



# Applications of various data-driven models for the prediction of groundwater quality index in the Akot basin, Maharashtra, India

Ahmed Elbeltagi<sup>1</sup> · Chaitanya B. Pande<sup>2,5</sup> · Saber Kouadri<sup>3</sup> · Abu Reza Md. Towfiqul Islam<sup>4</sup>

Received: 18 May 2021 / Accepted: 12 October 2021 / Published online: 20 October 2021  
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

## Abstract

Data-driven models are important to predict groundwater quality which is controlling human health. The water quality index (WQI) has been developed based on the physicochemical parameters of water samples. In this area, water quality is medium to poor and is found in saline zones; very high pH ranges are directly affected on the water quality in this study area. Conventional WQI computation demands more time and is often observed with enormous errors during the calculation of sub-indices. In the present work, four standalone methods such as additive regression (AR), M5P tree model (M5P), random subspace (RSS), and support vector machine (SVM) were employed to predict WQI based on variable elimination technique. The groundwater samples were collected from the Akot basin area, located in the Akola district, Maharashtra, in India. A total of nine different input combinations were developed in this study. The datasets were demarcated into two classes (ratio 80:20) for model construction (training dataset) and model verification (testing dataset) using a fivefold cross-validation approach. The models were assessed using statistical and graphical appraisal metrics. The best input combinations varied among the model, generally, the optimal input variables (EC, pH, TDS, Ca, Mg, and Cl) during the training and validation stages. Results show that AR outperformed the other data-driven models ( $R^2 = 0.9993$ ,  $MAE = 0.5243$ ,  $RMSE = 0.0.6356$ ,  $\%RAE = 3.8449$ , and  $RRSE\% = 3.9925$ ). The AR is proposed as an ideal model with satisfactory results due to enhanced prediction precision with the minimum number of input parameters and can thus act as the reliable and precise method in the prediction of WQI at the Akot basin.

**Keywords** Water quality index · Data-driven models · Groundwater · Additive regression

Responsible Editor: Xianliang Yi

✉ Chaitanya B. Pande  
chaitanay45@gmail.com

Ahmed Elbeltagi  
ahmedelbeltagy81@mans.edu.eg

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

<sup>1</sup> Agricultural Engineering Department, Faculty of Agriculture, Mansoura University, Mansoura 35516, Egypt

<sup>2</sup> Sant Gadge Baba Amravati University, Amravati, MS 444602, India

<sup>3</sup> Laboratory of Water and Environment Engineering in Sahara Milieu (GEEMS), Department of Civil Engineering, Hydraulics Faculty of Applied Sciences, Kasdi Merbah University Ouargla, Ouargla, Algeria

<sup>4</sup> Department of Disaster Management, Begum Rokeya University, Rangpur 5400, Bangladesh

<sup>5</sup> CAAST-CSAWM, MPKV Rahuri, Rahuri, India

## Introduction

Groundwater is among the most important fresh water resource, and it provides different nations domestic and irrigation demands (Kazakis et al., 2017). The unprecedented population growth, urban expansion, intensive use of chemical fertilizers, climate change, and poor management of groundwater resource have worsened groundwater quality all over the world (Li et al., 2015; Busico et al., 2020; Islam et al., 2017, 2019). Despite deterioration of groundwater quality, however, the absence of alternative sources more demand due to rising human population (Saha et al., 2020). Groundwater is a part of the hydrological system and freshwater resource; erratic rainfall put groundwater system under stress (Ahmed et al., 2019; Islam et al., 2020a). Thus, appraisal of water quality is of a thrust area of research in recent times. Horton (1965) developed the first water quality index (WQI) method in order to transform the several parameters containing water into one single number to describe the

overall water quality. Various WQIs have been developed by many researchers to assess groundwater and surface water suitability for drinking, irrigation, and industrial and biodiversity use (Islam et al., 2017, 2020b; 2018; Banerji and Mitra, 2019; Abbasnia et al., 2019; Kabir et al., 2021). One of the key challenges of this qualitative evaluation method is that it needs expert knowledge in the allocation of variable weights for calculating the WQI score, which means that the real result is unclear (Amiri et al., 2014; Gorgij et al., 2017). Several authors have studied, by assigning entropy-based weights to major ions, to reduce the subjectivity of the traditional WQI technique, which had shown to be a more precise, useful tool to the accurate weighing system (Fagbote et al., 2014; He and Wu, 2019). However, study of groundwater quality includes data collection and laboratory analysis, at a huge scale, testing, and data management (Tiyasha et al., 2020).

Meanwhile, due to the subjectivity of WQI's computation, it has contradictions in its result interpretation. It may be evident from the previous literature, but there is no ideal or universal WQI model. To address this issue, some research scholars have opted for a non-physical tool, successfully forecasting WQI using artificial intelligence (AI) models (Yassen et al., 2018; Leong et al., 2019; El Bilali et al., 2021). Therefore, it is crucial for reliable water quality evaluation to implement a potential and cost-effective technique. In such a case, AI-based models have reduced sub-index computations and rapidly generate a WQI value. Attention to AI approach is paid due to benefit that includes their non-linear frameworks, capability to forecast complicated events and to handle large scale datasets, and not sensitive to missing dataset (Bui et al., 2020). The predictive ability of AI approach depends on the model and precision of data acquisition and analytical procedure.

The AI technique has a potential and robust multi-functioning tool in water science-related fields (Babbar and Babbar 2017; Kisi et al., 2018; Leong et al., 2019; Bui et al., 2020; Abba et al., 2020; Singha et al., 2021; El Bilali et al., 2021; Adnan et al., 2021; Ahmadi et al., 2021; Babae et al., 2021; Bajirao et al., 2021; Elbeltagi et al., 2021, 2020a, 2020b, 2020c, 2020d; Jerin et al., 2021; Kumar et al., 2021; Mokhtar et al., 2021; Suryakant et al., 2021; Zerouali et al., 2021). Several research scholars have employed AI techniques including random forest (RF), support vector machine (SVM), and artificial neural network (ANN) worldwide in different water-related studies. The RF model was applied for groundwater quality prediction (Singha et al., 2021), flood susceptibility study (Islam et al., 2021), river water quality prediction (Bui et al., 2020), and others. Likewise, the SVM model was adopted for predicting marine water quality (Deng et al., 2021) and wastewater treatment plant monitoring (Nourani et al., 2018) at different precision levels (Islam et al., 2021; Zhu et al., 2020; Gazzaz et al.

2012). 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 3 years at seven sampling points in Mirim Lagoon. Based on parameters such as pH, dissolved oxygen, conductivity, turbidity, fecal coliform, and temperature, Hu et al. (2021) investigate 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 discussion of the previous literature review, it is apparent that different machine learning models have performed very well in various hydrogeological conditions; it is given a better accuracy levels. In this context, additive regression (AR) and random subspace (RSS) were applied on water quality index data for the estimation of WQ index value for semi-arid region; it is improving the reliability of water quality evaluation. However, MI models are scarcely used in the water field for the prediction of groundwater quality index and other water-related researches. Besides, after thoroughly reviewing earlier literatures, to the best of the author's knowledge, no previous studies have investigated and verified the performance of these above-mentioned models for the prediction of groundwater quality index. Thus, to fill this knowledge gap, the current study has been used four machine learning model based on the estimation of WQI prediction values within the Akot basin of the southwest India. Groundwater acts as a vital source of human used and consumption, and water quality may be affected by human-induced pollution; hence, this study analysis is scientific based on an evaluation of groundwater quality for this basin. In addition, such scientific investigation is not performed in these areas. The four ML models are a more robust tool than appraising it with any standalone tool for the estimation of WQI. Hence, the important goal of this study is to develop the ML models for WQI value prediction based on groundwater quality data; the proper groundwater sampling must be important for the development of models, and then the third and important objective is which model can be given a high accuracy and identify

the best prediction model for this study areas basis on performance metrics of models.

