



Proposed formulation of surface water quality and modelling using gene expression, machine learning, and regression techniques

Muhammad Izhar Shah¹ · Muhammad Faisal Javed¹ · Taher Abunama²

Received: 4 August 2020 / Accepted: 30 October 2020 / Published online: 11 November 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract

The rising water pollution from anthropogenic factors motivates further research in developing water quality predicting models. The available models have certain limitations due to limited timespan data and the incapability to provide empirical expressions. This study is devoted to model and derive empirical equations for surface water quality of upper Indus river basin using a 30-year dataset with machine learning techniques and then to determine the most reliable model capable to accurately predict river water quality. Total dissolve solids (TDS) and electrical conductivity (EC) were used as dependent variables, whereas eight parameters were used as independent variables with 70 and 30% data for model training and testing, respectively. Various evaluation criteria, i.e., Nash-Sutcliffe efficiency (NSE), root mean square error (RMSE), coefficient of determination (R^2), and mean absolute error (MAE), were used to assess the performance of models. The data is also validated with the help of k-fold cross-validation using R^2 and RMSE. The results indicated a strong correlation with NSE and R^2 both above 0.85 for all the developed models. Gene expression programming (GEP) outperformed both artificial neural network (ANN) and linear and non-linear regression models for TDS and EC. The sensitivity and parametric analyses revealed that bicarbonate is the most sensitive parameter influencing both TDS and EC models. Two equations were derived and formulated to represent the novel results of GEP model to help authorities in the effective monitoring of river water quality.

Keywords Surface water quality · Machine learning algorithms · Regression · Sensitivity and parametric analyses · k-fold cross-validation

Introduction

Surface water is a vital resource that is necessary for all aspects of life. The quality of water is affected by pollutants and its distribution with the flow (Kargar et al. 2020). Due to lack of facilities and infrastructure in developing countries, major

portion of the liquid waste is deposited to various surface water bodies. Moreover, the rapid industrialization and population growth adversely affect the quality of surface water bodies. The term water quality is used to define the condition of water covering its physical, chemical, and biological properties (Alizadeh et al. 2018). The quality of water gets contaminated due to some natural processes such as inputs from atmosphere or climatic conditions (Al-Mukhtar and Al-Yaseen 2019). Human activities are considered to cause major pollution to water (Azad et al. 2019), which needs quick mitigation actions.

The water quality assessment is considered a difficult task due to involvement of certain human and environmental factors (Tung and Yaseen 2020). The water pollution is a major issue in the effective management of aquatic environment (Salami et al. 2016). The substantial water quality parameters include total dissolved solids (TDS) and electrical conductivity (EC). TDS is composed of salts and organic matter percentage dissolved with soil coming as a source of pollutant in rainwater. Other sources of TDS are weathering of rocks

Responsible Editor: Marcus Schulz

✉ Muhammad Izhar Shah
mizhar@cuiatd.edu.pk

Muhammad Faisal Javed
arbabfaisal@cuiatd.edu.pk

Taher Abunama
tahera@dut.ac.za

¹ Department of Civil Engineering, COMSATS University Islamabad, Abbottabad Campus, Abbottabad 22060, Pakistan

² Institute for Water and Wastewater Technology, Durban University of Technology, PO Box 1334, Durban, South Africa

resulting in salts that released minerals over the time. The presence of TDS determine the use of water for specific purposes, and high TDS content adversely affect the natural resources and human health (Najah et al. 2013; Nasr and Zahran 2014). The EC is linked with water quality and the dilution effect of stream water. The alteration in EC is an indication of discharge to stream from some pollution source (Najah et al. 2013). Generally, both TDS and EC are directly linked with the salt content in water; therefore, high values of these parameters indicate poor water quality.

The effective management of river water quality is facing issues due to unavailability of reliable data. The conventional sample collection and laboratory testing are time-consuming, labor-intensive, and costly (Sattari et al. 2016). Moreover, lack of technological and financial resources restricts the water pollution control and management (Crocker and Bartram 2014). The water quality-related data is much challenging to design due to the complexity involved in terms of nonlinearity, imprecise properties, and non-stationary features. The underlying issue is due to human interferences and intermittent natural changes which consequently, resulting in noisy and poor quality data (Tung and Yaseen 2020). Therefore, advanced and reliable assessment techniques are needed to reduce the workload and overcome the underlying data problem for valuation of water quality (Aryafar et al. 2019; Bozorg-Haddad et al. 2017). Models with a wide range of parameters are desirable for the purpose to model multiple water quality variables (Gholampour et al. 2017). Previously, the water quality predictions were usually carried out with hard computing methods encompassing stochastic, statistical, deterministic, and numerical techniques. However, due to complex structure, insufficient accuracy, and detailed information compulsions, such models are difficult and costly to obtain and therefore leave a gap to try alternate methods (Khare and Warke 2014). The use of machine learning and artificial intelligence (AI) techniques is well-known in solving many environmental engineering problems (Liu and Lu 2014; Mustafa et al. 2014; Pal et al. 2014; Seyam et al. 2020; Shamshirband et al. 2019).

Different studies were conducted to estimate the water quality by various modeling techniques. Bozorg-Haddad et al. (2017) considered the combination of least square support vector regression (LSSVR) and genetic programming (GP) for estimating Na, K, Mg, SO₄, EC, pH, EC, and TDS in the Sefidrood River, Iran. The R² was above 0.9 for all the estimated parameters. (Al-Mukhtar and Al-Yaseen 2019) predicted TDS and EC with the help of adaptive neuro fuzzy inference system (ANFIS), artificial neural network (ANN), and multiple linear regression model (MLR) in Abu-Ziriq, Iraq. It was observed that chloride, nitrate, magnesium, calcium, sulfate, and total hardness were the most effective input parameters. The best results were obtained by using the ANFIS for estimating the water quality. (Sarkar and Pandey 2015) utilized ANN technique for assessing dissolved oxygen

(DO) concentration in river water at three different locations. Flow data, pH, temperature, DO, and biochemical oxygen demand (BOD) were used as the parameters for the analysis. The correlation value up to 0.9 was observed between predicted and measured data. Zhang et al. (2019) used hybrid ANN model developed from combination of ANN and GP to predict drinking water production from water treatment plants. The results revealed strong performance of the developed model in predicting the water treatment plant capacity. The performance of the model rose expressively by feeding more datasets during model training. Chen et al. (2020) used ten different machine learning models comprised of seven traditional and three ensemble models to compare the water quality prediction capacity using a large dataset. The results exposed that better performance of the models can be achieved by using a large dataset for water quality estimation.

Considering the aforementioned discussion, most of the modeling techniques have limited capacity to estimate different water quality parameters due to the availability of limited duration data. Moreover, no such techniques were adopted that provide empirical mathematical expressions for accurate prediction of water quality. Therefore, the main goal of this study was to apply various machine learning and regression techniques to predict surface water quality in Upper Indus Basin (UIB) at Bisham Qilla gauging station. Subsequently, to select the best model to be used in deriving empirical equations for forecasting the quality of water. The objectives of this study were achieved by applying gene expression programming (GEP), artificial neural network (ANN), and linear and non-linear regression (MLR and MNL) to model TDS and EC of the monthly water quality data available for almost 30 years. The developed models were challenged by computing various performance statistical indicators. The formulation of such a model that accurately estimates the concentration of TDS and EC by utilizing minimum number of parameters significantly reduces the time and cost required for water quality monitoring. As per author's knowledge, the water quality modeling is rarely performed previously by researchers in the study region. Therefore, it is imperative to select proper modeling techniques to derive representative and applicable equations using data collected for long-term periods.

Materials and methods

Models selection and development

Different data-driven approaches were employed to develop formulation for prediction of TDS and EC in surface water quality, including machine learning (GEP and ANN) and regression (MLR and MNL)

approaches. The proposed formulation of TDS and EC are meant to be presented as a function of the following predictors, as shown in Eqs. 1 and 2, respectively.

$$TDS = f(Ca, Mg, Na, Cl, SO_4, HCO_3, pH, T_{year}) \quad (1)$$

$$EC = f(Ca, Mg, Na, Cl, SO_4, HCO_3, pH, T_{year}) \quad (2)$$

Gene expression programming (GEP)

Genetic programming (GP) is a form of machine learning technique which is based on genetic evaluation process. GP works on the principle of neural and regression techniques. Due to empiric formulization, the neural network-based formulations are often too complex to develop (Gholampour et al. 2017). The GP algorithm provides a computer-based solution of complex problems considering the Darwin principle. The process for solving the required problems by genetic approach is given in Fig. 1.

The GP is a valuable technique due to its ability to create simple expressions without considering the base form. Initially, some regression-based functions need to be defined. GP has the capability to increase or delete some parameters or its combination given the fitness with experimental outcome (Abdollahzadeh et al. 2017). The enhanced and distinguished version of GP with encoded linear fixed chromosomes and parse tree-like structures is known as gene expression programming (GEP) (Azim et al. 2020). GEP uses simple conditions to develop genetic variety and solve the complex programs due to its multigene behavior. The various forms and sizes of non-linear entities are expressed as a parse tree and are termed as expression trees (ETs), as shown in Fig. 2.

Figure 3 presents a schematic diagram of GEP algorithm. The GEP process starts by creating a fixed length chromosome for a single individual. Afterwards, the chromosomes are represented by expression trees and fitness is evaluated. At last, the reproduction process starts and the assessment is done through fitness functions.

