Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2020.26108

200 (2020) 42–54
October

Assessment of water quality index in unmonitored river basin using multilayer perceptron neural networks and principal component analysis

Bachir Sakaa[a,b,*], Nabil Brahmia[c], Hicham Chaffai[b], Azzedine Hani[b]

[a]*Centre de Recherche Scientifique et Technique sur les Régions Arides CRSTRA, BP 1682 RP, 07000 Biskra, Algérie, email: sakaabachir@yahoo.fr (B. Sakaa)*
[b]*Laboratoire Ressource en Eau et Développement Durable, Faculté des Sciences de la Terre, Université Badji Mokhtar, BP 12, 23000 Annaba, Algérie, emails: hichamchaffai@yahoo.fr (H. Chaffai), haniazzedine@yahoo.fr (A. Hani)*
[c]*Laboratoire des Réservoirs Souterrains Pétroliers, Gaziers et Aquifères, Ouargla, Algeria, email: nabilbra@yahoo.fr (N. Brahmia)*

Received 26 November 2019; Accepted 12 May 2020

ABSTRACT

This study describes the combination of neural networks and multivariate methods to develop a proper model for the forecasting of water quality index (WQI) in the Saf-Saf River using water quality parameters. The main objectives of this work were to determine the importance of different input variables and to assess the spatial and temporal water quality variation. MLP models were trained using three different algorithms and tested, these models were compared in terms of efficiency criteria and goodness-of-fit for WQI modeling. The results show that MLP_{BFGS} model provide the best performance with small root mean square error value (RMSE = 0.007) and high coefficient of determination value (R^2 = 0.811) compared with the other types of MLP models. In the meantime, sensitivity analysis reveals that BOD_5 acts as the most contributor decreasing WQI. PCA/FA results show relatively spatial and seasonal changes in surface water quality, it generated three groups of sampling sites with similar characteristics. Group I (upstream sites), group II (midstream sites), and group III (downstream sites) correspond to a relatively low pollution, moderate pollution, and high pollution sites, respectively. Therefore, this approach can provide managers with the right tools to make decisions about the implementation of sustainable management practices.

Keywords: Water quality index; Multilayer perceptron; Principal component analysis; Factor analysis; Saf-Saf river basin

1. Introduction

Issues and challenges related to water quality deterioration have generated much debate and discussion on heightening awareness on water quality concerns, and the increasing demand to sustainably manage our water resources. In the last decades, a rapid industrial development without controlling discharges, the intensive use of fertilizers in agriculture and the over exploitation of water resources destroys river ecosystems and affect human health in different ways. This event produces a chemical modification of the water rending it unusable for other purposes and hence aggravates scarcity of water resources [1].

Surface water resources is unfortunately exposed more and more to pollution, in the form of discharges of industrial or domestic effluents, and are gradually becoming unfit for any use, without prior treatment. The latter is often complicated and expensive [2,3]. Guaranteeing a good water supply is not enough anymore, it must also be avoided that after-use water, known as wastewater, contaminates groundwater, rivers, and lakes, thus rendering them unfit for consumption and industrial use. It is, therefore, becoming

* Corresponding author.

increasingly necessary to contribute to a dual program of conserving and protecting water. For that, a better knowledge of the analytical level of the pollution of the rivers is essential [4].

Prevention of river pollution requires effective monitoring of physicochemical and biological parameters [5]. The water quality index (WQI) is a means of summarizing large amounts of water quality data into simple terms (e.g., good, fair, and poor) for reporting to policymakers and the public in a comprehensive, consistent manner [6]. The WQI is used to state the pollution status of hydro-systems, because it represents a single numeric score that describes the water quality condition at a particular location in a specific time [7,8]. There are different approaches of WQI were used by various countries and institutions around the world to assess the water quality status of their rivers like Argentina [9]; USA [10]; India [11]; Portugal [12]; Turkey [13]; and China [14,15]. In addition, water quality information becomes more easily and quickly interpretable than a list of numeric values.

Recently, several actors intervening at different levels in water resources management of the Saf-Saf river basin, (water quality specialists, other managers, legislators, or the general public), need to analyze and process this information in order to effectively fulfill their role. It becomes so imperative to simplify this perception of water quality so that the extension of water quality can serve a technical, social, and/or even political purpose [16,17]. The objectives of this study, therefore, are two-fold; the first objective is to develop WQI for assessing surface-water quality and defining water pollutants. The second objective is to establish a proper model based on artificial neural networks and multivariate statistical techniques; this model is aimed to assist planners and managers of water resources systems for solving surface water pollution problems. The ANN model provides a perfect knowledge and understanding about the relationship between water quality parameters and WQI, and defines the effective water quality parameter influencing the decreasing values of WQI through different sampling sites in Saf-Saf river basin.

Artificial neural networks (ANNs) have been successfully applied in a number of diverse fields including water resources [18,19]. In the context water quality prediction, ANNs may offer a promising alternative for water quality parameters [20–23]. There are many published works in the field of wastewater treatment plant performance using artificial intelligence methods such as neural networks [24,25]. Clair and Ehrman [26] used 10 y of data to examine the relationships between climate and geography on discharge and dissolved organic carbon (DOC) and dissolved organic nitrogen (DON) from 15 rivers in Canada's Atlantic region. Karul et al. [27] used a three-layer Levenberg–Marquardt feedforward learning algorithm to model the eutrophication process in three water bodies of Turkey (Keban Dam Reservoir, Mogan, and Eymir Lakes). Zhao et al. [28] developed three-layer feed-forward neural networks with back propagation (BP) for predicting biochemical oxygen demand (BOD) in Yuqiao Reservoir, China with a correlation coefficient of 0.8537 and average error of 2.56%.

Multivariate statistical techniques including principal component analysis (PCA) and factor analysis (FA) has been widely applied in environmental data reduction and interpretation of multiconstituent chemical, physical, and biological measurements [29,30]. Both PCA and FA are very powerful techniques whose main objective is to reduce the dimensions of a multivariate data set [2]. In addition, it allows to evaluate the relationship between variables, since they show the contribution of individual chemicals in several influence factors [31], also Helena et al. [32] using multivariate statistical techniques to characterize and evaluate groundwater quality, and it is useful in verifying temporal and spatial variations caused by natural and anthropogenic factors linked to seasonality.