## Materials and method

### Study area

The study area is located in the Akola district in Maharashtra (India). The study area latitudes and longitudes in between 20°54'30"N and 76°48'1"E (Fig. 1). Total area of this study area is 450 km<sup>2</sup>. The study area location is the mainstream of the Purna river basin which is a west flowing situated in the Maharashtra state. The mainstream basin area is the maximum under the alluvium zones with this formation most of the land directly affected on the soil than groundwater parameters. We have observed that a very high (pH = 8.4) value is found in study area. Most of the villages have been facing saline water problems (EC = 4332 ds/m), while this groundwater quality is continuing may be to affect the human health in the study area. Hence, the predication of the groundwater quality index can be more useful for planning of drinking, irrigation, and human health because in the future, groundwater should be a more important source for drinking purposes to living organism (Moharir et al. 2019). In the study area, most of the agriculture land is under dry-land zones and no conserved the rainfall water in the aquifer region. The normal yearly precipitation is 700–850 mm in the entire Akot basin. The basaltic rock is found in the upper side of basin (Pande et al. 2020). The purposive sampling strategy was applied to collect the groundwater samples and analyzed in the laboratory to determine the different ionic concentration through standard procedure.

### Sampling and analytical procedure

A total of thirty-five water samples were gathered from observation location. These water samples were collected as per the random sampling. The samples were taken in a fresh 2-l bottles covered polyethylene bottle, which was systematically washed by analytical ranking 1:1 HCl and then cleaned with purified water. Polyethylene containers were washed with water samples 2–3 times at the collection point. The polyethylenes were totally occupied to volume, and temperature and pH were reserved directly on the spot. The samples were chosen to take to the ice-cooled research lab and refrigerated to a temperature of 4 °C until other variables were determined. The TDS, conductivity, and other parameters were examined according to APHA (2005) standard procedure (Arun et 2021). All of these parameters were evaluated in water laboratory.

## Water quality index

WQI is described as a mathematical equation that displays the impact of each of the groundwater quality factors on all water quality for drinking to human society (Yidana et al. 2010; Zhang et al. 2021). For every chemical parameter, weight value between 1 and 5 is mainly determined in the first. Here, weight value 1 is allocated to the parameters that may be “will deteriorate” the water quality of drinking before the smallest, whereas 5 is allocated of weight value to the factors that may be “will affect” the drinking water quality (Table 2). Study of various water quality parameters such as EC, pH, TDS, Ca, Mg, Cl, SO<sub>4</sub>, Na, and K was given the greatest weight of “5” because these parameters are used to define the quality of fresh water (Pande et al. 2018; Panneerselvam et al. 2021; Sinha et al. 2021). The weight “1” is allocated to SO<sub>4</sub> parameter because it has the smallest significance in water quality estimation. Table 1 presents the assigned weights, relative weights, and limitations needed by the WHO. The relative weight values ( $W_i$ ) for each parameter were calculated using Eq. (1) (WHO 2011):

$$W_i = \frac{w_i}{\sum_{i=1}^n w_i} \quad (1)$$

where  $W_i$  is the relative weight,  $w_i$  is the weight of each factor, and  $n$  is the number of parameters.

A quality rating scale ( $q_i$ ) for every parameter is computed according to Eq. (2):

$$q_i = \left( \frac{C_i}{S_i} \right) \times 100 \quad (2)$$

where  $q_i$  is the excellence score,  $C_i$  is the chemical concentration of water sample, and  $S_i$  ( $i$  should an index,  $S_i$ ) is the WHO drinking water quality standard.

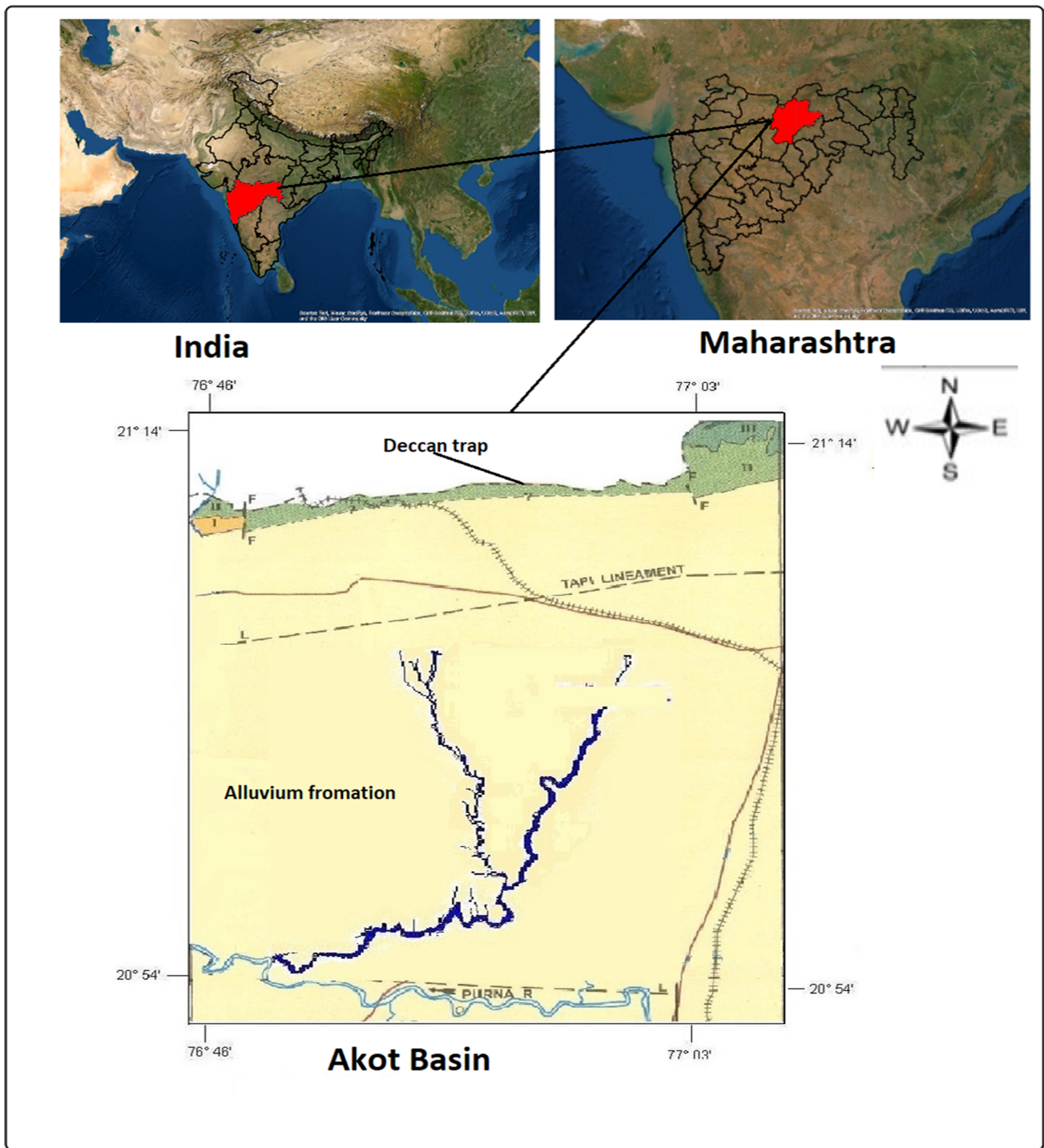
Lastly, the WQI is measured according to Eq. (3):

$$WQI = \sum Sli \quad (3)$$

where  $Sli$  is the sub-index of the  $i$ th parameter.

## Methodology

The suggested procedure and models are used for the estimation of water quality index; it is divided into four levels such as collection of groundwater samples, laboratory analysis, estimation of water quality index, and development of machine learning models. These four level results have been used for the predication of water quality index based on the machine learning models (Balamurugan Panneerselvam et al. 2021). The application of prediction models has involved the data standardization with splitting of data,



**Fig. 1** Location map of the study area showing the Akot basin, Maharashtra, India

which was further used for training and testing. Based on the WQI data the random subspace, support vector machine, M5P, and additive regression models were created. In model calibration, the optimization of ML models by well modification with training information and the established ML model's validation added for the process of the validation of

optimized ML models with testing dataset and model performance was assessed using statistical tool (*MAE*, *RMSE*, *RAE%*, *RRSE%*) and choice of the finest prediction model. The flow chart is illustrated as in Fig. 2, which shows the different stages that are used, while the complete material about the phases is defined in the subsequent sections.



