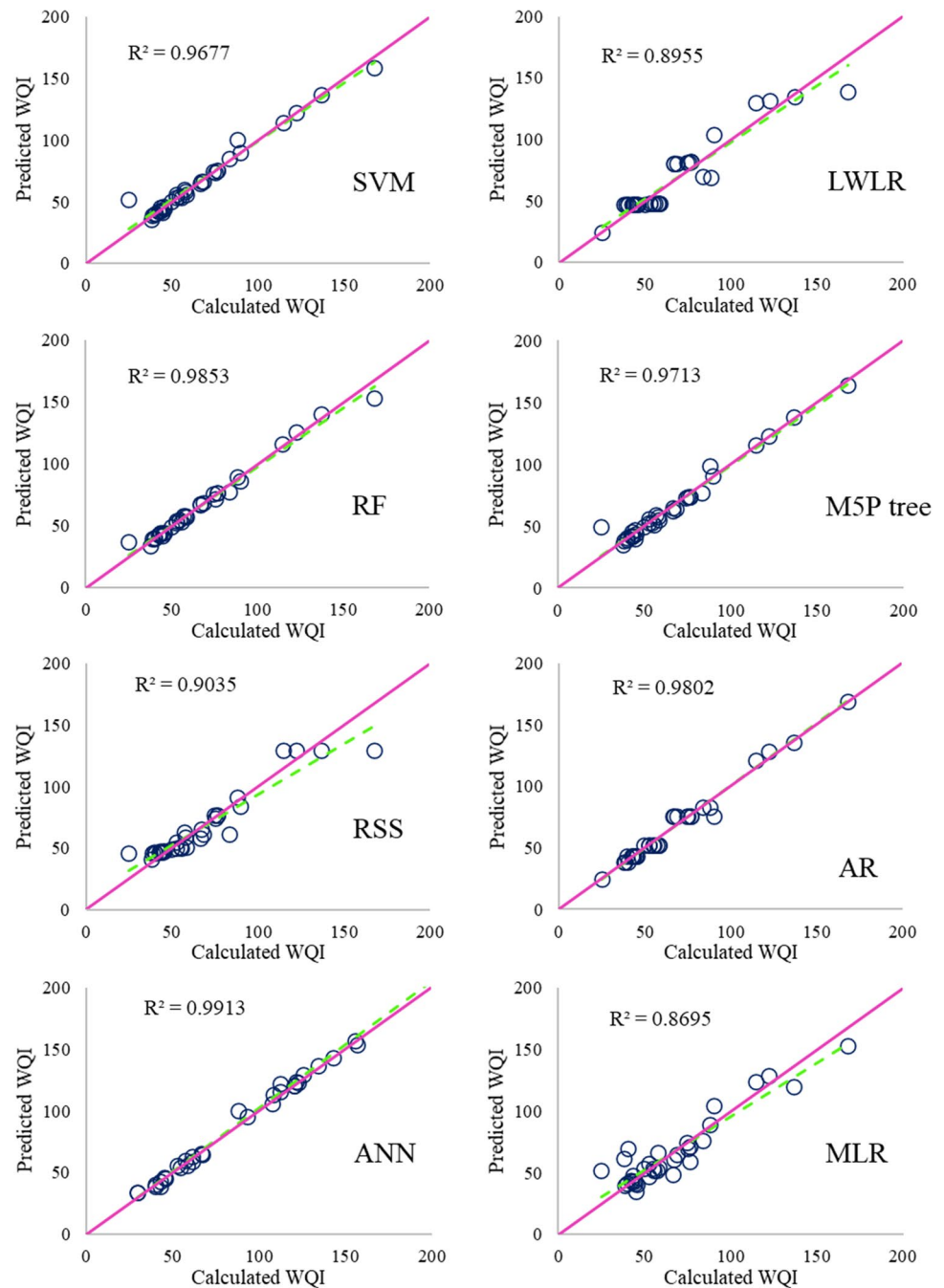
**Fig. 8** First input combination model predicted vs calculated WQI in testing phase



