



Prediction of irrigation water quality indices based on machine learning and regression models

Ali Mokhtar^{1,2} · Ahmed Elbeltagi³ · Yeboah Gyasi-Agyei⁴ · Nadhir Al-Ansari⁵ · Mohamed K. Abdel-Fattah⁶

Received: 13 November 2021 / Accepted: 8 February 2022 / Published online: 21 March 2022
© The Author(s) 2022

Abstract

Assessing irrigation water quality is one of the most critical challenges in improving water resource management strategies. The objective of this work was to predict the irrigation water quality index of the Bahr El-Baqr, Egypt, based on non-expensive approaches that requires simple parameters. To achieve this goal, three artificial intelligence (AI) models (Support vector machine, SVM; extreme gradient boosting, XGB; Random Forest, RF) and four multiple regression models (Stepwise Regression, SW; Principal Components Regression, PCR; Partial least squares regression, PLS; Ordinary least squares regression, OLS) were applied and validated for predicting six irrigation water quality criteria (soluble sodium percentage, SSP; sodium adsorption ratio, SAR; residual sodium carbonate, RSC; potential of salinity, PS; permeability index, PI; Kelly's ratio, KR). Electrical conductivity (EC), sodium (Na^+), calcium (Ca^{2+}) and bicarbonate (HCO_3^-) were used as input exploratory variables for the models. The results indicated the water source is not suitable for irrigation without treatment. A good soil drainage system and salinity control measures are required to avoid salt accumulation within the soil. Based on the performance statistics of the root mean square error (RMSE) and the scatter index (SI), SW emerged as the best (0.21% and 0.03%) followed by PCR and PLS with RMSE 0.22% and 0.21% for SAR, respectively. Based on the classification of the SI, all models applied having values less than 0.1 indicate good prediction performance for all the indices except RSC. These results highlight potential of using multiple regressions and the developed machine learning methods in predicting the index of irrigation water quality, and can be rapid decision tools for modelling irrigation water quality.

Keywords Irrigation water quality index · Machine learning · Support vector machine · Stepwise regression · Bahr El-Baqr drain

Introduction

Water resources are critical in the drinking, industrial, and agricultural sectors. As a result, improved water resource quality significantly reduces the cost of water treatment for irrigation and boosts agricultural yield (Kouadri et al.,

2021a, 2021b). Thus, water shortage is a global concern and is going to get worse as indicated by climate change projections (Pleguezuelo et al., 2018). This is particularly so for arid and semi-arid areas, like Egypt that depends on irrigated agriculture (Elbeltagi et al., 2021a, 2021b, 2020a, 2021c, 2021d, 2020b; Moharir et al., 2019). Irrigated agriculture

✉ Nadhir Al-Ansari
nadhir.alansari@ltu.se

✉ Mohamed K. Abdel-Fattah
mohammedkamal8@yahoo.com

¹ State of Key Laboratory of Soil Erosion and Dryland Farming On Loess Plateau, Institute of Soil and Water Conservation, Chinese Academy of Sciences and Ministry of Water Resources, Northwest Agriculture and Forestry University, Yangling 712100, China

² Department of Agricultural Engineering, Faculty of Agriculture, Cairo University, Giza 12613, Egypt

³ Agricultural Engineering Department, Faculty of Agriculture, Mansoura University, Mansoura 35516, Egypt

⁴ School of Engineering and Built Environment, Griffith University, Nathan, QLD 4111, Australia

⁵ Civil, Environmental and Natural Resources Engineering, Lulea University of Technology, 97187 Lulea, Sweden

⁶ Soil Science Department, Faculty of Agriculture, Zagazig University, Zagazig 44511, Egypt

needs an adequate supply of usable water. Water quality issues were frequently overlooked in the past due to widely available high-quality water supplies (Ayers and Westcott, 1985). Nowadays, water quality is becoming an issue because of the intensive and competitive use of water. This means that new irrigation projects, as well as the existing looking for additional or supplemental supplies, may have to rely on low-quality salt-laden water from unfavourable sources (Pande and Moharir, 2018). Unless proper strategies are put in place, the use of poor quality irrigation water could lead to problems with soil salinity, irrigation rate decline, plant growth toxicity, and other associated problems (Ayers and Westcott, 1985).

The most promising approach for increasing irrigation water availability is agricultural drainage water reuse (Assar et al., 2020). Due to Egypt's limited water supplies, agricultural drainage water must be reused for irrigation (Abdel-Fattah and Helmy, 2015; Abdel-Fattah et al., 2020). However, there are some concerns about the quality of the reuse water. The reuse of this drainage water without proper treatment may have negative impacts on the soil, crop, and irrigation system. A variety of metrics are used to assess the quality of irrigation water given by several organization and agencies (El Bilali and Taleb, 2020). These indices include soluble sodium percentage (SSP) (Todd and Mays, 2004), sodium adsorption ratio (SAR) (Ayers and Westcott, 1985), residual sodium carbonate (RSC) (Richards, 1954a), potential of salinity (PS) (Doneen, 1964), permeability index (PI) (Doneen, 1964; Gholami and Srikantaswamy, 2009) and Kelley's ratio (KR) (Kelley, 1963). Therefore, attempts are being made to develop a non-physical approach based on artificial intelligence (AI) to predict the water quality index (Gaya et al., 2020; Lu and Ma, 2020).

AI-based modelling is a useful tool for rapid prediction of the water quality indices. Some of the benefits of AI models include their nonlinear structure, capacity to anticipate complicated events, handling large datasets at diverse sizes, and handling missing data. Furthermore, AI systems have been demonstrated to be very capable of forecasting and monitoring water quality (Ahmed et al., 2019; Lu and Ma, 2020; Abdel-Fattah et al., 2020; Mokhtar et al., 2021b, 2021a). Also, AI is an appealing, rapid, and direct computing method for water quality modelling (Gaya et al., 2020; Tung and Yaseen, 2020; Yasin and Karim, 2020). For instance, support vector machine (SVM) (Hamzeh Haghbi et al., 2018), least square SVM (LSVM) (Leong et al., 2019) and artificial neural network (ANN) (Sakizadeh, 2016; Hameed et al., 2017; Machiwal et al., 2018) have successfully been used for water quality predictions. ANN was applied for prediction of the water quality index of the Langat River Basin, Malaysia (Juahir et al., 2004; Gazzaz et al., 2012). Mohammadpour et al. (2015) compared SVM, radial basis function neural network and backpropagation neural network techniques for

the forecast of the water quality in a wetland, and ANN was applied to predict the water quality index in the Red Sea State, Sudan (Ismael et al., 2021).

Currently, a few investigators are developing AI models to predict irrigation water quality index (IWQI). The groundwater quality for drinking purposes was assessed using statistical index of Akola and Buldhana districts, Maharashtra, India (Pande et al., 2020), also by using radial basic function (RBF) networks (Panneerselvam et al., 2021). ANN was used to forecast the suitability of groundwater for irrigation in India with 13 physicochemical parameters (Wagh et al., 2016). Interestingly, most of the previous studies have found good performance of AI algorithms for predicting water quality (Abdel-Fattah et al., 2020; Abba et al., 2020; Ahmed et al., 2019). Generally, multiple linear regression seeks to discover a link between a large number of independent or predictor factors (exploratory variables) and a dependent variable (Chenini and Khemiri, 2009). Multiple linear regression is regarded as a reliable approach for assessing groundwater quality since it creates a minimal dataset of indicators based on water's chemical composition (Doran et al., 1994). Using structural equation modelling, all of the predictor variables are combined in a single model to find potential interactions between them (Chenini and Khemiri, 2009). Many researchers, such as Charulatha et al. (2017); Yildiz and Degirmenci (2015) and Noori et al. (2010), have used regression analysis for water quality assessment. Multiple linear regression and structural equation modelling were used to assess the quality of groundwater by Chenini and Khemiri (2009). They found multiple linear regression as a useful tool for characterizing groundwater quality. Multiple linear regression and principal component analysis were used by Viswanath et al. (2015) and observed that the structural equation modelling allows for the simultaneous examination of the complete parameter system. Monitoring water quality and quantity of national watersheds in Turkey, Odemis and Evrendilek (2007) reported that multiple linear regression models provide a valuable assessment of controls that aid in the development of integrated and sustainable watershed management strategies. Yıldız and Karakuş (2019) explored the estimation of irrigation water quality index with the creation of an ideal model utilizing multiple regression and an artificial neural network (ANN) model. The approaches demonstrated to be effective ways for calculating irrigation water quality indexes by utilizing several water qualities measures. Assessment of water quality of Brahmani River using correlation and regression analysis carried out by Nayak (2020) showed that regression analysis might be a valuable approach for monitoring water quality and predicting trends in water quality variation.

The primary objective of this research is to employ artificial intelligence algorithms to predict the irrigation water quality index of the Bahr El-Baqr drain based on readily

observed and fewer data (EC, Na⁺, Ca²⁺ and HCO₃⁻). The Bahr El-Baqr drain, located on the eastern side of the Nile Delta area (Fig. 1), is one of Egypt's major drains, stretching for 106 kilometres. It is one of Egypt's most contaminated drains as stated by (Abdel-Shafy and Aly, 2002). The study's findings will assist farmers in arid/semi-arid nations manage irrigation water quality to boost agricultural productivity and policymakers make feasible water resource management decisions.