**Table 1** Details of physical and chemical factors, the WHO (2011) standard, and weight ( $w_i$ ) and relative weight ( $W_i$ ) based on the WHO standards

Sr. no.	Water quality parameters	WHO water quality standard (2011)	Weight ( $w_i$ )	Relative weight $W_i = \frac{w_i}{\sum_{i=1}^N w_i}$
1	pH	6.5-8.5	5	0.16
2	TDS	500	5	0.16
3	Ca	300	3	0.094
4	Mg	30	3	0.094
5	Cl	250	4	0.12
6	SO <sub>4</sub>	200	5	0.16
7	Na	200	4	0.12
8	K	10	2	0.062
<b>Total</b>			$\sum W_i=31$	$\sum W_i=1.00$

**Random subspace**

RSS is a popular ensemble technique, created by Ho (1998), which provides the different classifier with a pseudo-randomly selected subset of features and combines their results with voting. RSS is an entity classification forest construction method to progress the recital of the weak individual classifications (Pham et al. 2017). The RSM may be supported from random subspaces for two structures and merging the classifiers. When the sum of training occurrences is moderately minor as associated to the dimension of data, by structure classifiers in random subspaces, the little sample size issue can be explained. The subspace dimension will be a smaller number than the unique element space; the number of incidences of training is the same. This increases the sample size of the relative training. The better classifier can be identified in random subspaces when the data has several redundant features than in the unique feature space (Skurichina et al. 2002). The integration of training sample S and classification systems for the development of hybrid models voted in the rule of the final decision. The RSS algorithm can be read accordingly in Eq. (4):

$$\gamma(s) = \operatorname{argmax} \sum_d \delta_{sng}(C^d(s)), y \tag{4}$$

where  $\delta_{i,j}$  is the Kronecke sign,  $y = (-1, 1)$  is a class label of the classifier, and  $C^d(s)$  is the classifier ( $d = 1, 2, 3, \dots, D$ ).

**Support vector machine**

SVMs are used mainly for classification purposes but may be used for the regression analysis. SVMs describe a hyper-plane between groups and increase the range to ensure maximum a difference in between both the classes by displaying data points presented on the plane, resulting in reduced close

errors. The training data would be separated non-linearly (Tong et al. 2001). A non-linear separable boundary must then be built. In order to create a non-linear border, the original space needs to be mapped to a high dimension. A function of the kernel defines how to map the space in a particular input space. A penalty factor ( $c$ ) for misclassification was added for the optimization of the model. The all punish in plotting is determined by totaling the drawbacks on every misclassification. The various helpful applications of SVM method have been identified in the groundwater and hydrological engineering (Asefa et al., 2005; Raghavendra et al. 2014; Nguyen 2017).

The deficiency of the finest result due to the curved nature of the target purpose in the SVM model is some limitations. The SVM work depending on the principle of basic hazard-less was carried out to mitigate the simplification slightly in the training errors.

Equation (5) presented the SVM algorithm:

$$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\} \tag{5}$$

$T$  is the training dataset and is considered in this equation.

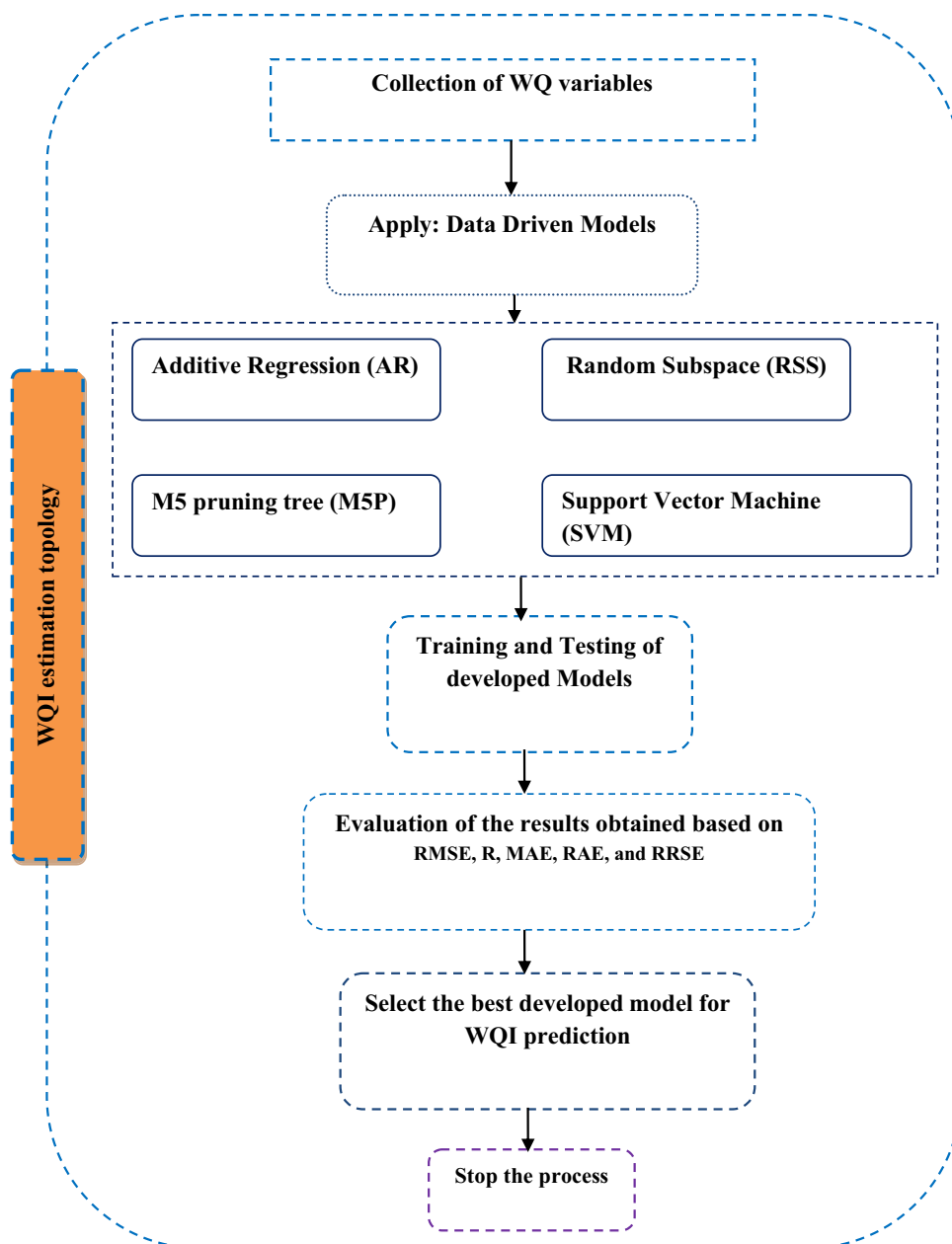
**M5P model**

M5P-tree has first introduced a genetic algorithm to regression issues (Mohammed et al. 2020). In a multivariate linear regression model, this decision tree model defines a linear regression at the terminal node and fits for every sub-location by categorizing or splitting various data areas into several various areas. Error assessment is shown with data on the M5P-tree algorithm tree separation norms on each node. Errors were calculated by the default variance range of the class involving the node. The attribute maximizes the predicted error reduction to approximate certain features of this node. The data is captured based on error calculations per node on the M5P-tree model tree separating criteria. The M5P error is determined by the default class variation in the node. The function is selected which maximizes the expected error reduction by evaluating every attribute for this node. After evaluating all possible structures, select a device with the highest potential error reduction. This division also creates a large tree-like structure leading to overlap. The second stage cuts the huge tree and replaces the trimmed sub-trees with linear regression functions. Equation (6) shows the M5P tree model:

$$SDR = sd(T) \sum \left| \frac{T_i}{T} \right| \times sd(T_i) \tag{6}$$

where  $T$  are sets of examples that reaches the node, “SD” denotes standard deviation, and  $T_i$  are the sets following from the separated node as per the given attribute and fragmented value.