inadequate sampling. In addition to this, the uncertainty problem of the physical-based models in water quality modeling is inevitable and has been discussed in many studies (Bui et al. 2020; Kisi et al. 2018; Singha et al. 2021). Future research may add the use of different input physicochemical parameters to predict the WQI based on WHO guidelines, to compare with other standard indexes. The model presented here should be also appraised for other similar climatic and hydrological settings. However, given the noisy characteristics of this dataset, there was

still a threat that the models did not fit the data well, which might undermine the outcomes of the scenario forecasting. Besides, adding more influential physicochemical factors could also improve model fitting. For example, there may be other factors affecting TDS concentration besides climate and hydrogeological features (Islam et al. 2017). As the new development of machine learning models, it is promising for further work to predict contaminant concentration under the future pollution scenarios if the machine learning algorithm fits data well.

**Fig. 9** Second input combination model predicted vs calculated WQI in testing phase



As mentioned in previous studies, a key gap in water quality studies has been a lack of consideration of cross effects between explanatory variables, such as the cross-correlation between land covers and the cross-correlation between land cover and climate in influencing stream water quality (Islam et al. 2021). Machine learning models can use input variables and improving model predictive accuracy, which is an advantage over conventional statistical models. For example, it is likely that physicochemical factors showed effects with environmental variables and

groundwater pollution on groundwater water quality and the predictive accuracy can therefore be improved.

## Conclusion

In this work, the effectiveness of a group of artificial intelligence methods in predicting the water quality parameter in a dry desert environment was examined based on the 114 samples collected from six aquifers at different time