Materials and methods

Irrigation water samples and chemical composition analysis

A total of 105 water samples were collected during July 2020 from the Bahr El-Baqr drain. Figure 1 shows the location of the sampling sites which are uniformly spread out along the whole drain. At each location, 1 litre of water was collected at 1 m depth. The samples were immediately filtered and prepared for chemical composition analysis based

on the standard methods described by APHA (1998) and (Richards, 1954a).

An EC-meter and a pH-meter (with a combined glass/reference Ag/AgCl electrode) were used to determine the electrical conductivity (EC, dSm⁻¹) and the pH of the water on site. Sodium (Na⁺) and potassium (K⁺) concentrations were determined using a flame photometer, while calcium (Ca²⁺) and (Mg²⁺) were volumetrically determined by titration with ethylene diamine tetra acetic acid disodium salt (EDTA-2Na). Chloride (Cl⁻) was determined by titration with silver nitrate solution in the presence of potassium chromate indicator. The carbonate (CO₃²⁻) and bicarbonate (HCO₃⁻) compositions were determined by titration with a standard solution of sulphuric acid using phenolphthalein as an indicator for former and methyl-orange for latter. The sulphate (SO₄²⁻) composition was calculated by the difference between total cations and anions.

Irrigation water quality's criteria

The three principal problems that can arise from poor quality irrigation water are salinity hazard, sodicity hazard and toxicity hazard (Ayers and Westcot, 1985). The water from

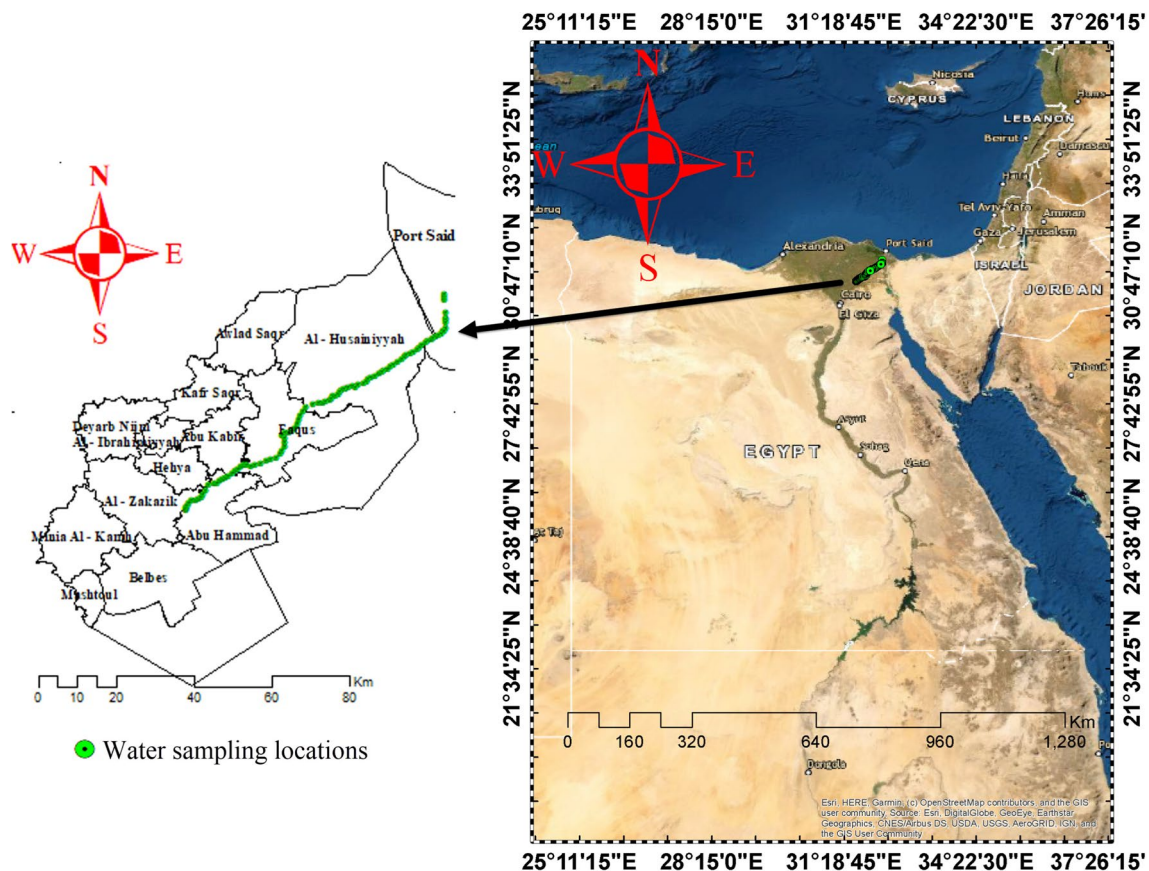


Fig. 1 Location of the Bahr El-Baqr drain and sampling locations in green along the drain

Bahr El-Baqr drain is used for agricultural irrigation. Thus, it is urgent to monitor and predict the chemical composition of this irrigation water source. The chemical composition of water (i.e., pH, EC, Ca^{2+} , Mg^{2+} , Na^+ , K^+ , CO_3^{2-} , HCO_3^- , Cl^- and SO_4^{2-}) from the drain was used to calculate the water quality Criteria indicated in Table 1. Water with SSP less than 60 is safe with little sodium accumulations that will cause a breakdown of the soil's physical properties (Fipps, 2003). SAR places the irrigation water into four categories; low (<10), medium (10–18), high (18–26), and very high (>26) (Richards, 1954a). Based on RSC criterion, the irrigation water is classified into three; no-hazard (<1.25 mmolc l^{-1}), medium hazard (1.25–2.5 mmolc l^{-1}) and extreme hazard (>2.5 mmolc l^{-1}). The water quality criteria of PI have three classes; excellent (>75%), good (25–75%) and unsuitable (<25%) (Al-Amry, 2008). The PS criterion divides the water quality into three classes; safely used in fine, medium and coarse textured soils (1–3 mmolc l^{-1}), safely used in medium and coarse textured soils (3–15 mmolc l^{-1}), and safely used only in coarse textured soils (15–20 mmolc l^{-1}). The irrigation water quality criteria based on KR have two classes; safe (< 1 mmolc l^{-1}) and unsuitable (>1 mmolc l^{-1}).

Multiple regression and machine learning models applied

In this study, we used seven models (machine learning and multiple regressions) to predict irrigation water quality criteria defined in Figure 2. The water quality criteria of SSP, SAR, RSC, PS, PI and KR were considered as the dependant variables and EC, Na^+ , Ca^{2+} and HCO_3^- were used as input variables in Table 1. To facilitate the regression task, the input data were normalized to the range from 0 to 1 as:

$$X_n = \frac{X_0 - X_{\min}}{X_{\max} - X_{\min}} \quad (1)$$

where X_n is the normalized data, X_0 is the original data, while X_{\min} and X_{\max} are the minimum and maximum values

of the original data. The datasets were divided into 75% for training and 25% for testing. Scikit-learn 0.22.1, a Python computer language package, was used to create the machine learning models. The computations were performed on Google Cloud Platform virtual software. For each model, the hyper-parameter tuning was carried out using a grid search strategy in order to obtain the best score as well as the optimum parameter settings that gave the lowest prediction errors in the testing stages. Below is a brief description of the models.

Machine learning models

Support vector machine (SVM) SVM algorithm was developed by Vapnik (Vapnik, 2013). SMV is a supervised learning algorithm that can be used for both regression and classification. SVR uses a similar theory as SVM for classification, with a few minor changes. The main aim is to minimize the errors by individualizing the hyperplane which increases the limit of tolerance. In contrast to an ANN model, which typically has several local minima, the SVM provides a unique solution due to the convex nature of the optimality issue (Chen et al., 2013; Kouadri et al., 2021b). The estimated function in the SVM method is shown as follows:

$$f(x) = \omega\varphi(x) + b \quad (2)$$

where $\varphi(x)$ refers to the higher-dimensional feature space translated from input vector x . ω and b correspond the weights vector and a threshold, respectively, which may be determined by minimizing the following regularized risk function:

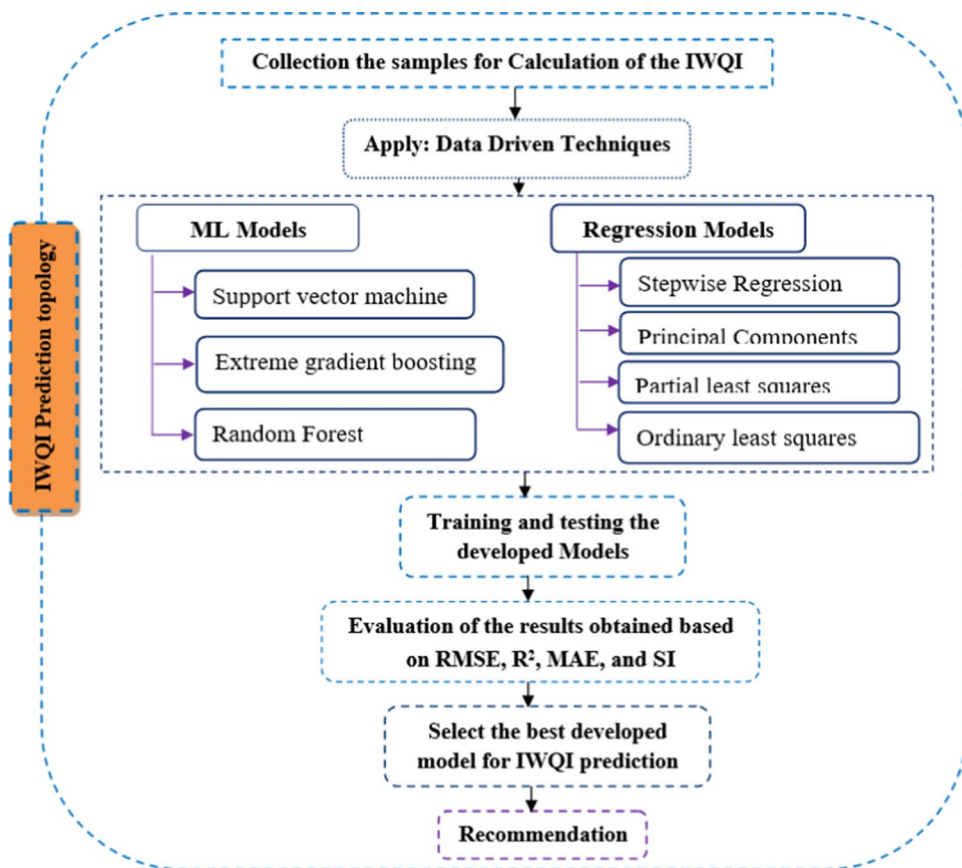
$$R(C) = C \frac{1}{n} \sum_{i=1}^n L(d_i, y_i) + \frac{1}{2} \|\omega\|^2 \quad (3)$$

where C represents the error's penalty parameter, d_i represents the intended value, n is the observations number, and

Table 1 The equation applied to calculate the water quality Criteria

Criterion	Equation	Unit	References
Soluble sodium percentage	$SSP = \frac{Na^+}{Ca^{2+} + Mg^{2+} + Na^+ + K^+} \times 100$	%	Richards, (1954a, b)
Sodium adsorption ratio	$SAR = \frac{Na^+}{\sqrt{\frac{Ca^{2+} + Mg^{2+}}{2}}}$	–	Ayers and Westcot, (1985)
Residual sodium carbonate	$RSC = (CO_3^{=} + HCO_3^-) - (Ca^{2+} + Mg^{2+})$	mmolc l^{-1}	Richards, (1954a, b)
Permeability index	$PI = \frac{Na^+ + \sqrt{HCO_3^-}}{Ca^{2+} + Mg^{2+} + Na^+} \times 100$	%	Doneen, (1964) and Siamak and Srikantaswamy, (2009)
Kelly ratio	$KR = \frac{Na^+}{Ca^{2+} + Mg^{2+}}$	mmolc l^{-1}	Kelley, (1963)
Potential of salinity	$PS = Cl^- + \frac{1}{2}SO_4^{=}$	mmolc l^{-1}	Doneen, (1964)

Fig. 2 Flow chart of the methodology of IWQI prediction



$C_n^{\frac{1}{2}} \sum_{i=1}^n L(d_i, y_i)$ is the empirical error, in which the function L_{ϵ} is determined as:

$$L_{\epsilon}(d, y) = |d - y| - \epsilon \text{ if } |d - y| \geq \epsilon \text{ or } 0 \text{ otherwise} \quad (4)$$

where $\frac{1}{2} \|\omega\|^2$ refers to the so-called regularization term and ϵ presents the tube size. Finally, the estimated function in Eq. (1) is represented explicitly by using Lagrange multipliers and exploiting the optimality constraints as:

$$f(x, \alpha_i, \alpha_i^*) = \sum_{i=1}^n (\alpha_i - \alpha_i^*) K(x, x_i) + b \quad (5)$$

where $k(x, x_i)$ corresponds the kernel function. (Vapnik, 2013; Fan et al., 2018) provided detailed information and SVM algorithm computation techniques. We applied two different kernels (radial basis function and linear) and regularization parameter C from the set (1, 2, 3, 4, 5) and maintained the remaining hyper-parameters default values. The best score was achieved by setting C=5 and kernel='linear'.

Extreme gradient boosting (XGB) Chen and Guestrin, (2016) developed the XGB algorithm as a unique implementation approach for the gradient boosting machine based on regression trees. The method is built on the con-

cept of "boosting," which combines all of the predictions of a group of "weak" learners to create a "strong" learner using additive training procedures. XGB reduces overfitting and under-fitting issues and can reduce computing expenses (Mokhtar et al., 2021a, b). The general function for predicting at step t is as follows:

$$f_i^{(t)} = \sum_{k=1}^t f_k(x_i) = f_i^{(t-1)} + f_t(x_i) \quad (6)$$

where $f_t(x_i)$ denotes the learner at each step t, $f_i^{(t)}$ and $f_i^{(t-1)}$ denote the predictions at steps t and t-1, and x_i represents the input variable. To prevent the over-fitting problem while maintaining the model's computing speed, the XGB uses the analytic formula below to evaluate the "goodness" of the model from the original function:

$$Obj^{(t)} = \sum_{k=1}^n l(\bar{y}_i, y_i) + \sum_{k=1}^t \Omega(f_i) \quad (7)$$

where l denotes the loss function, n refers to the observations number and Ω represents the regularization term described as:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2 \quad (8)$$

where ω denotes the leaves scores vector, λ is the regularization parameter, and γ is the lowest loss required to further divide the leaf node. More details regarding the XGB algorithm's computing techniques may be found in Chen and Guestrin (Chen and Guestrin, 2016). We used the XGB with 400 trees, 10 maximum depths, a learning rate of 0.1, with the other hyper-parameters set to their default levels. The following hyper-parameter settings were used: n estimators (number of trees) (100, 200, 300, 400, and 500); max depth (1, 2, 5, 10, and 12); and learning rate (0.05, 0.1 and 0.5).

Random forest (RF) Breiman, (2001) created the RF model, which is a set of decision trees with controlled variation. It is commonly used to solve regression and classification issues. A random forest regression is a subset of a bootstrap assembly. It is concerned with random binary trees, which use a portion of the observations using the bootstrapping approach, in which a random subset of the training dataset is picked from the raw dataset and used to create the model. This inquiry gave a full explanation of the RF model as well as the computing procedure (Breiman, 2001; Ferreira and da Cunha, 2020; Mokhtar et al., 2021a). To get the highest possible score, an RF was trained with 400 trees, a maximum depth of ten, and the other hyper-parameters set to their default levels. During the hyper-parameter tuning phase, the following hyper-parameter sets and values were evaluated: number of trees (100, 200, 300, 400, and 500), and max depth (1, 2, 5, 10, 12).

Multiple regressions

Stepwise regression The predictive variables are selected automatically in the stepwise regression method (Hocking, 1975; Draper and Smith, 1981). Stepwise regression involves three main techniques: forward selection, backward elimination, and bidirectional elimination (Jia et al., 2016). The commonly utilized method was initially proposed by (Efroymson, 1960). It is an automated approach for statistical model selection where there are a large number of potential explanatory variables and no underlying theory on which to base the model selection. The stepwise process is most commonly employed in regression analysis; however, the basic idea is adaptable to many types of model selection. A test is run to see if any variables can be removed without significantly raising the residual sum of squares (RSS). The technique ends when the measure is (locally) maximized or when the available improvement falls below a crucial value. The selection procedure begins by including the variable that makes the greatest contribution to the model (the criteria employed is the student's t statistic). If the probability

associated with the t statistic of a second variable is smaller than the "probability for entrance," it is added to the model. The procedure is repeated with the third and remaining variables, analysing the impact of deleting each component from the model (still using the t statistic). The variable is eliminated if the likelihood is larger than the "probability of removal." The technique is repeated until there are no more variables that can be added or deleted.

Ordinary least squares regression (OLS) The most commonly used statistical method in regression is the OLS. A distinction is made between simple linear regression and multiple linear regression, the first one contains only one explanatory variable while the second contains several explanatory variables (Addinsoft, 2019). OLS is used to predict an outcome (Y, a quantitative dependent variable) through predictor variables (X_1, X_2, \dots, X_p , the quantitative explanatory variables) (Addinsoft, 2019). The model with p explanatory variables is written as:

$$y_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + \varepsilon_i \quad (9)$$

where y_i denotes the dependent variable value for observation i, x_{ij} refers to the value assigned to variable j for observation i, and ε_i is the random error with mean 0 and variance s^2 of the model for observation i, β_j being the parameters of the model.

Principal component regression (PCR) Multicollinearity is a big problem with multiple linear regression analysis due to the presence of a strong correlation between the explanatory variables, resulting in an increase in the regression parameter estimators. This makes the results of OLS unreliable since it is based on the assumption of no multicollinearity between the explanatory variables. PCR, first suggested by (Pearson, 1901), is used to address the multicollinearity problem, and it is based on principal component analysis (PCA) (Addinsoft, 2019). PCR application has three steps: (1) runs a PCA to address multicollinearity problem, (2) performs an OLS regression on the selected components, and (3) computes the model parameters that denotes the input variables.

