Environment, Development and Sustainability (2021) 23:18252–18277
https://doi.org/10.1007/s10668-021-01437-6

# Remote sensing-based water quality index estimation using data-driven approaches: a case study of the Kali River in Uttar Pradesh, India

**Saif Said**[1] · **Shadab Ali Khan**[1]

Received: 4 April 2020 / Accepted: 13 April 2021 / Published online: 21 April 2021


**Abstract**
The present study evaluates the water quality status of 6-km-long Kali River stretch that passes through the Aligarh district in Uttar Pradesh, India, by utilizing high-resolution IRS P6 LISS IV imagery. In situ river water samples collected at 40 random locations were analyzed for seven physicochemical and four heavy metal concentrations, and the water quality index (WQI) was computed for each sampling location. A set of 11 spectral reflectance band combinations were formulated to identify the most significant band combination that is related to the observed WQI at each sampling location. Three approaches, namely multiple linear regression (MLR), backpropagation neural network (BPNN) and gene expression programming (GEP), were employed to relate WQI as a function of most significant band combination. Comparative assessment among the three utilized approaches was performed via quantitative indicators such as $R^2$, RMSE and MAE. Results revealed that WQI estimates ranged between 203.7 and 262.33 and rated as "*very poor*" status. Results further indicated that GEP performed better than BPNN and MLR approaches and predicted WQI estimates with high $R^2$ values (i.e., 0.94 for calibration and 0.91 for validation data), low RMSE and MAE values (i.e., 2.49 and 2.16 for calibration and 4.45 and 3.53 for validation data). Moreover, both GEP and BPNN depicted superiority over MLR approach that yielded WQI with $R^2 \sim 0.81$ and 0.67 for calibration and validation data, respectively. WQI maps generated from the three approaches corroborate the existing pollution levels along the river stretch. In order to examine the significant differences among WQI estimates from the three approaches, one-way ANOVA test was performed, and the results in terms of $F$-statistic ($F = 0.01$) and $p$-value ($p = 0.994 > 0.05$) revealed WQI estimates as "*not significant*," reasoned to the small water sample size (i.e., $N = 40$). The study therefore recommends GEP as more rational and a better alternative for precise water quality monitoring of surface water bodies by producing simplified mathematical expressions.

**Keywords** Kali River · WQI · Spectral reflectance · MLR · ANN · GEP

✉ Saif Said
saif_said@rediffmail.com

Shadab Ali Khan
shadab7856gc@gmail.com

[1] Civil Engineering Department, Aligarh Muslim University, (AMU), Aligarh, India

## 1 Introduction

İn recent years, water quality of major rivers, lakes and ponds in India has alarmingly deteriorated due to significant population increase leading to rapid urban development and industrialization. Increased anthropogenic activities including direct discharge of untreated industrial effluents, domestic sewage and agricultural waste have severely degraded the quality of surface water bodies. In India, the management strategies for cleaning up of rivers are often not optimally prioritized and therefore, spatiotemporal monitoring of pollution levels becomes essential to devise effective measures for reclamation of the degraded urban water bodies (Farhad et al., 2013; Abba et al., 2015). In situ measurement and monitoring of water quality at point locations is exhaustive and time taking (Song et al., 2012). Mathematical models integrated with geospatial techniques form a reliable time-saving solution towards controlling and sustainably managing the surface water resources (Mondal and Satpaty, 2020). Geospatial techniques offer uninterrupted scaled monitoring of several water quality parameters (WQPs) over large water bodies at spatiotemporal scales (Fulazzaky et al., 2010; Prabu et al., 2011).

In the last two decades, water quality monitoring of the urban water bodies has been the focus of research for researchers across the globe. The qualitative assessment of river water quality is carried out in terms of its physical, chemical and biological parameters and involves the analysis of complicated data matrix with large number of water quality attributes. Many studies concentrated on evaluating pollution levels in terms of individual WQPs, namely electrical conductivity (EC), turbidity, dissolved oxygen (DO), total dissolved solids (TDS), biochemical oxygen demand (BOD), chemical oxygen demand (COD), alkalinity, total suspended sediment (TSS), chlorophyll-a (Chl-a), and heavy metals such as Iron (Fe), magnesium (Mg), chromium (Cr) and lead (Pb) by utilizing remote sensing data in geographical information system (GIS) framework (Milanović Pešić et al., 2020; Nas et al., 2010; Sharma et al., 2018; Waxter, 2014; Yao et al., 2020). To reduce the number of WQPs in the analysis, a lot of consideration has been given to the development of single numerical indicators to ascertain the overall water quality trends with respect to the threshold limits. The water quality index (WQI) is a numeric indicator of the degree of severity in the quality of water for practical usage within the prescribed range and is computed by considering several significant quality parameters (Bordalo et al., 2006; Dunca, 2018; Markogianni et al., 2014; Mohamed et al., 2019; Said & Hussain, 2019; Sharaf, 2017; Sharma et al., 2018; Syahreza et al., 2012; Zhu, 2013). To classify the degree of severity, WQI is grouped into broad classes, i.e., excellent, good, moderate, poor, etc. For assessing the quality of any water body, numerous water quality indices have been proposed. Most commonly utilized WQIs are weighted arithmetic index method (Brown et al., 1970), national sanitation foundation water quality index (NSFWQI) (Hoseinzadeh et al., 2014), overall index of pollution (OIP) (Sargaonkar & Deshpande, 2003), etc. The OIP furnishes an in-depth understanding of the water quality status of the surface water sources, especially under Indian conditions (Sargaonkar & Deshpande, 2003). Remote sensing of water quality involves visible and infrared portion of the electromagnetic spectrum to explore the sensitivity of spectral band combinations by utilizing advanced computing techniques. Several data-driven approaches have been implemented to quantify the relationship between actual and modeled WQPs for qualitative modeling of water quality and requires input data, model parameters, and other relevant information (Bordalo et al., 2006). Many studies employed statistical approaches to explore linear correlations, such as MLR, logarithmic relation and exponential relation, while others concentrated on more

efficient, nonlinear analytical methods, viz. artificial neural network (ANN), genetic programming (GP), group method of data handling (GMDH), GEP, etc., in conjunction with geospatial techniques (Akbal et al., 2011; Avdan et al., 2019; Boyacioglu, 2010; Chapagain et al., 2010; Hussain et al., 2008; Lotfinasabasl et al., 2018).

In recent years, ANN modeling has been widely utilized to quantify the severity of water quality issues due to its fast training process and ability to solve linear and nonlinear complex problems (Bonansea et al., 2015; Nasri, 2010; Nathan et al., 2017). Many studies utilized the BPNN and radial basis function (RBF) neural network for evaluating water quality and provided favorable outcomes through modeling complex nonlinear response functions, such as spectral reflectance values and WQP estimates (Ekercin, 2007; Gürsoy & Atun, 2019; Marquez et al., 2018; Zhang et al., 2003; Zhao et al., 2014). In river management programs, ANNs have effectively been used to evaluate the WQI levels to simulate wetland processes (Reynolds & Maberly, 2002; Kuo et al., 2007; Li et al., 2009; Song et al., 2012; Wang et al., 2012). Chu et al. (2013) developed ANN model that could effectively predict the quality of the surface water bodies and introduced the factor analysis technique to identify significant water quality parameters. In another study conducted by Hafeez et al. (2018), four machine learning approaches, namely artificial neural network (ANN), random forest (RF), cubist regression (CB) and support vector regression (SVR), were compared for retrieval of water quality indicators (i.e., Chl-a, SS and turbidity) over the coastal waters of Hong Kong by employing water reflectance values acquired from hand-held spectroradiometer and satellite data. Results revealed ANN as the best performer than other three approaches. More recent studies conducted by Wang et al. (2019, 2020) inferred deep learning process as a promising tool for formulating environmental property prediction models for screening of green solvents. Several studies successfully applied GEP, along with GP, to a variety of water resources issues (Azamathulla & Ghani, 2011; Ghavidel & Montaseri, 2014; Liu & Wang, 2019; Zakaria et al., 2010). Furthermore, these techniques have been considered as substantial tools in solving complex environmental and river engineering problems (Aras et al., 2007; Chen et al., 2008; Mohammadpour et al., 2015). Ni et al. (2012) effectively evaluated the water fluctuations in the wetlands by utilizing the GP approach. Xu and Qin (2013) measured the agricultural water quality through the combined application of GA and fuzzy simulation. In a significant study by Martí et al. (2013), comparison of three approaches, namely ANN, GEP and MLR for estimation of outlet dissolved oxygen in micro-irrigation, was carried out, and the outcomes revealed GEP as the most effective approach. In a recent study carried out by Li and Wang (2019), a reliable turbidity model was developed to predict reservoir turbidity based on Landsat-8 satellite imagery by utilizing an MLR and GEP approach. Results revealed GEP to be more rational and accurate for turbidity simulation. Quantification of pollution levels in water bodies during the lockdown period worldwide forms a crucial aspect for researchers to interpret the short and long-term effect of the coronavirus disease 2019 (COVID-19) on the river dynamics. It has been reported in few recent studies that the pollution level has exceedingly reduced and most water bodies have completely been restored (Clifford, 2020; Häder et al., 2020; Stone, 2020).

Kali River, a major source of irrigation in western Uttar Pradesh, India, has completely deteriorated due to ever increasing disposal of municipal and industrial waste from adjoining cities. Some earlier studies suggested the river water quality as safe for irrigation purposes, whereas later studies revealed river water to be severely polluted with heavy metal concentrations exceeding far beyond the permissible limits (Mishra et al., 2015; Maurya & Malik, 2016). The Kali River has been identified as the most critically contaminated after Markanda River (in Haryana State) in terms of BOD levels (CPCB, 2012). Spatial

monitoring of the water quality of Kali River by employing reliable data-driven approaches is a prerequisite to conserve and manage the river restoration process. Therefore, the main objective of the study is to evaluate and map WQI estimates along a 6-km-long stretch of the Kali River passing through the Aligarh district in Uttar Pradesh, India, by utilizing high-resolution IRS P6 LISS IV imagery. Eleven spectral reflectance band combinations were formulated to identify the most significant band combination associated with the observed WQI at the sampling locations. Three approaches, namely MLR, BPNN and GEP, were employed to relate WQI as a function of most significant band combination. The performance of three approaches was assessed by via quantitative indicators such as coefficient of determination ($R^2$), root mean square error (RMSE) and mean absolute error (MAE). A one-way ANOVA (analysis of variance) test was also performed to assess significant differences among WQI estimates from the three approaches at a confidence level of 0.05. Maps depicting spatial variation of WQI levels in the river stretch were generated in GIS framework. The present study configures the basis for policy makers and environmentalists to devise effective and sustainable strategies and policies to reclaim the completely degraded river ecosystems.