In our present study, we have applied artificial neural networks ANNs to forecasting WQI in the Saf-Saf river basin based on a cause–effect relationship. Here, we have investigated the possibility of building a relationship between water quality parameters (independent variables) with WQI (dependent variable). What's more, the ANNs and decision-makers opinion are used in the characterizing and prioritizing of the most effective variable. The selected variables have been classified using the multivariate statistical techniques including principal components analysis and factor analysis.

2. Materials and methods

2.1. Study area and data description

Saf-Saf river basin is located on North–East of Algeria between parallels 6°40′–7°10′ East and 36°25′–36°53′ North, this basin covers a surface of 1,158 km^2, limited by the Guebli River basin from the west, the Hajar Mountain from the south, Kebir West river basin from the east, and finally the Mediterranean Sea from the north (Fig. 1). The climate is sub-humid with average annual rainfall varies from 636 mm in South to 750 mm in North, and the average monthly temperatures (minimal and maximal) varied between 12°C and 36°C [16].

The components of water resources balance for the Saf-Saf river basin has been developed based on the estimates of all water inputs and outputs to the river basin. Table 1 shows that the present net water balance in the Saf-Saf river basin is negative (–6.28 hm^3 y^{-1}) which indicates that there is a water deficit. The negative balance leads to decreasing the volume of freshwater in the river basin and the degradation of water quality [17].

A total of 35 samples of surface water were collected at various sampling sites along Saf-Saf river basin (Fig. 1). During April and September, 2015; all the water samples were sampled from a depth of 15 cm below the surface and preconditioned high density polyethylene bottles. They were conditioned by washing initially with five percent (5%) nitric acid, and then rinsing several times with distilled water. This was carried out to ensure that the sampling bottles were free from contaminants. Each of the surface water samples was analyzed for various physicochemical and biochemical parameters such as water temperature (WT, °C), the potential of hydrogen (pH), oxygen saturation (OS, %), total dissolved solids (TDS, mg L^{-1}), turbidity (NTU), nitrate (NO_3^-, mg L^{-1}), phosphate (PO_4^{3-}, mg L^{-1}), 5 d BOD_5 (mg L^{-1}), chemical oxygen demand (COD, mg L^{-1}), and chloride (Cl^-, mg L^{-1}).

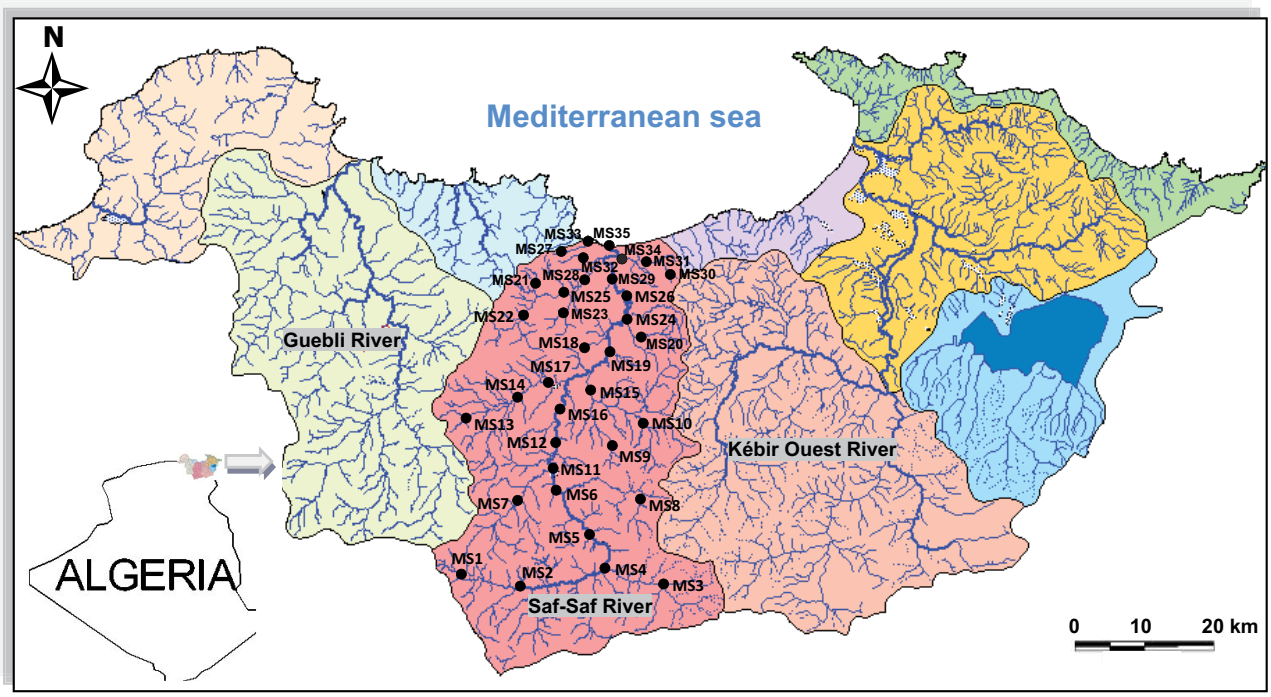

Fig. 1. Geographical projection of sampling sites in Saf-Saf river basin.