**Fig. 2** Flowchart of WQI estimation methodology for a study area



### Additive regression

Methodology trees are created for regression models in which the portion of the data set entering the leaf is identified by a linear regression model in each branch of the tree. It is a mixture of the conventional tree of judgment with the probability of linear regression at the nodes. Pruning, evacuation, and restoration of trees are included in the process. The generalized additive model (GAM) was implemented by Hastie and Tibshirani (1986) and considered an extension of the generalized linear model (GLM). GLM model is based on the clear assumption that the parameters are linear, but GAM assumes no dependence, and that the relationship is

not always linear (Laanaya et al., 2017). In that model, linear dependence is replaced by more general features of dependency (Fu et al., 2019). The equation used for this algorithm is written as (Eq. 7)

$$Y = \beta_0 + \sum_{i=1}^p f_j(X_j) + \epsilon \quad (7)$$

In this study, four latest machine models were selected for the prediction of water quality index for the study area. Various parameters of machine learning models were used to the development of water quality index modeling. Random subspace algorithm presented the various parameters, viz., Batch size-100, Classifier = REP Tree, random seed-1,

subspace size = 0.5, numbers of executions slots = 1, and number of iterations= 10. Support vector machine, M5P, and additive regression algorithms showed that the different parameters are Batch size-100, C=0.1, kernel used= poly kernel; Batch size-100, Minimum number of instances = 4; and Batch size-100, Classifier = Decision-stump, shrinkage=1, number of iterations= 10 (Table 2). The developed models of the various variables combination are shown in Table 3.

**Performance metrics**

Mean absolute error (MAE), relative absolute error (RAE), root relative squared error (RRSE), and root means square error (RMSE) values were utilized to evaluate the capability of the above-stated machine learning modeling approaches. Therefore, four general statistical calculations, MAE, RAE, RMSE, and RRSE, were utilized to the assessment to obtain the usefulness of machine learning methods. The four errors such as MAE values (Eq. 4), RAE (Eq. 5), RMSE (Eq. 6), and RRSE (Eq. 7) values display well model accuracy:

$$MAE = \frac{\sum_{i=1}^n |y_i - x_i|}{N} \tag{8}$$

$$RAE = \frac{\sum_{i=1}^n |p_i - a_i|}{\sum_{i=1}^n |\bar{a} - a_i|} \tag{9}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - x_i)^2}{N}} \tag{10}$$

$$RRSE = \sqrt{\frac{\sum_{j=1}^n (P_{(ij)} - T_j)^2}{\sum_{j=1}^n (T_j - \bar{T}_{jj})^2}} \tag{11}$$

where  $P_{(ij)}$  is the value predicted by the single algorithm  $i$  for the reported  $j$  (out of  $n$  data) and  $T_j$  is the target value for reported  $j$ .

**Table 3** The developed models at the different combinations of variables

Models	Input variables	Output
1	EC, pH, TDS, Ca, Mg, Cl, SO <sub>4</sub> , Na, K	WQI
2	EC, pH, TDS, Ca, Mg, Cl, SO <sub>4</sub> , Na	WQI
3	EC, pH, TDS, Ca, Mg, Cl, SO <sub>4</sub>	WQI
4	EC, pH, TDS, Ca, Mg, Cl	WQI
5	EC, pH, TDS, Ca, Mg	WQI
6	EC, pH, TDS, Ca	WQI
7	EC, pH, TDS	WQI
8	EC, pH	WQI
9	EC	WQI

**Results and discussion**

**Water quality index and statistical analysis**

The minimum and maximum values of WQI ranged between 47.50 and 100; the various water quality parameters have been included for the estimation of WQI values; these values are used to which observation wells are suitable for the irrigation and drinking entire study area. Based on the high WQI values shown the more water suitable in the winter and summer periods of semi-arid region (Arun et al. 2021). The more the land is under clay soil (alluvium formation) with flat surface, the land acts as filter to reduce the pollution level in groundwater after the infiltration. The observed WQI data has been used and processed in the machine learning models and models in which suitable for these areas in various conditions (Pande et al. 2020). In this context, we have chosen four best models for the estimation of future WQI values particular in this area. Based on these models, the results will prepare a suitable plan for the development of groundwater quality and maintain the water quality status in this area.

The statistical study of water quality parameters and water quality index datasets used during the process of collected, training, and testing results is presented in Tables 4–6, which includes various statistical parameters like mean, standard error of mean (SE), mean standard deviation (St.

**Table 2** The parameters of the machine learning algorithms used for water quality index modeling

Model name	Description of parameters
Random subspace (RSS)	Batch size-100, Classifier = REPTree, random seed-1, subspace size = 0.5, numbers of executions slots = 1, number of iteration= 10
Support vector machine (SVM)	Batch size-100, C=0.1, kernel used=poly kernel
M5P	Batch size-100, Minimum number of instances = 4
Additive regression (AR)	Batch size-100, Classifier = Decision-stump, shrinkage=1, number of iterations= 10

**Table 4** Descriptive statistics of the all collected data

Variable	Mean	SE Mean	St. Dev	Minimum	Q1	Median	Q3	Maximum	BIS
EC	1647	191	1211	323	687	1342	1883	5432	1000
pH	7.48	0.049	0.31	7.0	7.2	7.5	7.6	8.4	8.5
TDS	988	116	735	198	409	847	1159	3265	500
Ca	46.08	7.00	44.25	8.00	18.00	37.50	51.25	227.00	75
Mg	93.0	10.6	66.8	14.0	48.3	70.0	125.0	358.0	30
Cl	302.1	40.6	256.7	14.0	148.3	223.0	398.0	1384.0	250
SO <sub>4</sub>	29.18	4.17	26.34	7.00	14.50	19.50	37.50	158.00	200
Na	72.67	8.36	52.85	2.44	28.25	70.00	100.50	209.00	200
K	12.96	1.74	11.01	0.30	6.00	10.50	17.00	58.00	12
WQI	82.44	2.26	14.31	47.50	70.63	82.50	95.00	100.00	

Dev.), minimum and maximum ranges, first quartile (Q1), and third quartile (Q3). These statistical factors present the unpredictability of information over the period. The average value of EC ( $1647 \pm 1211$   $\mu\text{g}/\text{cm}$ ), TDS ( $988 \pm 735$   $\text{mg}/\text{L}$ ), Cl ( $302.1 \pm 256.7$   $\text{mg}/\text{L}$ ), Mg ( $93.0 \pm 66.8$   $\text{mg}/\text{L}$ ), and K ( $12.96 \pm 11.01$   $\text{mg}/\text{L}$ ) surpasses the BIS (2012) threshold limit for human consumption in the groundwater of the Akot basin. It can be noticed from Table 4 that the mean and standard error values of all variables except for EC, TDS, and Cl showed a marginal variation. The pH value varied from 7 to 8.5 with an average value of  $7.4 \pm 0.312$  which implies that most groundwater of the study basin is slight alkaline in characteristics. In the current basin, mostly high contents of EC, TDS, Mg, and K are noticed in groundwater (Table 4). The high content of EC and TDS in the groundwater of the study basin is attributed to the runoff of domestic waste into groundwater and salt leaching out via fertile soil layers (Islam et al., 2020b). Groundwater in the Akot basin is mostly enriched with high Mg and K contents, which are mostly ascribed to the natural source (Islam et al. 2018). WQI value ranged from 47.5 to 100 with a mean and standard deviation value of 82.44 and 14.31 respectively, within the Akot basin (Fig. 3). The calculated quality indices of the groundwater samples are demarcated into five subclasses: 0–50 (excellent), 51–100 (good), 101–150 (moderate), 151–200 (poor), and > 200 (very Poor) (Islam et al., 2017). The groundwater belongs to good to poor quality category in the basin.

When separating the water quality index dataset into training and testing subgroups, it is essential to cross-validate the whole data to have the similar statistical population. The observed water quality dataset shows the Q1 and Q3 quartiles are 70.63 and 95.00, while training data shows Q1 and Q3 quartiles are 80.00 and 95.00, and testing dataset shows quartiles of Q1 and Q3 is 70.00 and 97.50. The testing datasets show very good water quality index predicting according to the statistical analysis. This is correct for water quality index assessment and predicted at the Akot basin area. The standard deviation for the datasets shows that the

data variability is higher by values which are further from zero. The variation from the mean value of the data is therefore higher.