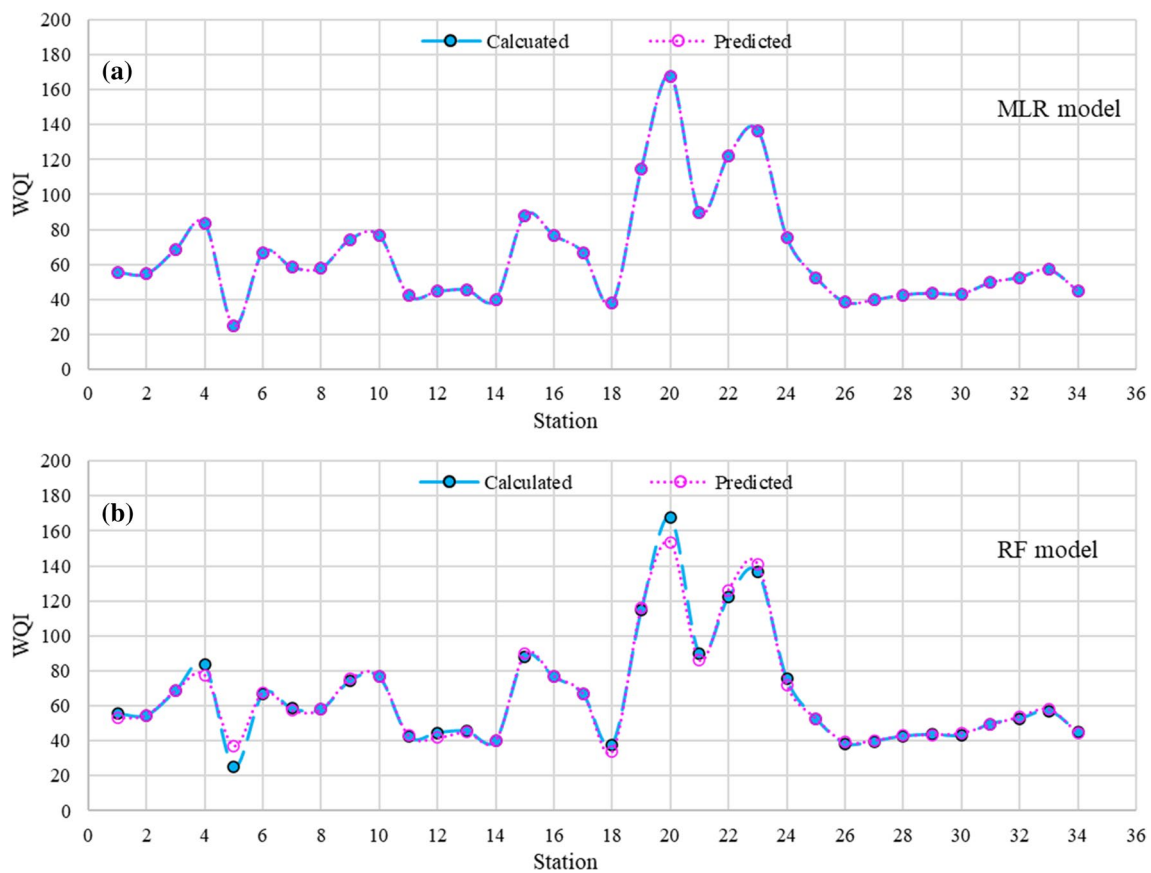


Fig. 10 Scatter plot of calculated and predicted WQI values in testing period using best models, a MLR model and b RF model

Table 9 Wilcoxon rank-sum test results (P values) of the optimal models in predicting WQI using first and second combination inputs

Model	First combination inputs	Second combination inputs
MLR	<b>0.9951</b>	0.8588
SVM	0.6992	0.9072
ANN	0.9755	0.8685
RSS	0.4730	0.5519
RF	0.9365	<b>0.9951</b>
AR	<b>0.9951</b>	0.7585
LWLR	0.6150	0.5894
M5P tree	0.8205	0.9462

P values at the significance level of 0.05

periods in Illizi state, southeast Algeria. Eight artificial intelligence models, namely MLR, ANN, SVM, M5P tree, RSS, RF, AR, and LWLR, were used, and their ability to predict was tested based on two scenarios and 2 different input combinations. The proposed two scenarios aim to solve two main problems. First, the classical computational method is replaced with modeling approach.

Second, when there is a lack or unavailability of data in critical cases, this study provides an alternative solution. The first set of inputs included all the chemical elements present in the water and used in calculating the WQI, while the second combination contained the controlling parameters of the water quality changes which were determined using the sensitivity analysis.

The sensitivity analysis shows that all the subset performed well as predictors in modeling WQI, where the selection of only two parameters as input in the second scenario was developed in order to propose an alternative solution for monitoring the WQI in the study area in critical cases. In second scenario, the modeling procedure showed that TDS and TH concentrations were the most vital determinants of WQI. The MLR model was performed perfectly in the first scenario because the calculation procedures of the WQI was linear, which make the task executed perfectly using MLR model with all the parameters as inputs. The reduction of the number of inputs affects directly the performance of models, where the aim in second scenario was constructing which model performed well in such conditions. RF models observed to be the best model in predicting WQI based on TH and TDS as parameters in the study area.

It is worth noting that MLR and RF algorithms generate robust results using a dataset covering the longer periods based on two scenarios. Thus, these algorithms might be useful for developing places that have very limited well-bore. Our results recommend that the RF algorithms could be a robust and cost-effective model to enhance groundwater quality management plans in an arid region in southeast Algeria. It is possible that this model is more applicable in developing countries where the costs of estimating several water quality variables are high and might be commonly restrictive. These outcomes could not be generalized and employed to other regions or other hydrogeological datasets, and these algorithms might not be optimal (i.e., most reliable) in all areas and under all conditions.

**Author contribution** SK, SK, and AE had the idea of this study and contributed to conceptualization and formal analysis. AE and SK implemented the modeling process; SK wrote the original draft; ARTI wrote the discussion and improved other sections; all co-authors performed review and editing and accepted the final draft.