## 2 Materials and methods

### 2.1 Study area

The study area, illustrated in Fig. 1, covers 6-km-long stretch of Kali River (meaning "black" in the local language) that passes through the Aligarh district in Uttar Pradesh, India. Study area is confined within latitude 28.11°N to 28.15°N and longitude 78.14°E to 78.18°E at an elevation of 213 m above the mean sea level. The river had been a major source of water for domestic as well as irrigation requirements in the past two decades. The Kali River originates from the village of Antwada, in the Muzaffarnagar district, Uttar Pradesh, passes through many important cities and joins the Ganges River at the city of Kannauj in the Farrukhabad district. The river covers a total span of almost 300 km. Large cities, including Meerut, Hapur and Bulandshahr, accommodate numerous small- and large-scale industries along the river banks, such as sugar mills, paper mills, textile industries, slaughterhouses and distilleries. The current status of the river justifies its name, owing to the excessive discharge of domestic sewage and untreated industrial effluents into the river thus, conveying more than 60 per cent of the pollution load (CPCB, 2012). Over the years, the river has completely transformed into a highly toxic flow of chemicals, harmful for human consumption, and offers a restricted use for irrigation or any other purpose. Toxic water from the Kali River is widely consumed for fulfilling the irrigation requirements of surrounding areas. The present condition of the river is pity and demands immediate attention for its reclamation.

### 2.2 Data collection and analysis

River water samples were collected from the midstream at a depth of 0.5 m on April 27, 2018, concurrent to the date of satellite overpass. Grab sampling procedure was adopted for the analysis of various WQPs as recommended by the standard methods of analysis (APHA, 1998). Water samples from the Kali River were analyzed in the laboratory of Environmental Engineering, Civil Engineering Department, AMU, Aligarh, and the WQI

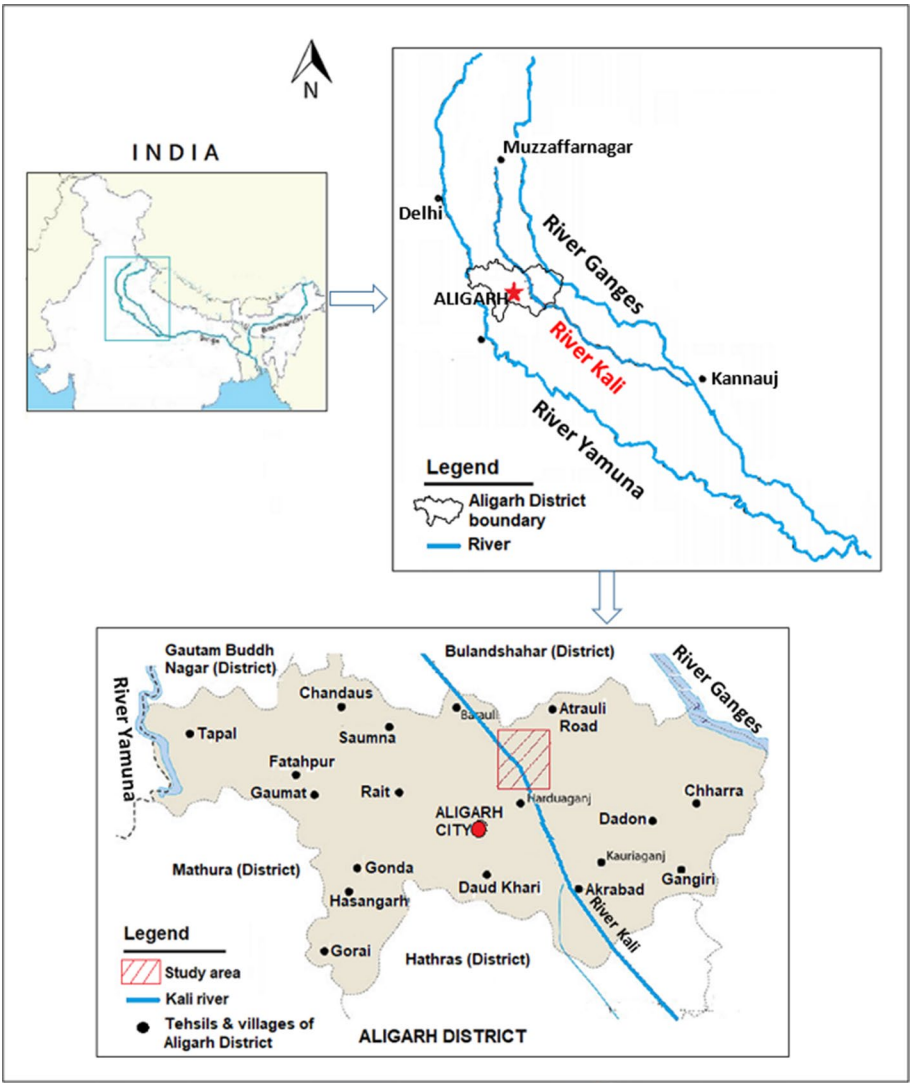



**Fig. 1** Location map of the study area (map not to scale)

for each sampling location was estimated from 11 physicochemical parameters and heavy metals, namely pH, EC, DO, TDS, BOD, COD, alkalinity, Fe, Mn, Cr and Pb. The heavy metal concentration was measured by adopting American Society for Testing and Materials (ASTM, 2000) procedure involving the digestion of water samples with concentrated HNO3 and employing an atomic absorption spectrophotometer (AAS).

WQI values for 40 water samples were computed by following a three-step procedure (Water programme, 2007). The first step assigns weight (*wi*) to all the WQPs ranging from 1 to 5 in accordance with their relative significance towards the overall quality grading of the water for irrigation purposes. The relative significance among WQPs was decided

on the basis of collective expert opinions taken from different published studies (Ramakrishnaiah et al., 2009; Nabizadeh et al., 2013; Suneetha et al., 2015). The highest weight value, i.e., 5, was assigned to two heavy metals, i.e., Pb and Cr, on account of their prominence towards rendering severity to the water quality. Lower rank of 1 was assigned to pH, and 2 was assigned to COD and BOD. Ranks 3 and 4 were appropriately assigned to alkalinity, TDS, DO, EC, Fe and Mn on the basis of their relative severity (Srinivasamoorthy et al., 2008). The second step computes the relative weight ($W_i$) as per the equation below.

$$W_i = {}^{w_i}\Big/\sum_{i=1}^{n} w_i, \tag{1}$$

where $W_i$ is the relative weight, $w_i$ is the individual parameter weight, and $n$ is the number of parameters. In the third step, a quality rating scale ($q_i$) for each parameter was evaluated by dividing its concentration levels for every water sample by its corresponding standard concentration, as per the Bureau of Indian Standards (BIS, 1986).

$$q_i = {}^{C_i}\Big/_{S_i} \times 100, \tag{2}$$

where $q_i$ is the quality rating in percent, $C_i$ is the concentration of each chemical parameter in each water sample in mg/L, and $S_i$ is the irrigation water quality standard for each chemical parameter in mg/L. Finally, the WQI for each sampling location was computed as per Brown et al. (1970) expressed as Eq. 3, where $SL_i$ is the product of $W_i$ and $q_i$.

$$\mathrm{WQI} = \sum_{i=1}^{n} SL_i \tag{3}$$

The WQI values corresponding to the sampling locations were evaluated by following the above procedure and scaled for quality rating in accordance with BIS (1986) specifications, provided in Table 1.

### 2.3 Remote sensing data used

Image from IRS P6 Resourcesat-2 LISS IV sensor of April 27, 2018, was utilized in the present study for evaluating and mapping the water quality of Kali River in terms of WQI measures. The study area was delineated, and a subset image was created using the Erdas Imagine software, shown in Fig. 2. IRS LISS IV sensor produces a high-resolution multispectral image in three bands (i.e., green, red and near Infrared) with 5.8 m spatial resolution in the multispectral mode at nadir. Corresponding to the sampling locations, pixel values with reference to digital numbers (DN) from three spectral bands were extracted and converted into physical quantities (e.g., radiance) and then

**Table 1** WQI and corresponding water quality rating as per the BIS (1986) specifications

| S no | WQI | Status | Possible usages |
|---|---|---|---|
| 1 | 0–50 | Excellent | Drinking, irrigation and industrial |
| 2 | 50–100 | Good | Domestic, irrigation and industrial |
| 3 | 100–200 | Poor | Irrigation |
| 4 | 200–300 | Very poor | Restricted use for all purposes |
| 5 | > 300 | Severe | Proper treatment required before use |

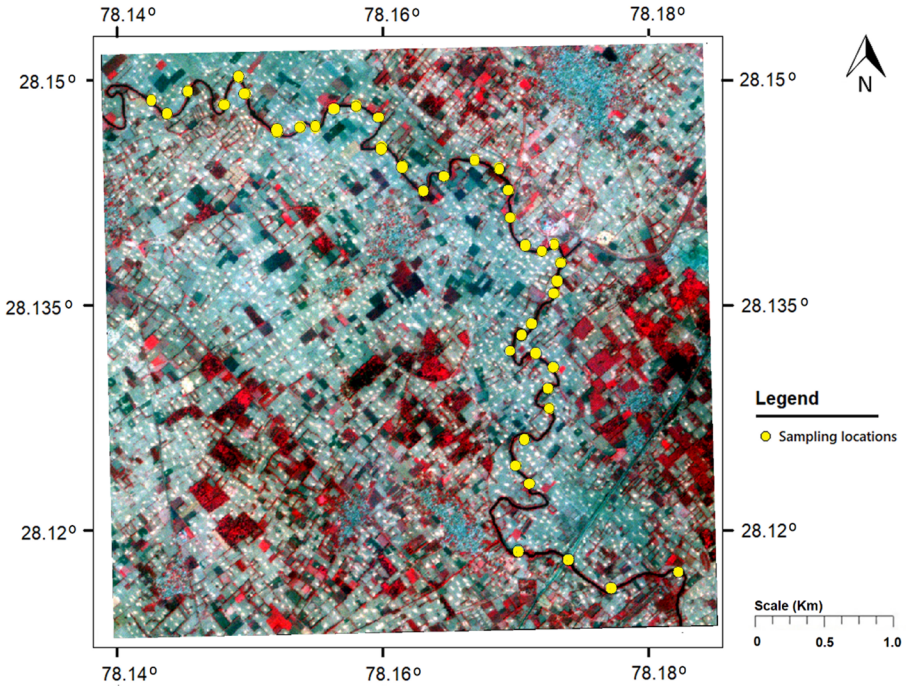



**Fig. 2** Subset image of study area with sampling locations along the river stretch

into spectral reflectance. The process takes into account the terrain and atmospheric corrections. The conversion involved the utilization of the radiometric "gain and offset" extracted from the image metadata and employed Eqs. 4 and 5 for radiance and reflectance, respectively, proposed by Chander and Markham (2003)