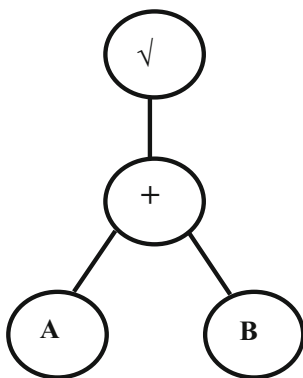


Fig. 1 Genetic programming (GP) model representation

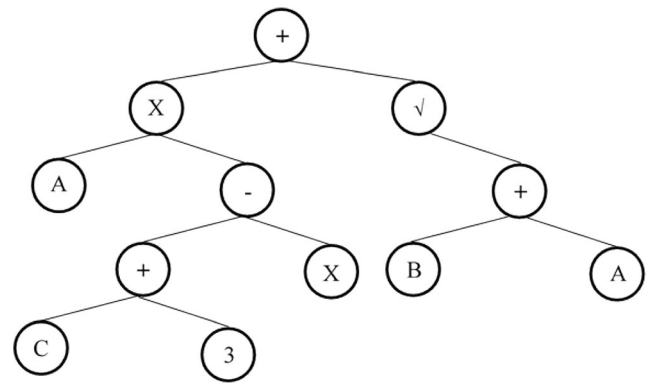


Fig. 2 Example of expression tree (ET)

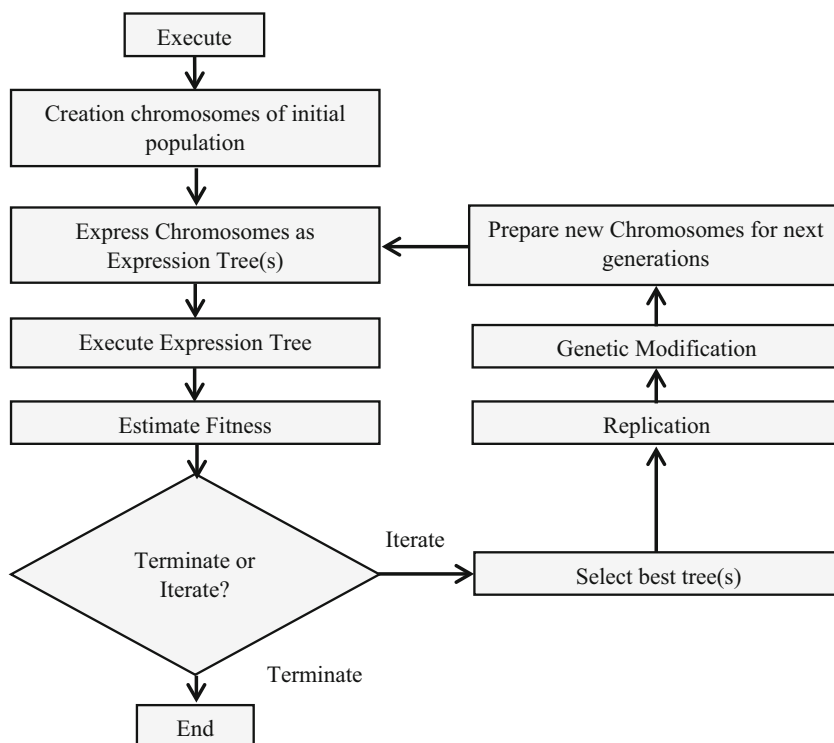
Artificial neural networks (ANN)

ANNs are algorithms that roughly replicate and simulate the microstructures of the biological nervous system, where the artificial neuron is the basic building step of ANN. The non-linear and complex functions can simply be represented by ANN with various parameters or variables that are trained in a way where the output of ANN match the measured output based on a known data set. The three forms of layers, i.e., an input layer, hidden layers, and output layers, are incorporated in each network. A broad network of hidden layers is distributed between the input and output layers. Figure 4 presents a typical neural network architecture with an input layer, output layer, and hidden layers. A significant amount of data is required for training ANN models that have the ability to replicate output from previously unseen inputs. Recently, these neural operations have been extended to environmental engineering applications of groundwater prediction, atmospheric temperature prediction, hydrological processes, and water quality prediction and monitoring (Azamathulla et al. 2018; Najah et al. 2013).

Multiple linear and non-linear regression

Linear and non-linear regression models are able to generate regression equations that can further be used in various engineering fields. The multiple linear regression (MLR) model provides a linear association between independent and dependent variables, while nonlinear relationship is assumed in multiple nonlinear regression (MNL) based on a single or more predictor variables (Adamowski et al. 2012). Different functions such as quadratic, cubic, exponential, and logarithmic were used to determine the optimum relation in MNL. In the present study, the statistical package for social sciences (SPSS) was used to develop MLR and MNL models. A range of statistical analysis can be easily performed from basic to complex one using SPSS. All the acquired datasets were evaluated prior to perform linear and non-linear regression.

Fig. 3 Schematic diagram of GEP algorithm (Ferreira 2006)



Study area and water quality data

Description of the study area

Indus River is among the major rivers in Asia with a length of 2880 km. It drains an area of about 912,000 km² and also spreads in areas of China, Afghanistan, Pakistan, and India (Ali et al. 2015). The upstream portion of Indus River above Tarbela Dam is called the Upper Indus Basin (UIB). The total length and drainage area of UIB are nearly 1150 km and 165,400 km², respectively (Khan et al. 2014; Khan and Koch 2018; Shah et al. 2020). Large part of UIB is glacial

ice with 2174 km³ ice reserves, and the elevation varies from 455 to 8611 m (Tahir et al. 2011). The effect of the summer monsoon is least, and almost 90% area lies under the rain domination of Himalayas (Khan and Koch 2018). The annual precipitation ranges from 100 to 200 mm (Ali et al. 2015). About 70–80% snow is collected in winter, and only 10–15% remains in melting period (Tahir et al. 2011). Both the glaciers melt and storm runoff are responsible to generate the water flow in the basin. Being a mountainous and glacierized region with a great variation in altitude, the direct and field measurement of water quality parameters such as TDS and EC are challenging. Therefore, it is imperative to adopt the

Fig. 4 Structure of a multi-layer feed forward ANN model (Sarkar and Pandey 2015)

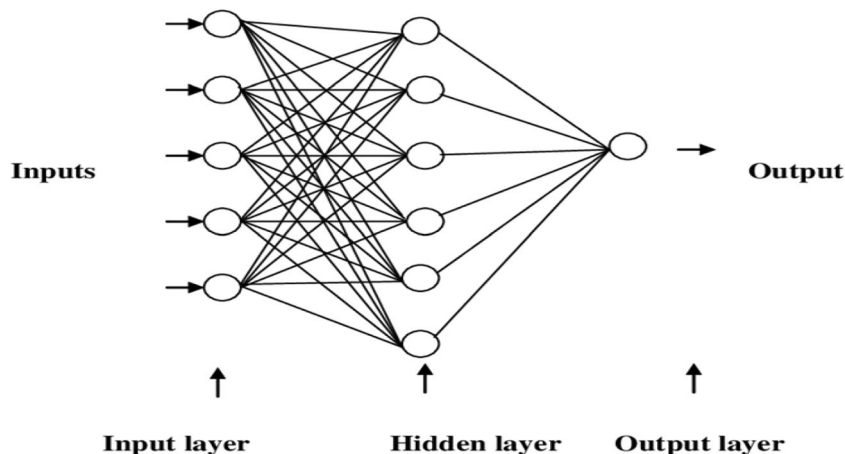
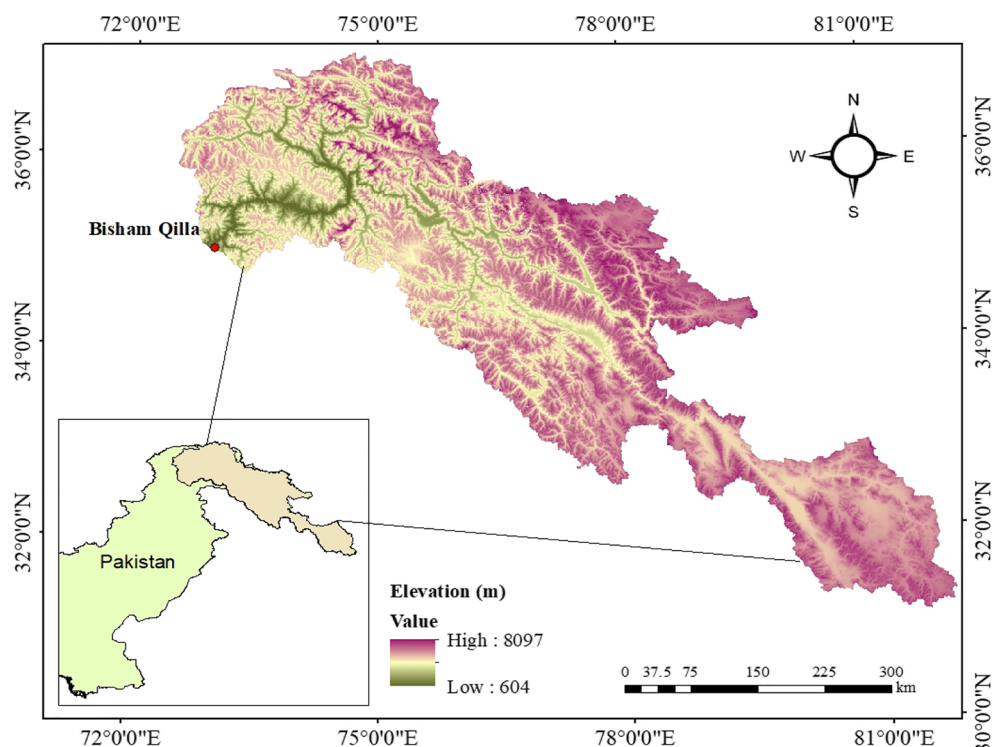


Fig. 5 Brief description of the study area and outlet station



modeling techniques in order to measure water quality accurately and easily. A graphical description of the study area is given in Fig. 5.

Water quality data

A large and continuous water quality dataset of 30 years contained 321 monthly records measured at Bisham Qilla outlet from year 1975 to year 2005 is used in this study. The data was collected from water and power development authority (WAPDA), Pakistan. Each record of the collected data included ten variables, namely, calcium (Ca^{2+}), magnesium (Mg^{2+}),

sodium (Na^+), chloride (Cl^-), sulfate (SO_4^{2-}), bicarbonates (HCO_3^-), pH, TDS, EC, and year effect (T_{year}). A statistical summary of the water quality variables is shown in Table 1.

It is known that the performance of the proposed model considerably depends on the number of data points (Gholampour et al. 2017). For this purpose and to check the suitability of the used datasets, Frank and Todeschini (1994) proposed that 5 is a reasonable ratio between the number of data points and the selected input variables. In the current study, this ratio is 40.1 (321/8) which surpasses the reasonable criteria set.

The input parameters were selected based on Pearson correlation. The correlation matrix among the water quality variables is presented in Table 2. It is evident from literature survey that including too much inputs that may have low correlation with the targeted output reduces the model performance and increases its complexity (Abunama et al. 2019; Ansari et al. 2018). Therefore, eight parameters were used for the model development as predictors of the dependent variables (EC and TDS). In the modeling process, the complete dataset (321 monthly samples \times 8 variables) was divided as training (221 monthly samples \times 8 variables) and testing (100 monthly samples \times 8 variables). Therefore, the total data covers 70 and 30% for the model training and testing, respectively.