**Funding** This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**Data Availability** The datasets generated during and/or analyzed during the current study are available from corresponding author based on reasonable request.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

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

## References

- Abba SI, Hadi SJ, Sammen SS, Salih SQ, Abdulkadir RA, Pham QB, Yaseen ZM (2020) Evolutionary computational intelligence algorithm coupled with self-tuning predictive model for water quality index determination. *J Hydrol* 587:124974
- Abbasi T, Abbasi SA (2012) Water quality indices. Elsevier
- Adnan RM, Khosravinia P, Karimi B, Kisi O (2021) Prediction of hydraulics performance in drain envelopes using Kmeans based multivariate adaptive regression spline. *Appl Soft Comput* 100:107008. <https://doi.org/10.1016/j.asoc.2020.107008>
- Ahmed U, Mumtaz R, Anwar H, Shah AA, Irfan R, García-Nieto J (2019) Efficient water quality prediction using supervised Machine Learning. *Water* 11(11):2210. <https://doi.org/10.3390/w11112210>
- Alam MJ, Islam MR, Muyen Z, Mamun M, Islam S (2007) Water quality parameters along rivers. *Int J Environ Sci Technol* 4(1):159–167
- 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* 9:104599. <https://doi.org/10.1016/j.jece.2020.104599>
- Babae M, Maroufpoor S, Jalali M, Zarei M, Elbeltagi A (2021) Artificial intelligence approach to estimating rice yield\*. *Irrig Drain*. <https://doi.org/10.1002/ird.2566>
- Babbar R, Babbar S (2017) Predicting river water quality index using data mining techniques. *Environ Earth Sci*. <https://doi.org/10.1007/s12665-017-6845-9>
- Bilali AE, Taleb A, Brouziyne Y (2021) Groundwater quality forecasting using machine learning algorithms for irrigation purposes. *Agricul Water Manage* 245:106625
- Boudjema, A., 1987, Evolution structurale du bassin petrolier «Triasique» du Sahara Nord Oriental (Algerie): Thèse a l'Université de Paris-Sud, Centre d'Orsay, 290 p.
- Bournas N, Galdeano A, Hamoudi M, Baker H (2003) Interpretation of the aeromagnetic map of Eastern Hoggar (Algeria) using the Euler deconvolution, analytic signal and local wavenumber methods. *J African Earth Sci* 37:191–205. <https://doi.org/10.1016/j.jafrearsci.2002.12.001>
- Bruneau, B. and Grégoire, F., 2011. Étude de la distribution spatiale des données d'abondance de maquereau bleu (*Scomber scombrus*) et de capelan (*Mallotus villosus*) des relevés d'hiver aux poissons de fond des Divisions 4VW de l'OPANO à l'aide de modèles additifs généralisés. Rapport technique canadien des sciences halieutiques et aquatiques, 2930, vi + 22.
- Bui DT, Khosravi K, Tiefenbacher J et al (2020) Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Sci Total Environ*. <https://doi.org/10.1016/j.scitotenv.2020.137612>
- Castrillo M, García AL (2020) Estimation of high frequency nutrient concentrations from water quality surrogates using machine learning methods. *Water Res* 172:115490. <https://doi.org/10.1016/j.watres.2020.115490>
- Chen K, Chen H, Zhou C, Huang Y, Qi X, Shen R, Liu F, Zuo M, Zou X, Wang J, Zhang Y (2020) Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water Res* 171:115454
- Claeskens G, Hjort N (2008) Model selection and model averaging. Cambridge University Press
- Cleveland WS, Devlin SJ (1988) Locally weighted regression: an approach to regression analysis by local fitting. *J Am Stat Assoc* 83:596–610. <https://doi.org/10.1080/01621459.1988.10478639>
- Deng T, Chau KW, Duan HF (2021) Machine learning based marine water quality prediction for coastal hydro-environment management. *J Environ Manage* 284:112051
- Dos Santos S, Adams EA, Neville G, Wada Y, de Sherbinin A, Mulin Bernhardt E, Adamo SB (2017) Urban growth and water access in sub-Saharan Africa: progress, challenges, and emerging research directions. *Sci Total Environ* 607–608:497–508. <https://doi.org/10.1016/j.scitotenv.2017.06.157>
- Egbueri JC (2020) Groundwater quality assessment using pollution index of groundwater (PIG), ecological risk index (ERI) and hierarchical cluster analysis (HCA): a case study. *Groundw Sustain Dev* 10:100292. <https://doi.org/10.1016/j.gsd.2019.100292>
- Elbeltagi A, Deng J, Wang K, Hong Y (2020a) Crop Water footprint estimation and modeling using an artificial neural