$$L_\lambda = \text{Gain}_\lambda \times \text{DN}_\lambda + \text{offset}_\lambda, \tag{4}$$

where $\lambda$ is the specific spectral band of the image; $L_\lambda$ is the spectral radiance for band $\lambda$ at the sensor's aperture (mW/cm$^2$/µm/str); gain$_\lambda$ is the radiometric calibration gain (mW/cm$^2$/µm/str/DN) for band $\lambda$ from product metadata (gain values for three bands were considered: $G=52$, $R=47$ and NIR = 31); DN$_\lambda$ is digital number value for band $\lambda$ of the image; and offset$_\lambda$ is the radiometric calibration (mW/cm$^2$/µm/str) for band $\lambda$ from product metadata, which is zero for the three bands

$$\rho_P = \frac{\pi \times L_\lambda \times d^2}{E_{SUN} \times \text{Cos}\theta_S}, \tag{5}$$

where $\rho_P$ is the dimensionless planetary reflectance, d is the Earth–Sun distance (astronomical units, $1-(0.01674 \cos (0.9856 (\text{JD-4})))^2$, where JD is Julian Day), $E_{SUN\lambda}$ is the average solar exo-atmospheric spectral irradiances (mW/cm$^2$/µm) at 1 astronomical unit (AU) distance between the Earth and Sun, $\theta_s$ is the Sun's zenith angle (~67.337461° from product metadata), and $L_\lambda$ is the spectral radiance for band $\lambda$ at the sensor's aperture (mW/cm2/µm/str).

## 2.4 Modeling approaches

### 2.4.1 Multiple linear regression (MLR)

MLR analysis predicts the unknown variable from two or more known variables that are termed as the predictors. In other words, a multiple regression analysis aids in predicting the $Y$ value for given $X_1$, $X_2$, …, $X_k$ values. The multiple regression equation of $Y$ with known $X_1$, $X_2$, …, $X_k$ is given by

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

where $b_0$ is the intercept and $b_1$, $b_2$, $b_3$, …, $b_k$ are the regression coefficients that correspond to the slope in a linear regression equation. An MLR was employed to examine the most appropriate formulated spectral reflectance band combination, producing WQI estimates with high $R^2$ values and low RMSE and MAE values.

### 2.4.2 Artificial neural network (ANN)

The feed forward backpropagation neural network (FF-BPNN) algorithm looks for the least error function in weight space by employing the gradient descent method. The learning process resolves the complexity of the problem through randomly assigning weights that produce the least error function. The entire process is executed in two phases. In the first phase, assigned weights to the network architecture are initialized randomly to propagate forward, along with input data, to compute the target value. In the second phase, the error between the actual and estimated targets is compared and the error value that is higher than the threshold value is rolled backward through the network. The weight values are recalculated, and the process is continued until the minimum error is attained. During the training process, the errors for both training and testing data decrease with number of iterations until a constant minimum error value is attained. Training is stopped at a point when, the least difference between training and testing data errors is observed so as to avoid overtraining of the network (Said et al., 2008). The most general neural network architecture consists of three layers, i.e., input, hidden and output layers, as illustrated in Fig. 3.

Every unit in a layer is connected with units in the adjoining layer with a unique weight value. Variables in the input layer, along with connected weights, propagate to every unit of the next hidden layer. The end product of every unit forming an output is compounded with weights of preceding connecting units and is advanced to the successive layer before finally being subjected to the sigmoid activation function. The output value from the $j^{th}$ unit of layer $m$ is represented as

$$6O_j^m = S\left(l_j^m\right), \tag{7}$$

where $S$ is the sigmoid activation function, as proposed by Rumelhart et al. (1986), and

$$f(x) = \frac{1}{1 + \exp(-x)}. \tag{8}$$

The function $f(x)$ acquires values from zero to unity for the entire range of inputs; $x$ is the input value, viz. $l_j^m$ obtained for of layer $m$, as

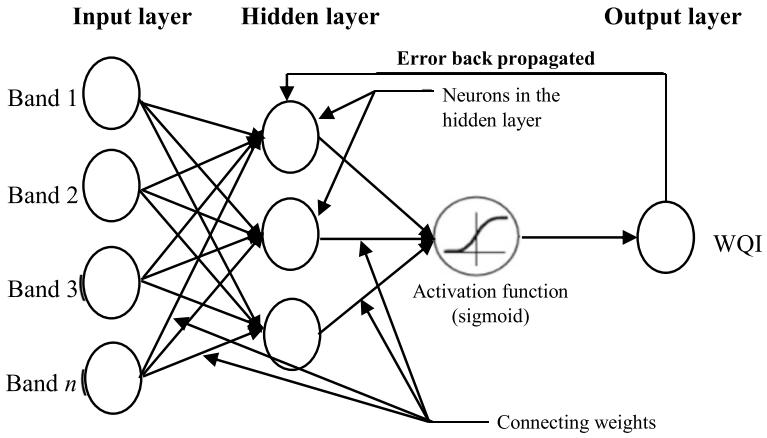



**Fig. 3** Neural network architecture with input variables as bands/band combinations and WQI as target variable

$$I_j^m = \sum O_i^{m-1} w_{ij} + b_j^m, \tag{9}$$

where $b_j^m$ is the threshold value of the $j$th unit of layer $m$. $O_j^{m-1}$ and $w_{ij}$ are the outputs of the $i$th unit of layer $m-1$ and the weight of the connection between $i$th and $j$th units of layers $m-1$ and $m$, respectively. The error function is expressed as

$$E = 0.5 \sum_k \left(T_k - O_k\right)^2, \tag{10}$$

where $T_k$ is the desired target value and $O_k$ is the corresponding output value determined for $k$ training samples.

### 2.4.3 Gene expression programming (GEP)

GEP, proposed by Ferreira (2001), is an evolutionary technique that has the advantage of solving complex nonlinear problems based on the GP approach developed by Koza (1999). GEP is an improved version of GA and GP that overcomes premature convergence and a 100 times higher evolution rate. GEP undergoes a continuous evolution process with the random propagation of an initial population comprising of individual chromosomes of predefined length containing one, or more than one, gene. The structure of genes comprises a head and a tail. The head consists of both functions and terminals, whereas the tail holds only terminals. For reaching an optimal solution to the defined problem, the head length $h$ is selected; further, the tail length $t$ is related to $h$, and the function is evaluated by using Eq. 11 below:

$$t = h \times (n-1) + 1, \tag{11}$$

where $n$ is the number of arguments of the function. Ferreira (2001) represented the encoded genetic information in the gene in the form of an expression tree (ET). With the help of the unequivocal Karva language, the gene composition of a given ET can be generalized on the basis of simple rules of top–down and right–left (Li and Wang, 2019). An

example of a gene is shown in the form of an ET in Fig. 4, for which an equivalent mathematical expression is encoded as $[(b\times a)\times(b+a)]+[(a/b)\times(b-a)]$.

The fitness of every chromosome $i$ in the initial population is computed by utilizing the fitness function $f_i$ expressed as Eq. 12, proposed by Ferreira (2001).

$$f_i = \sum_{j=1}^{C_i} \left( M - \left| C_{ij} - T_j \right| \right) \tag{12}$$

where $M$ is the selection range, $C_{i,j}$ is the value recalled by the $i$th chromosome for the $j$th fitness case, and $T_j$ is the target value for the $j$th fitness case. It is to be noted that, for a perfect fit, $C_{i,j}=T_j$ and $f_i=f_{\max}=C_t\times M$. Fitness function resolves the selection of the optimal chromosomes for the next generation level through modifications achieved by genetic operators such as mutation, inversion, transposition and recombination.

Mutation is the most effective genetic operator that represents the probability of a function or a variable (symbol) to get mutated in each generation. Any symbol in the gene heads can be replaced by a terminal function; however, in the gene tails, terminals can be replaced by variables only, since there is no function in the tail. Inversion chooses a random starting as well as ending symbol in a gene, which is then reversed in order. Transposition involves actuating a sequence of symbols from one position to another within a gene or from one gene to another gene in the same chromosome. In a recombination stage, two new chromosomes are developed by the exchange of genetic information through random selection. The process is analogous to the breeding of two biological species that produces a new offspring sharing genetic material from both parents. Figure 5 illustrates the generalized process of GEP model building in the form of a flowchart.

Table 2 depicts 11 spectral reflectance bands/band combinations (including three inherent single bands, i.e., green, red and infrared) formulated to explore the most significant band combination related to the observed WQI estimates. As described in the preceding sections, WQI estimates as a function of most sensitive spectral band combination were examined via three approaches and the performance were compared using $R^2$, RMSE and MAE (quantitative indicators). Out of 40 data samples in total, 80% were used for training and testing or calibration and the remaining (20%) were used for validation. Neural network architectures were developed in accordance with the band combinations, i.e., 2, 3, 4, 5 and 6 spectral bands/combinations as input variables. The same band combinations were analyzed for MLR and GEP approaches, keeping WQI as target variable. For BPNN and