Further, the fit goodness test was conducted using the normal probability curve as presented in Fig. 6 for TDS and EC concentrations, respectively. A distribution is said to be normal when the probability curve is symmetrical and positioned

Table 1 Statistical summary of the water quality dataset ($n = 321$ months)

Parameters	Unit	Range	Min.	Max.	Mean	SD
Ca^{2+}	meq/l	1.84	0.61	2.45	1.46	0.32
Mg^{2+}	meq/l	2.61	0.03	2.64	0.63	0.33
Na^+	meq/l	8.95	0.05	9.0	0.53	0.69
Cl^-	meq/l	4.15	0.05	4.2	0.28	0.28
SO_4^{2-}	meq/l	3.1	0.1	3.2	0.55	0.37
HCO_3^-	meq/l	7.1	0.3	7.4	1.73	0.63
TDS	ppm	200	60	260	139.87	38.64
EC	$\mu\text{S/cm}$	358	92	450	242.65	67.49
pH	-	1.22	7.08	8.3	7.83	0.65
T_{year}	-	30	0	30	-	-

Table 2 Correlation matrix among the water quality parameters

Parameters	Ca ²⁺	Mg ²⁺	Na ⁺	HCO ₃ ⁻	Cl ⁻	SO ₄ ²⁻	TDS	EC	pH	T _{year}
Ca ²⁺	1									
Mg ²⁺	0.019	1								
Na ⁺	-0.003	0.471	1							
HCO ₃ ⁻	0.036	0.532	0.741	1						
Cl ⁻	0.023	0.503	0.704	0.529	1					
SO ₄ ²⁻	0.021	0.541	0.485	0.274	0.369	1				
TDS	0.745	0.700	0.862	0.817	0.741	0.629	1			
EC	0.641	0.665	0.842	0.855	0.692	0.574	0.962	1		
pH	0.002	0.073	0.041	0.054	0.056	-0.044	0.032	0.034	1	
T _{year}	0.004	0.005	0.045	0.008	0.056	0.034	0.085	0.076	0.067	1

around the mean of a data series (Abunama et al. 2019; Ramzan et al. 2013).

Cross-validation using k-fold method

The k-fold cross-validation is performed for the purpose to reduce the random sampling biases and overfitting problem. Kohavi (1995) suggested that the tenfold cross-validation algorithm provided the enhanced computational time along with reliable variance. In the current study, the k-fold algorithm was adopted in order to judge the performance of the models which distributed the data sample into ten subclasses. Among all the ten rounds of model formation and validation, it considers a separate subclass for training and testing the models with other datasets. The test subclasses are then utilized to check the accuracy of the model as given in Fig. 7. The accuracy of the final algorithm is further expressed as a mean accuracy attained by the ten models in ten validation rounds.

Models evaluation criteria

Various performance evaluation indicators were selected for the assessment of the model including Nash-Sutcliffe efficiency (NSE), coefficient of determination (R²), mean absolute error (MAE), and root mean square error (RMSE) (Abunama et al. 2019; Montaseri et al. 2018). These statistical parameters were used to differentiate the model accuracy and performance. NSE values range between -∞ to 1, where 1 is a perfect match. NSE value greater than 0.65 depicts a very good correlation (Ansari et al. 2018). R² values lie from 0 and 1, and the higher values indicate less errors. RMSE and MAE are error index parameters commonly used for evaluating the yielded modeling errors. Lower values for both criteria indicating better modeling performance. The mathematical expressions of NSE, R², RMSE and MAE are shown in Eqs. 3–6, respectively. The above selected indicators are frequently used in various studies (Abunama et al. 2019; Montaseri et al. 2018).

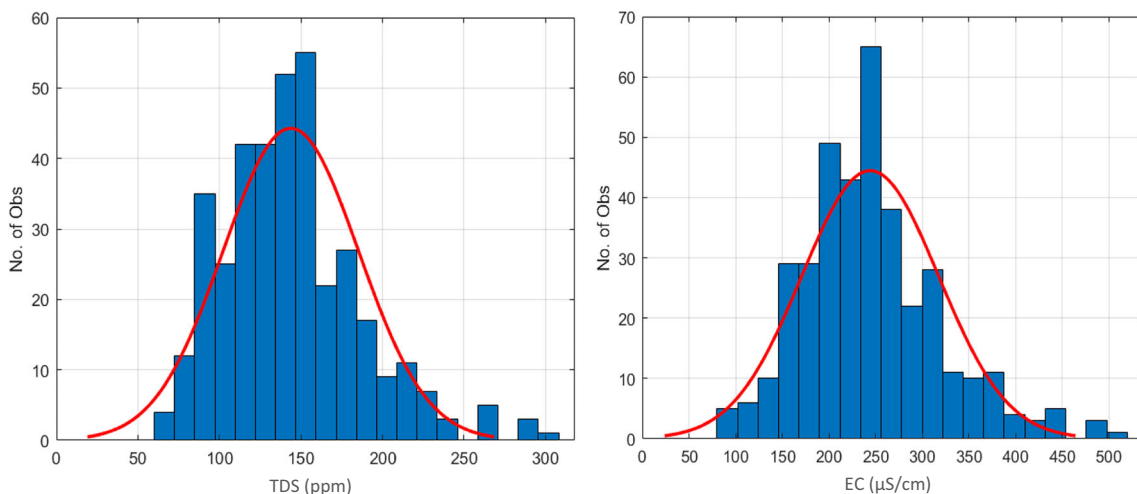


Fig. 6 Normal probability curves of EC and TDS data

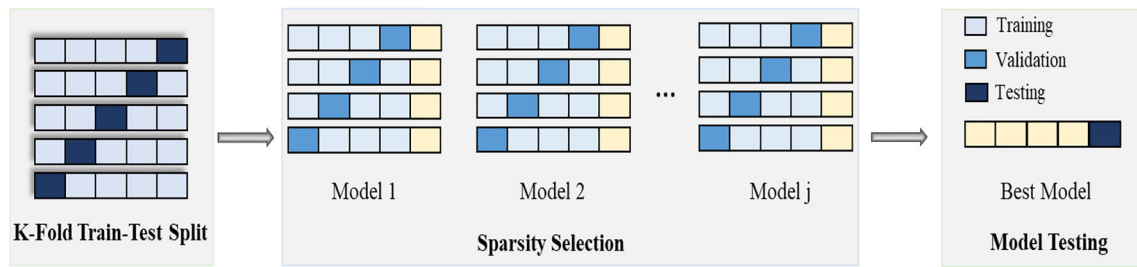


Fig. 7 k-fold cross-validation algorithm

$$NSE = 1 - \frac{\sum_{i=1}^n (M_i - P_i)^2}{\sum_{i=1}^n (M_i - \bar{M}_i)^2} \quad (3)$$

$$R^2 = \frac{\sum_{i=1}^n (M_i - \bar{M}_i)(P_i - \bar{P}_i)}{\sqrt{\sum_{i=1}^n (M_i - \bar{M}_i)^2 \sum_{i=1}^n (P_i - \bar{P}_i)^2}} \quad (4)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (P_i - M_i)^2}{N}} \quad (5)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |P_i - M_i| \quad (6)$$

where n = number of inputs, M_i = measured values, P_i = predicted values, \bar{M}_i = average of measured values, and \bar{P}_i = average of predicted values.

Results and discussion

GEP-based model development for TDS and EC

GEP modeling was employed to develop TDS and EC models. The optimal setting of the general parameters, genetic operators, and numerical constants used in GEP modeling are given in Table 3. Basic function sets such as addition, subtraction, and division were set along with structural association of chromosomes prior to GEP algorithm. The GEP models for TDS and EC were selected after running a set of GEP algorithms started from the smallest head size with a single gene chromosome. Upon choosing basic operators, smallest head size and lowest number of chromosomes, one can get simplest final mathematical expressions. Moreover, the modeling process becomes simple and less time-consuming. Table 3 lists

the optimal setting parameters that led to the best possible GEP structure.

Initially the GEP process starts by creating a population of the most viable solutions. Afterwards, the best possible solution could be attained through an iterative process from one to another generation. GEP iterations continued until no changes occurred between the fitness function and the associated correlation values. The results of developed GEP modeling for both TDS and EC are presented in Fig. 8 (a) and (b) respectively. The proposed GEP models successfully simulated the TDS and EC records, with high R^2 results of above 0.90 for both training and testing phases as shown in both sub-figures.

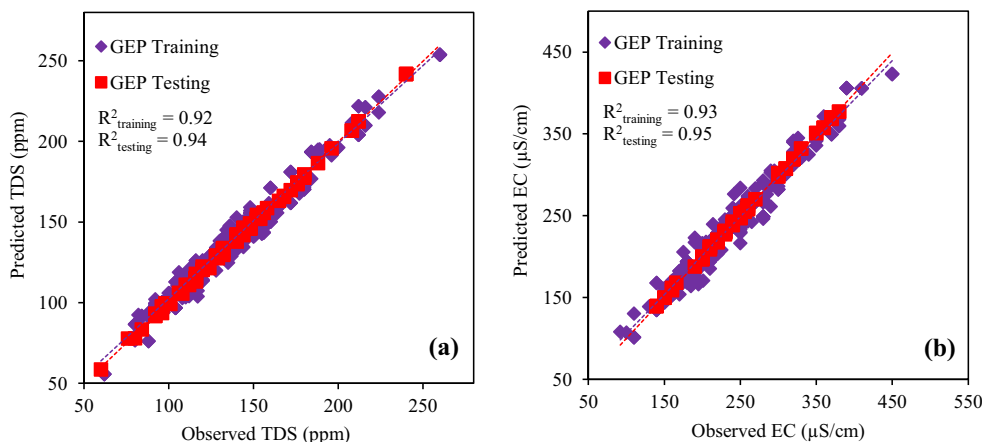
ANN-based model development for TDS and EC

In ANN technique, the feed forward propagation algorithm is used to train the TDS and EC models. As there is no common rule to get the optimum ANN structure and lowest error, so an

Table 3 Summary of the general, genetic operators, and numerical constants setting for GEP

Parameter	Setting
No. of chromosomes	30
No. of Genes	4
Head size	10
Gene size	26
Linking function	Addition
Function sets	$+$, $-$, \times , \div , \wedge^2 , $\sqrt[3]{}$
Mutation rate	0.0138
Inversion rate	0.00546
IS transposition rate	0.00546
RIS transposition rate	0.00546
One point recombination rate	0.00277
Two point recombination rate	0.00277
Gene recombination rate	0.00755
Gene transposition rate	0.00277
Constants per gene	10
Data type	Floating type
Upper and lower bounds	± 10

Fig. 8 Observed versus predicted values using GEP for (a) TDS (b) EC



optimization routine was adopted to search for optimum number of neurons by simultaneously changing the neurons from hidden layers (Basant et al. 2010; Najah et al. 2013; Ouma et al. 2020). The performance of the network is highly influenced by the number of neurons. The desired results cannot be attained by neural network using a limited number of neurons. Similarly, too many neurons make the process lengthier and sometime result in overfitting of the model (Najah et al. 2013). Figures 9 (a) and (b) graphically demonstrate the results of ANN model prediction against the measured values of TDS and EC, respectively. In both sub-figures, there is a strong correlation of the actual versus predicted water quality values as depicted from statistical indicators. In TDS model, the yielded R^2 results for training and testing phases were 0.89 and 0.86, respectively, while, in EC model, both values were 0.88 and 0.82, respectively.