- network approach in the Nile Delta. *Egypt Agric Water Manag* 235:106080. <https://doi.org/10.1016/j.agwat.2020.106080>
- Elbeltagi A, Deng J, Wang K, Malik A, Maroufpoor S (2020b) Modeling long-term dynamics of crop evapotranspiration using deep learning in a semi-arid environment. *Agric Water Manag* 241:106334. <https://doi.org/10.1016/j.agwat.2020.106334>
- Elbeltagi A, Rizwan M, Malik A, Mehdinejadani B, Srivastava A, Singh A, Deng J (2020c) The impact of climate changes on the water footprint of wheat and maize production in the Nile Delta. *Egypt Sci Total Environ* 743:140770. <https://doi.org/10.1016/j.scitotenv.2020.140770>
- Elbeltagi A, Zhang L, Deng J, Juma A, Wang K (2020d) Modeling monthly crop coefficients of maize based on limited meteorological data : a case study in Nile Delta. *Egypt Comput Electron Agric* 173:105368. <https://doi.org/10.1016/j.compag.2020.105368>
- Elbeltagi A, Kumari N, Dharpure JK, Mokhtar A, Alsafadi K, Kumar M, Mehdinejadani B, Ramezani Etedali H, Brouziyne Y, Tawfiqul Islam ARM, Kuriqi A (2021a) Prediction of combined terrestrial evapotranspiration index (Ctei) over large river basin based on machine learning approaches. *Water (switzerland)* 13:1–18. <https://doi.org/10.3390/w13040547>
- Elbeltagi A, Pande CB, Kouadri S, Islam ARM (2021) Applications of various data-driven models for the prediction of groundwater quality index in the Akot basin, Maharashtra, India. *Environ Sci Pollut Res*, pp 1–15
- García-Pedrajas N, Ortiz-Boyer D (2008) Boosting random subspace method. *Neural Netw* 21(9):1344–1362
- Gazzaz NM, Yusoff MK, Aris AZ, Juahir H, Ramli MF (2012) Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors. *Mar Pollut Bull* 64:2409–2420
- Gilmour SG (1996) The interpretation of Mallows's Cp-statistic. *J Royal Statist Soc: D (The Statistician)* 45(1):49–56
- Hayder G, Kurniawan I, Mustafa HM (2021) Implementation of machine learning methods for monitoring and predicting water quality parameters. *Biointerf Res Appl Chem* 11(2):9285–9295
- Helena B, Pardo R, Vega M, Barrado E, Fernandez JM, Fernandez L (2000) Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principal component analysis. *Water Res* 34(3):807–816
- Ho TK (1998) The random subspace method for constructing decision forests. *IEEE Trans Pattern Anal Mach Intell* 20(8):832–844
- Hossain M, Patra PK (2020) Water pollution index – A new integrated approach to rank water quality. *Ecol Indic* 117:106668. <https://doi.org/10.1016/j.ecolind.2020.106668>
- Hu C, Zhao D, Jian S (2021) Corrected Proof 1 © 2021 1–20. <https://doi.org/10.2166/ws.2021.082>
- Islam ARMT, Ahmed N, Bodrud-Doza M, Chu R (2017) Characterizing groundwater quality ranks for drinking purposes in Sylhet district Bangladesh, Using Entropy Method, Spatial Autocorrelation Index and Geostatistics. *Environ Sci Pollut Res* 24(34):26350–26374. <https://doi.org/10.1007/s11356-017-0254-1>
- Islam ARMT, Shen S, Haque MA et al (2018) Assessing groundwater quality and its sustainability in Joypurhat district of Bangladesh using GIS and multivariate statistical approaches. *Environ Dev Sustain* 20(5):1935–1959. <https://doi.org/10.1007/s10668-017-9971-3>
- Islam ARMT, Bodrud-doza M, Rahman MS, Amin SB, Chu R, Mamun HA (2019) Sources of trace elements identification in drinking water of Rangpur district, Bangladesh and their potential health risk following multivariate techniques and Monte-Carlo simulation. *Groundw Sustain Dev* 9:100275. <https://doi.org/10.1016/j.gsd.2019.100275>