Table 1
Estimated water balance of Saf-Saf river basin

Inflows ($hm^3\ y^{-1}$)	Minimum	Maximum	Outflows ($hm^3\ y^{-1}$)	Minimum	Maximum
Groundwater	29.45	31.38	Municipal mobilization	25.35	26.75
Surface water	22.55	25.75	Agriculture mobilization	23.45	25.15
Non-conventional water	1.62	3.56	Industrial mobilization	7.75	7.95
Inflow from other basin	12.20	13.50	Discharge to the sea	15.55	20.08
Totals	65.82	74.19		72.10	79.93
Net balance	–6.28	–7.83			

The WQI values are calculated using the software CCME calculator version 1.0 developed by Canadian Council of Ministers of the Environment [6]. The CCME WQI was originally developed as the Canadian Water Quality Index (CWQI). It comprises of three factors and is well-documented [6]:

- *Factor 1 (scope)*: represents the percentage of variables that do not meet their objectives at least once during the time period under consideration (failed variables), relative to the total number of variables measured:

$$F_1 = \left(\frac{\text{Number of failed variables}}{\text{Total number of variables}}\right)\times 100 \quad (1)$$

- *Factor 2 (frequency)*: represents the percentage of individual tests that do not meet objectives ("failed tests"):

$$F_2 = \left(\frac{\text{Number of failed tests}}{\text{Total number of tests}}\right)\times 100 \quad (2)$$

- *Factor 3 (amplitude)*: represents the amount by which failed test values do not meet their objectives. F_3 is calculated in three steps.
- The number of times by which an individual concentration is greater than (or less than, when the objective is a minimum) the objective is termed an "excursion" and is expressed as follows. When the test value must not exceed the objective:

$$\text{Excursion}_i = \left(\frac{\text{Failed testvalue}_i}{\text{Objective}_j}\right) - 1 \quad (3)$$

For the cases in which the test value must not fall below the objective:

$$\text{Excursion}_i = \left(\frac{\text{Objective}_i}{\text{Failed testvalue}_i}\right) - 1 \quad (4)$$

- The amount by which individual tests are out of compliance is calculated by summing the excursions of

individual tests from their objectives and dividing by the total number of tests (both those meeting objectives and those not meeting objectives). This variable referred to as the normalized sum of excursions, or nse, is calculated as:

$$\text{nse} = \left(\frac{\sum_{i=1}^{n} \text{Excursion}_i}{\text{Number of tests}} \right) \quad (5)$$

- F_3 is then calculated by an asymptotic function that scales the normalized sum of the excursions from objectives (nse) to yield a range between 0 and 100.

$$F_3 = \left(\frac{\text{nse}}{0.01\text{nse} + 0.01} \right) \quad (6)$$

Once the factors have been obtained, the index itself can be calculated by summing the three factors as if they were vectors. The sum of the squares of each factor is therefore equal to the square of the index. This approach treats the index as a three-dimensional space defined by each factor along one axis. With this model, the index changes in direct proportion to changes in all three factors [6].

$$\text{WQI} = \left(\frac{\sqrt{F_1^2 + F_2^2 + F_3^2}}{1.732} \right) \quad (7)$$

The divisor 1.732 normalizes the resultant values to a range between 0 and 100, where 0 represents the worst water quality and 100 represents the best water quality. Once the WQI value has been calculated, water quality is ranked by relating it to one of the following class (Table 2).

2.2. Methodology

In this research, ANNs, decision-makers opinion and judgment, descriptive statistics and multivariate statistical techniques were used in the characterizing of WQI [33–35]. The analysis plan may be decomposed into four major steps which again are decomposed into many tasks (Fig. 2). The contents of the four steps are:

- *Step 1*: the first step aims to create a neural networks model, characterize, and prioritize the effective water quality parameters, and to establish a relationship between water quality parameters and WQI.
- *Step 2*: this step expresses the analysis of the questionnaire data to examine the decision-makers opinion and judgment of various stakeholders using descriptive statistics. The results of step 2 were compared with the results of the ANNs in step 1 to explore the understanding and knowledge of the local decision-makers about the health status of surface water in Saf-Saf river basin.
- *Step 3*: the purpose of this step is to transform the variables that were not normally distributed and to calculate the correlation matrix the variables selected from step 1.
- *Step 4*: Two methods of multivariate statistical techniques (PCA and FA) were used in step (4) for the selected water quality parameters, to classify them with the different sampling sites during wet and dry season.

Table 2
Water quality class, index, and water status

Class	Sampling sites	WQI value	Water status and observations
I	/	95–100	*Excellent*: water quality is protected with a virtual absence of threat or impairment; conditions very close to natural levels
II	/	80–94	*Good*: water quality is protected with only a minor degree of threat or impairment; conditions rarely depart from natural or desirable levels
III	MS1, MS3, MS5, MS10	65–79	*Fair*: water quality is usually protected but occasionally threatened or impaired; conditions sometimes depart from natural or desirable levels
IV	**MS1**, MS2, **MS2**, **MS3**, MS4, **MS4**, **MS5**, MS6, **MS6**, MS7, **MS7**, MS8, MS9, **MS10**, MS11, **MS11**, MS12, **MS12**, MS13, MS14, MS14, MS15, MS15, MS16, MS17, MS18, MS19, MS20, MS21, MS22, MS23, MS24, MS25, MS26, MS27.	45–64	*Marginal*: water quality is frequently threatened or impaired; conditions often depart from natural or desirable levels
V	**MS8**, **MS9**, **MS13**, **MS16**, **MS17**, **MS18**, **MS19**, **MS20**, **MS21**, **MS22**, **MS23**, **MS24**, **MS25**, **MS26**, **MS27**, MS28, **MS28**, MS29, **MS29**, MS30, **MS30**, MS31, **MS31**, MS32, **MS32**, MS33, **MS33**, MS34, **MS34**, MS35, **MS35**.	0–44	*Poor*: water quality is almost always threatened or impaired; conditions usually depart from natural or desirable levels

MS1: sampling site during wet season and **MS1**: sampling site during dry season