MLR and MNL model development for TDS and EC

Both MLR and MNL models were developed to simulate TDS and EC, and their results are graphically shown in Figs.

10 and 11, respectively. Comparing with actual data, both models showed good estimation for both water quality parameters. For EC, R^2 results were low in the testing dataset compared with the training one, while, for TDS model, R^2 results were above 0.80 in both training and testing phases.

Models cross-validation results

In order to evaluate the performance and ensure the desired accuracy of any model, the validation is of utmost important. The cross-validation is performed to enhance the robustness of the developed models with the help of k-fold cross-validation algorithm. The cross-validation is applied to all the models in each tenfold, and representation of results is illustrated in Fig. 12 and Fig.13 for TDS and EC models, respectively. A fluctuation in the results for individual can be observed, although it maintained a high level of accuracy. For TDS models, the mean R^2 obtained values through cross-validation are 0.82, 0.71, 0.60, and 0.67 for GEP, ANN, MLR, and MNL, respectively. The minimum and maximum R^2 value for TDS models are attained as 0.72 and 0.92, respectively. Similarly, the mean RMSE values, i.e., 6.29, 9.92, 9.35, and 11.43, are

Fig. 9 Observed versus predicted values using ANN for (a) TDS (b) EC

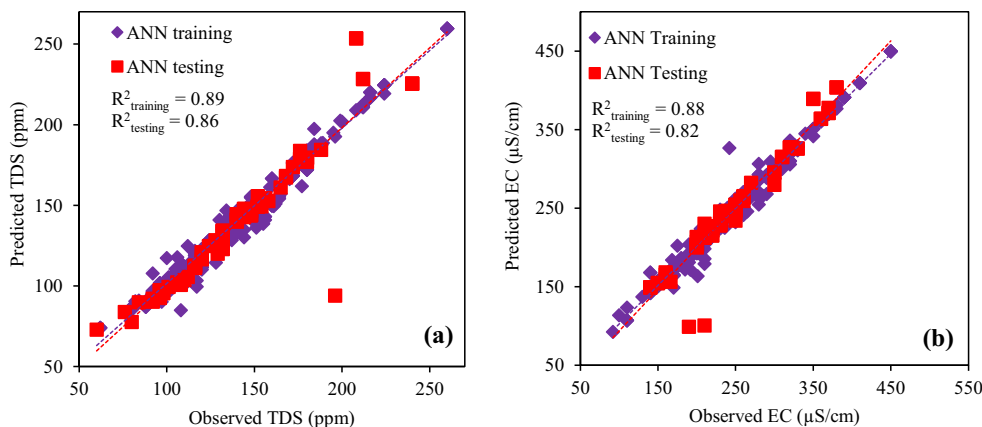
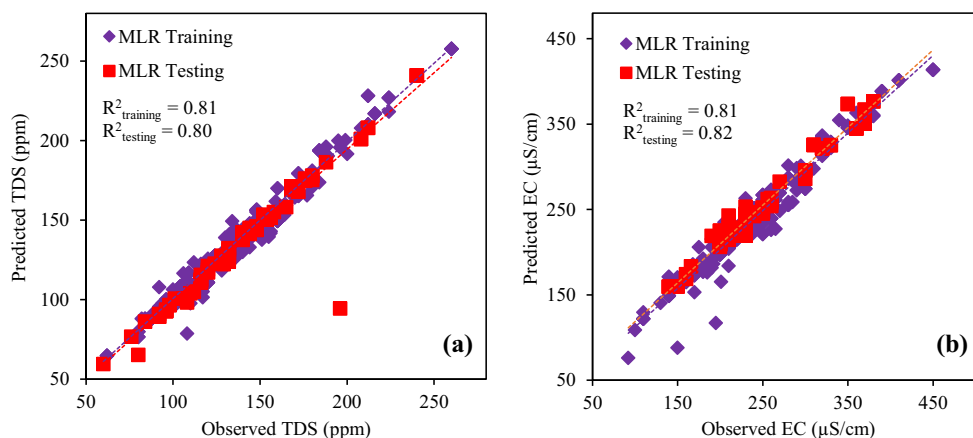


Fig. 10 Observed versus predicted values using MLR for (a) TDS (b) EC



observed for GEP, ANN, MLR, and MNLR, respectively. The method is also applied to EC developed models. The cross-validation results for EC give the mean R^2 values of 0.8, 0.85, 0.56, and 0.65 for GEP, ANN, MLR, and MNLR models, respectively. Meanwhile, the k-fold output for EC revealed lowest and highest RMSE values ranges from 10 to 13 and 30 to 35, respectively. Summarizing the results of k-fold cross-validation, the aforementioned statistics confirm the generalized capability and accurate performance of the developed models. As discussed in last sections, the cross-validation outcome also recognized the superior and accurate performance of GEP model.

Model comparison

The comparison based on R^2 results is not sufficient to distinguish and identify the optimum performance. Therefore, the above-mentioned developed models were challenged by various statistical indicators to analyze their robustness. Table 4 lists the results of these performance measure indicators including NSE, MAE, and RMSE. The RMSE errors are

squared which means a much larger weight is assigned to the larger errors.

In both TDS and EC modeling, GEP technique showed the lowest error values represented by MAE and RMSE values. GEP outperformed the other modeling methods with RMSE of only 6.82 and 9.65 for both TDS and EC models, respectively. The superior performance of GEP was reported in various research studies (Azamathulla et al. 2011; Liu and Wang 2019; Martí et al. 2013; Mehdipour et al. 2017). Furthermore, from the previous figures, it was clear that in both TDS and EC prediction, the performance of ANN models during training was superior as compared with testing phase. This can be referred to the modeling with overfitting, which is considered one of the drawbacks of ANN. The ANN is considered a black box model as it adopts the numerical approach only without taken into account the underlying principles and mechanism (Juditsky et al. 1995). The neural networks have limited ability to clearly identify and portray the possible relationship. Moreover, the neural network models are very prone to overfitting problem due to complexity of the network structure (Tu 1996).

Nevertheless, GEP results showed the lowest MAE errors as well as the highest values of R^2 and NSE for the overall

Fig. 11 Observed versus predicted values using MNLR for (a) TDS (b) EC

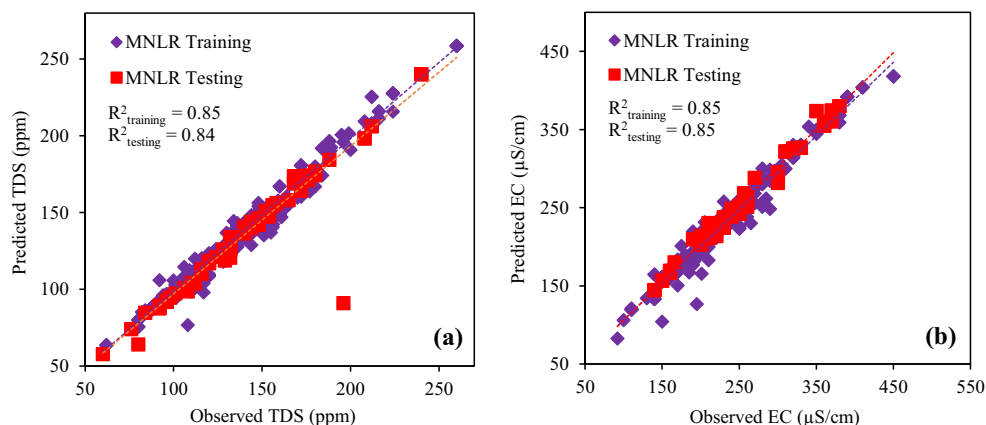
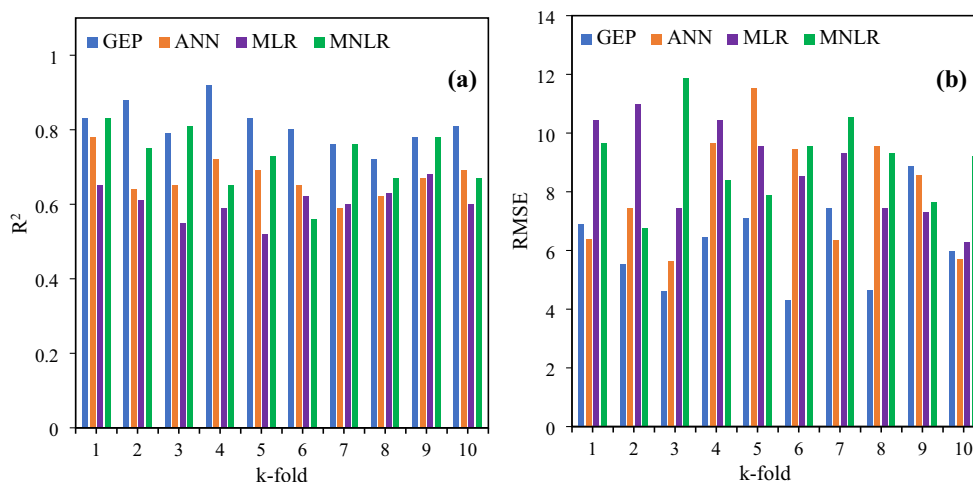


Fig. 12 k-fold cross-validation results of TDS models based on (a) R² (b) RMSE



dataset. For TDS model, both R² and NSE values were 0.92 and 0.91, respectively, while in EC model they were same (0.93). These results indicate that the developed models using GEP modeling were better than the other techniques. The most reliable and accurate results can be adequately obtained by the GEP.