SE mean, standard error of mean; St Dev, standard deviation; Q1, first quartile; median, middle number; Q3, third quartile

### Evaluation of results from various models

In the selection process of the best model, several trials have been performed on a single output. The trials of all models were performed based on the different input combination. The results of each model based on deferent input combination have been listed in Table 5 parts A–D based on testing results. The results of the fourth and the fifth inputs combinations (Table 5 part A); the fifth input combination (Table 5 part B); the fourth input combination (Table 5 part C); and the fifth input combination (Table 5 part D) were found to be more promising than the other input combinations. Out of these, a total of five models have been imposed based on techniques and input combinations and further evaluated to find the optimal one for water quality index estimation at the study site (Table 6).

### Quantitative and qualitative evaluation of results

The model performance was classified as very good ( $\text{PCC} > 0.95$ ), good ( $0.85 \leq \text{PCC} \leq 0.95$ ), satisfactory ( $0.70 \leq \text{PCC} \leq 0.85$ ), and unsatisfactory ( $\text{PCC} \leq 0.70$ ), as stated by Moriasi et al. (2012). After considering all the model's best performance from different combination inputs, it was noted that the additive regression model performed better than the M5P tree, random subspace, and SVM models based on quantitative performance evaluation indicators. It was also observed that the performance of the additive regression model was optimal based on the fourth and fifth combinations input. The value of  $R^2$ ,  $RAE$ ,  $RMSE$ , and  $RRSE$  were obtained as 0.9993, 0.5243, 0.6356, 3.8449, and 3.9925, respectively, for additive regression. However,



**Table 5** Performance metrics analysis for the developed WQ models

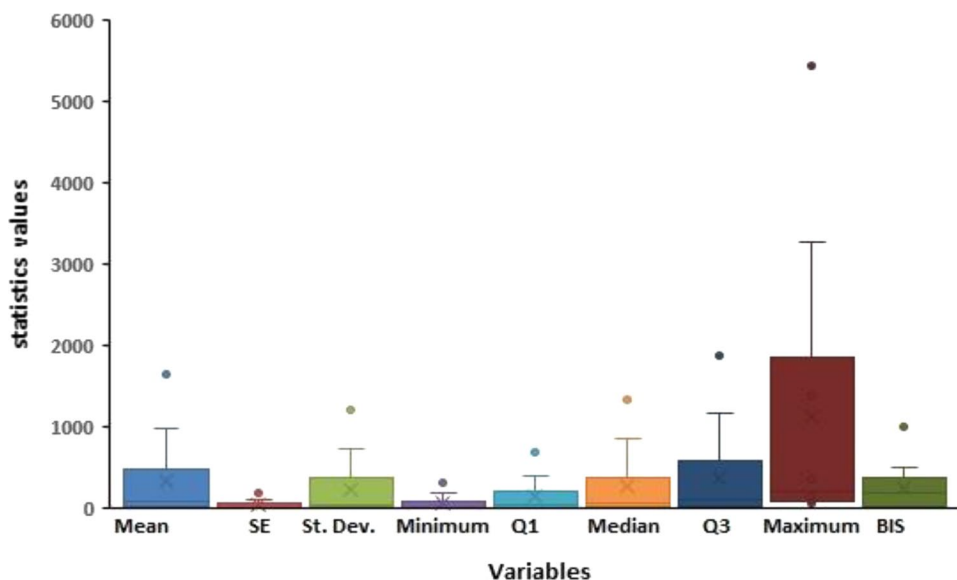
Model	Input's combination	Performance metrics				
		Correlation coefficient	MAE	RMSE	RAE (%)	RRSE (%)
<b>(A) Additive regression</b>						
Additive regression	1	0.9992	0.6319	0.7506	4.6337	4.7151
	2	0.9992	0.6319	0.7506	4.6337	4.7151
	3	0.9977	0.868	1.0879	6.3654	6.8337
	4	0.9993	0.5243	0.6356	3.8449	3.9925
	5	0.9993	0.5243	0.6356	3.8449	3.9925
	6	0.9958	0.9427	1.473	6.913	9.2532
	7	0.9953	1.3481	1.5764	9.8857	9.9026
	8	0.9895	1.7725	2.23268	12.9981	14.6167
	9	0.9341	4.2312	5.7658	31.0285	36.2202
<b>(B) M5P tree</b>						
M5P tree	1	0.9205	5.2735	6.2214	38.6723	39.0821
	2	0.9205	5.2735	6.2214	38.6723	39.0821
	3	0.9205	5.2735	6.2214	38.6723	39.0821
	4	0.9576	4.0343	4.5879	29.5849	28.8205
	5	0.9746	3.1484	3.6027	23.0886	22.6317
	6	0.7469	8.7703	10.9883	64.3156	69.0272
	7	0.7469	8.7703	10.9883	64.3156	69.0272
	8	0.7469	8.7703	10.9883	64.3156	69.0272
	9	0.7469	8.7703	10.9883	64.3156	69.0272
<b>(C) Random subspace</b>						
Random subspace	1	0.9076	7.2689	9.5808	53.3056	60.1852
	2	0.8529	9.6434	12.0887	70.7183	75.9397
	3	0.9051	7.6818	10.1172	56.3333	63.5546
	4	0.9315	7.7294	10.0311	56.6825	63.0141
	5	0.8834	6.6355	9.0226	48.6603	56.679
	6	0.8397	8.7192	11.2932	63.9405	70.9422
	7	0.8715	8.0181	10.2079	58.7992	64.1248
	8	0.8315	9.6282	12.1411	70.6071	76.2687
	9	0.8397	8.7192	11.2932	63.9405	70.9422
<b>(D) SVM</b>						
SVM	1	0.9647	2.1645	4.2147	15.8733	26.476
	2	0.9604	2.9101	4.8359	21.3404	30.3783
	3	0.9604	2.9814	4.7539	21.8635	29.8631
	4	0.9614	2.9296	4.7492	21.4836	29.834
	5	0.9685	2.7179	4.1187	19.9312	25.8733
	6	0.791	7.2721	10.3813	53.3288	65.2139
	7	0.6583	9.2646	12.1108	67.9401	76.0783
	8	0.6363	9.4802	12.3172	69.5215	77.375
	9	0.6101	9.9555	12.6892	73.0066	79.7116

0.9746, 3.1484, 3.6027, 23.0886, and 22.6317 for M5P tree, 0.9315, 7.7294, 10.0311, 56.6825 and 63.0141 for Random Subspace. Meanwhile for SVM it was as follows 0.9685, 2.7179, 4.1187, 19.9312 and 25.8733, respectively. The order of model performance based on PCC from very good to unsatisfactory was attained as additive regression (0.9993) > M5P tree (0.9746) > SVM (0.9685) > random

subspace (0.9315). The order of model performance based on RMSE from best to inferior was obtained as additive regression (0.6356) > M5P tree (3.6027) > SVM (4.1187) > random subspace (10.0311). The order of model performance based on the MAE from best to inferior was found as additive regression (0.5243) > SVM (2.7179) > M5P tree (3.1484) > random subspace (7.7294). The