Partial least squares regression (PLS) PLS is a regression method that combines principal component analysis (PCA) and multiple linear regression theories (Wold, 1995). PLS overcomes multicollinearity and over-fitting problems through variable transformation to new orthogonal factors (Huang et al., 2004). The PLS approach is rapid, efficient, and optimum for a covariance-based criteria. It is advised when the number of variables is large and the explanatory factors are likely to be associated. The

PLS regression model with components has the following equation:

$$\begin{aligned}
 Y &= T_c C'_h + E_h \\
 Y &= XW_k^* C_h + E_h \\
 Y &= XW_k^* (P'_h W_h)^{-1} C'_h + E_h
 \end{aligned}
 \tag{10}$$

where Y denotes the matrix of the dependent variables and X denotes the matrix of the explanatory variables. T_h , C_h , W_h^* and P_h are the matrices produced by the PLS method while E_h is the matrix of the residuals. The regression coefficients of Y on X are represented by the matrix B, which has h components created by the PLS regression process as follows:

$$B = W_h (P'_h W_h)^{-1} C'_h
 \tag{11}$$

Performance statistics for model evaluation

The mean absolute error (MAE), root mean square error (RMSE), and scatter index (SI) were used to evaluate the models in this work which are presented as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^n |O_i - P_i|
 \tag{12}$$

$$RMSE = \sqrt{\frac{i}{n} \sum (P_i - O_i)^2}
 \tag{13}$$

$$SI = \frac{RMSE}{\bar{O}}
 \tag{14}$$

where \bar{O} represents the average values of the observed IWQI, O_i and P_i are the actual and foreseen IWQI, respectively, and i represents the observations number. SD denotes the standard deviation between the anticipated and observed IWQI values.

Results and discussions

Chemical composition of the Bahr El-Baqr drain water

The chemical composition of the Bahr El-Baqr Drain is summarized in Table 2. The chemical composition of the Bahr El-Baqr Drain clearly varies substantially. The values ranged from 6.9 to 8.31 with an average of 7.64 ± 0.03 for pH, 1.25–2.70 dSm⁻¹ with an average of 1.59 ± 0.02 for EC, 1.62–3.42 mmolc l⁻¹ with an average of 2.52 ± 0.04 for Ca²⁺, 0.77–2.09 mmolc l⁻¹ with an average of 1.46 ± 0.02 for Mg²⁺, 7.04–14.25 mmolc l⁻¹ with an average of 10.31 ± 0.13 for Na⁺, 0.09–8.30 mmolc l⁻¹ with an average of 1.59 ± 0.17 for K⁺, 1.52–4.77 mmolc l⁻¹ with an average 3.36 ± 0.05 for HCO₃⁻, 5.45–12.65 mmolc l⁻¹ with an average of 9.07 ± 0.13 for Cl⁻ and 0.01–12.23 dSm⁻¹ with an average of 3.43 ± 0.22 for SO₄²⁻. These results agree with those of Abdel-Fattah and Helmy (2015) and Abdel-Fattah et al. (2020). The acceptable level of irrigation water pH ranges between 6.5 and 8.4 (Ayers and Westcot, 1985). Therefore, the pH values of the Bahr El-Baqr drain are within the acceptable limits for irrigation purposes. According to Richards (1954a), the water of Bahr El-Baqr drain is of high salinity and in agreement with the findings of Abdel-Fattah and Helmy (2015) and Abdel-Fattah et al. (2020). Accordingly, the water should not be used for irrigation process unless the soil has good drainage and a special management strategy for salinity control is put in place, or salt tolerant plants are being irrigated (Richards, 1954a). It is observed from the results that the dominant cation in the water is sodium and the concentration of the cations are in the following order; Na⁺ > Ca²⁺ > Mg²⁺ > K⁺. According to Ayers and Westcot (1985), the cations (Na⁺, K⁺, Ca²⁺, and Mg²⁺) are within the acceptable limits of irrigation water. Concerning the anions, the dominant is chloride followed by sulphate and then bicarbonate. Also, the concentration of the anions is within the acceptable limits (Ayers and Westcot, 1985).

Table 3 shows that the water quality criteria of the Bahr El-Baqr drain vary greatly. The values ranged from 41.49 to

Table 2 Chemical composition of the Bahr El-Baqr drain water

	pH	EC dS/m	Cations, mmolc l ⁻¹				Anions, mmolc l ⁻¹		
			Ca ²⁺	Mg ²⁺	Na ⁺	K ⁺	HCO ₃ ⁻	Cl ⁻	SO ₄ ²⁻
Min	6.90	1.25	1.62	0.77	7.04	0.09	1.52	5.45	0.01
Max	8.31	2.70	3.42	2.09	14.25	8.30	4.77	12.65	12.23
Mean	7.64	1.59	2.52	1.46	10.31	1.58	3.36	9.07	3.43
SD	0.27	0.24	0.39	0.25	1.33	1.74	0.51	1.31	2.25
SE	0.03	0.02	0.04	0.02	0.13	0.17	0.05	0.13	0.22

Table 3 Summary of the water quality criteria

	SSP %	SAR	RSC mmolc l ⁻¹	PI %	KR mmolc l ⁻¹	PS mmolc l ⁻¹
Min	41.49	4.63	-2.99	72.52	1.40	7.84
Max	74.35	9.95	1.07	93.06	3.66	16.23
Mean	65.51	7.34	-0.62	84.93	2.62	10.79
SD	6.82	0.96	0.68	3.27	0.41	1.45
SE	0.67	0.09	0.07	0.32	0.04	0.14

74.35 with an average of 65.51 ± 0.67 for SSP, 4.63 to 9.95 with an average of 7.34 ± 0.09 for SAR, -2.99 to 1.07 with an average of -0.62 ± 0.07 for RSC, 72.52–93.06 with an average of 84.93 ± 0.32 for PI, 1.40–3.66 with an average of 2.62 ± 0.04 for KR and 7.84–16.32 with an average of 10.79 ± 0.14 for PS. According to the SSP average (>60%), use of Bahr El-Baqr drain water may result in sodium accumulation that could cause a breakdown of the soil's physical properties (Todd and Mays, 2004; Fipps, 2003). The use of this polluted water for irrigation should be restricted to a reasonable degree. Regarding SAR, the Bahr El-Baqr drain water has low values. Gupta and Gupta (1997) and Richards (1954a) reported that low SAR water (low sodicity) can be utilized for irrigation on most soils with little chance of hazardous amounts of exchangeable salt developing. The low RSC values (<1.25) indicate that the Bahr El-Baqr drain water is safe for irrigation process without alkalinity hazard development. With PI values greater than 75% (with an average 84.9), the water can be used for irrigation without soil permeability impairment (Al-Amry, 2008; Doneen, 1964; Raghunath, 1987). Long-time use of irrigation water containing high levels of Na⁺ could affect the physical properties of soil and impair soil permeability (Doneen, 1964). Meanwhile, KR values greater than one indicate that the water is unsuitable for irrigation (Kelley, 1963). The average PS value of 10.79 is an indication that the water can be safely used in medium and coarse textured soils. Doneen (1964) outlined possible salinity difficulties with irrigation water and pointed out that the appropriateness of irrigation water is reliant on more than just the percentage of soluble salts. It has been observed that following consecutive irrigation whereas the concentration of highly soluble salts increases the salinity of the soil (Gholami and Srikantaswamy, 2009). The Bahr El-Baqr Drain water is characterized as high salinity-medium sodicity based on SAR and salinity measurements, and is considered acceptable (usable) for irrigation purposes (Richards, 1954b).

Abdel-Fattah et al. (2020) reported that the chemical composition of irrigation water plays a crucial role in its quality. For identifying the basic criteria for evaluating the water quality, i.e., salinity hazard, sodicity hazard, alkalinity hazard and toxicity hazard, there is a need to determine the chemical composition of irrigation water (i.e., ECw, soluble cations and anions) (Zaman et al. 2018; Abdel-Fattah and

Ayman, 2015). EC, SAR, KAR, RSC, SSP, and PI criteria were used to assess the appropriateness of water for agricultural irrigation purposes by Kumar et al. (2016). Prunty et al. (1991) reported that SAR of irrigation water correlated with crop yield and quality. SSP is a key parameter for assessing agricultural water quality (Sarker et al., 2000). It reflects the possibility of degradation of the soil physical properties that influence plant growth. Salts build up in the soil (Longenecker et al. 1969), causing soil structure dispersion that decreases the infiltration rate, (Agassi et al., 1981). Aboukarima et al. (2018) demonstrated that the rate of infiltration is sensitive to the SAR of the applied water. Sadick et al. (2017) observed a negative correlation between SAR, as well as KR, with Ca, Mg, HCO₃ and CO₃ which implies that high values of SAR are associated with decrease in these chemical parameters and vice versa. Aboukarima et al. (2018) and Aggag (2016) mentioned a positive correlation between EC and SAR of water. Raiham and Alam (2008) reported that there is negative correlation between RSC and Ca+Mg concentration a positive one with CO₃+HCO₃. Xu et al. (2019) reported that high values of PI were correlated with high Na and HCO₃ ions in water. Agarwal et al. (1982) demonstrated a highly significant positive correlation between EC and the concentration of salts that may have an impact on irrigation water quality due to the salinity hazard.