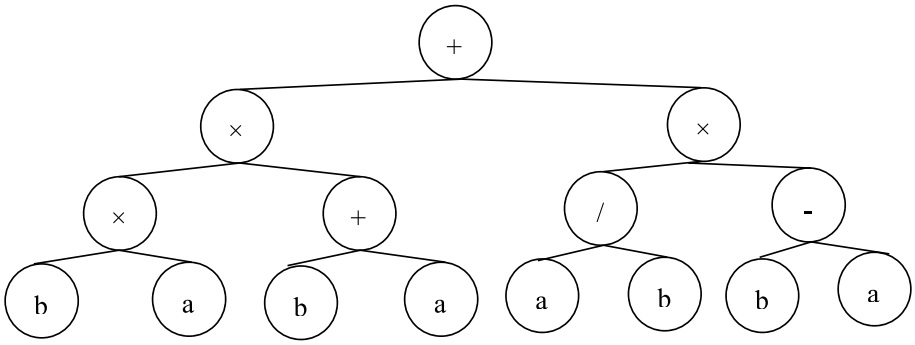



**Fig. 4** An example of gene ET

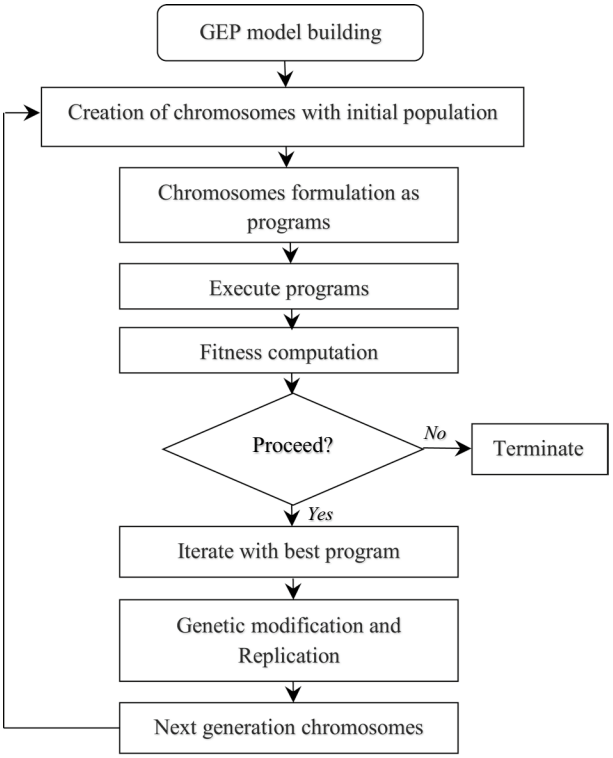



**Fig. 5** Flowchart illustrating the process of GEP model building

**Table 2** Formulated band combinations with details of BPNN architectures

| Band combination cases | Input/independent variables | | Target variable | Network architecture I–H–O[a] | Learning rate |
|---|---|---|---|---|---|
| | No | Description | | | |
| 1 | 2 | G, R | WQI | 2–2-1 | 0.058 |
| 2 | 2 | G, NIR | | 2–2-1 | 0.072 |
| 3 | 2 | R, NIR | | 2–3-1 | 0.069 |
| 4 | 3 | G, R, NIR | | 3–4-1 | 0.087 |
| 5 | 4 | G, R, NIR, G/R | | 4–5-1 | 0.055 |
| 6 | 4 | G, R, NIR, G/NIR | | 4–3-1 | 0.047 |
| 7 | 4 | G, R, NIR, R/NIR | | 4–6-1 | 0.025 |
| 8 | 5 | G, R, NIR, G/R, G/NIR | | 5–4-1 | 0.065 |
| 9 | 5 | G, R, NIR, G/R, R/NIR | | 5–6-1 | 0.046 |
| 10 | 5 | G, R, NIR, G/NIR, R/NIR | | 5–7-1 | 0.095 |
| 11 | 6 | G, R, NIR, G/R, G/NIR, R/NIR | | 6–8-1 | 0.075 |

[a]I–H–O: input–hidden layer neurons–output

GEP analysis, the entire data set was normalized to lie within 0 to 1 range by using Eq. 13 below (Rajurkar et al., 2004), to ensure that data are logically structured and proportionally scaled.

$$X_{\text{norm}} = 0.1 + 0.8 \times \left( \frac{X_i}{X_{\text{max}}} \right), \tag{13}$$

where $X_{\text{norm}}$ is the normalized, unitless variable; $X_i$ is the observed variable; and $X_{\text{max}}$ is the maximum value in the data range. The optimal count of neurons in the hidden layer was ascertained by a hit-and-trial procedure. The learning rate for BPNN was gradually varied within the defined range of 0.01 to 0.5. The final values of the learning rate and the optimum count of neurons in the hidden layer obtained by the trial process are provided in Table 2.

Further, for building the optimal GEP model, the number of chromosomes or population size after many trials was selected as 50, the gene head length was selected as 14, and the number of genes per chromosome was selected as 8. Seven necessary function operators, i.e., $+, -, \times, \div, 1/a, -a, a^2$, were adopted for building the simplified GEP model with a reduced iteration process as well as nonconvergence occurrences. Furthermore, subgene ETs were linked by an addition function. The parameters adopted for the optimal GEP model for precise evaluation of WQI levels are illustrated in Table 3.

## 3 Results

In situ water samples collected were analyzed for 11 WQPs in the laboratory, and the basic descriptive statistics of the samples are summarized in Table 4. The physicochemical and heavy metal concentrations ranged far beyond the permissible limits prescribed under BIS

**Table 3** Parameters adopted for the optimal GEP model

| Parameter | Value |
|---|---|
| Population size | 50 |
| Genes per chromosomes | 8 |
| Gene head length | 14 |
| Maximum generations | 5000 |
| Fitness function | $R^2$ |
| Precision (hit tolerance) | 0.01 |
| Mutation rate | 0.054 |
| Inversion rate | 0.1 |
| Computational functions | $+, -, \times, /, 1/a, a^2, -a$<br>Addition, subtraction, multiplication, division, inverse, square, negation |
| Linking function | addition |
| IS transposition rate | 0.1 |
| RIS transposition rate | 0.1 |
| Gene transposition rate | 0.1 |
| Recombination one-point rate | 0.3 |
| Recombination two-point rate | 0.3 |

**Table 4** Descriptive statistics of the measured WQPs

| S. no | Water quality parameters (mg/l) | Range min–max | Mean | Standard deviation ( σ) | Population variance ($s^2$) | Sample variance ( $\sigma^2$) | SEM[a] | BIS standard (BIS, 1986) |
|---|---|---|---|---|---|---|---|---|
| 1 | pH | 6.93–7.56 | 7.33 | 0.12 | 0.0147 | 0.0141 | 0.0242 | 6.5–8.5 |
| 2 | EC (µs/cm) | 1083–1852 | 1644 | 190.61 | 36,334 | 34,881 | 38.12 | 300 |
| 3 | TDS | 754–851 | 800.32 | 20.06 | 402.56 | 386.53 | 4.013 | 500 |
| 4 | Alkalinity | 540–680 | 616.68 | 40.39 | 1631.56 | 1566.3 | 8.078 | 200 |
| 5 | COD | 72.15–91.20 | 81.55 | 6.10 | 57.38 | 49.32 | 1.515 | 250 |
| 6 | DO | 2.08–6.74 | 5.85 | 1.04 | 1.089 | 0.882 | 0.211 | 5 |
| 7 | BOD | 23.20–38.50 | 29.61 | 4.02 | 16.124 | 15.48 | 0.803 | 5 |
| 8 | Cr | 0 | 0 | 0 | 0 | 0 | 0 | 0.05 |
| 9 | Pb | 0.19–0.24 | 0.21 | 2.21 | $1.4\times10^{-4}$ | $1.1\times10^{-4}$ | 0.002 | 0.1 |
| 10 | Fe | 0.01–0.03 | 0.02 | 0.01 | $8.7\times10^{-5}$ | $8.4\times10^{-5}$ | 0.0018 | 0.3 |
| 11 | Mn | 0 | 0 | 0 | 0 | 0 | 0 | 0.1 |

[a]*SEM* standard error of means

specifications, although there were no traces of Cr and Mn in all the measured samples. The WQI values computed from nine WQPs (excluding Cr and Mn) for 40 water samples collected along the Kali River stretch ranged between 203.7 and 262.33, and rated under "*very poor*" category on the basis of BIS criteria provided in Table 1. The WQI range indicates restricted use of river water almost for all purposes including irrigation. The results of the WQI estimates from the three employed approaches, i.e., MLR, BPNN and GEP, are illustrated in Table 5.

Results from the MLR analysis indicate that, out of 11 band combination cases analyzed, a combination of 4 bands, i.e., G, R, NIR and G/R (band combination case no. 5), exhibited strong correlation with the observed WQI yielding $R^2 \sim 0.81$ and low RMSE and MAE values (i.e., 4.36 and 4.64, respectively) for calibration data. However, the same band combination yielded WQI estimates with $R^2 \sim 0.6$, and relatively high RMSE and MAE values (i.e., 6.3 and 4.64) for validation data. Regression coefficients for the most significant band combination are provided in Table 6, and the formulated regression equation is expressed as Eq. 14. Scatter plot between the observed and estimated WQI for calibration and validation data is illustrated in Fig. 7(a), depicting estimated values of the WQI within $\pm 20\%$ error lines. The regression equation formulated for the most significant band combination was utilized in the generation of spatially distributed WQI map of the river segment.

$$\mathrm{WQI} = -183.98 + (2309.744 \times \mathrm{GREEN}) + (297.18 \times \mathrm{RED}) + (200.93 \times \mathrm{NIR}) + \left(35.84 \times \frac{\mathrm{GREEN}}{\mathrm{RED}}\right) \quad (14)$$

Neural network architectures for all band combinations were trained using the TRAINGD function and FF-BPNN algorithm. Optimal architectures were obtained during the training process by adopting the number of neurons in the hidden layer from 2 to 10 and varying the learning rate in the defined range of 0.001 to 0.5. It was observed that neural network architectures trained with 3, 4 and 6 neurons in the hidden layer yielded much better WQI estimates in terms of $R^2$, RMSE and MAE values (Table 6).

Results further reveal that neural network architecture trained with 3 input bands, i.e., G, R and NIR, and 4 neurons in the hidden layer (i.e., 3-4-1) produced WQI estimates with highest accuracy than the rest of combinations, yielding $R^2 \sim 0.95$ and 0.87, RMSE as 2.36 and 4.48, and MAE as 2.15 and 3.61 for calibration and validation data, respectively. Scatter plot between the observed and estimated WQIs as shown in Fig. 7b depicted WQI estimates within $\pm 10\%$ error lines. It was also observed that almost all neural network architectures with different band combinations conceded WQI estimates with considerable accuracies for calibration data, i.e., $R^2$ ranging from 0.92 to 0.79, respectively. Table 7 illustrates the final weight matrix for the most optimal neural network architecture (i.e., 3-4-1) producing highest WQI retrieval accuracies.