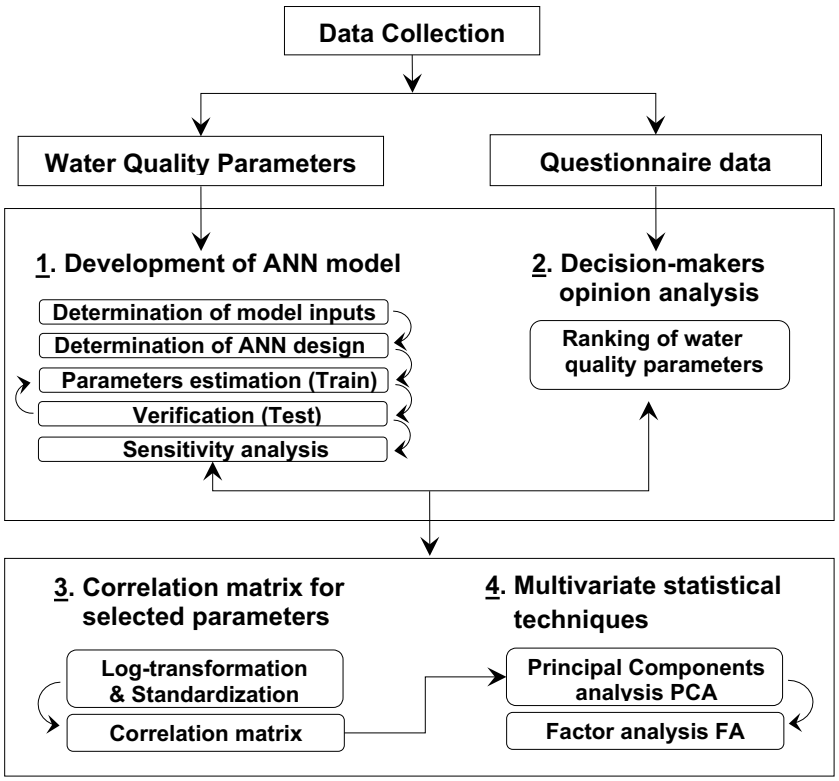

Fig. 2. Proposed water quality evaluation model.

2.2.1. Artificial neural networks

ANNs are able to map input–output relationships for natural complex phenomena and were developed to model the brain's interconnected system of neurons so that computers could using to imitate the brain's ability to sort patterns and learn from trial and error, thus observing relationships in data [36]. The main differences between various types of ANNs involve network architecture, method for determining the weights and transfer function as a creator of output value [37]. Feed-forward neural networks with back propagation are successfully applied to environmental problems. Multilayer perceptron (MLP) is perhaps the most popular network architecture in use today, due originally to Rumelhart et al. [38] and discussed at length in most neural network textbooks [39].

In this study, three-layer feed-forward MLP neural networks with gradient descent (GD), Broyden–Fletcher–Goldfarb–Shanno (BFGS), and conjugate gradient(CG) back-propagation learning were developed for the relationship between water quality parameters and WQI. The variables representing the water quality parameters were considered as the possible input variables including water temperature (WT, °C), potential of hydrogen (pH), oxygen saturation (OS, %), total dissolved solids (TDS, mg L^{-11}), turbidity (NTU), nitrate (NO_3^-, mg L^{-1}), phosphate (PO_4^{3-}, mg L^{-1}), 5 d BOD_5 (mg L^{-1}), COD (mg L^{-1}), and chloride (Cl^-, mg L^{-1}), while the target output variable was the WQI, which is the major means of assessing the levels of pollution of the Saf-Saf river. The MLP Neural network can be represented by the following compact form:

$$\{\mathrm{WQI}\} = \mathrm{ANN}_{\mathrm{MLP}}\begin{pmatrix}\mathrm{pH},\mathrm{WT},\mathrm{OS},\mathrm{TDS},\mathrm{Turbidity},\\ \mathrm{NO_3^-},\mathrm{PO_4^{3-}},\mathrm{BOD_5},\mathrm{COD},\mathrm{Cl^-}\end{pmatrix} \tag{8}$$

A schematic diagram of MLP neural network is given in Fig. 3. It shows a typical feed forward MLP structure with signals flow from input nodes, forward through hidden nodes, eventually reaching the output node.

Each hidden node (*j*) receives signals from every input node (*i*) which carries standardized values ($\bar{X}_i$) of an input variable where various input variables have different measurement units and span different ranges. $\bar{X}_i$ is expressed as:

$$\bar{X}_i = \frac{X_i - X_{\min}(i)}{X_{\max}(i) - X_{\min}(i)} \tag{9}$$

Each signal comes via a connection that has a weight (W_{ij}). The net integral incoming signals to a receiving hidden node (Net_j) is the potential of the neuron, $\bar{X}_i$ and the corresponding weights, (W_{ij}) plus a constant reflecting the node threshold value (TH_j):

$$\mathrm{Net}_j = \sum_{i=1}^{n} \bar{X}_i W_{ij} + \mathrm{TH}_j \tag{10}$$

The net incoming signals of a hidden node (Net_j) is transformed to an input (O_j) from the hidden node by using a non-linear transfer function (*f*) of sigmoid type, given by the following equation form:

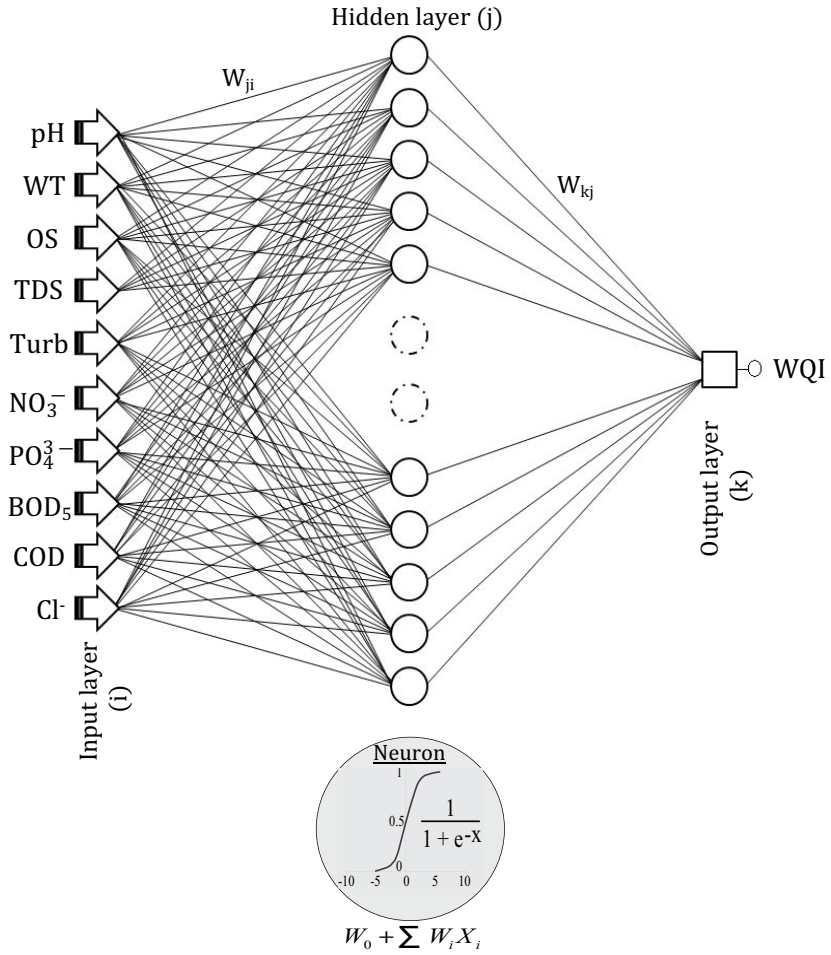

Fig. 3. Schematic diagram of a three-layer feed forward MLP neural network.

$$O_j = f\left(\text{Net}_j\right) = \frac{1}{1+e^{-\text{Net}_j}} \tag{11}$$

where (O_j) passes as a signal to the output node (k). The net entering signals of an output node (Net_k):

$$\text{Net}_k = \sum_{i=1}^{n} O_j W_{jk} + \text{TH}_k \tag{12}$$

The net incoming signals of an output node (Net_k) transformed using the sigmoid type function to a standardized or scaled output ($\bar{O}_k$) that is:

$$\bar{O}_k = f\left(\text{Net}_k\right) = \frac{1}{1+e^{-\text{Net}_k}} \tag{13}$$

Then, ($\bar{O}_k$) is standardized to produce the target output:

$$O_k = \bar{O}_k\left[O_{\max(k)} - O_{\min(k)}\right] + O_{\min(k)} \tag{14}$$

According Rumelhart et al. [38], the sigmoid function must be continuous, differentiable, and bounded from above and below in the range [0,1]. The calculated error between the observed value and the simulated value of the dependent variable is back propagated through the network and the weights are adjusted. Liu et al. [40] confirmed that the cyclic process of feed forward and error back propagation are repeated until the validation error is minimal.

The performance of each of the selected models (MLP_{DG}, MLP_{CG}, and MLP_{BFGS}) was determined using the criteria, such as the root mean square error (RMSE), the coefficient of determination (R^2), and the accuracy factor (A_f) computed from the measured and model predicted values of the dependent variables [41,42]. Values of the criteria parameters were calculated for all the two sets (training and test) as:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}\left(\text{WQI} - \text{W}\hat{\text{Q}}\text{I}\right)^2}{N}} \tag{15}$$

$$R^2 = \frac{\sum_{i=1}^{N}\left(\text{WQI} - \text{W}\dot{\text{Q}}\text{I}\right)^2}{\sum_{i=1}^{N}\left(\text{WQI} - \text{W}\bar{\text{Q}}\text{I}\right)^2} \tag{16}$$

Table 3
Basic statistics of water quality parameters during wet and dry season (N = 35)

	Wet season					Dry season				
	Mean	Minimum	Maximum	SD	Skew	Mean	Minimum	Maximum	SD	Skew
pH	7.50	7.24	7.86	0.14	0.80	7.61	7.24	8.30	0.24	1.43
WT	20.27	15.50	27.00	2.46	0.48	24.76	19.90	28.00	2.94	–0.32
OS	84.81	21.40	120.00	28.30	–1.55	99.18	21.40	159.21	40.58	–0.69
TDS	369.1	60.00	710.00	131.88	0.57	573.7	390.00	810.00	136.65	0.24
Turbidity	59.00	5.00	150.00	34.60	0.95	94.91	60.00	350.00	47.67	4.79
NO_3^-	2.20	0.16	6.00	1.23	0.85	3.99	0.43	8.00	2.15	0.24
PO_4^-	0.94	0.003	4.10	1.32	1.22	2.69	0.50	6.10	1.93	0.67
BOD_5	14.58	1.80	32.00	9.42	0.58	23.79	5.00	41.00	11.84	–0.28
COD	23.75	2.96	50.00	14.67	0.48	43.67	10.67	85.40	19.89	–0.13
Cl^-	95.21	25.00	200.00	55.61	0.83	154.0	25.00	350.00	104.23	0.38
WQI	54.54	29.00	65.00	10.89	–1.09	36.45	18.00	56.00	13.34	–0.21

$$A_f = 10\left[\sum_{i=1}^{N}\frac{\left|\log\left(\frac{\mathrm{W\hat{Q}I}}{\mathrm{W\bar{Q}I}}\right)\right|}{N}\right] \quad (17)$$

where WQI is the observed output value; WQ̂I is the simulated output value; WQ̄I is the mean value of WQI values; N is the total number of data sets. The RMSE, a measure of the goodness-of-fit, best describes an average measure of the error in predicting the dependent variable. R^2 value is an indicator of how well the network fits the data and accounts for the variability with the variables specified in the network [43]. A value of R^2 above 90% refers to a very satisfactory model performance. The A_f is a simple multiplicative factor showing the spread of simulation results. The larger value of A_f, the less accurate is the average estimate. A value of 1 indicates that there is a perfect agreement between all the predicted and the measured values. Finally, the goodness-of-fit of the selected models (MLP_{DG}, MLP_{CG}, and MLP_{BFGS}) were also checked through the analysis of the residuals.