Evaluation of the machine learning and regression models

The chemical composition 105 samples were used for the training and test stage. Table 4 presents the regression equations established for the different regression models (i.e., OLS, PCR and SW, PLS), and Fig. 3 displays the performance statistics for all models. As judged by all of the performance statistics, SW emerged as the best model for predicting the water quality criteria followed by PCR. The highest RMSE was recorded by RF for SSP as 3.27%.

Moreover, the highest MAE was found in PI and SSP as 2.62% and 2.13%, respectively, for RF model. The R² ranged from 0.53 to 0.98 (Fig. 2). SW recorded the highest R² values (0.87–0.98), followed by PCR, and the lowest by RF that ranged from 0.53 to 0.78. With the regression models, the highest R² values were similarly recorded for SSP by all regression models applied; this is also true among the

Table 4 The regression equations established between the water quality indices and the chemical composition variables

Model	Parameter	Equation
OLS	SSP	$SSP = 63.73 - 34.87 * EC + 5.64 * Na - 0.53 * Ca + 0.063 * HCO_3$
	SAR	$SAR = 2.70 - 0.199 * EC + 0.73 * Na - 0.95 * Ca - 0.046 * HCO_3$
	RSC	$RSC = -1.11 - 0.34 * EC + 0.048 * Na - 1.06 * Ca + 0.95 * HCO_3$
	PI	$PI = 85.22 - 2.61 * EC + 1.39 * Na - 6.24 * Ca + 1.57 * HCO_3$
	KR	$KR = 1.91 - 0.12 * EC + 0.26 * Na - 0.67 * Ca - 0.032 * HCO_3$
	PS	$PS = 4.19 + 6.05 * EC - 0.044 * Na + 0.15 * Ca + 2 - 0.85 * HCO_3$
SW	SSP	$SSP = 63.41 - 36.03 * EC + 5.74 * Na$
	SAR	$SAR = 2.67 - 0.40 * EC + 0.76 * Na - 0.98 * Ca$
	RSC	$RSC = -0.71 - 1.17 * Ca + 0.91 * HCO_3$
	PI	$PI = 85.42 - 4.37 * EC + 1.73 * Na - 6.46 * Ca + 1.47 * HCO_3$
	KR	$KR = 1.899 - 0.27 * EC + 0.28 * Na - 0.67 * Ca$
	PS	$PS = 4.27 + 6.23 * EC - 0.99 * HCO_3$
PCR	SSP	$SSP = 62.19 - 35.15 * EC + 5.73 * Na - 0.44 * Ca + 0.31 * HCO_3$
	SAR	$SAR = 2.70 - 0.24 * EC + 0.75 * Na - 0.99 * Ca - 0.047 * HCO_3$
	RSC	$RSC = -1.08 - 0.39 * EC + 0.074 * Na - 1.16 * Ca + 0.96 * HCO_3$
	PI	$PI = 84.71 * EC + 1.63 * Na - 6.83 * Ca + 1.69 * HCO_3$
	KR	$KR = 1.90 - 0.15 * EC + 0.27 * Na - 0.70 * Ca - 0.034 * HCO_3$
	PS	$PS = 4.46 + 6.68 * EC - 0.09 * Na - 0.011 * Ca - 0.98 * HCO_3$
PLS	SSP	$SSP = 62.01 - 34.12 * EC + 5.62 * Na - 0.32 * Ca + 0.157 * HCO_3$
	SAR	$SAR = 2.95 - 0.041 * EC + 0.70 * Na - 1.02 * Ca - 0.051 * HCO_3$
	RSC	$RSC = -0.86 - 0.12 * EC + 0.0075 * Na - 1.13 * Ca + 0.95 * HCO_3$
	PI	$PI = 85.83 - 1.35 * EC + 1.169 * Na - 6.53 * Ca + 1.67 * HCO_3$
	KR	$KR = 2.08 - 0.016 * EC + 0.24 * Na - 0.72 * Ca - 0.036 * HCO_3$
	PS	$PS = 3.73 + 5.68 * EC + 0.019 * Na + 0.29 * Ca - 0.85 * HCO_3$

machine learning models. Based on the classification of the SI index, the SVR, XGB and RF values were less than 0.1 which suggest excellent models for all water quality indices except RSC. XGB and RF having an SI value of 0.52 for RSC, indicate poor models. This may be related to the significant correlation between the input and output variables. Therefore, one of the most significant aspects of machine learning models for improving performance is the selection of input variables. Significantly, the SVR model emerged as the best model for predicting water quality index followed by XGB and then RF.

This finding is consistent with Leong et al. (2019) who used SVR model in forecasting the index of water quality at Perak state Malaysia. Furthermore, our results are similar to those reported by El Bilali and Taleb (2020) who used the RF method in the prediction of the irrigation water quality index in Nfifikh watershed in Morocco. Their RMSE and R² values were 1.88 and 0.5 for KR, and 6 (mmolc l⁻¹) and 0.6 for SAR. Based on Figs. 4 and 5 and Table 4, which confirm the results from Fig. 2, the SW model is superior for prediction of IWQI. SW model reported the lowest RMSE, MAE and SI values as 0.21%, 0.17 and 0.03, respectively, and the highest R² value of 0.98 for SAR equation, and this supports the findings of (Li et al., 2013). Finally, boxplot was

developed to compare the performance of the AI models for PS and PI IWQI (Fig. 6).

Positive and negative estimate errors denote under-estimation and over-estimation, respectively. Some parameters of the boxplot are the first quartile (Q1), third quartile (Q3), and inter-quartile range (IQR), and the median shown as a vertical line in the box. The SVR model having the lowest median error is selected as the best model. In error analysis, Q3 is more relevant than Q1 since it contains 75% of the error. It was observed that the SW model, with a Q3 difference of 0.2 compared to SVR’s value of 0.11 and XGB value has the highest accuracy. Moreover, SW has a lower IQR than the other two models, indicating that the error distribution is close to zero. In addition, the median line in the centre of the rectangle indicates the error distribution’s normalcy.

In general, the model’s prediction accuracy varies over the AI models and also with the IWQI. This may be related to the models’ structure, and the inputs applied for each model. Our findings agreed with the results of El Bilali and Taleb (2020). In contrast, our results disagree with Wang et al. (2020) who used only 17 samples in predicting anaerobic digestion performance. Although, increasing the data size of the model and applying the ensemble models play a critical role in improving the prediction accuracy of SVM (Chen et al., 2020; Zhou and Feng, 2019). Moreover, our