The optimal GEP model was achieved through many trials (Table 6), comprising a chromosomal architecture with 50 chromosomes, head length at 14 and number of genes at 8, and 4 spectral bands as input, viz. G, R, NIR and G/R (band combination case no. 5). The optimized GEP model produced WQI estimates with considerably high accuracies, yielding $R^2 \sim 0.94$ and 0.91, RMSE as 2.49 and 4.45, and MAE as 2.16 and 3.53 for calibration and validation data, respectively. As observed from the results, GEP model performs substantially well with validation data as compared with BPNN and MLR models, thus indicating significant rationality in the optimized GEP model. The optimal GEP model constitutes four subordinate expression trees (i.e., sub-ET1, sub-ET2, sub-ET3 and sub-ET4), developed in accordance with the selection of the number of input

**Table 5** Coefficient of determination ($R^2$), RMSE and MAE between the observed and estimated WQIs from three approaches

| Band combination cases | | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Inputs/variables | | $X^a$ | 2 | 2 | 2 | **3** | **4** | 4 | 4 | 5 | 5 | 5 | 6 |
| Iterations | | $I^b$ | 35,234 | 28,731 | 32,186 | **33,174** | 36,871 | 44,528 | 47,351 | 51,277 | 49,545 | 55,285 | 57,368 |
| MLR | Cal (80%) | $R^2$ | 0.46 | 0.66 | 0.65 | 0.72 | **0.81** | 0.77 | 0.56 | 0.59 | 0.51 | 0.32 | 0.27 |
| | | RMSE | 18.51 | 15.67 | 9.45 | 6.85 | **4.36** | 7.31 | 10.52 | 12.35 | 11.68 | 17.77 | 21.45 |
| | | MAE | 16.77 | 15.21 | 8.63 | 5.37 | **3.00** | 6.82 | 8.86 | 10.73 | 11.24 | 15.52 | 18.96 |
| | Val (20%) | $R^2$ | 0.34 | 0.52 | 0.47 | 0.51 | **0.60** | 0.55 | 0.57 | 0.49 | 0.43 | 0.15 | 0.09 |
| | | RMSE | 16.52 | 8.96 | 13.76 | 9.63 | **6.30** | 7.55 | 7.37 | 11.52 | 13.93 | 28.79 | 36.21 |
| | | MAE | 15.23 | 8.44 | 10.73 | 8.68 | **4.64** | 7.15 | 6.97 | 9.38 | 12.51 | 23.48 | 31.92 |
| BPNN | Cal Tr Te[c] | $R^2$ | 0.84 | 0.88 | 0.85 | **0.95** | 0.92 | 0.89 | 0.92 | 0.87 | 0.79 | 0.87 | 0.93 |
| | | RMSE | 3.78 | 3.64 | 3.32 | **2.36** | 2.58 | 3.38 | 3.76 | 4.81 | 5.24 | 4.95 | 3.17 |
| | | MAE | 3.34 | 3.46 | 2.98 | **2.15** | 2.42 | 3.15 | 3.28 | 4.47 | 4.89 | 4.33 | 2.87 |
| | Val (20%) | $R^2$ | 0.76 | 0.79 | 0.84 | **0.87** | 0.85 | 0.81 | 0.83 | 0.69 | 0.73 | 0.79 | 0.83 |
| | | RMSE | 5.21 | 6.87 | 5.91 | **4.48** | 5.85 | 6.76 | 6.62 | 7.23 | 7.17 | 6.95 | 6.34 |
| | | MAE | 4.85 | 5.54 | 5.73 | **3.61** | 4.97 | 6.16 | 5.65 | 7.12 | 6.78 | 6.23 | 5.92 |
| GEP | Cal (80%) | $R^2$ | 0.78 | 0.84 | 0.91 | 0.88 | **0.94** | 0.92 | 0.89 | 0.83 | 0.91 | 0.79 | 0.82 |
| | | RMSE | 6.89 | 6.34 | 5.21 | 5.96 | **2.49** | 3.16 | 3.67 | 7.78 | 3.41 | 7.39 | 6.57 |
| | | MAE | 5.74 | 5.22 | 4.78 | 5.17 | **2.16** | 2.87 | 2.83 | 6.41 | 2.75 | 5.53 | 5.27 |
| | Val (20%) | $R^2$ | 0.63 | 0.79 | 0.82 | 0.86 | **0.91** | 0.87 | 0.81 | 0.76 | 0.82 | 0.71 | 0.79 |
| | | RMSE | 8.87 | 7.21 | 6.87 | 6.34 | **4.45** | 5.92 | 6.54 | 7.11 | 7.08 | 7.87 | 8.29 |
| | | MAE | 7.34 | 6.88 | 6.12 | 5.43 | **3.53** | 4.58 | 5.71 | 6.51 | 5.24 | 6.19 | 7.33 |

[a] $X$ no. of input/independent variables, [b] $I$ total no. of iterations for BPNN, [c] *Tr* training data (60%), *Te* testing data (20%), *Cal* calibration, *Val* validation

'Bold' indicates the optimal WQI model achieved

**Table 6** MLR coefficients for calibration data for the most appropriate band combination

Multiple regression equation
$Y = b_o + b_1 \times X_1 + b_2 \times X_2 + b_3 \times X_3 + \cdots + e$
where $Y$ = WQI

| BC[a] case Nn | $X$[b] | Parameter distribution | Regression coefficients | $R^2$ (Cal[c]) | RMSE | MAE | $R^2$ (Val[d]) | RMSE | MAE |
|---|---|---|---|---|---|---|---|---|---|
| 5 | 04: G, R, NIR, G/R | $X_1$ = G<br>$X_2$ = R<br>$X_3$ = NIR<br>$X_4$ = G/R | $b_0 = -183.98$<br>$b_1 = 2309.74$<br>$b_2 = 297.18$<br>$b_3 = 200.93$<br>$b_4 = 35.84$ | 0.81 | 4.36 | 3.0 | 0.60 | 6.3 | 4.64 |

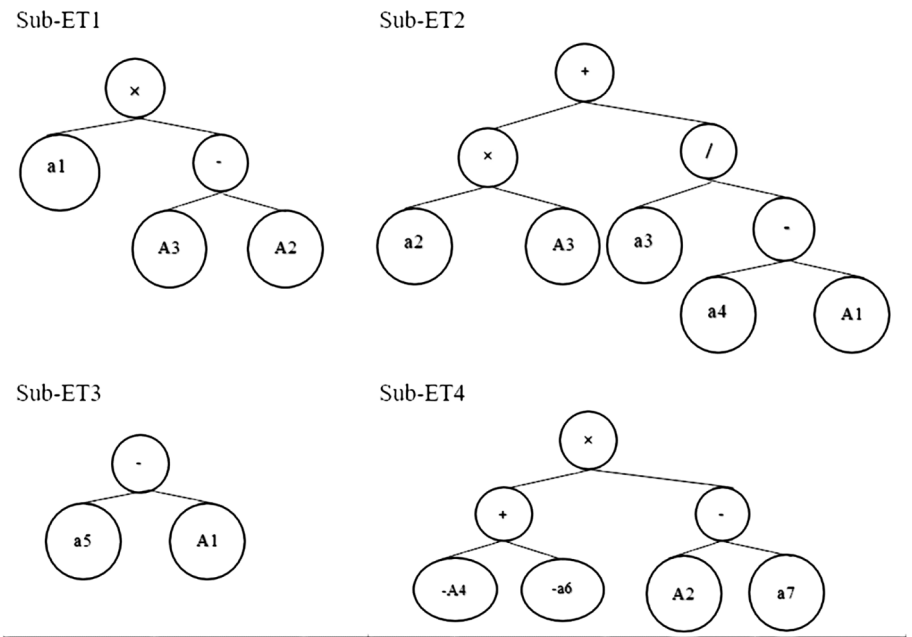
[a]*BC* band combination,[b]*X* no. of input/independent variables, [c]*Cal* calibration, [d]*Val* validation

**Table 7** Final weight matrix of the trained BPNN model with 3 input variables

| Predictor variables | BPNN prediction model (hidden layer) | | | |
|---|---|---|---|---|
| | Connecting weights of 4 neurons | | | |
| Input layer | N1 | N2 | N3 | N4 |
| G | −0.234 | 0.046 | 0.824 | 0.147 |
| R | −0.463 | −0.042 | 0.568 | −0.221 |
| NIR | 0.173 | −0.386 | −0.034 | −0.579 |
| Bias | 0.251 | −0.366 | −0.721 | 0.632 |
| Target variable (WQI) output layer | 0.437 | −0.254 | −0.022 | 0.771 |

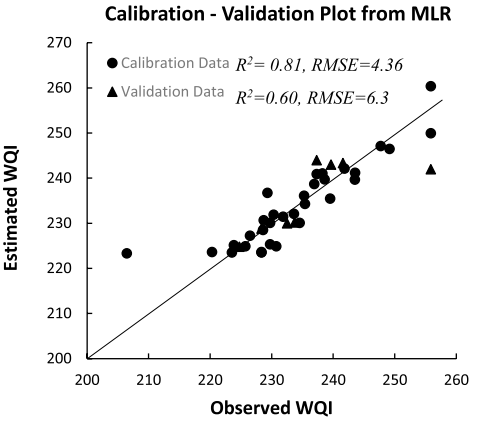
*N1* neuron 1, *N2* neuron 2, *N3* neuron 3, *N4* neuron 4

variables and function operators during model-building process. Sub-ETs were linked together by an addition function to finally form the mathematical expression that was further simplified to obtain more generalized form for estimating the WQI, expressed as Eq. 15. The developed subgene ETs are shown as in Fig. 6, and a scatter plot between the observed and estimated WQI is shown in Fig. 7c, depicting estimated WQI values within ± 10% error lines.