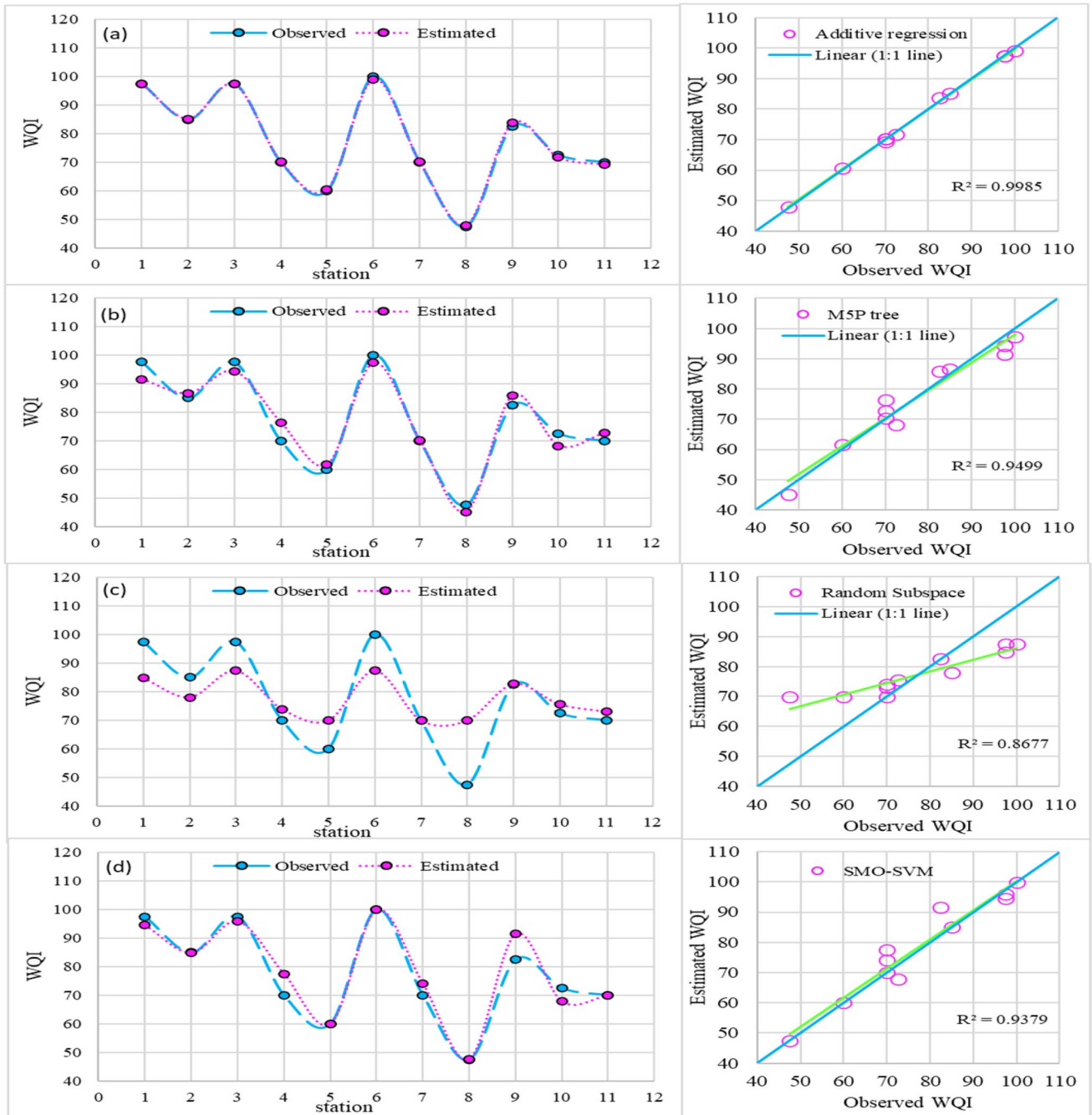
**Table 6** Comparison of the best developed models

Model	Input's combination	Performance metrics				
		Correlation coefficient	MAE	RMSE	RAE (%)	RRSE (%)
AR	4	0.9993	0.5243	0.6356	3.8449	3.9925
	5	0.9993	0.5243	0.6356	3.8449	3.9925
M5P tree	5	0.9746	3.1484	3.6027	23.0886	22.6317
RSS	4	0.9315	7.7294	10.0311	56.6825	63.0141
SVM	5	0.9685	2.7179	4.1187	19.9312	25.8733

**Fig. 3** Box plot of statistics analysis

comparison of results in Table 6 confirmed the superiority of the additive regression model with the fourth and fifth input combination having the lowest value of  $RMSE = 0.6356$  and the highest value of  $PCC = 0.9993$ . The results of the four optimal models were plotted between the observed and estimated WQI value in the form of time variation and scatter plots through Fig. 4; it shows the complete match of the points representing the calculated WQI value with the estimated value when using the additive regression model. However, underestimated value of WQI was observed using the M5P tree and SVM models; the underestimated value becomes much clear in using random subspace model. Based on  $R^2$  value, the order of the model performance varies from very satisfactory to unsatisfactory and was found as additive regression (0.9985) > M5P tree (0.9499) > SVM (0.9379) > random subspace (0.8677). The main cause for the better performance of the AR model in input combinations can be related to the physicochemical parameters, which was identified by the higher concentration of Mg, K, and Cl in the study area. This observation was analogous to the results reported by Zhu and Heddad (2019). 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. According to Bui et al. (2020), AR works very accurately even with a poorly structured dataset and can generate a higher precision as compared to the RF model. Research results are well satisfactory compared to the findings of Bui et al. (2020) who improved the prediction of WQ using hybrid machine learning. Their findings showed that hybrid model (bagging random tree) was the best model ( $R^2 = 0.941$ ,  $RMSE = 2.71$ ,  $MAE = 1.87$ ,  $NSE = 0.941$ ,  $PBIAS = 0.500$ ) compared to others. Furthermore, our results are in agreement with the outcomes of Al-Adhaileh and Alsaade (2021) who used the adaptive neuro-fuzzy inference system (ANFIS) algorithm and feed-forward neural network (FFNN) and stated that the accuracy during the testing phase, with a regression coefficient of 96.17% and the FFNN model, achieved the highest accuracy (100%) for WQ. Moreover, the developed models in this study achieved good results, and it is in line with model outputs of Aldhyani et al. (2020) who used several algorithms and found the highest accuracy was 97.01%. As well, these outcomes coincide with Asadollah et al. (2021) who applied a new ensemble machine learning model called extra tree regression (ETR) for predicting



**Fig. 4** Observed versus estimated WQI of best **a** additive regression, **b** M5P tree, **c** random subspace, and **d** SVM models during the testing period

monthly WQI values and compared its performance with the classic standalone models, support vector regression (SVR), and decision tree regression (DTR). Their results agreed with the current study. In addition, our findings are higher than the results of Li et al. (2021) who used four models: random forest, SVM, partial least squares regression (PLSR), and PLSR-SVM in WQ prediction and

found that the highest performance was  $R^2 = 0.87$  using PLSR-SVM.

Overall, the analysis shows that the application of a proposed AR model for forecasting the variation in groundwater quality is mostly realistic. In other words, the applied AR can accurately detect the significance of input parameters for predicting groundwater quality, and it can be used in

other regions instead of conventional tools. The proposed AR may be applied as the optimal predictive model for not only groundwater quality evaluation works but also for other environmental-associated studies. However, the only drawback of this study is that a single and limited dataset is regarded in the prediction methods. A large dataset can lead to a higher understanding of the groundwater quality of this study. Such consideration will probably guide the higher acceptance of AI models in the field of sustainable groundwater resources studies.

### Limitation of this research

Despite the fact that this study is based on a scientific study of a variety of environmental elements, it has a number of flaws that must be addressed. The important limitations are defined below:

1. Most of the previous water quality data of this study are collected and provided by various departments. However, large amounts of data and maps are extremely comprehensive in the environment. An excessive variation in whole water quality parameters may be observed at the regional and local scale under various climate conditions. Therefore, due to this issues, the developed best machine models and water quality analysis may not be actual precise at a small scale or various climate conditions.
2. Machine learning models are a programming or scientific method where the training and testing of data is completed based on the capability of models. While in this investigation it displays better accurateness, a large precise water quality index may be predicted using the additive regression model by testing and training data.
3. This study is completely fixated on machine learning models and gives a comprehensive water quality index prediction of groundwater. However, for highly accurate valuation, ground-based collection of water quality data is very suitable for the prediction of water quality index. However, the estimation of groundwater water index and traditional methods is not high cost but also very time-consuming process as compare to machine learning models.