Summarizing the above discussion and the performance comparison of all the developed models, the performance of GEP model outclass all other models. Therefore, GEP models were employed to formulate an applicable equations which can be easily used to estimate TDS and EC values, as described in the following section. The comparative results of

the developed models for TDS and EC parameters are graphically presented in Fig. 14.

Proposed formulation for TDS and EC

Using the results of the developed models by GEP, the following Es. 7 and 8 are proposed to predict TDS and EC, respectively. These equations were derived from the developed expression trees (ETs), as shown in Appendix. The results are TDS and EC estimation in both equations, respectively.

$$\begin{aligned}
 \text{TDS} = & \left(\left(\frac{23895}{Ca} \right)^{\frac{1}{3}} - 45HCO_3 \right) \times (SO_4 - 45HCO_3)^{\frac{1}{3}} + \frac{1}{83HCO_3^{\frac{1}{3}} \times \ln(8.14Cl - 1.11)^2} \\
 & + (4.15 + Na) \times 25 - Na \times 76HCO_3 - \frac{28}{Ca} \left(Mg \times Cl \times 1.17 + \frac{SO_4}{1.03} \right) \times (51 - 7.33Cl)
 \end{aligned} \tag{7}$$

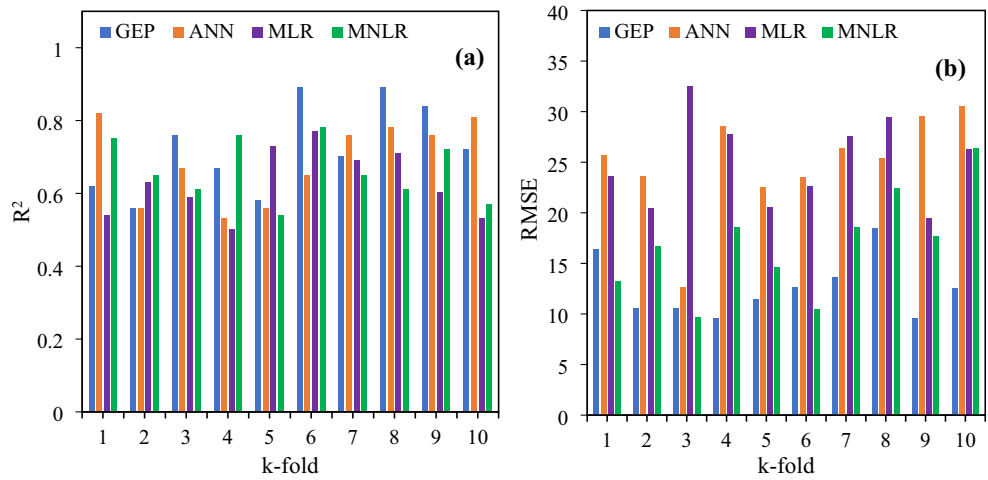
$$\begin{aligned}
 \text{EC} = & (9.6Cl + 5.1SO_4 - \ln HCO_3) \times (4.9 - H45CO_3)^2 + \left\{ \left(SO_4 \times PH \times 2.6 - \frac{32.7}{7CA} \right) - 5.8 \right\} \times 4HCO_3 \\
 & + (N65a + HCO_3 \times 10.4 - SO_4 \times Cl) \times 45.63
 \end{aligned} \tag{8}$$

Sensitivity and parametric analyses

The sensitivity analysis is carried out for the purpose to know the influence of the inputs on the targeted output, since there are uncertainties associated with model inputs, model parameters, or model structure (Chen and Chau 2019). A model can provide accurate results for training and testing data, but its accuracy is not certain on different datasets. Therefore, sensitivity analysis is

essential for the optimization of input parameters and relative contribution of each parameter on the models outputs (TDS and EC). In this study, the developed method by Gandomi et al. (2013) and Javed et al. (2020) was adopted. This method considered the effect of a single parameter on model output. By this method, it is very easy to elaborate and verify the results with actual data. The same method has been adopted in various research studies (Azim et al. 2020; Iqbal et al.

Fig. 13 k-fold cross-validation results of EC models based on (a) R² (b) RMSE



2020). The following Eqs. 9 and 10 were used to find out the contribution by each input variable to the model output.

$$N_i = f_{max}(x_i) - f_{min}(x_i) \tag{9}$$

$$S_i = \frac{N_i}{\sum_{j=1}^n N_j} \times 100 \tag{10}$$

where $f_{max}(x_i)$ and $-f_{min}(x_i)$ is the maximum and minimum of the estimated output over i^{th} output.

The sensitivity of input parameters essential for modeling the water quality, i.e., TDS and EC, was identified as graphically shown in Fig. 15. The results indicated that bicarbonates (HCO₃) is the most sensitive parameter followed by magnesium (Mg) for TDS

concentration in water, with 26.10 and 18.92% relative contribution, respectively. In contrast, pH has no effects on TDS concentration. Similarly, the sensitivity analysis results for EC were mostly affected by HCO₃ content. The second most sensitive parameter for EC estimation was SO₄. However, both magnesium (Mg) and pH have a little or no effect on EC estimation. The effect of year (T_{year}) is least in both TDS and EC models. The relative contribution of T_{year} is 1.01 and 1.92% to TDS and EC models, respectively.

Secondly, parametric analysis is performed aiming at further verifying the robustness of the proposed models. This test was performed by changing the values of a single input parameter while keeping the rest of the variables constant in order to enhance the modeling accuracy. This process measures the competence of the model and helps to recognize the performance of the system being modeled. The values of the input variable were changed with a specific increment for all input variables. Respectively, Figs. 16 and 17 show the

Table 4 Comparison of the developed models' goodness of fit for all dataset

Parameter	Technique	R ²	NSE	MAE	RMSE	R ²	NSE	MAE	RMSE
		Training				Testing			
TDS	GEP	0.92	0.91	5.67	6.53	0.94	0.91	4.45	7.12
	ANN	0.89	0.87	8.17	11.47	0.86	0.81	9.56	11.76
	MLR	0.81	0.80	9.42	8.92	0.80	0.82	8.22	8.25
	MNLR	0.85	0.80	8.76	10.56	0.84	0.82	6.67	12.76
EC	GEP	0.93	0.93	7.55	10.54	0.95	0.94	4.45	8.76
	ANN	0.88	0.84	12.80	39.62	0.82	0.80	13.65	33.56
	MLR	0.81	0.80	14.44	21.23	0.82	0.80	18.76	15.56
	MNLR	0.85	0.82	10.96	19.31	0.85	0.79	9.11	16.54