- Islam ARMT, Mamun AA, Rahman MM, Zahid A (2020a) Simultaneous comparison of modified-integrated water quality and entropy weighted indices: Implication for safe drinking water in the coastal region of Bangladesh. *Ecol Ind* 113:106229. <https://doi.org/10.1016/j.ecolind.2020.106229>
- Islam ARMT, Siddiqua MT, Zahid A, Tasnim SS, Rahman MM (2020b) Drinking appraisal of coastal groundwater in Bangladesh: An approach of multi-hazards towards water security and health safety. *Chemosphere* 255:126933. <https://doi.org/10.1016/j.chemosphere.2020.126933>
- Islam ARMT, Talukdar S, Mahato S et al (2021) Machine learning algorithm-based risk assessment of riparian wetlands in Padma River Basin of Northwest Bangladesh. *Environ Sci Pollut Res*. <https://doi.org/10.1007/s11356-021-12806-z>
- Kabir MM, Akter S, Ahmed FT, Mohinuzzaman M, Didar-ul-Alam M, Mostofa KMG, Islam ARMT, Niloy NM (2021) Salinity-induced fluorescent dissolved organic matter influence co- contamination, quality and risk to human health of tube well water, southeast coastal Bangladesh. *Chemosphere* 275:130053. <https://doi.org/10.1016/j.chemosphere.2020.130053>
- Kawo NS, Karuppannan S (2018) Groundwater quality assessment using water quality index and GIS technique in Modjo River Basin, central Ethiopia. *J African Earth Sci* 147:300–311. <https://doi.org/10.1016/j.jafrearsci.2018.06.034>
- Khan N (2011) Eruption time of permanent teeth in Pakistani children. *Iran J Public Health* 40(4):63
- Kim J, Han H, Johnson LE, Lim S, Cifelli R (2019) Hybrid machine learning framework for hydrological assessment. *J Hydrol*. <https://doi.org/10.1016/j.jhydrol.2019.123913>
- Kisi O, Ozkan C (2017) A new approach for modeling sediment-discharge relationship: local weighted linear regression. *Water Resour Manag* 31:1–23. <https://doi.org/10.1007/s11269-016-1481-9>
- Kisi O, Azad A, Kashi H, Saeedian A, Ali S, Hashemi A, Ghorbani S (2018) Modeling groundwater quality parameters using hybrid neuro-fuzzy methods. *Water Resour Manage*. <https://doi.org/10.1007/s11269-018-2147-6>
- Kouadri S, Samir K (2021) Hydro-chemical study with geospatial analysis of groundwater Quality Illizi Region, South-East of Algeria. *Iran J Chem Chemical Eng (IJCCCE)* 40(4):1315–1333. <https://doi.org/10.30492/ijcce.2020.39800>
- Kouadri S, Kateb S, Zegait R (2021) Spatial and temporal model for WQI prediction based on back-propagation neural network, application on EL MERK region (Algerian southeast). *J Saudi Soci Agricul Sci* 20(5):324–336
- Kuncheva LI, Plumpton CO (2010) Choosing parameters for random subspace ensembles for fMRI classification. In *International Workshop on Multiple Classifier Systems* (pp. 54–63). Springer, Berlin, Heidelberg.
- Lai C, Reinders MJ, Wessels L (2006) Random subspace method for multivariate feature selection. *Pattern Recogn Lett* 27(10):1067–1076
- Li P, He S, Yang N, Xiang G (2018) Groundwater quality assessment for domestic and agricultural purposes in Yan'an City, northwest China: implications to sustainable groundwater quality management on the Loess Plateau. *Environ Earth Sci* 77:1–16. <https://doi.org/10.1007/s12665-018-7968-3>
- Liu P, Wang J, Sangaiah AK, Xie Y, Yin X (2019) Analysis and prediction of water quality using LSTM deep neural networks in IoT environment. *Sustainability*. <https://doi.org/10.3390/su11072058>
- Loecke TD, Burgin AJ, Riveros-Iregui DA, Ward AS, Thomas SA, Davis CA, Clair MAS (2017) Weather whiplash in agricultural regions drives deterioration of water quality. *Biogeochemistry* 133(1):7–15
- Luo X, Lin F, Chen Y, Zhu S, Xu Z, Huo Z, Peng J (2019) Coupling logistic model tree and random subspace to predict the landslide