2.2.2. Correlation matrix

Correlation matrix is a table evaluating the relationship between water quality variables. It calculates the direction and strength of the relationship between any two variables in the data set. A correlation coefficient near −1 or 1 means the strongest negative or positive relationship between two variables and its value closet to 0 means no linear relationship between them at a significant level of $p < 0.05$ [44]. The most commonly used measure of correlation is Pearson's r, it is called the linear correlation coefficient because r measures the linear association between two variables. Pearson's r assumes that the data follow bivariate normal distribution [46]. The correlation coefficient can be used to estimate the population Pearson correlation r between X and Y, it is written as:

$$r_{xy} = \frac{\sum x_i y_i - n\overline{xy}}{(n-1)S_x S_y} = \frac{n\sum x_i y_i - \sum x_i y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}} \quad (18)$$

2.2.3. Multivariate statistical techniques (PCA and FA)

Principal component analysis is a multivariate technique focused on a particular collection of variables. It is a powerful tool for pattern recognition that explains the variance of a large set of inter-correlated variables and transforms them into a smaller set of independent principal components [2,32]. These PCs provides information on the most meaningful parameters, which describe the whole data set through data reduction with a minimum loss of original information. Each principal component (PC) is a linear combination of the original variables and describes a different source of variation. PC is expressed as:

$$\mathrm{PC}_i = w_1 x_1 + w_2 x_2 + \ldots + w_n x_n \quad (19)$$

where x_i and w_i are the original variable and the component weight, respectively. The principal component weights are used as measures of the correlation between the variables and the principal components. The special feature of PCA is the graphics that provide a visual aid for the classification of variables and cases.

Factor analysis (FA) is formulated to transform the original variables into new uncorrelated variables called factors, which are linear combinations of the original variables. In addition, during the computation of FA, the most researchers performed a varimax rotation (raw) of the principal components coming from the original standardized variables, in order to reduce the contribution of variables with minor significance. The Varimax rotation was done taking into account previous works using FA for the evaluation of temporal and spatial changes in water quality [31,32]. The FA can be expressed as:

$$Z_{ij} = a_1 f_{1j} + a_2 f_{2j} + \ldots + a_m f_{mj} + e_{ij} \quad (20)$$

where Z is the measured variable, f is the factor score and e is the residual term accounting for errors or other source of variation.

3. Results and discussion

3.1. Summary descriptive statistics of water quality parameters and WQI

In this research, the data sets that we analyzed from 35 sampling sites in the study area were processed. The selected parameters for the estimation of surface water quality characteristics were: water temperature (WT, °C), potential of hydrogen (pH), oxygen saturation (OS, %), total dissolved solids (TDS, mg L^{-1}), turbidity (NTU), nitrate (NO_3^-, mg L^{-1}), phosphate (PO_4^{3-}, mg L^{-1}), 5 d BOD_5 (mg L^{-1}), COD (mg L^{-1}), chloride (Cl^-, mg L^{-1}), and water quality index (WQI, %). The summarized basis statistics of these parameters (minimum, maximum, standard deviation, and skewness) are presented in Table 3.

3.2. Artificial neural networks

The main dataset is divided into two sub-datasets: training (used 80% of the total available sets for finding the appropriate weight for each input) and the test dataset (20% of the total available sets for the evaluation of actual model performance). The test datasets were extracted randomly. Different MLP models (MLP_{DG}, MLP_{CG}, and MLP_{BFGS}) were created and tested in order to determine the optimum number of nodes in the hidden layer. According to Fletcher and Goss [45], the appropriate number of nodes in a hidden layer ranges from $(2n^{1/2} + m)$ to $(2n + 1)$, where n is the number of input nodes and m is the number of output nodes. Regarding the results obtained from the 100 MLP models created using three different algorithms, it can be concluded that the best optimal MLP model found is MLP_{BFGS} with 14 hidden nodes and a minimal root mean square error RMSE of 0.007 in testing data sets compared with the other types of MLP networks (Table 4). The MLP_{BFGS} model has very good performance in the two data sets (training and testing) with standard deviation of 15.030 and the 13.459, respectively (Table 5). The respective values of coefficient of determination (R^2) values and accuracy factor (A_f) for the two data sets are 0.929 and 1.420 for the training phase, and 0.811 and 1.210 for the testing phase (Table 4).

Fig. 4 presents a scatter plot of the MLP_{BFGS}-simulated vs. the observed WQI values while Fig. 5, compares between the MLP_{BFGS}-simulated and the observed WQI values for each sampling site. Error graphs indicate the contrast of the observed and simulated WQI value (Fig. 5). The error values for each observation were ranged between –14.54% and 11.23%. Both figures show that the overall agreement between the observed and simulated WQI values was satisfactory.

Fig. 6 shows a scatter-plot of MLP_{BFGS}-simulated WQI values and residuals corresponding to the training, testing, and all data sets. The observed relationship between MLP_{BFGS}-simulated WQI values and residuals for all the two sets shows complete independence and random distribution. It is further supported by the negligible small correlations (R^2 = 0.000 for training, R^2 = 0.119 for testing, and R^2 = 0.014 for all data sets). Fig. 6 explains that the points are well distributed on both sides of the horizontal line of zero ordinate representing the average of the residuals suggesting that the model fits the data well [46].

In order to identify the effect of input variables (water quality parameters) toward the output (WQI), the MLP_{BFGS} neural network sensitivity analysis was calculated in both training and testing phases. Table 6 indicates that the fifth most effective water quality parameters for WQI decreasing, in descending order, are BOD_5, Cl^-, NO_3^-, DCO, and TDS. The remaining water quality parameters according to their ranking in the testing phase are: oxygen saturation, phosphates, turbidity, pH, and water temperature. In light of these findings, the water quality monitoring agency may give priority consideration to these fifth water quality parameters.

The results of the MLP_{BFGS} neural network and expert opinion (Table 7) are similar only in ranking the first, second, and third priority water quality parameters which are BOD_5, chloride, and NO_3^-, whilst they differ in ranking the remaining water quality parameters.