### Conclusion

This study considered the groundwater quality of the Akot basin area, located in the Akola district, Maharashtra, in India. Four AI models such as random subspace (RSS), support vector machine, M5 pruning tree, and additive

regression were applied to predict WQI based on nine scenarios of input variables. Water quality data were divided into two sections 80% and 20% for training and testing the developed models. Our findings demonstrated that additive regression is the best prediction model with the high correlation coefficient and fewer statistical errors comparing to other AI models with the combination of optimal inputs, i.e., EC, pH, TDS, Ca, Mg, and Cl during the training and testing phases. Moreover, input variable importance computed by prediction models highlights that machine learning models are the more reliable method in the prediction of WQI. This study can further be improved by using optimized AR model's prediction capability against other AI models by taking into consideration different physicochemical input parameters. The physicochemical parameters chosen in the current study may also pose a drawback, however, due to possible inadequate sampling. Future research may add the use of different input physicochemical parameters to predict the WQI based on the WHO guidelines, to compare with other standard indexes. AI models can use input variables and improving model predictive accuracy, which is an advantage over conventional statistical models. As the new development of AI models, it is promising for further work to predict contaminant concentration under the future pollution scenarios if the AI algorithm fits data well. This study recommends using the best developed model in WQI predicting, especially in the Akola district, Maharashtra, in India.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s11356-021-17064-7>.

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

**Author contribution** Ahmed Elbeltagi: methodology, development of ML models, validation, and formal analysis and writing (review and editing)

Chaitanya B. Pande: methodology, original draft writing, writing editing, plotting, supervision, data collection and analysis for modeling purpose, and investigation

Saber Kouadri: writing the Results section and development of graphs

Abu Reza Md. Towfiqul Islam: writing review and editing

### Declarations

**Ethics approval** Not applicable.

**Consent to participate** Not applicable.

**Consent for publication** Not applicable.

**Competing interests** The authors declare no competing interests.



## 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
- Abbasnia A, Yousefi N, Mahvi AH, Nabizadeh R, Radfard M, Yousefi M, Alimohammadi M (2019) Evaluation of groundwater quality using water quality index and its suitability for assessing water for drinking and irrigation purposes: case study of Sistan and Baluchistan province (Iran). *Hum. Ecol. Risk Assess* 25(4):988–1005. <https://doi.org/10.1080/10807039.2018.1458596>
- Adnan, R.M., Jaafari, A., Mohanavelu, A., Kisi, O., Elbeltagi, A., 2021. Novel ensemble forecasting of streamflow using locally weighted learning algorithm. *Sustain*.
- Ahmadi M, Etedali HR, Elbeltagi A (2021) Evaluation of the effect of climate change on maize water footprint under RCPs scenarios in Qazvin plain. *Iran. Agric. Water Manag.* 254:106969. <https://doi.org/10.1016/j.agwat.2021.106969>
- Al-Adhaileh MH, Alsaade FW (2021) Modelling and prediction of water quality by using artificial intelligence. *Sustain.* 13:1–18. <https://doi.org/10.3390/su13084259>
- Aldhyani THH, Al-Yaari M, Alkahtani H, Maashi M (2020) Water quality prediction using artificial intelligence algorithms. *Appl. Bionics Biomech.* 2020. <https://doi.org/10.1155/2020/6659314>
- 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>
- 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>
- Ajmera TK, Goyal MK (2012) Development of stage discharge rating curve using model tree and neural networks: an application to Peachtree Creek in Atlanta. *Expert Syst. Appl.* 39(5):5702–5710
- Asefa T, Kemblowski M, Urroz G, McKee M (2005) Support vector machines (SVMs) for monitoring network design. *Ground Water* 43:413–422
- APHA, American Public Health Association (2005) Standard methods for the examination of water and waste water, 21st edn. APHA, Washington
- Arun Pratap Mishra, Harish Khali, Sachchidanand Singh, Chaitanya B Pande, Raj Singh, Shardesh K Chaurasia, (2021) An assessment of in-situ water quality parameters and its variation with Landsat 8 level 1 surface reflectance datasets, *Int J Environ Anal Chem*, pp. 1–23, <https://doi.org/10.1080/03067319.2021.1954175>.
- Babae M, Maroufpoor S, Jalali M, Zarei M, Elbeltagi A (2021) Artificial intelligence approach to estimating rice yield\*. *Irrig. Drain.* 1–11. <https://doi.org/10.1002/ird.2566>
- Bajirao TS, Kumar P, Kumar M, Elbeltagi A, Kuriqi A (2021) Superiority of hybrid soft computing models in daily suspended sediment estimation in highly dynamic rivers. *Sustain.* 13:1–29. <https://doi.org/10.3390/su13020542>
- Babbar, R., Babbar, S., (2017), Predicting river water quality index using data mining techniques, *Environ Earth Sci* (2017) 76:504 <https://doi.org/10.1007/s12665-017-6845-9>
- Banerji S, Mitra D (2019) Geographical information system-based groundwater quality index assessment of northern part of Kolkata, India for drinking purpose. *Geocarto Int.* 34:943e958. <https://doi.org/10.1080/10106049.2018.1451922>
- Panneerselvam B, Muniraj K, Pande C, Ravichandran N (2021a) Prediction and evaluation of groundwater characteristics using the radial basic model in semi-arid region. *India, International Journal of Environmental Analytical Chemistry*, pp 1–17. <https://doi.org/10.1080/03067319.2021.1873316>
- BIS (Bureau of Indian Standards) (2012) Indian standard drinking water-specification, 1st rev., pp 1–8
- Brown, A., & Matlock, M. D. (2011) A review of water scarcity indices and methodologies. White paper106, 19.
- Brown, R.M., McClelland, N.I., Deininger, R.A., Tozer, R.G., 1970. A water quality index do we dare.
- Bui DT, Khosravi K, Tiefenbacher J et al (2020a) Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Sci Total Environ* 721:137612. <https://doi.org/10.1016/j.scitotenv.2020.137612>
- Busico, G., Kazakis, N., Cuoco, E., Colombani, N., Tedesco, D., Voudouris, K., Mastrociccio, M., 2020. A novel hybrid method of specific vulnerability to anthropogenic pollution using multivariate statistical and regression analyses.
- Bui DT, Khosravi K, Tiefenbacher J, Nguyen H, Kazakis N (2020b) Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Sci. Total Environ.* 721:137612. <https://doi.org/10.1016/j.scitotenv.2020.137612>
- Chen W, Pradhan B, Li S, Shahabi H, Rizeei HM, Hou E, Wang S (2019) Novel hybrid integration approach of bagging-based Fisher's linear discriminant function for groundwater potential analysis. *Nat. Resour. Res.* 28:1239–1258
- Deng T, Chau KW, Duan HF (2021) Machine learning based marine water quality prediction for coastal hydro-environment management. *Journal of Environmental Management* 284:112051
- El Bilali A, Taleb A, Brouziyne Y (2021) Groundwater quality forecasting using machine learning algorithms for irrigation purposes. *Agricultural Water Management* 245:106625
- Elbeltagi A, Azad N, Arshad A, Mohammed S, Mokhtar A, Pande C, Ramezani H, Ahmad S, Reza A, Islam T, Deng J (2021) Applications of Gaussian process regression for predicting blue water footprint: case study in Ad Daqahliyah. *Egypt. Agric. Water Manag.* 255:107052. <https://doi.org/10.1016/j.agwat.2021.107052>
- 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, Mehdinejadiani 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>
- Fagbote EO, Olanipekun EO, Uyi HS (2014) Water quality index of the ground water of bitumen deposit impacted farm settlements using entropy weighted method. *Int. J. Environ. Sci. Technol.* 11:127e138. <https://doi.org/10.1007/s13762-0120149-0>
- Fu JC, Huang HY, Jang JH, Huang PH (2019) River stage forecasting using multiple additive regression trees. *Water Resour. Manag.* 33:4491–4507. <https://doi.org/10.1007/s11269-019-02357-x>
- 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. *Marine Pollut Bull* 64:2409–2420
- Gorgij AD, Kisi O, Moghaddam AA, Taghipour A (2017) Groundwater quality ranking for drinking purposes, using the entropy method and the spatial autocorrelation index. *Environ Earth Sci* 76(7):269