- susceptibility areas with considering the uncertainty of environmental features. *Sci Rep* 9(1):1–13
- Malone BP, Styc Q, Minasny B, McBratney AB (2017) Digital soil mapping of soil carbon at the farm scale: a spatial downscaling approach in consideration of measured and uncertain data. *Geoderma* 290:91–99. <https://doi.org/10.1016/j.geoderma.2016.12.008>
- McCullagh P, Nelder JA (1989) *Generalized linear models*. CRC Press, London
- McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* 5:115–133. <https://doi.org/10.1007/BF02478259>
- Montgomery S (1993) Ghadames Basin of north central Africa. *Stratigraphy, Geologic History, and Drilling Summary: Petroleum Frontiers* 10(3):51
- Mukate S, Wagh V, Panaskar D, Jacobs JA, Sawant A (2019) Development of new integrated water quality index (IWQI) model to evaluate the drinking suitability of water. *Ecol Indic* 101:348–354. <https://doi.org/10.1016/j.ecolind.2019.01.034>
- Nourani V, Elkiran G, Abba SI (2018) Wastewater treatment plant performance analysis using artificial intelligence—an ensemble approach. *Water Sci Technol* 78(10):2064–2076
- Ongley ED (2000) Water quality management: design, financing and sustainability considerations-II. In: *Invited Presentation at the World Bank's Water Week Conference: towards a Strategy for Managing Water Quality Management*, pp. 1e16
- Patil VBB, Pinto SM, Govindaraju T, Hebbalu VS, Bhat V, Kannanur LN (2020) Multivariate statistics and water quality index (WQI) approach for geochemical assessment of groundwater quality—a case study of Kanavi Halla Sub-Basin, Belagav India. *Environ Geochem Health* 42(9):2667–2684
- Peng, R.D. and Dominici, F., 2008. *Statistical methods for environmental epidemiology with R. R: A Case Study in Air Pollution and Health* (Springer). doi:<https://doi.org/10.1007/978-0-387-78167-9>
- Peterson JA (1985) *Geology and petroleum resources of north-central and northeastern Africa: U.S. Geological Survey Open-File Report 85–709*, 54 p
- Pham BT, Tien Bui D, Prakash I, Dholakia MB (2017) Hybrid integration of multilayer perceptron neural networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS. *CATENA* 149:52–63. <https://doi.org/10.1016/j.catena.2016.09.007>
- Rahman MM, Bodrud-Doza M, Siddique T, Zahid A, Islam ARMT (2020) Spatiotemporal distribution of fluoride in drinking water and associated probabilistic human health risk appraisal in the coastal region, Bangladesh. *Sci Total Environ* 724:138316. <https://doi.org/10.1016/j.scitotenv.2020.138316>
- Saha N, Bodrud-doza M, Islam ARMT et al (2020) Hydrogeochemical evolution of shallow and deeper aquifers in central Bangladesh: arsenic mobilization process and health risk implications from the potable use of groundwater. *Environ Earth Sci* 79(20):477. <https://doi.org/10.1007/s12665-020-09228-4>
- Salam R, Islam ARMT (2020) Potential of RT, Bagging and RS ensemble learning algorithms for reference evapotranspiration prediction using climatic data-limited humid region in Bangladesh. *J Hydrol* 590:125241. <https://doi.org/10.1016/j.jhydrol.2020.125241>
- Sihag P, Mohsenzadeh Karimi S, Angelaki A (2019) Random forest, M5P and regression analysis to estimate the field unsaturated hydraulic conductivity. *Appl Water Sci* 9:1–9. <https://doi.org/10.1007/s13201-019-1007-8>