3.3. Correlation matrix

The water temperature (WT) has significant environmental effects by influencing the physical, chemical, and biochemical. It was positively correlated with pH (r = 0.61) and oxygen saturation (r = 0.62) (Table 8). Table 8 shows a

Table 5
Regression statistical parameters for the target output (MLP_{BFGS} WQI)

Data sets	Data mean	Data SD	RMSE	Correlation
Training	43.590	15.030	0.009	0.975
Testing	53.343	13.459	0.007	0.911

Table 4
Performance criteria in various MLP neural networks

		Training data sets			Testing data sets		
ANN	Architecture	RMSE	R^2	A_f	RMSE	R^2	A_f
MLP (CG 45)	10–12–1	0.033	0.814	1.595	0.035	0.784	1.412
MLP (CG 39)	10–10–1	0.021	0.895	1.571	0.024	0.795	1.398
MLP (BFGS 60)	10–14–1	0.009	0.929	1.420	0,007	0.811	1.210

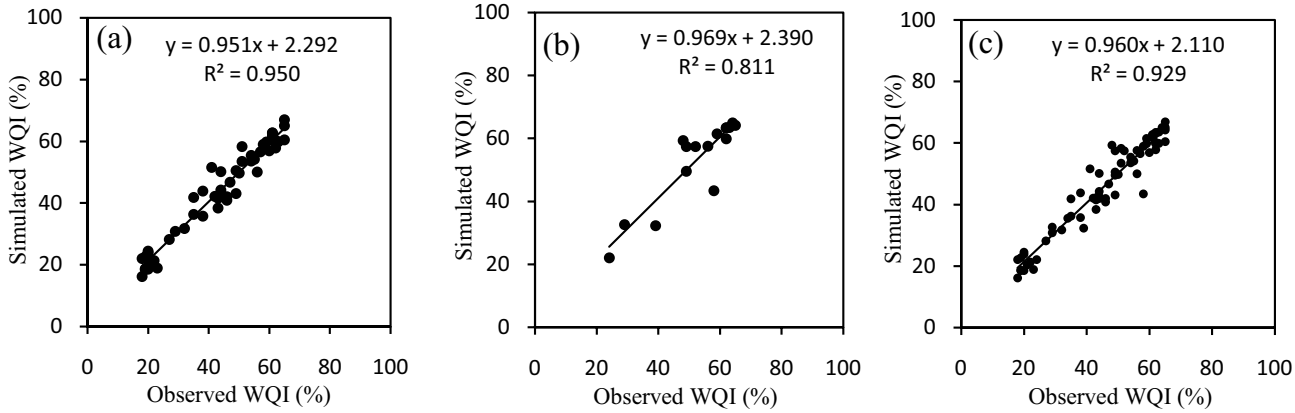

Fig. 4. Observed WQI vs. simulated WQI, (a) training, (b) testing, and (c) all data sets.

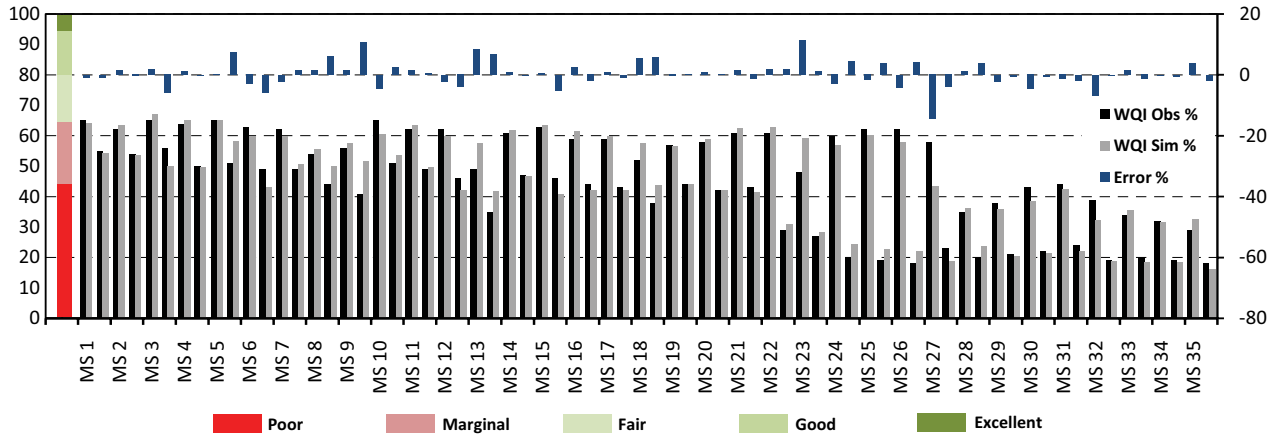

Fig. 5. Simulated WQI vs. observed WQI (in %).

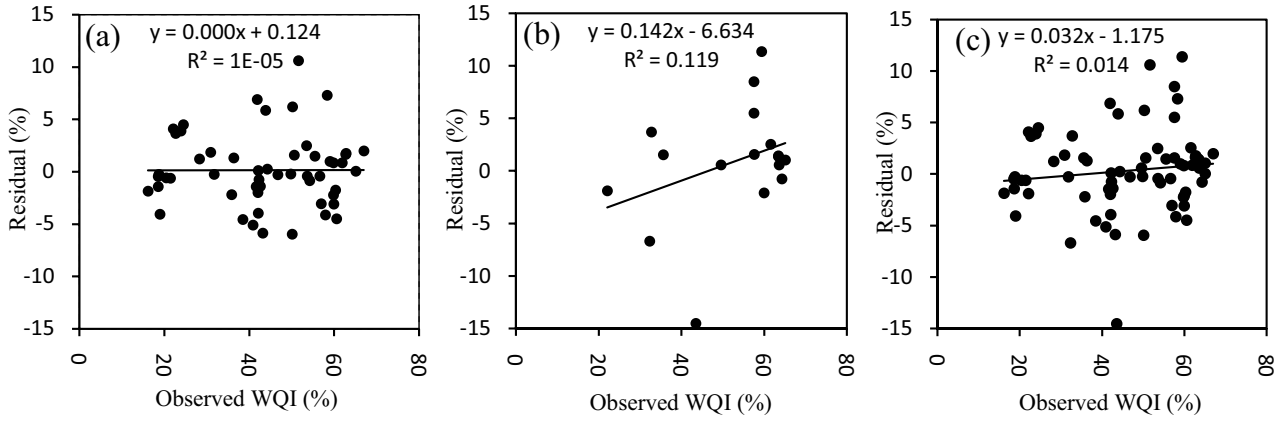

Fig. 6. Plot of the residuals vs. MLP_{BFGS} values of WQI, (a) training, (b) testing, and (c) all data sets.