Fig. 14 Comparison of actual data versus the developed models for (a) TDS (b) EC

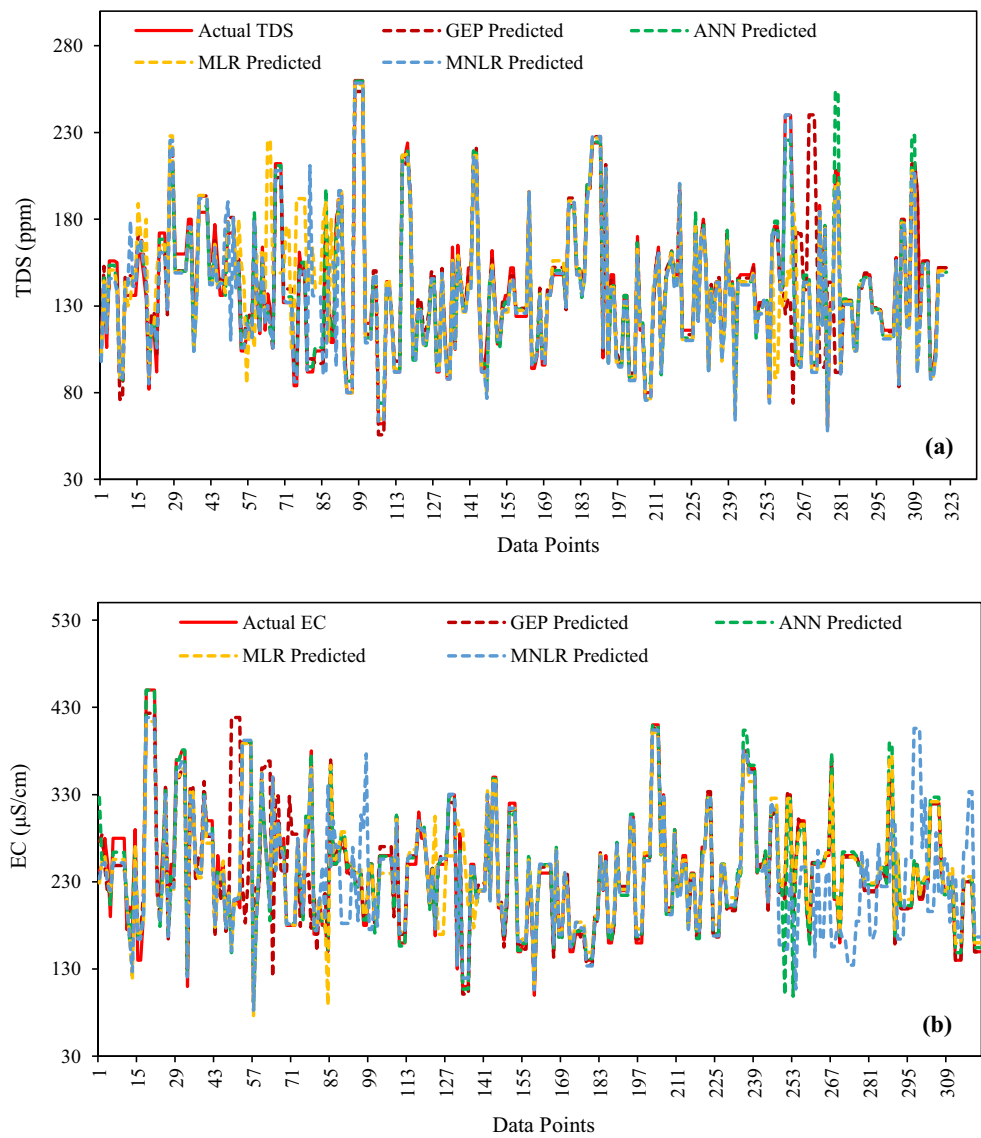


Fig. 15 Results of the sensitivity analysis (a) for TDS and (b) for EC

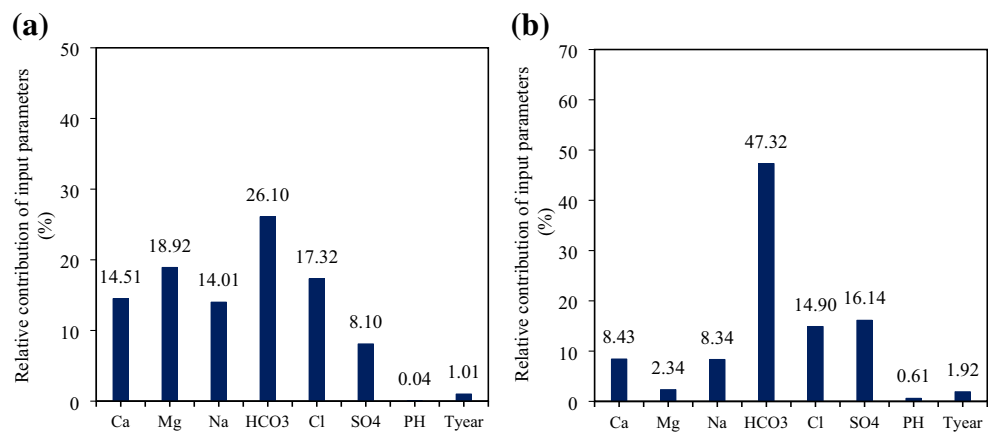


Fig. 16 Results of the parametric analysis for TDS simulation with (a) Ca, (b) Mg, (c) Na, (d) HCO₃, (e) Cl, (f) SO₄, (g) pH, and (h) T_{year}

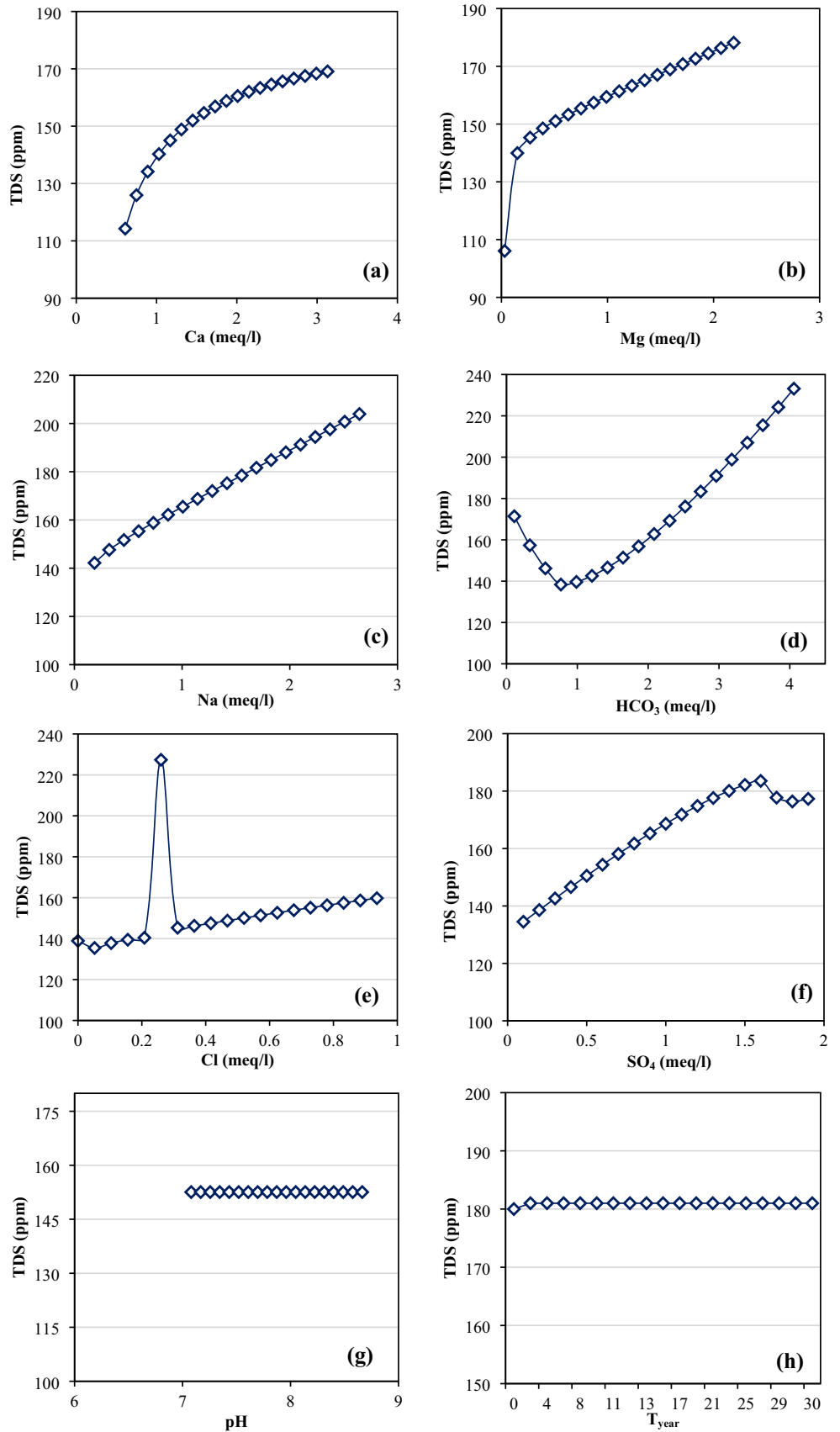
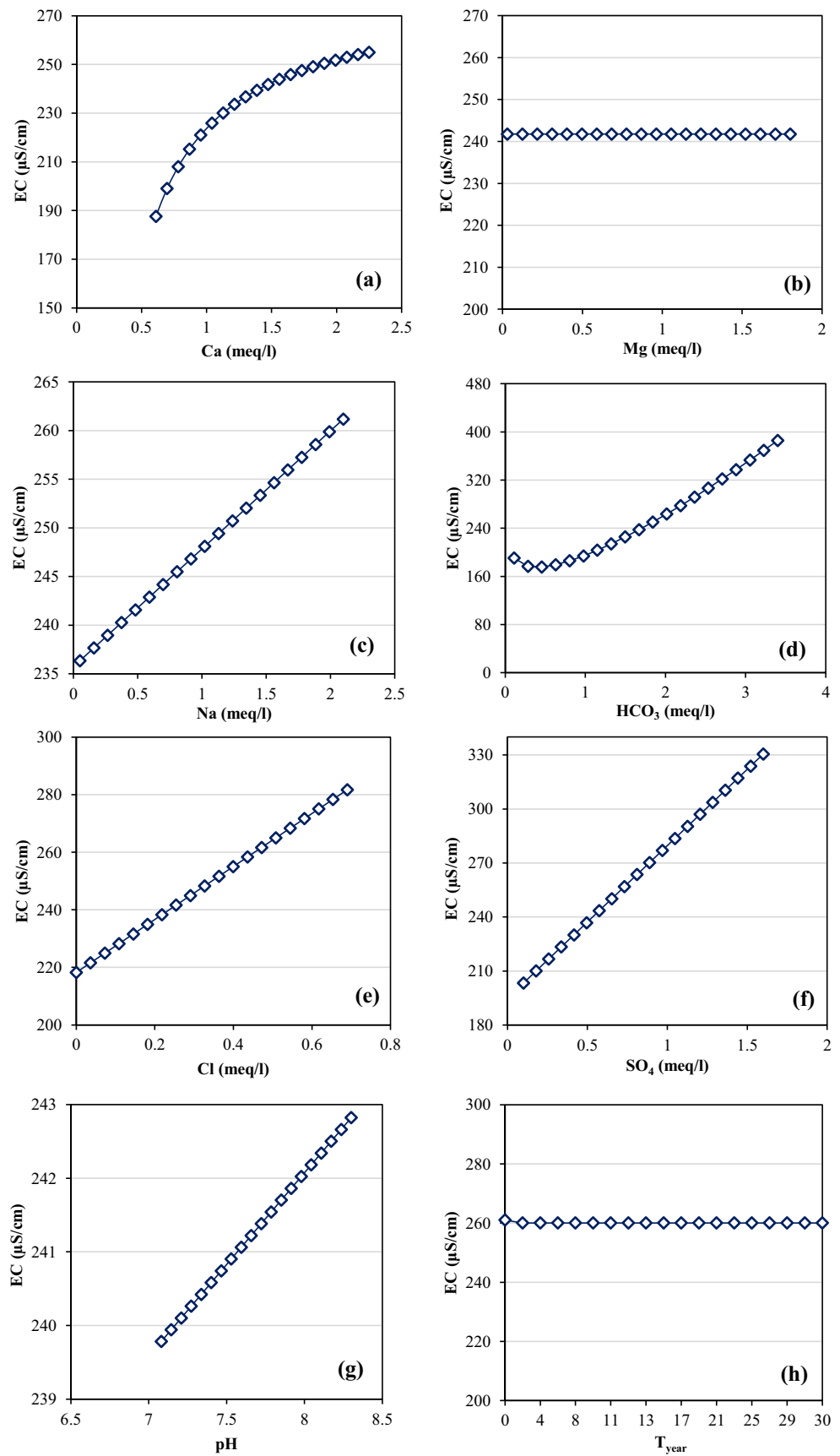


Fig. 17 Results of the parametric analysis for EC simulation with (a) Ca, (b) Mg, (c) Na, (d) HCO₃, (e) Cl, (f) SO₄, (g) pH, and (h) T_{year}



prediction capacity of GEP models for simulating TDS and EC with the variation in input variables, i.e., Ca^{2+} , Mg^{2+} , Na^+ , Cl^- , SO_4^{2-} , HCO_3^- , pH and T_{year} . The results of the parametric analysis revealed that the concentration of TDS and EC varies linearly (with an increasing trend) with all the input variables, where TDS concentration is constant with the increase of pH values. It can also be observed from Figs. 16 and 17 that the TDS and EC values remain the same with addition of successive years (T_{year}). As it is known that both TDS and EC are affected by ion and salt concentration in water. The input variables, i.e., Ca, Mg, Na, HCO_3^- , Cl, and SO_4 , are basically ion and salt concentration. Therefore, any change in the input variables directly affects TDS and EC levels. Various research studies reported the influence on TDS and EC with a variation in ions and salts (Al-Mukhtar and Al-Yaseen 2019; Montaseri et al. 2018). Maedeh et al. (2013) reported that the TDS concentration can be controlled by limiting the amount of ion and salt contents in water. The aforementioned studies are much in line with results of this study, which justified the modeling outcome of the current study.