- Sihag P, Angelaki A, Chaplot B (2020) Estimation of the recharging rate of groundwater using random forest technique. *Appl Water Sci* 10:1–11. <https://doi.org/10.1007/s13201-020-01267-3>
- Singh KP, Basant N, Gupta S (2011) Support vector machines in water quality management. *Anal Chim Acta* 703(2):152–162
- Singh B, Sihag P, Singh K (2017) Modelling of impact of water quality on infiltration rate of soil by random forest regression. *Model Earth Syst Environ* 3:999–1004. <https://doi.org/10.1007/s40808-017-0347-3>
- Singha S, Pasupuleti S, Singha SS, Singh R, Kumar S (2021) Prediction of groundwater quality using efficient machine learning technique. *Chemosphere* 276:130265
- Tiyasha TM, Yaseen ZM (2020) A survey on river water quality modelling using artificial intelligence models: 2000–2020. *J Hydrol*. <https://doi.org/10.1016/j.jhydrol.2020.124670>
- Towfiqul Islam ARM, Talukdar S, Mahato S, Kundu S, Eibek KU, Pham QB, Kuriqi A, Linh NTT (2021) Flood susceptibility modelling using advanced ensemble machine learning models. *Geosci Front*. <https://doi.org/10.1016/j.gsf.2020.09.006>
- Trevett AF, Carter RC, Tyrrel SF (2005) Mechanisms leading to post-supply water quality deterioration in rural Honduran communities. *Int J Hyg Environ Health* 208(3):153–161
- Uddin MG, Nash S, Olbert AI (2021) A review of water quality index models and their use for assessing surface water quality. *Ecol Indic* 122:107218. <https://doi.org/10.1016/j.ecolind.2020.107218>
- 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* 3:1–11. <https://doi.org/10.1007/s42452-020-04005-1>
- Wang G, Zhang Z, Sun J, Yang S, Larson CA (2015) POS-RS: A Random Subspace method for sentiment classification based on part-of-speech analysis. *Inf Process Manage* 51(4):458–479
- Wang X, Zhang F, Ding J (2017) Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed China. *Sci Rep*. <https://doi.org/10.1038/s41598-017-12853-y>
- Wang Q, Xu W, Zheng H (2018) Combining the wisdom of crowds and technical analysis for financial market prediction using deep random subspace ensembles. *Neurocomputing* 299:51–61
- van de Weerd AA, Ware PLG (1994) A review of the East Algerian Sahara oil and gas province (Triassic, Ghadames and Illizi Basins): *First Break*, 12(7):363–373
- Yaseen Z, Ehteram M, Sharafati A, Shahid S, Al-Ansari N, El-Shafie A (2018) The integration of nature-inspired algorithms with least square support vector regression models: Application to modeling river dissolved oxygen concentration. *Water* 10(9):1124
- Yaseen ZM, Sulaiman SO, Deo RC, Chau K-W (2019) An enhanced extreme learning machine model for river flow forecasting: State-of-the-art, practical applications in water resource engineering area and future research direction'. *J Hydrol* 569:387–408
- Zhu S, Heddad S (2019) Prediction of dissolved oxygen in urban rivers at the three gorges reservoir, China: Extreme learning machines (ELM) versus artificial neural network(ANN)', *Water Qual. Res J* 55(1):1–13
- Zhu S, Hrnjica B, Ptak M, Choinski A, Sivakumar B (2020) Forecasting of water level in multiple temperate lakes using machine learning models. *J Hydrol* 585:124819

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