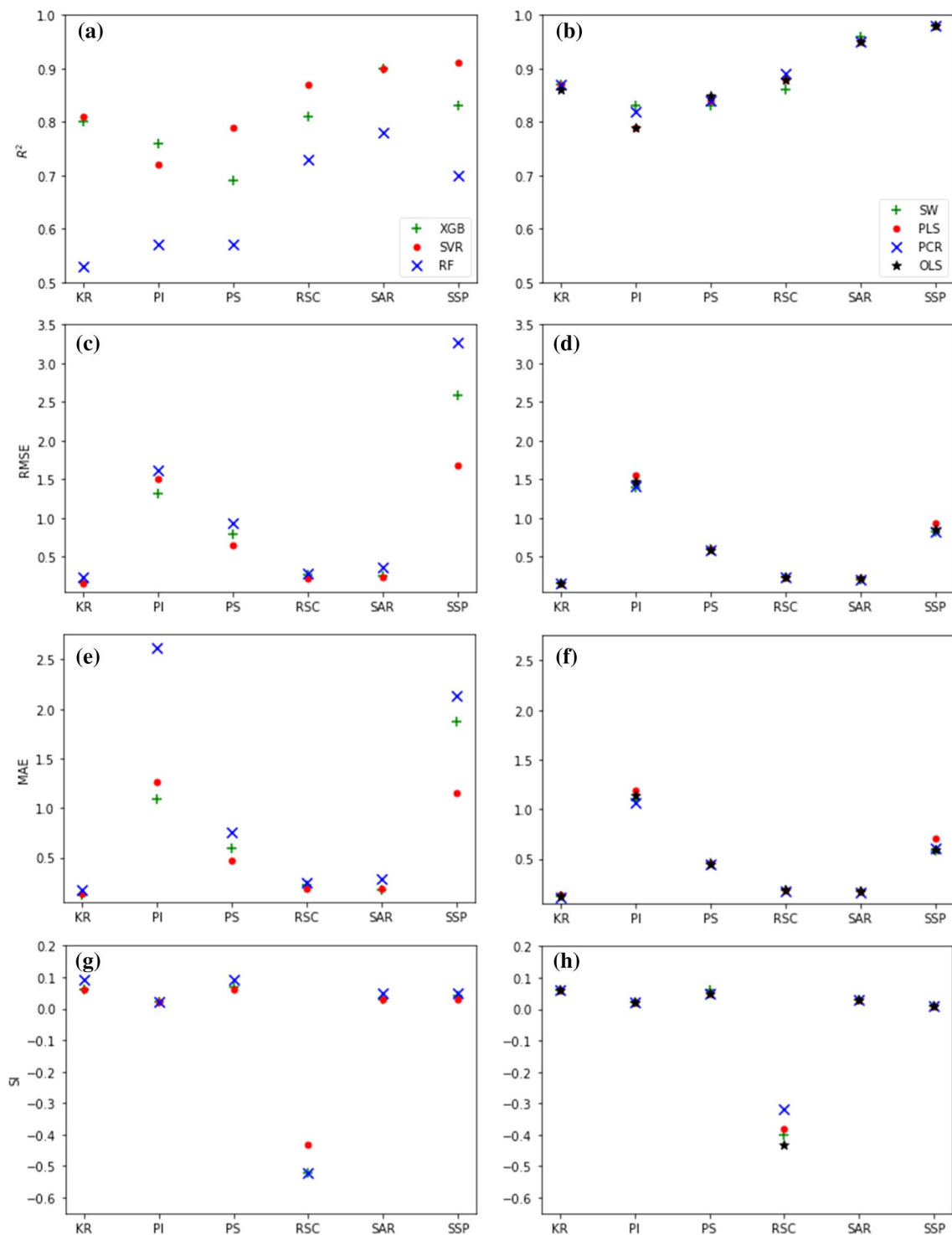


Fig. 3 R² (a, b), RMSE (c, d), MAE (e, f) and SI (j, h) values for the seven applied models

findings show a better predict of the PI index compared with El Bilali and Taleb (2020). Furthermore, it is discovered that stronger correlation between input and output variables reflects better model performance.

Conclusions

The study explored the capabilities of three machine learning algorithms (SVR, XGB and RF) and four multiple regressions (SW, PCR), PLS and OLS) for predicting six

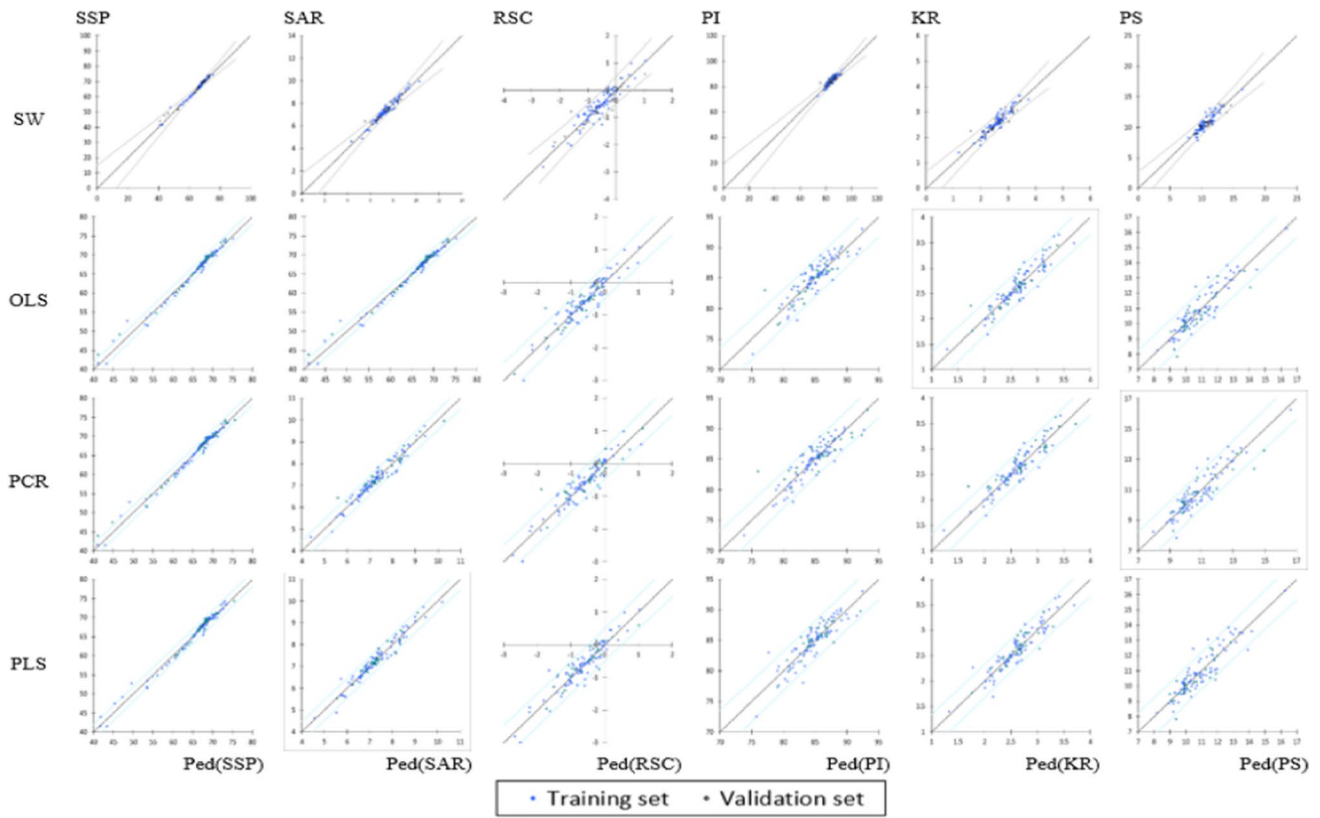


Fig. 4 Scatterplots of the estimated and calculated values of the IWQIs for the applied models (SW, OLS, PCR and PLS)

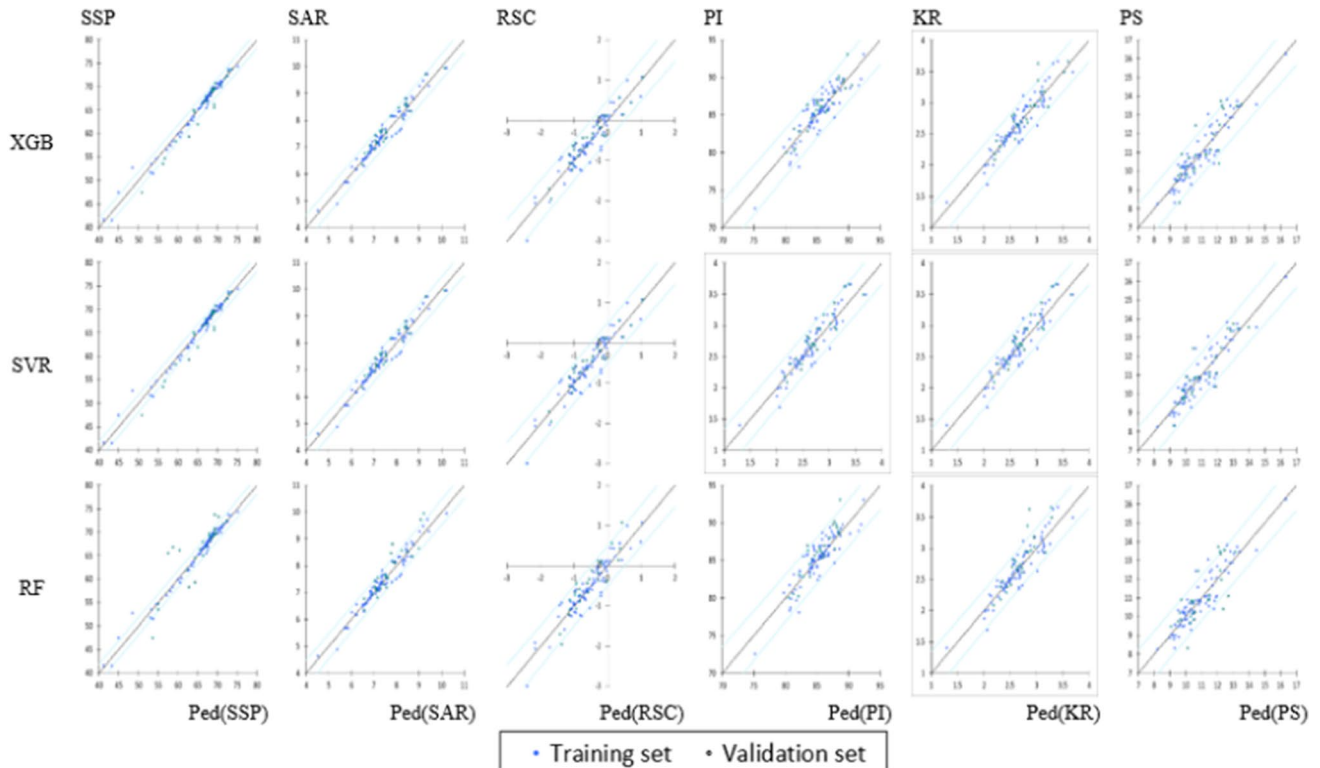


Fig. 5 Scatterplots of the estimated and calculated values of the IWQIs for the applied models (XGB, SVR and RF)