Conclusion

This study presents the application of data-driven models, i.e., GEP, ANN, MLR, and MNL, for estimating the TDS and EC in the upper Indus river basin. Despite the largely unknown factors responsible for the variation of the river's water quality, the developed models were trained and tested on a monthly data set of TDS and EC measured over a period of almost 30 years (i.e., 1975–2005). The performance of the models was challenged using NSE, R^2 , MAE, and RMSE. The data is also validated with k-fold cross-validation using R^2 and RMSE. All the models exhibited an excellent correlation for observed and simulated data. It was found that the GEP model is superior and outperformed all the other techniques. The developed GEP empirical equations for both TDS and EC could be confidently used for estimating water quality parameters. The novel GEP technique evaluates suitable connections portrayed the physical processes and does not assume prior solution, thus making it superior to others. Both the GEP and ANN models are capable of estimating TDS and EC in river water for a given set of inputs. However, the performance of ANN reduced during model testing and may be

due to data overfitting, limited ability of neural networks, and complexity of the network structure. Moreover, the proposed formulated equations for TDS and EC could assist and help policy makers and engineers to devise a strategy for successful and sustainable management of the water quality.

The work presented in this study has certain limitations. An extensive dataset is essential for modeling studies particularly for data-driven models. The dataset included in this study was for larger duration, i.e., 30 years, but limited up to 2005. Indeed, research on more recent data should be done to know the situation that is important for environmental perspective. Furthermore, the temporal variation is not included in the current study.

It is recommended that further research studies should be spread to surrounding catchments with extensive databank. The spatiotemporal analysis should also be taken in to account. Additionally, some deep learning machine learning algorithm should be considered such as convolution neural network (CNN), multi expression programming (MEP), and recurrent neural network.

Acknowledgments The authors acknowledge the support of water and power development authority (WAPDA), Pakistan, for providing the water quality data of Indus River.

Author's contributions Conceptualization, data collection, and writing original draft preparation: Muhammad Izhar Shah; data analysis, modeling, review, and editing: Muhammad Faisal Javed; validation check, data curation, and manuscript revision: Taher Abunama. All authors approved the final manuscript.

Funding This research received no external funding.

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

Compliance with ethical standards

Competing interests The authors declare that they have no competing interests.

Ethical approval Not applicable

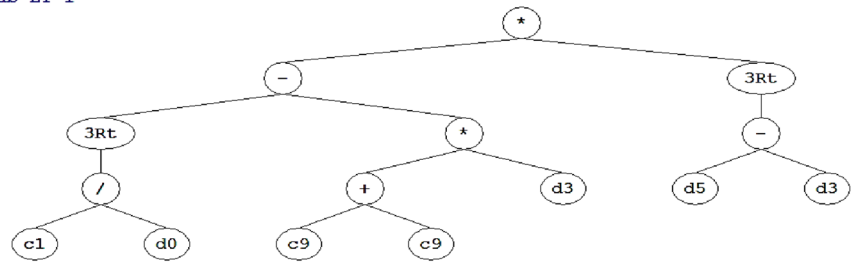
Consent to participate Not applicable

Consent to publish Not applicable

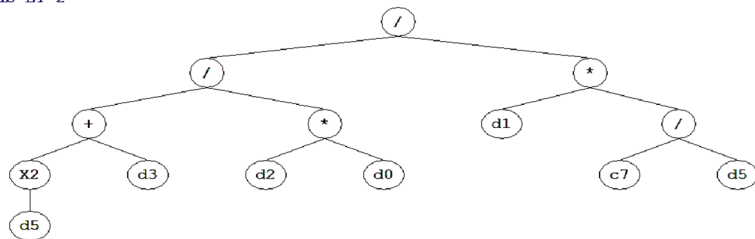
Appendix. A Expression tree diagrams

Fig. 18 Expression tree of the developed GEP model for TDS

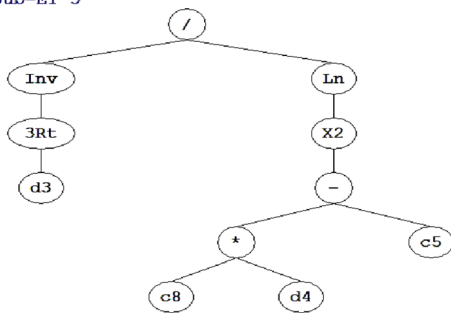
Sub-ET 1



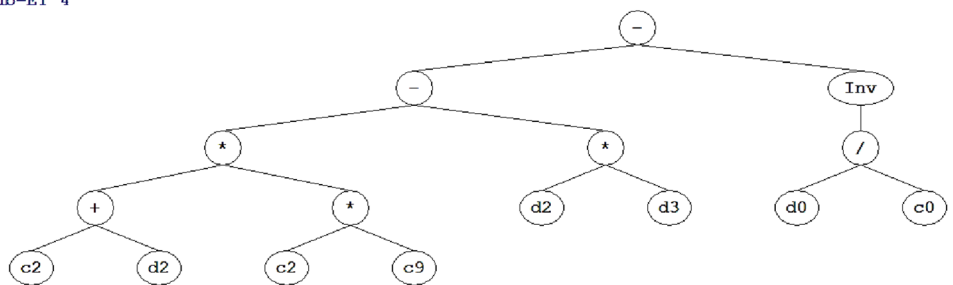
Sub-ET 2



Sub-ET 3



Sub-ET 4



Sub-ET 5

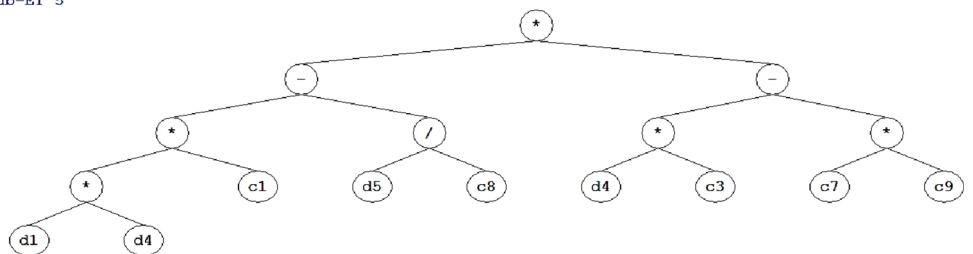
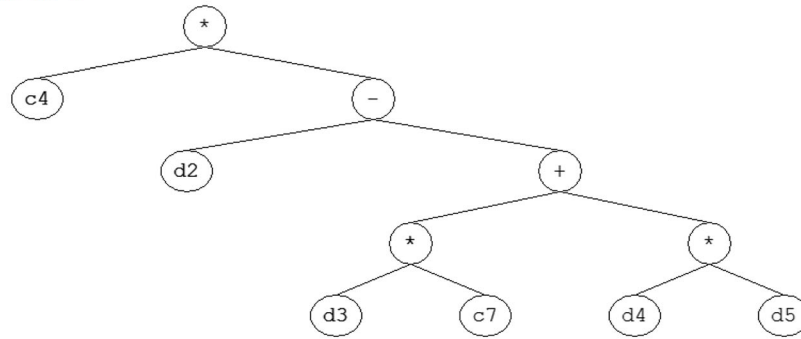
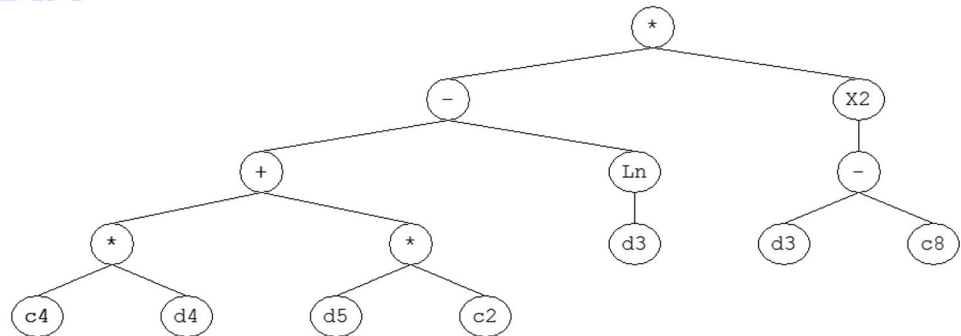


Fig. 19 Expression tree of the developed GEP model for EC

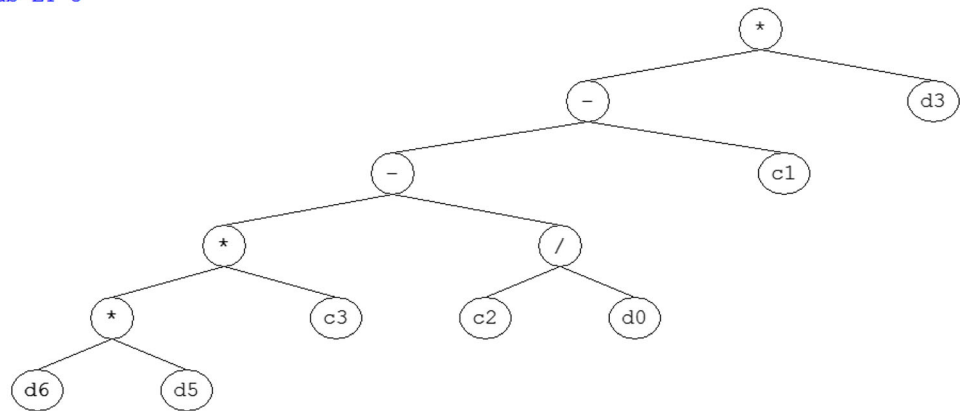
Sub-ET 1



Sub-ET 2



Sub-ET 3



References

- Abdollahzadeh G, Jahani E, Kashir Z (2017) Genetic programming based formulation to predict compressive strength of high strength concrete. *Civil Eng Infrastructures J* 50(2):207–219
- Abunama T, Othman F, Ansari M, El-Shafie A (2019) Leachate generation rate modeling using artificial intelligence algorithms aided by input optimization method for an MSW landfill. *Environ Sci Pollut Res* 26(4):3368–3381
- Adamowski J, Fung Chan H, Prasher SO, Ozga-Zielinski B, Sliusarieva A (2012) Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. *Water Resour Res* 48(1): W01528
- Ali S, Li D, Congbin F, Khan F (2015) Twenty first century climatic and hydrological changes over Upper Indus Basin of Himalayan region of Pakistan. *Environ Res Lett* 10(1):014007. <https://doi.org/10.1088/1748-9326/10/1/014007>
- Alizadeh MJ, Kavianpour MR, Danesh M, Adolf J, Shamshirband S, Chau K-W (2018) Effect of river flow on the quality of estuarine and coastal waters using machine learning models. *Eng Appl Computational Fluid Mech* 12(1):810–823
- Al-Mukhtar M, Al-Yaseen F (2019) Modeling water quality parameters using data-driven models, a case study Abu-Ziriq marsh in south of Iraq. *Hydrology* 6(1):24
- Ansari M, Othman F, Abunama T, El-Shafie A (2018) Analysing the accuracy of machine learning techniques to develop an integrated influent time series model: case study of a sewage treatment plant, Malaysia. *Environ Sci Pollut Res* 25(12):12139–12149