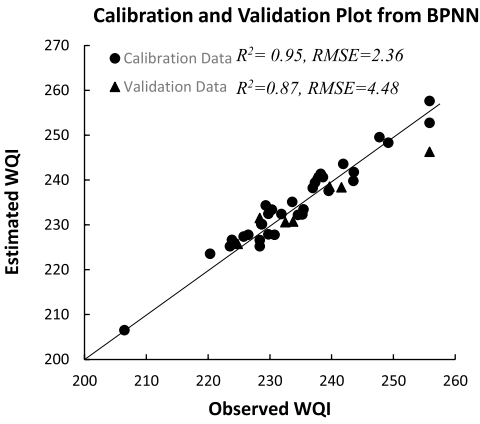



| | | | | | |
|---|---|---|---|---|---|
| a1 = 479.153 | a2 = -52.948 | a3 = 0.00186 | a4 = 0.13093 | a5 = -256.825 | a6 = 1.755 |
| a7 =130.50571 | A1= Green | A2 = Red | A3 = NIR | A4 = G/R | - |
| WQI= (a1×(A3-A2))+(((a2) ×A3)+(a3/(a4-A1)))+(a5-A1)+((-(A4+a6)) × (A2-a7))<br>WQI= (sub-ET1)+(sub-ET2)+(sub-ET3)+(sub-ET4) | | | | | |

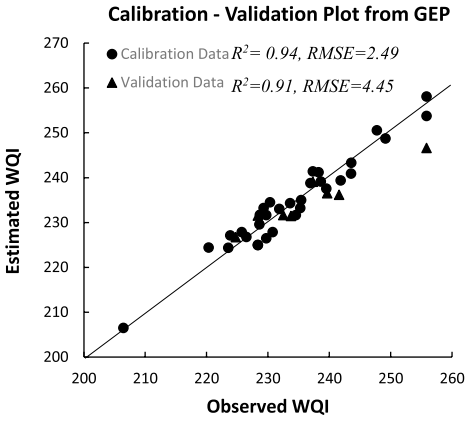
**Fig. 6** Expression trees for the optimal GEP model with 4 spectral bands




(a)




(b)




(c)

**Fig. 7** Scatter plots between observed and estimated WQIs from **a** MLR approach for band combination 5; 4 inputs, **b** BPNN approach for band combination 4; 3 inputs and **c** GEP approach for band combination 5; 4 inputs

$$WQI = \begin{cases} \left(\dfrac{1.86 \times 10^{-3}}{1.31 \times 10^{-1} - GREEN}\right) & \text{Sub ET - 1} \\ \left[RED\left(477.4 - \dfrac{GREEN}{RED}\right)\right] + & \text{Sub ET - 2} \\ (426.2NIR - GREEN) + & \text{Sub ET - 3} \\ \left[\left(130.5\dfrac{GREEN}{RED}\right) - 27.78\right] & \text{Sub ET - 4} \end{cases} \quad (15)$$

## 4 Discussion

The severe contamination of River Kali stretch assessed through the laboratory analysis of 11 physicochemical parameters and heavy metals as well as WQI estimates is mainly attributed to the unrestricted toxic waste disposal from numerous small- and large-scale industries. Although several studies on Kali River water quality have predicted the similar outcomes (CPCB, 2012; Mishra et al., 2015; Singh et al., 2020; Sirohi et al., 2014), a comprehensive monitoring of WQI levels by formulating spectral band combinations has been lacking. Results from the three approaches further reveal that GEP outperforms the other two approaches in terms of WQI estimates for validation data (i.e., $R^2 \sim 0.91$, 0.87 and 0.60; RMSE ~ 4.45, 4.48 and 6.30 for GEP, ANN and MLR, respectively), suggesting a higher measure of explanatory power possessed by this approach. Moreover, the GEP approach is simple and produces reliable WQI measures and reduces substantial time and effort by optimizing the computations to generate simplified prediction expressions. This technique is highly recommended by many researchers (Hashmi et al., 2011; Mohammadpour et al., 2016; Liu & Wang, 2019) for the water quality evaluation of wetlands and other surface water bodies. In addition, the ANN approach is relatively time-consuming and does not furnish any governing equations of the optimized models, which is considered as one of its major disadvantages. The WQI estimates predicted by MLR model were of insufficient accuracy when tested with validation data, since this approach utilizes the method of least squares and is linear in nature. However, MLR is still practicable for its fast predicting ability. Figure 8 depicts comparative line plots of WQI estimates for calibration and validation data, along with the observed WQI measures.

The contamination levels throughout the Kali River stretch exhibited consistency which lead to similar spectral distribution of remotely-sensed signal above the water surface. Therefore, WQI maps created in the GIS framework (Fig. 9) from the three approaches corroborate to the actual severity in WQI levels, exhibited by the darker spectral tones covering the entire length of the river segment. This severe contamination in the river is majorly attributed to the addition of industrial effluents, agricultural runoff, natural matter and nutrients in the water body (Jindal & Sharma, 2011).

A one-way ANOVA test for means and variance was applied to further ascertain the spatial variability of WQI estimates from the three approaches. The null hypothesis "$H_0$" stated "no significant difference between means of WQI estimates from the three approaches," whereas alternate hypothesis "$H_a$" stated "significant difference between means of WQI estimates from three approaches." The test results unveiled $F$-statistic (i.e., $F = 0.01$ and $p$-value, i.e., $p = 0.994$) as exceedingly higher than the significance level $\alpha = 0.05$ (Table 8), implying that there were no critical differences in the mean values and variances of WQI estimates. The

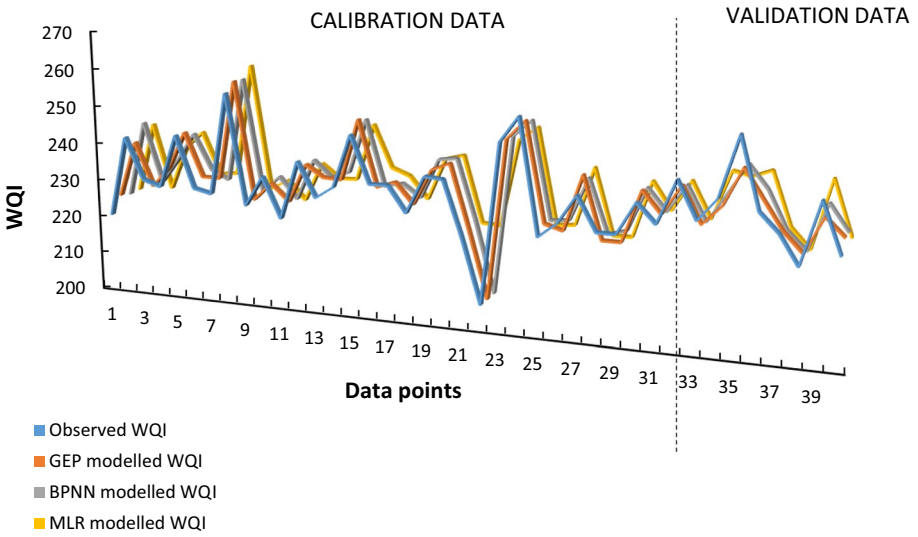



**Fig. 8** Comparative line plot of observed WQI and estimated WQI from the three employed approaches

ANOVA test results therefore fail to reject the null hypothesis inferring that the WQI estimates from the three approaches are statistically "*not significant*." The data set may, however, be consistent with the differences of practical importance. Moreover, failing to reject the null hypothesis does not necessarily imply that no potential difference in the data set exists, rather; an increased sample size could bring out the difference. Thus, larger sample sizes allow hypothesis tests to detect effects that are statistically significant. Further, to visually summarize and compare the results, box plot of WQI estimates shown in Fig. 10, were analyzed. İt was observed that, the respective medians of each box plot laid at the same level (i.e., 233.23 for GEP, 232.94 for ANN and 231.96 for MLR) suggesting no likely difference between the three estimated WQI groups. The median line of the three box plots further indicates symmetric data representation with no right or left skewness within each of the three WQI groups. Upon comparing the interquartile ranges, the relatively longer box corresponding to MLR revealed slight dispersion in WQI estimates.

Overall comparison of the results indicate that GEP is much superior to MLR and ANN approaches. Furthermore, despite the restrictive spectral resolution of IRS P6 LISS IV sensor (i.e., comprising three bands), a combination of 4 bands (i.e., G, R, NIR, G/R) is identified as the most effective for modeling WQI levels through GEP approach. The methodology adopted and the WQI maps generated can be of immense help in the decision making to impose corrective conservation measures for improvement in the Kali River water quality so that the river may regain its historical importance. Moreover, the methodology can be implemented to other contaminated surface water bodies to generalize the GEP model prediction ability.

## 5 Conclusions

The present study evaluates WQI levels along 6-km-long Kali River segment from three approaches, namely MLR, BPNN and GEP, by utilizing spectral reflectance values from high-resolution IRS P6 LISS IV image. The water samples were collected from 40 random

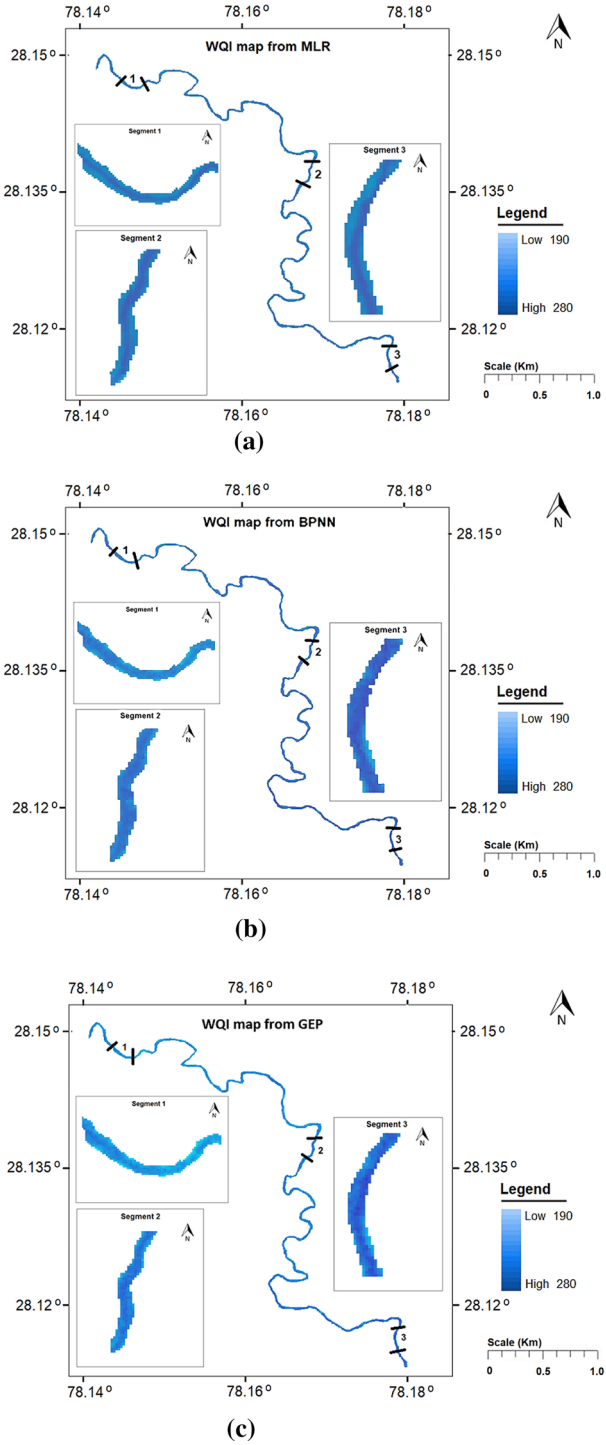



**Fig. 9** WQI maps of the river stretch generated from **a** MLR, **b** BPNN and **c** GEP analysis

**Table 8** One-way ANOVA test for WQI estimates from the three employed approaches

| | Square summation | Degree of freedom | Mean of square | *F* | *P*-value | *F* crit |
|---|---|---|---|---|---|---|
| Between groups | 0.9329 | 2 | 0.4665 | 0.01 | 0.994 | 3.08 |
| Within groups | 9830.53 | 117 | 84.021 | | | |
| Total | 9831.47 | 119 | | | | |