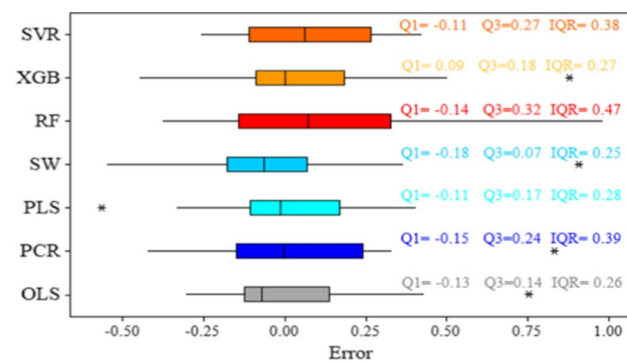


Fig. 6 Boxplots depicting the distribution of SAR estimate errors in the test section for the models under consideration. Q25: lower quartile of errors, Q75: upper quartile of errors, IQR: inter-quartile range for each model

different IQWI (KR, PI, PS, RSC, SAR and SSP) of the Bahr El-Baqr drain irrigation water source. Chemical composition of 105 water samples, collected during July 2020 at locations uniformly spread along the Bahr El-Baqr drain, was determined in the laboratory. The main conclusions are as follows:

- The pH of the Bahr El-Baqr drain water is within the acceptable limits for irrigation. The EC values were high rendering the water unsuitable for irrigation process unless the soil has good drainage, a special management plan for salinity control is put in place, and/or salt tolerant plants are used.
- According to the SSP and SAR, the water can be used for irrigation on most soils with little risk of dangerous amounts of exchangeable salt developing (low sodicity). Furthermore, the water from the Bahr El-Baqr Drain is suitable for agriculture without alkalinity hazard development and impairment of soil permeability.
- SW emerged as the optimal regression model for predicting the IWQI. For the AI models, SVR was the best, although SW marginally performed better.
- The outcome of this study that modelled the IWQI in Bahr El-Baqr drain, Egypt, using SW is satisfactory. Hence, the SW model is a useful decision tool for agricultural policy decision-makers to help improve irrigation water quality.

Author contributions AM and MK involved in conceptualization; AM and MK involved in methodology; AM and MK involved in software; AM, MK and YGA involved in formal analysis; AM, MK and YGA involved in investigation; MK involved in resources; AM and MK involved in data curation; AM, MK, NA, AE and YGA involved in writing—original draft preparation; AM, MK and YGA involved in

writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding None.

Declaration

Conflict of interest The authors declare 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 S, Hadi SJ, Sammen SS, Salih SQ, Abdulkadir R, Pham QB, Yaseen ZM (2020) Evolutionary computational intelligence algorithm coupled with self-tuning predictive model for water quality index determination. *J Hydrol* 587:124974
- Abdel-Fattah MK, Mokhtar A, Abdo AI (2020) Application of neural network and time series modeling to study the suitability of drain water quality for irrigation: a case study from Egypt. *Environ Sci Pollut Res* 28(1):898–914
- Abdel-Fattah M, Helmy A (2015) Assessment of water quality of wastewaters of Bahr El-Baqr, Bilbies and El-Qalyubia drains in east delta, Egypt for irrigation purposes. *Egypt J Soil Sci* 55:287–302
- Abdel-Shafy HI, Aly RO (2002) Water issue in Egypt: resources, pollution and protection endeavors. *Central Eur J Occupat Environ Med* 8:3–21
- Aboukarima AM, Al-Sulaiman MA, El Marazky MS (2018) Effect of sodium adsorption ratio and electric conductivity of the-applied water on infiltration in a sandy-loam soil. *Water SA* 44(1):105–110
- Addinsoft (2019) XLSTAT statistical and data analysis solution. Addinsoft Long Island: New York
- Agarwal RR, Yadav JSP, Gupta RN (1982) Saline soils of India, Indian Council of Agricultural Research, New Delhi
- Agassi M, Shainberg I, Morin J (1981) Effect of electrolyte concentration and soil sodicity on infiltration rate and crust formation. *Soil Sci Soc Am J* 45:848–861
- Aggag AM (2016) Evaluation of water quality and heavy metal indices of some water resources at Kafr El-Dawar region, Egypt. *Alexandria Sci Exchange J* 37:337–348
- Ahmed AN, Othman FB, Afan HA, Ibrahim RK, Fai CM, Hossain MS, Ehteram M, Elshafie A (2019) Machine learning methods for better water quality prediction. *J Hydrol* 578:124084
- Al-Amry, A. S. (2008) Hydrogeochemistry and groundwater quality assessment in an arid region: a case study from Al Salameh Area, Shabwah, Yemen. In: the 3rd international conference on water resources and arid environments, the 1st Arab water Forum
- Apha (1998) American Public Works Association and Water Environment Federation, 1998. Standard methods for the examination

- of water and wastewater, 20th Edition: American Public Health Association, Washington, DC, pp. 9–26
- Assar W, Ibrahim MG, Mahmud W, Allam A, Tawfik A, Yoshimura C (2020) Effect of water shortage and pollution of irrigation water on water reuse for irrigation in the Nile Delta. *J Irrig Drain Eng* 146:05019013
- Ayers, R. & Westcot, D. (1985) Water quality for agriculture. FAO Irrigation and drainage paper 29 Rev. 1. Food and Agricultural Organization. Rome, 1: 74
- Breiman L (2001) Random forests. *Mach Learn* 45:5–32
- Charulatha G, Srinivasulu S, Maheswari OU, Venugopal T, Giridharan L (2017) Evaluation of ground water quality contaminants using linear regression and artificial neural network models. *Arab J Geosci* 10:128
- Chen J-L, Li G-S, Wu S-J (2013) Assessing the potential of support vector machine for estimating daily solar radiation using sunshine duration. *Energy Convers Manage* 75:311–318
- Chen, T. & Guestrin, C (2016) Xgboost: a scalable tree boosting system. In: proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016. 785–794
- Chen K, Chen H, Zhou C, Huang Y, Qi X, Shen R, Liu F, Zuo M, Zou X, Wang J (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
- Chenini I, Khemiri S (2009) Evaluation of ground water quality using multiple linear regression and structural equation modeling. *Int J Environ Sci Technol* 6:509–519
- Doneen LD (1964) Notes on water quality in agriculture. Department of Water Science and Engineering, University of California, Davis
- Doran J, Coleman D, Bezdicsek D, Stewart B (1994) A framework for evaluating physical and chemical indicators of soil quality. *Defining Soil Qual Sustain Environ* 35:53–72
- Draper NR, Smith H (1981) Applied regression analysis. Wiley, NY
- EL Bilali A, Taleb A (2020) Prediction of irrigation water quality parameters using machine learning models in a semi-arid environment. *J Saudi Soc Agric Sci* 19:439–451
- Efroymsen, M. (1960) Multiple regression analysis. *Mathematical methods for digital computers*, pp. 191–203
- Elbeltagi A, Kumari N, Dharpure JK, Mokhtar A, Alsafadi K, Kumar M, Mehdinejadani B, Ramezani Etedali H, Brouziyne Y, Islam T (2021c) Prediction of combined terrestrial evapotranspiration index (CTEI) over large river basin based on machine learning approaches. *Water* 13:547
- Elbeltagi A, Deng J, Wang K, Hong Y (2020) Crop water footprint estimation and modeling using an artificial neural network approach in the Nile Delta, Egypt. *Agric Water Manag* 235:106080
- Elbeltagi A, Zhang L, Deng J, Juma A, Wang K (2020) Modeling monthly crop coefficients of maize based on limited meteorological data: a case study in Nile Delta, Egypt. *Comput Electr Agric* 173:105368
- Elbeltagi A, Aslam MR, Mokhtar A, Deb P, Abubakar GA, Kushwaha N, Venancio LP, Malik A, Kumar N, Deng J (2021) Spatial and temporal variability analysis of green and blue evapotranspiration of wheat in the Egyptian Nile Delta from 1997 to 2017. *J Hydrol* 594:125662
- Elbeltagi A, Azad N, Arshad A, Mohammed S, Mokhtar A, Pande C, Etedali HR, Bhat SA, Islam ARMT, Deng J (2021) Applications of Gaussian process regression for predicting blue water footprint: case study in Ad Daqahliyah, Egypt. *Agric Water Manag* 255:107052
- 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*. <https://doi.org/10.1007/s11356-021-17064-7>
- Fan J, Wang X, Wu L, Zhou H, Zhang F, Yu X, Lu X, Xiang Y (2018) Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: a case study in China. *Energy Convers Manage* 164:102–111
- Ferreira LB, Da Cunha FF (2020) Multi-step ahead forecasting of daily reference evapotranspiration using deep learning. *Comput Electr Agric* 178:1057
- Fipps, G. (2003) Irrigation water quality standards and salinity management strategies. Texas FARMER Collection
- Gaya MS, Abba SI, Abdu AM, Tukur AI, Saleh MA, Esmaili P, Wahab NA (2020) Estimation of water quality index using artificial intelligence approaches and multi-linear regression. *Int J Artif Intell ISSN* 2252:8938
- 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
- Gholami S, Srikantaswamy S (2009) Analysis of agricultural impact on the Cauvery river water around KRS dam. *World Appl Sci J* 6:1157–1169
- Gupta SK, Gupta I (1997) Management of saline soils and waters. Scientific Publishers, Rajasthan
- Hameed M, Sharqi SS, Yaseen ZM, Afan HA, Hussain A, Elshafie A (2017) Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia. *Neural Comput Appl* 28:893–905
- Hamzeh Haghbi A, Nasrolahi A, Parsaie A (2018) Water quality prediction using machine learning methods. *Water Qual Res J* 53:3–13
- Hocking RR (1976) A Biometrics invited paper. The analysis and selection of variables in linear regression. *Biometrics* 32:1–49
- Huang Z, Turner BJ, Dury SJ, Wallis IR, Foley WJ (2004) Estimating foliage nitrogen concentration from HYMAP data using continuum removal analysis. *Remote Sens Environ* 93:18–29
- Ismael M, Mokhtar A, Farooq M, Lü X (2021) Assessing drinking water quality based on physical, chemical and microbial parameters in the Red Sea State, Sudan using a combination of Water quality index and artificial neural network model. *Groundwater Sustain Dev* 14:100612
- Jia R, Fang S, Tu W, Sun Z (2016) Driven factors analysis of China's irrigation water use efficiency by stepwise regression and principal component analysis. *Discrete Dynamics in Nature and Society*. <https://doi.org/10.1155/2016/8957530>
- Juahir H, Zain SM, Toriman ME, Mokhtar M, Man HC (2004) Application of artificial neural network models for predicting water quality index. *Malaysian J Civ Eng* 16:42–55
- Kelley W (1963) Use of saline irrigation water. *Soil Sci* 95:385–391
- Kouadri S, Elbeltagi A, Islam ARM, Kateb S (2021a) Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast). *Appl Water Sci* 11:1–20
- Kouadri, S., Pande, C. B., Panneerselvam, B., Moharir, K. N. & Elbeltagi, A. 2021b. Prediction of irrigation groundwater quality parameters using ANN, LSTM, and MLR models. *Environmental Science and Pollution Research*, 1–25.
- Kumar VS, Amarendra B, Dhakate R, Sankaran S, Kumar KR (2016) Assessment of groundwater quality for drinking and irrigation use in shallow hard rock aquifer of Pudunagaram, Palakkad District Kerala. *Appl Water Sc* 6(2):149–167
- Leong, W. C., Bahadori, A., Zhang, J. & Ahmad, Z. 2019. Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM). *International Journal of River Basin Management*, 1–8.