- Hastie T, Tibshirani R (1986) Generalized additive models. *Stat. Sci.* 6:15–51
- He S, Wu J (2019) Relationships of groundwater quality and associated health risks with land use/land cover patterns: a case study in a loess area, northwest China. *Hum. Ecol. Risk Assess.* 25(1e2):354–373
- Heddam S, Kisi O (2018) Modelling daily dissolved oxygen concentration using least square support vector machine, multivariate adaptive regression splines and M5 model tree. *J. Hydrol.* 559:499–509
- Horton RK (1965) An index number system for rating water quality. *J. Water Pollut. Control Fed.* 37:300–306
- 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 Poll Res.* <https://doi.org/10.1007/s11356-021-12806-z>
- Islam ARMT, Mamun AA, Rahman MM, Zahid A (2020b) Simultaneous comparison of modified-integrated water quality and entropy weighted indices: implication for safe drinking water in the coastal region of Bangladesh. *Ecological Indicators* 113:106229. <https://doi.org/10.1016/j.ecolind.2020.106229>
- Islam ARMT, Siddiqua MT, Zahid A, Tasnim SS, Rahman MM (2020a) 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, 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. *Groundwater Sustain Dev* 9:100275. <https://doi.org/10.1016/j.gsd.2019.100275>
- 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 Poll Res* 24(34):26350–26374. <https://doi.org/10.1007/s11356-017-0254-1>
- Jerin JN, Islam HMT, Islam T, Shahid S, Zhenghua H, Mehnaz B, Ronghao C, Ahmed E (2021) Spatiotemporal trends in reference evapotranspiration and its driving factors in Bangladesh. *Theor. Appl. Climatol.* <https://doi.org/10.1007/s00704-021-03566-4>
- 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(6):431–447
- 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>
- Kazakis N, Mattas C, Pavlou A, Patrikaki O, Voudouris K (2017) Multivariate statistical analysis for the assessment of groundwater quality under different hydrogeological regimes. *Environ Earth Sci* 76(9):349
- Khosravi K, Pham B, Chapi K, Shirzadi A, Shahabi H, Revhaug I, Bui D (2018) A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran. *Sci. Total Environ.* 627:744–755
- Khosravi K, Shahabi H, Pham BT, Adamowski J, Shirzadi A, Pradhan B, Dou J, Ly H-B, Gróf G, Ho HL et al (2019) A comparative assessment of flood susceptibility modeling using multi-criteria decision-making analysis and machine learning methods. *J. Hydrol.* 573:311–323
- Khozani Z, Khosravi K, Pham B, Kløve B, Mohtar W, Yaseen Z (2019) Determination of compound channel apparent shear stress: application of novel data mining models. *J. Hydro. inform.* 21:798–811
- 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 Manag.* <https://doi.org/10.1007/s11269-018-2147-6>
- Kumar M, Kumari A, Kumar D, Al-ansari N, Ali R, Kumar R, Kumar A, Elbeltagi A, Kuriqi A (2021) The superiority of data-driven techniques for estimation of daily pan evaporation. *Atmosphere (Basel)*:1–23
- Laanaya F, St-Hilaire A, Gloaguen E (2017) Water temperature modeling: comparison between the generalized additive model, logistic, residuals regression and linear regression models. *Hydrol. Sci. J.* 62:1078–1093. <https://doi.org/10.1080/02626667.2016.1246799>
- Leong WC, 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). *Intl. J. River Basin Manag.*:1–8. <https://doi.org/10.1080/15715124.2019.1628030>
- Li X, Ding J, Ilyas N (2021) Machine learning method for quick identification of water quality index (WQI) based on Sentinel-2 MSI data: Ebinur Lake case study. *Water Sci. Technol. Water Supply* 21:1291–1312. <https://doi.org/10.2166/ws.2020.381>
- Li PY, Wu JH, Qian H (2010) Groundwater quality assessment based on entropy weighted osculating value method. *Int. J. Environ. Sci.* 1(4):621e630
- Mokhtar A, Jalali M, Elbeltagi A, Al-Ansari N, Alsafadi K, Abdo HG, Sammen SS, Gyasi-Agyei Y, Rodrigo-Comino J, He H (2021) Estimation of SPEI meteorological drought using machine learning algorithms. *IEEE Access* XX. <https://doi.org/10.1109/ACCESS.2021.3074305>
- Moriassi DN, Wilson BN, Douglas-Mankin KR, Arnold JG, Gowda PH (2012) Hydrologic and water quality models: use, calibration, and validation. *Trans. ASABE* 55:1241–1247
- Nguyen L (2017) Tutorial on support vector machine. *Appl. Comput. Math.* 6:1–15
- Ongley, E.D., 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.
- Pham BT, Bui DT, Prakash I, Dholakia M (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
- 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* 8:106. <https://doi.org/10.1007/s13201-018-0754-2>
- Pande CB, Moharir KN, Singh SK et al (2020) Groundwater evaluation for drinking purposes using statistical index: study of Akola and Buldhana districts of Maharashtra, India. *Environ Dev Sustain* 22:7453–7471. <https://doi.org/10.1007/s10668-019-00531-0>
- Panneerselvam B, Muniraj K, Thomas M, Ravichandran N (2021b) GIS-based legitimatic evaluation of groundwater's health risk and irrigation susceptibility using water quality index, pollution index, and irrigation indexes in semiarid region. In: Pande CB, Moharir KN (eds) *Groundwater resources development and planning in the semi-arid region*. Springer, Cham. [https://doi.org/10.1007/978-3-030-68124-1\\_13](https://doi.org/10.1007/978-3-030-68124-1_13)
- Raghavendra NS, Deka PC (2014) Support vector machine applications in the field of hydrology: a review. *Appl. Soft Comput.* 19:372–386

- 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>
- Sharafati A, Khosravi K, Khosravinia P, Ahmed K, Salman SA, Yaseen ZM (2019) The potential of novel data mining models for global solar radiation prediction. *Int. J. Environ. Sci. Technol.* 16:7147–7164
- Singha S, Pasupuleti S, Singha SS, Singh R, Kumar S (2021) Prediction of groundwater quality using efficient machine learning technique. *Chemosphere* 276:130265
- Sinha MK, Rajput P, Baier K, Azzam R (2021) GIS-based assessment of urban groundwater pollution potential using water quality indices. In: Pande CB, Moharir KN (eds) *Groundwater resources development and planning in the semi-arid region*. Springer, Cham. [https://doi.org/10.1007/978-3-030-68124-1\\_15](https://doi.org/10.1007/978-3-030-68124-1_15)
- Skurichina M, Duin RPW (2002a) Bagging, boosting and the random subspace method for linear classifiers. *Pattern Anal Appl.* 5(2):121–135
- Skurichina M, Duin RP (2002b) Bagging, boosting and the random subspace method for linear classifiers. *Pattern Anal. Appl.* 5:121–135
- Suryakant T, Pravendra B, Manish K, Ahmed E, Alban K (2021) Potential of hybrid wavelet - coupled data - driven - based algorithms for daily runoff prediction in complex river basins. *Theor. Appl. Climatol.* 21. <https://doi.org/10.1007/s00704-021-03681-2>
- Ho TK, Baird HS (Apr. 1998) Pattern classification with compact distribution maps. *Computer vision and image understanding* 70(1):101–110
- Tiyasha Tung TM, Yaseen ZM (2020) A survey on river water quality modelling using artificial intelligence models: 2000e2020. *J. Hydrol.* 585:124670. <https://doi.org/10.1016/j.jhydrol.2020.124670>
- Tong S, Koller D (2001) Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.* 2:45–66
- 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.* 12. <https://doi.org/10.1016/j.gsf.2020.09.006>
- 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>
- Water Res. 171, 115386, Buja A, Hastie T, Tibshirani R (1989) Linear smoothers and additive models. *Ann Stat* 17(2):453–555 JSTOR 2241560
- WHO (World Health Organization) (2011) *Guidelines for drinking water quality*, 4th edn. World Health Organization, Geneva
- 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
- Yidana SM, Yidana A (2010) Assessing water quality using water quality index and multivariate analysis. *Environ Earth Sci* 59(7):1461–1473
- Zerouali B, Al-ansari N, Chettih M, Mohamed M, Abda Z, Santos C, Zerouali B, Elbeltagi A (2021) An enhanced innovative triangular trend analysis of rainfall based on a spectral approach. *Water (Switzerland)*:13. <https://doi.org/10.3390/w13050727>
- Zhang Q, Qian H, Xu P, Hou K, Yang F (2021) Groundwater quality assessment using a new integrated-weight water quality index (IWQI) and driver analysis in the Jiaokou Irrigation District, China. *Ecotoxicol Environ Saf* 212:111992
- Zhu S, Heddam 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.* 124819

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