As the *P*-value is more than 0.05, it is not significant at 2.5% level

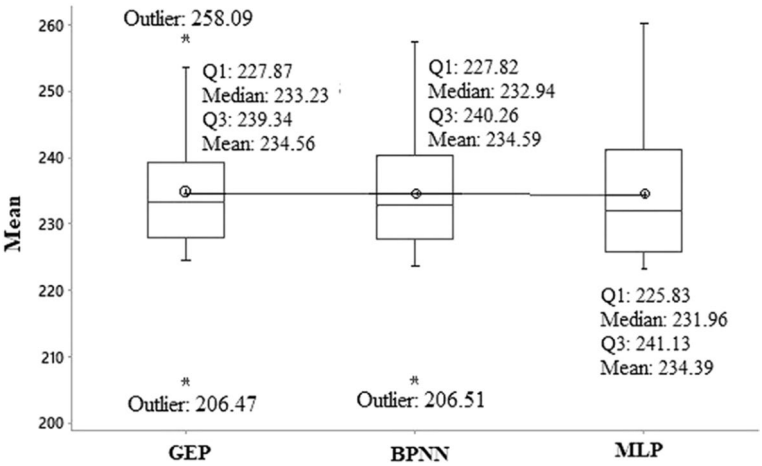



**Fig. 10** Boxplot of WQI estimates from the three employed approaches

locations along the river stretch and analyzed for seven physicochemical and four heavy metal concentrations (i.e., 11 WQPs in total). All measured WQP concentrations ranged beyond the permissible limits as per BIS specifications, except for Cr and Mn, that were found to be absent in the water samples. Further, the WQI values computed from nine WQPs were found to range between 203.7 and 262.33, thus, designating the river condition as unfit for all purposes. Eleven spectral reflectance band combinations (including three inherent single bands) were considered to explore the sensitivity of the most significant band/band combination with the observed WQI. The analyses of the results revealed that GEP approach outperformed both BPNN and MLR approaches with considerably high WQI retrieval accuracies, yielding $R^2 \sim 0.94$ and 0.91, RMSE as 2.49 and 4.45 and MAE as 2.16 and 3.53 for calibration and validation data, respectively. Results further revealed that both GEP and MLR approaches identified the combination of 4 spectral bands (i.e., G, R, NIR, G/R) as the most significant band combination for estimating WQI levels, whereas BPNN recognized 3 band combination (i.e., G, R, NIR) as the most significant. The results are also suggestive of the fact that machine learning approaches, viz. ANN and GEP, yield promising potential for water quality monitoring by utilizing spectral band combinations, wherein GEP proved to be superior. The ANOVA test revealed statistically insignificant difference among WQI estimates from the three approaches at a confidence level of 0.05, attributed to small river water sample size. The spatial distribution maps of WQI levels exhibited uniform spectral tones in the entire river stretch, signifying the severity

of pollution concentrations in the river water. The study showcases the river condition as extremely critical, requiring immediate attention of the decision makers involved in the task of its reclamation. Future research can be focused on using hyperspectral satellite data along with integrated approaches such as fuzzy optimal model, GP, support vector machine (SVM) and RBF along with an increased water sample size.

**Acknowledgements** Authors acknowledge anonymous reviewers for their constructive comments and suggestions that have substantially improved the quality of the manuscript.

## References

Abba, S. I., Said, Y. S., & Bashir, A. (2015). Assessment of water quality changes at two location of Yamuna River using the National Sanitation Foundation of water quality. *Journal of Civil Engineering and Environmental Technology, 2*(8), 730–733

Akbal, F., Gürel, L., Bahadır, T., Güler, İ, Bakan, G., & Büyükgüngör, H. (2011). Multivariate statistical techniques for the assessment of surface water quality at the mid-Black Sea coast of Turkey. *Water Air Soil Pollution, 216*, 21–37

APHA. (1998). *Standard methods for the examination of water and waste water*. (20th ed., p. 1998). American Public Health Association.

Aras, E., Togan, V., & Berkun, M. (2007). River water quality management model using genetic algorithm. *Environmental Fluid Mechanics, 7*, 439–450

ASTM. (2000). *American society for testing and materials*. (p. 20402). Published by United States Environmental Protection Agency.

Avdan, Z. Y., Kaplan, G., Goncu, S., & Avdan, U. (2019). Monitoring the water quality of small water bodies using high-resolution remote sensing data. *International Journal of Geo-Information (MDPI), 8*, 553

Azamathulla, H. M., & Ghani, A. A. (2011). Genetic programming for predicting longitudinal dispersion coefficients in streams. *Water Resources Management, 25*, 1537–1544

BIS. (1986). Indian standard specification for irrigation water. IS: 11624. *Indian Standard Institute,* India.

Bonansea, M., María, C. R., Lucio, P., & Susana, F. (2015). Using multi-temporal landsat imagery and linear mixed models for assessing water quality parameters in Río Tercero Reservoir (Argentina). *Remote Sensing of Environment, 158*, 28–41

Bordalo, A. A., Teixeira, R., & Wiebe, W. J. (2006). A water quality index applied to an international shared river basin: The case of the Douro River. *Environmental Management, 38*, 910–920

Boyacioglu, H. (2010). Utilization of the water quality index methods: A classification tool. *Environmental Monitoring and Assessment, 167*, 115–124

Brown, R. M., McClelland, N. I., Deininder, R. A., & Tozer, R. G. (1970). A water quality index- do we dare? *Water Sewage Works, 117*(10), 339–343

Chander, G., & Markham, B. (2003). Revised Landsat-5 TM radiometric calibration procedures and post calibration dynamic ranges. *IEEE Transactions on Geoscience and Remote Sensing, 41*, 2674–2677

Chapagain, S. K., Pandey, V. P., Shrestha, S., Nakamura, T., & Kazama, F. (2010). Assessment of deep groundwater quality in Kathmandu valley using multivariate statistical techniques. *Water Air Soil Pollution, 210*, 277–288

Chen, L., Tan, C. H., Kao, S. J., & Wang, T. S. (2008). Improvement of remote monitoring on water quality in a subtropical reservoir by incorporating grammatical evolution with parallel genetic algorithms into satellite imagery. *Water Resources, 42*, 296–306

Chu, H. B., Lu, W. X., & Zhang, L. (2013). Application of artificial neural network in environmental water quality assessment. *Journal of Agriculture Science and Technology, 15*(2), 343–356

Clifford, C. (2020). The Water in Venice, Italy's Canals Is Running Clear amid the COVID-19Lockdown—Take a Look. Retrieved 17 April 2020 from https://www.cnbc.com/2020/03/18/photos-water-in-venice-italys-canals-clear-amid-covid-19lockdown.html.

CPCB. (2012). *Reconnaissance survey of pollution load of River Kali*. Central Pollution Control Board.

Dunca, A. M. (2018). Water pollution and water quality assessment of major transboundary rivers from Banat (Romania). *Journal of Chemistry* (Article ID 9073763).

Ekercin, S. (2007). Water quality retrievals from high resolution IKONOS multispectral imagery: A case study in Istanbul, Turkey. *Water Air Soil Pollution, 183*, 239–251

Farhad Yousefabadi, L.O., Shariati, F., & Mardookhpour, A. (2013). A Comparison of water quality indices for Haraz River. *Department of Environmental Engineering Lahijan Branch, Islamic*, *3*(3), 30–36.

Ferreira, C. (2001). Gene expression programming: a new adaptive algorithm for solving problems. *Complex Systems, 13*(2), 87–129

Fulazzaky, M. A., Seong, T. W., & Masirin, M. I. M. (2010). Assessment of water quality status for the Selangor River in Malaysia. *Water Air Soil Pollution, 205*, 63–77

Ghavidel, Z. Z. S., & Montaseri, M. (2014). Application of different data-driven methods for the prediction of total dissolved solids in the Zarinehroud basin. *Stochastic Environmental Research and Risk Assessment, 28*, 2101–2118

Gürsoy, Ö., & Atun, R. (2019). Investigating surface water pollution by integrated remotely sensed and field spectral measurement data: A case study. *Polish Journal of Environmental Studies, 28*, 2139–2144

Häder, D. P., Banaszak, A. T., Villafañe, V. E., Narvarte, M. A., González, R. A., & Helbling, E. W. (2020). Anthropogenic pollution of aquatic ecosystems: Emerging problems with global implications. *Science of Total Environment, 713*, 136586

Hafeez, S., Wong, M. S., Ho, H. C., Nazeer, M., Nichol, J., et al. (2018). Comparison of machine learning algorithms for retrieval of water quality indicators in case-II waters: A case study of Hong Kong. *Remote Sensing, 11*(684), 1–26

Hashmi, M. Z., Shamseldin, A. Y., & Melville, B. W. (2011). Statistical downscaling of watershed precipitation using gene expression programming (GEP). *Environmental Modelling and Software, 26*, 1639–1646

Hoseinzadeh, E., Khorsandi, H., Wei, C., & Alipour, M. (2014). Evaluation of aydughmush river water quality using the national sanitation foundation water quality index (NSFWQI), river pollution index (RPI), and forestry water quality index (FWQI). *Desalination and Water Treatment, 54*, 2994–3002

Hussain, M., Ahmed, S. M., & Abderrahman, W. (2008). Cluster analysis and quality assessment of logged water at an irrigation project, eastern Saudi Arabia. *Journal of Environment Mangement, 86*(1), 297–307

Jindal, R., & Sharma, C. (2011). Studies on water quality of Sutlej River around Ludhiana with reference to physicochemical parameters. *Environmental Monitoring and Assessment, 174*(1–4), 417–425

Koza, J. R. (1999). *Genetic programming: On the programming of computers by means of natural selection*. The MIT Press.