- Li M-F, Tang X-P, Wu W, Liu H-B (2013) General models for estimating daily global solar radiation for different solar radiation zones in mainland China. *Energy Convers Manage* 70:139–148
- Longenecker DE, Thaxton Jr EL, Lyerly PI (1969) Salt concentration in soils furrow irrigated with saline waters, Texas Agric. Exp. Stn. Misc. Publ. MP. 939, College Station
- Lu, H. & Ma, X. 2020. Hybrid decision tree-based machine learning models for short-term water quality prediction. *Chemosphere*, 249, 126169.
- Machiwal D, Cloutier V, Güler C, Kazakis N (2018) A review of GIS-integrated statistical techniques for groundwater quality evaluation and protection. *Environmental Earth Sciences* 77:681
- Mohammadpour R, Shaharuddin S, Chang CK, Zakaria NA, Ab Ghani A, Chan NW (2015) Prediction of water quality index in constructed wetlands using support vector machine. *Environ Sci Pollut Res* 22:6208–6219
- Moharir K, Pande C, Singh SK, Choudhari P, Kishan R, Jeyakumar L (2019) Spatial interpolation approach-based appraisal of groundwater quality of arid regions. *J Water Supply Res Technol AQUA* 68:431–447
- Mokhtar, A., He, H., He, W., Elbeltagi, A., Maroufpoor, S., Azad, N., Alsafadi, K. & Gyasi-Agyei, Y. 2021a. Estimation of the rice water footprint based on machine learning algorithms. *Computers and Electronics in Agriculture*, 191, 106501.
- Mokhtar, A., Jalali, M., Elbeltagi, A., Al-Ansari, N., Alsafadi, K., Abdo, H. G., Sammen, S. S., Gyasi-Agyei, Y., Rodrigo-Comino, J. & He, H. 2021b. Estimation of SPEI Meteorological Drought using Machine Learning Algorithms. *IEEE Access*.
- Prunty L, Montgomery BR, Sweeney MD (1991) Water quality effects on soils and alfalfa: 1. Water use, yield, and nutrient concentration. *Soil Sci Soc Am J* 55:196–202
- Nayak, S. 2020. Assessment of Water Quality of Brahmani River using Correlation and regression Analysis.
- Noori R, Sabahi MS, Karbassi AR, Baghvand A, Zadeh HT (2010) Multivariate statistical analysis of surface water quality based on correlations and variations in the data set. *Desalination* 260:129–136
- Odemis B, Evrendilek F (2007) Monitoring water quality and quantity of national watersheds in Turkey. *Environ Monit Assess* 133:215–229
- Pande CB, Moharir K (2018) Spatial analysis of groundwater quality mapping in hard rock area in the Akola and Buldhana districts of Maharashtra, India. *Appl Water Sci* 8:1–17
- Pande CB, Moharir KN, Singh SK, Dzwairo B (2020) Groundwater evaluation for drinking purposes using statistical index: study of Akola and Buldhana districts of Maharashtra, India. *Environ Dev Sustain* 22:7453–7471
- Panneerselvam B, Muniraj K, Pande C, Ravichandran N (2021) Prediction and evaluation of groundwater characteristics using the radial basic model in Semi-arid region, India. *Int J Environ Anal Chem*. <https://doi.org/10.1080/03067319.2021.1873316>
- Pearson K (1901) On lines of closes fit to system of points in space, London, E dinb. Dublin Philos Mag J Sci 2:559–572
- Pleguezuelo CRR, Rodríguez BC, Tejero IFG, Ruíz BG, Tarifa DF, Martínez JRF & Zuazo VHD (2018) Irrigation strategies for mango (*Mangifera indica* L.) Under water-scarcity scenario in the Mediterranean subtropical environment. Water scarcity and sustainable agriculture in semiarid environment. Elsevier
- Raghunath, H. M. 1987. *Ground water*, New Age International.
- Raihan F, Alam B (2008) Assessment of groundwater quality in sunamladesh. Iran. *J Environ Health Sci Eng* 5(3):155–156
- Richards, L. (1954a) Diagnosis and improvement of saline and alkaline soils-United States Department of Agriculture Handbook No. 60. United States Government Printing Office, Washington DC
- Richards LA (1954) Diagnosis and improvement of saline and alkali soils. *Soil Sci* 78(2):154
- Sadick A, Asante PC, Dugan E, Asaana J (2017) Correlation analysis of irrigation water quality parameters from Lake Bosomtwe in the Ashanti Region of Ghana. *Sci Res J Agric* 2(2)
- Sakizadeh M (2016) Artificial intelligence for the prediction of water quality index in groundwater systems. *Model Earth Syst Environ* 2:8
- Sarker BC, Hara M, Zaman MW (2000) Suitability assessment of natural water in relation to irrigation and soil properties. *Soil Sci Plant Nutr* 46(4):773–786
- Todd DK, Mays LW (2004) *Groundwater hydrology*. John
- Tung TM, Yaseen ZM (2020) A survey on river water quality modelling using artificial intelligence models: 2000–2020. *J Hydrol* 585:124670
- Vapnik V (2013) *The nature of statistical learning theory*. Springer, Berlin
- Viswanath NC, Kumar PD, Ammad K (2015) Statistical analysis of quality of water in various water shed for Kozhikode City, Kerala, India. *Aquatic Procedia* 4:1078–1085
- Wagh VM, Panaskar DB, Muley AA, Mukate SV, Lolage YP, Aamalar ML (2016) Prediction of groundwater suitability for irrigation using artificial neural network model: a case study of Nanded tehsil, Maharashtra, India. *Model Earth Syst Environ* 2:1–10
- Wang L, Long F, Liao W, Liu H (2020) Prediction of anaerobic digestion performance and identification of critical operational parameters using machine learning algorithms. *Bioresour Technol* 298:1224
- Wold, S. 1995. PLS for multivariate linear modeling. *Chemometric methods in molecular design*, 195–218
- Xu P, Feng W, Qian H, Zhang Q (2019) Hydrogeochemical characterization and irrigation quality assessment of shallow groundwater in the central-western Guanzhong basin, China. *Int J Environ Res Public Health* 16(9):1492
- Yasin MI, Karim SAA (2020) A new fuzzy weighted multivariate regression to predict water quality index at Perak Rivers. In: Karim SAA, Kadir EA, Nasution AH (eds) Optimization based model using fuzzy and other statistical techniques towards environmental sustainability. Springer, Berlin
- Yildiz S, Karakuş CB (2019) Estimation of irrigation water quality index with development of an optimum model: a case study. *Environ Dev Sustain* 22(5):4771–4786
- Yildiz S, Degirmenci M (2015) Estimation of oxygen exchange during treatment sludge composting through multiple regression and artificial neural networks (estimation of oxygen exchange during composting). *Int J Environ Res* 9:1173–1182
- Zaman M, Shahid SA, Heng L (2018) Irrigation water quality. Guideline for salinity assessment, mitigation and adaptation using nuclear and related techniques. Springer, Cham, pp 113–131
- Zhou Z-H, Feng J (2019) Deep forest. *Natl Sci Rev* 6:74–86

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