- Aryafar A, Khosravi V, Zarepourfard H, Rooki R (2019) Evolving genetic programming and other AI-based models for estimating groundwater quality parameters of the Khezri plain, Eastern Iran. *Environ Earth Sci* 78(3):69
- Azad A, Karami H, Farzin S, Mousavi S-F, Kisi O (2019) Modeling river water quality parameters using modified adaptive neuro fuzzy inference system. *Water Sci Eng* 12(1):45–54
- Azamathulla HM, Ghani AA, Leow CS, Chang CK, Zakaria NA (2011) Gene-expression programming for the development of a stage-discharge curve of the Pahang River. *Water Resourc Manag* 25(11):2901–2916
- Azamathulla HM, Rathnayake U, Shatnawi A (2018) Gene expression programming and artificial neural network to estimate atmospheric temperature in Tabuk, Saudi Arabia. *Appl Water Sci* 8(6):184
- Azim I, Yang J, Javed MF, Iqbal MF, Mahmood Z, Wang F, and Liu Q-F. (2020). Prediction model for compressive arch action capacity of RC frame structures under column removal scenario using gene expression programming. Paper presented at the Structures.
- Basant N, Gupta S, Malik A, Singh KP (2010) Linear and nonlinear modeling for simultaneous prediction of dissolved oxygen and biochemical oxygen demand of the surface water—a case study. *Chemom Intell Lab Syst* 104(2):172–180
- Bozorg-Haddad O, Soleimani S, Loáiciga HA (2017) Modeling water-quality parameters using genetic algorithm–least squares support vector regression and genetic programming. *J Environ Eng* 143(7): 04017021
- Chen X-Y, Chau K-W (2019) Uncertainty analysis on hybrid double feedforward neural network model for sediment load estimation with LUBE method. *Water Resourc Manag* 33(10):3563–3577
- Chen K, Chen H, Zhou C, Huang Y, Qi X, Shen R, 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
- Crocker J, Bartram J (2014) Comparison and cost analysis of drinking water quality monitoring requirements versus practice in seven developing countries. *Int J Environ Res Public Health* 11(7):7333–7346
- Ferreira C (2006). *Gene expression programming: mathematical modeling by an artificial intelligence* (Vol. 21): Springer.
- Frank IE, and Todeschini R (1994). *The data analysis handbook*: Elsevier.
- Gandomi AH, Yun GJ, Alavi AH (2013) An evolutionary approach for modeling of shear strength of RC deep beams. *Mater Struct* 46(12): 2109–2119
- Gholampour A, Gandomi AH, Ozbakkaloglu T (2017) New formulations for mechanical properties of recycled aggregate concrete using gene expression programming. *Constr Build Mater* 130:122–145
- Iqbal MF, Liu Q-F, Azim I, Zhu X, Yang J, Javed MF, Rauf M (2020) Prediction of mechanical properties of green concrete incorporating waste foundry sand based on gene expression programming. *J Hazard Mater* 384:121322
- Javed MF, Amin MN, Shah MI, Khan K, Iftikhar B, Farooq F, Aslam F, Alyousef R, Alabduljabbar H (2020) Applications of gene expression programming and regression techniques for estimating compressive strength of bagasse ash based concrete. *Crystals* 10(9):737
- Juditsky A, Hjalmarsson H, Benveniste A, Delyon B, Ljung L, Sjöberg J, Zhang Q (1995) Nonlinear black-box models in system identification: Mathematical foundations. *Automatica* 31(12):1725–1750
- Kargar K, Samadianfard S, Parsa J, Nabipour N, Shamshirband S, Mosavi A, Chau K-W (2020) Estimating longitudinal dispersion coefficient in natural streams using empirical models and machine learning algorithms. *Eng Appl Computational Fluid Mech* 14(1): 311–322
- Khan AJ, Koch M (2018) Correction and informed regionalization of precipitation data in a high mountainous region (Upper Indus Basin) and its effect on SWAT-modelled discharge. *Water* 10(11): 1557
- Khan A, Richards KS, Parker GT, McRobie A, Mukhopadhyay B (2014) How large is the Upper Indus Basin? The pitfalls of auto-delineation using DEMs. *J Hydrol* 509:442–453
- Khare MJK, Warke A (2014) Selection of significant input parameters for water quality prediction—a comparative approach. *Int J Res Advent Technol* 2(03):81–90
- Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings Ijcai*, 14th edn. Montreal, Canada, pp 1137–1145
- Liu M, Lu J (2014) Support vector machine—an alternative to artificial neuron network for water quality forecasting in an agricultural non-point source polluted river? *Environ Sci Pollut Res* 21(18):11036–11053. <https://doi.org/10.1007/s11356-014-3046-x>
- Liu L-W, Wang Y-M (2019) Modelling reservoir turbidity using Landsat 8 Satellite Imagery by gene expression programming. *Water* 11(7): 1479
- Maedeh A, Mehrdadi N, Bidhendi G, Abyaneh HZ (2013) Application of artificial neural network to predict total dissolved solids variations in groundwater of Tehran plain: Iran. *Int J Environ Sustain* 2(1):10–20
- Martí P, Shiri J, Duran-Ros M, Arbat G, De Cartagena FR, Puig-Bargués J (2013) Artificial neural networks vs. gene expression programming for estimating outlet dissolved oxygen in micro-irrigation sand filters fed with effluents. *Comput Electron Agric* 99:176–185
- Mehdipour V, Memarianfard M, Homayounfar F (2017) Application of Gene Expression Programming to water dissolved oxygen concentration prediction: *Int. J Hum Cap Urban Manag* 2(1):1–10
- Montaseri M, Ghavidel SZZ, Sanikhani H (2018) Water quality variations in different climates of Iran: toward modeling total dissolved solid using soft computing techniques. *Stoch Env Res Risk A* 32(8): 2253–2273
- Mustafa YA, Jaid GM, Alwared AI, Ebrahim M (2014) The use of artificial neural network (ANN) for the prediction and simulation of oil degradation in wastewater by AOP. *Environ Sci Pollut Res* 21(12): 7530–7537. <https://doi.org/10.1007/s11356-014-2635-z>
- Najah A, El-Shafie A, Karim OA, El-Shafie AH (2013) Application of artificial neural networks for water quality prediction. *Neural Comput & Applic* 22(1):187–201
- Nasr M, Zahran HF (2014) Using of pH as a tool to predict salinity of groundwater for irrigation purpose using artificial neural network. *Egypt J Aqua Res* 40(2):111–115
- Ouma YO, Okuku CO, Njau EN (2020) Use of artificial neural networks and multiple linear regression model for the prediction of dissolved oxygen in rivers: case study of hydrographic basin of River Nyando, Kenya. *Complexity* 2020:9570789 1–23
- Pal S, Mukherjee S, Ghosh S (2014) Estimation of the phenolic waste attenuation capacity of some fine-grained soils with the help of ANN modeling. *Environ Sci Pollut Res* 21(5):3524–3533. <https://doi.org/10.1007/s11356-013-2315-4>
- Ramzan S, Zahid FM, Ramzan S (2013) Evaluating multivariate normality: a graphical approach. *Middle-East J Sci Res* 13(2):254–263
- Salami E, Salari M, Ehteshami M, Bidokhti N, Ghadimi H (2016) Application of artificial neural networks and mathematical modeling for the prediction of water quality variables (case study: southwest of Iran). *Desalin Water Treat* 57(56):27073–27084
- Sarkar A, Pandey P (2015) River water quality modelling using artificial neural network technique. *Aqua Proc* 4:1070–1077
- Sattari MT, Joudi AR, Kusiak A (2016) Estimation of water quality parameters with data-driven model. *J-Am Water Works Assoc* 108(4):E232–E239
- Seyam MS, Alagha J, Abunama T, Mogheir Y, Affam AC, Heydari M, Ramlawi K (2020) Investigation of the influence of excess pumping on groundwater salinity in the Gaza Coastal Aquifer (Palestine) using three predicted future scenarios. *Water* 12(8):2218

- Shah MI, Khan A, Akbar TA, Hassan QK, Khan AJ, Dewan A (2020) Predicting hydrologic responses to climate changes in highly glacierized and mountainous region Upper Indus Basin. *R Soc Open Sci* 7(8):191957
- Shamshirband S, Jafari Nodoushan E, Adolf JE, Abdul Manaf A, Mosavi A, Chau, K.-w. (2019) Ensemble models with uncertainty analysis for multi-day ahead forecasting of chlorophyll a concentration in coastal waters. *Engineering Applications of Computational Fluid Mechanics* 13(1):91–101
- Tahir AA, Chevallier P, Arnaud Y, Neppel L, Ahmad B (2011) Modeling snowmelt-runoff under climate scenarios in the Hunza River basin, Karakoram Range, Northern Pakistan. *J Hydrol* 409(1-2):104–117
- Tu JV (1996) Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J Clin Epidemiol* 49(11):1225–1231
- Tung TM, Yaseen ZM (2020) A survey on river water quality modelling using artificial intelligence models: 2000–2020. *J Hydrol* 585: 124670
- Zhang Y, Gao X, Smith K, Inial G, Liu S, Conil LB, Pan B (2019) Integrating water quality and operation into prediction of water production in drinking water treatment plants by genetic algorithm enhanced artificial neural network. *Water Res* 164:114888

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