Kuo, J., Hsieh, M., Lung, W., & She, N. (2007). Using artificial neural network for reservoir eutrophication prediction. *Ecological Modelling, 200*, 171–177

Li, H., Liu, C. G., Fan, J., et al. (2009). Application of back-propagation neural network for predicting chlorophyll-A concentration in rivers. *China Water and Waste Water, 25*(5), 75–79

Liu, L. W., & Wang, Y. M. (2019). Modelling reservoir turbidity using Landsat 8 satellite imagery by gene expression programming. *Water (MDPI), 11*, 1479. https://doi.org/10.3390/w11071479

Lotfinasabasl, S., Gunale, V. R., & Khosroshahı, M. (2018). Applying geographic information systems and remote sensing for water quality assessment of mangrove forest. *ActaEcologicaSinica, 38*, 135

Markogianni, V., Dimitriou, E., & Karaouzas, I. (2014). Water quality monitoring and assessment of an urban Mediterranean lake facilitated by remote sensing applications. *Environmental Monitoring and Assessment, 186*(8), 5009–5026

Marquez, L. C. G., Bejarano, F. M. T., Espınoza, A. C. T., & Rodríguez, I. R. H. (2018). Use of LANDSAT 8 images for depth and water quality assessment of El Guájaro reservoir, Colombia. *Journal of South American Earth Sciences, 82*, 231

Martí, P., Shiri, J., Duran-Ros, M., Arbat, G., De Cartagena, F. R., & Puig-Bargués, J. (2013). Artificial neural networks vs. gene expression programming for estimating outlet dissolved oxygen in micro-irrigation sand filters fed with effluents. *Computers and Electronics in Agriculture, 99*, 176–185

Maurya, P. K., & Malik, D. S. (2016). Distribution of heavy metals in water, sediments and fish tissue (Heteropneustisfossilis) in Kali River of western UP India. *International Journal of Fisheries and Aquatic Studies, 4*(2), 208–215

MilanovićPešić, A., Brankov, J., & MilijaševićJoksimović, D. (2020). Water quality assessment and populations' perceptions in the National park Djerdap (Serbia): Key factors affecting the environment. *Environment Development and Sustainability, 22*, 2365–2383. https://doi.org/10.1007/s10668-018-0295-8

Mishra, S., Kumar, A., Yadav, S., & Singhal, M. K. (2015). Assessment of heavy metal contamination in Kali river, Uttar Pradesh, India. *Journal of Applied and Natural Science, 7*(2), 1016–1020

Mohamed, E., Ioannis, G., Anas, O., Jarbou, B., & Petros, G. (2019). Assessment of water quality parameters using temporal remote sensing spectral reflectance in arid environments Saudi Arabia. *Water, 11*(3), 556

Mohammadpour, R., Shaharuddin, S., Chang, C., Zakaria, N., Ghani, A. A., & Chan, N. (2015). Prediction of water quality index in constructed wetlands using support vector machine. *Environmental Science and Pollution Research, 22*, 6208–6219

Mohammadpour, R., Shaharuddin, S., Zakaria, N.A., Ghani, A. A., Vakili, M., & Chan, N. W. (2016). Prediction of water quality index in free surface constructed wetlands. *Environmental Earth Sciences, 75*, 139. https://doi.org/10.1007/s12665-015-4905-6.

Mondal, M., & Satpati, L. (2020). Human intervention on river system: a control system—A case study in Ichamati River, India. *Environment Development and Sustainability, 22*, 5245–5271. https://doi.org/10.1007/s10668-019-00423-3

Nabizadeh, R., Amin, M. V., Alimohammadi, M., Naddafi, K., Mahvi, A. H., & Yousefzadeh, S. (2013). Development of innovative computer software to facilitate the setup and computation of water quality index. *Journal of Environmental Health Science and Engineering, 11*, 1

Nas, B., Ekercin, S., Karabork, H., Berktay, A., & Mulla, D. J. (2010). An application of Landsat-5TM image data for water quality mapping in lake Beysehir, Turkey. *Water Air Soil Pollution, 212*, 183–197

Nasri, M. (2010). Application of Artificial Neural Networks (ANNs) in prediction models in risk management. *World Applied Science Journal, 10*(12), 1493–1500

Nathan, N. S., Saravanane, R., & Sundararajan, T. (2017). Application of ANN and MLR models on groundwater quality using CWQI at Lawspet, Puducherry in India. *Journal of Geoscience and Environment Protection, 5*, 99–124. https://doi.org/10.4236/gep.2017.53008

Ni, Q., Wang, L., Zheng, B., & Sivakumar, M. (2012). Evolutionary algorithm for water storage forecasting response to climate change with small data sets: The Wolonghu Wetland, China. *Environmental Engineering and Science, 29*, 814–820

Prabu, P. C., Wondimu, L., & Tesso, M. (2011). Assessment of water quality of Huluka and Alaltu Rivers of Ambo, Ethiopia. *Journal of Agricultural Science and Technology, 13*(1), 131–138

Rajurkar, M. P., Kothyari, U. C., & Chaube, U. C. (2004). Modelling of Daily Rainfall runoff relationship with artificial neural network. Journal of Hydrology, 285, 96–113.

Ramakrıshnaıah, C. R., Sadashıvaıah, C., & Ranganna, G. (2009). Assessment of water quality index for the groundwater in Tumkur Taluk, Karnataka State, India. *E-Journal of Chemistry, 6*(2), 523–530

Reynolds, C. S., & Maberly, S. C. (2002). A simple method for approximating the supportive capacities and metabolic constraints in lakes and reservoirs. *Freshwater Biology, 47*(6), 1183–1188

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back propagating errors. *Nature, 323*, 533–536

Said, S., & Hussain, A. (2019). Pollution mapping of Yamuna river segment passing through Delhi using high resolution GeoEye-2 imagery. *Applied Water Science, 9*, 46. https://doi.org/10.1007/s13201-019-0923-y

Said, S., Kothyari, U. C., & Arora, M. K. (2008). ANN-based soil moisture retrieval over bare and vegetated areas using ERS-2 SAR data. *Journal of Hydrologic Engineering, 13*(6), 461–475

Sargaonkar, A., & Deshpande, V. (2003). Development of an overall index of pollution for surface water based on a general classification scheme in Indian context. *Environmental Monitoring and Assessment, 89*, 43–67

Sharaf Essam, E. D., Zhang, Y., & Suliman, A. (2017). Mapping concentrations of surface water quality parameters using a novel remote sensing and artificial intelligence framework. *International Journal of Remote Sensing, 38*(4), 1023–1042

Sharma, G., Said, S., & Hussain, A. (2018). Water quality mapping of Yamuna River stretch passing through Delhi sate using high resolution GeoEye-2 imagery. *International Journal of Applied Geospatial Research, 9*(4), 23–35

Singh, G., Patel, N., Jindal, T., Srivastava, P., & Bhowmik, A. (2020). Assessment of spatial and temporal variations in water quality by the application of multivariate statistical methods in the Kali River, Uttar Pradesh, India. *Environmental Monitoring and Assessment, 192*, 394

Sirohi, S., Sirohi, S. P. S., & Tyagi, P. K. (2014). Impact of industrial effluents on water quality of Kali River in different locations of Meerut, India. *Journal of Engineering Technology and Research, 6*, 4347

Song, K. S., Li, L., Li, S., Tedesco, L., Hall, B., & Li, L. H. (2012). Hyperspectral remote sensing of total phosphorus (TP) in three central Indiana water supply reservoirs. *Water Air Soil and Pollution, 223*, 1481–1502

Srinivasamoorthy, K., Chidambaram, M., Prasanna, M. V., Vasanthavigar, M., John Peter, A., & Anandhan, P. (2008). Identification of major sources controlling Groundwater Chemistry from a hard rock terrain—A case study from Mettur taluk, Salem district, Tamilnadu. *India. Journal of Earth System Sciences, 117*(1), 49–58

Stone, M. (2020). Carbon emissions are falling sharply due to coronavirus. But not for long. Retrieved 17 April 2020 from https://www.nationalgeographic.com/science/2020/04/co-ronavirus-causing-carbon-emissions-to-fall-but-not-for-long/.

Suneetha, M., SyamaSundar, B., & Ravindhranath, K. (2015). Calculation of water quality index (WQI) to assess the suitability of groundwater quality for drinking purposes in Vinukonda Mandal, Guntur District, Andhra Pradesh, India. *Journal of Chemical and Pharmaceutical Research, 7*(9), 538–545

Syahreza, S., MatJafri, M. Z., & Lim, H. S. (2012). Water quality assessment in Kelantan delta using remote sensing technique. *Proceedings of SPIE 8542, Electro-Optical Remote Sensing, Photonic Technologies and Applications VI*, 85420X.https://doi.org/https://doi.org/10.1117/12.978931

Wang, L., Li, X., & Cui, W. (2012). Fuzzy neural networks enhanced evaluation of wetland surface water quality. *International Journal of Computation and Applied Technology, 44*, 235–240

Wang, Z., Su, Y., Jin, S., Shen, W., Ren, J., Zhang, X., & Clark, J. H. (2020). A novel unambiguous strategy of molecular feature extraction in machine learning assisted predictive models for environmental properties. *Green Chemistry, 22*, 3867–3876

Wang, Z., Su, Y., Shen, W., Jin, S., Clark, J. H., Ren, J., & Zhang, X. (2019). Predictive deep learning models for environmental properties: the direct calculation of octanol–water partition coefficients from molecular graphs. *Green Chemistry, 21*, 4555–4565

Water Programme. (2007). Global drinking water quality index development and sensitivity analysis. In *Report of United Nations Environment Programme and Global Environment Monitoring System*. (GEMS)/Water Programme.

Waxter, M. T. (2014). *Analysis of Landsat satellite data to monitor water quality parameters in Tenmile Lake, Oregon*. MSc. Thesis, Portland State University.

Xu, T. Y., & Qin, X. S. (2013). Solving water quality management problem through combined genetic algorithm and fuzzy simulation. *Journal of Environmental Information, 22*, 39–48

Yao, H., Ni, T., & Zhang, T. (2020). Estimation of phosphorus flux into the sea through one reversing river using continuous turbidities and water quality modeling. *Environment Development and Sustainability, 22*, 4251–4265. https://doi.org/10.1007/s10668-019-00382-9

Zakaria, N. A., Azamathulla, H. M., Chang, C. K., & Ghani, A. A. (2010). Gene expression programming for total bed material load estimation—A case study. *Science of the Total Environment, 408*, 5078–5085

Zhang, Y., Pulliainen, J. T., Koponen, S. S., & Hallikainen, M. T. (2003). Water quality retrievals from combined Landsat TM Data and ERS-2 SAR data in the Gulf of Finland. *IEEE Transactions on Geoscience and Remote Sensing, 41*, 622–629

Zhao, F., Zhu, F. Q., & Feng, Z. K. (2014). Study on water body information extraction method based on ZY-3 Imagery. *Bulletin of Surveying and Mapping, 3*, 007

Zhu, L. (2013). *Water quality analysis and evaluation of current situation in the riparian of west lake Taihu in Yixing*. Nanjing Forestry University.