significant and positive correlation between TDS, NO_3^-, PO_4^{3-}, BOD_5, COD, and Cl^- (r = 0.69–0.77), which are responsible for water contamination. The NO_3^- concentration showed a significant positive correlation with PO_4^{3-} (r = 0.88), BOD_5 (r = 0.78), COD (r = 0.69), and chloride (r = 0.82), and negative correlation with WQI (r = –0.79). PO_4^{3-} correlated reasonably well with BOD_5, COD, Chloride (r = 0.72–0.89), and WQI (r = –0.90) suggesting that PO_4^{3-} originated from anthropogenic sources. BOD_5 and COD are two parameters used to estimate the organic contamination load [47]. BOD_5 and COD showed a positive correlation between them, indicating contamination of organic matter. As also

Table 6
Sensitivity analysis of independent input variables (training and testing datasets)

	BOD_5	Cl^-	NO_3^-	COD	TDS	OS	PO_4^{3-}	Turbidity	pH	WT
Rang	1	2	3	4	5	6	7	8	9	10
Ratio	3.531	3.256	3.032	2.937	1.799	1.602	1.398	1.038	0.893	0.746
Rang	1	2	3	4	5	6	7	8	9	10
Ratio	3.884	3.565	2.865	2.672	2.409	1.996	1.598	1.239	1.079	0.878

Table 7
Ranking of input variables via decision-makers opinion and judgment

	Rang
BOD_5	1
Cl^-	2
NO_3^-	3
COD	5
TDS	4
OS	7
PO_4^{3-}	8
Turbidity	6
pH	10
WT	9

shown in Table 8, both BOD_5 and COD present a significant positive correlation with chloride and a significant negative correlation with WQI.

3.4. Principal component analysis

PCA module is applied to reduce the dimensionality of a data set consisting of a large number of interrelated variables while retaining as much as possible the variability present in data set and to evaluate the relationship between variables, since they show the participation of individual chemicals in several influence factors [2].

Table 9 shows that there are 11 variables in the analysis, the number of principal components was chosen in accordance with Kaiser's criterion, and Cattell's scree test. It shows that the principal components with eigenvalues closed to or greater than 1 were considered for interpretation. Therefore, two principal components were chosen for analysis with a cumulative variance of 73.28%. The remaining eigenvalues each account for less than 10% of the total variance.

The principal components loading of the different water quality parameters and WQI are presented in Table 9. The first principal component PC_1 explains 55.55% of the total variances and corresponds to the largest eigenvalue (6.11), PC_1 is highly correlated with TDS, nitrates, phosphates, BOD_5, COD, and chloride (negative correlation) and WQI (positive correlation). These are the parameters that primarily affect the Saf-Saf rivers' water quality. The second principal component PC_2 corresponding to the second eigenvalue (1.94) accounts for 17.70% of the total variance. It is highly correlated to pH, water temperature, and oxygen saturation (negative correlation). These parameters could provide insights into the effect of seasonal change. They are secondary parameters that affect surface water quality of Saf-Saf river basin.

The projection of the cases on the factor plane ($PC_1 \times PC_2$) shows that sampling sites were grouped into three main groups (Fig. 7b). The group I gathers the sampling sites which are typical by the average WQI and characterized by low values of chloride, TDS, NO_3^-, PO_4^{3-}, BOD_5, and COD. The sampling sites of group I are located in upstream of Saf-Saf river basin and correspond to a relatively low pollution during two seasons (wet and dry season). The group II includes the sampling sites of Saf-Saf valley during dry season and the sampling sites which are located in downstream of Saf-Saf river basin during wet season. This group represents waters with marginal quality based on the WQI. The group III gathers the sampling sites (during dry season) located in downstream area which are characterized by the high values of chloride, TDS, NO_3^-, PO_4^{3-}, BOD_5, and COD and very low WQI which showed evidence of surface water quality deterioration.

3.5. Factor analysis

Factor analysis was carried out 11 variables to identify the various varifactors that influence each of them. Three varifactors VFs were obtained through FA performed on the PCs and it explaining more than 90% of the total variance (Fig. 8). The corresponding VFs, variable loadings are presented in Table 10. Varifactor 1, which explained 59.03% of the total variance, had strong negative loadings (=–0.94) on WQI, a positive loading on TDS, NO_3^-, PO_4^{3-}, BOD_5, COD, and chloride. This varifactor can be interpreted as anthropogenic effects on surface water of Saf-Saf river basin. The VF_2 accounts for 21.07% of the total variance and had positive loadings on pH, water temperature, and oxygen saturation; it represents the effect of seasonal change on surface water Saf-Saf river basin. The third VF accounts for 10.62% of the total variance and had strong positive loadings on turbidity. This VF represents the effects of agricultural runoff and erosion in river basin. In comparison with the PCA results for water quality parameters and WQI (Table 9), the FA introduced a new water quality parameter which is turbidity.

4. Conclusion

In this paper, we developed a new methodology based on a combination of artificial neural networks and multivariate

Table 8
Correlation matrix – water quality parameters and WQI

	pH	WT	OS	TDS	Turbidity	NO_3^-	PO_4^{3-}	BOD_5	COD	Cl^-	WQI
pH	1.00										
WT	0.61	1.00									
OS	0.27	0.62	1.00								
TDS	–0.10	0.53	0.22	1.00							
Turbidity	0.13	0.20	0.30	0.15	1.00						
NO_3^-	0.17	0.48	0.35	0.69	0.12	1.00					
PO_4^{3-}	–0.04	0.38	0.19	0.77	0.15	0.88	1.00				
BOD_5	–0.13	0.27	0.16	0.71	0.09	0.78	0.83	1.00			
COD	–0.09	0.30	0.12	0.67	0.13	0.69	0.72	0.93	1.00		
Cl^-	–0.14	0.29	0.19	0.74	0.08	0.82	0.89	0.87	0.78	1.00	
WQI	0.09	–0.42	–0.12	–0.86	–0.19	–0.79	–0.90	–0.84	–0.78	–0.87	1.00

Underlined correlations are significant at $p < 0.0500$

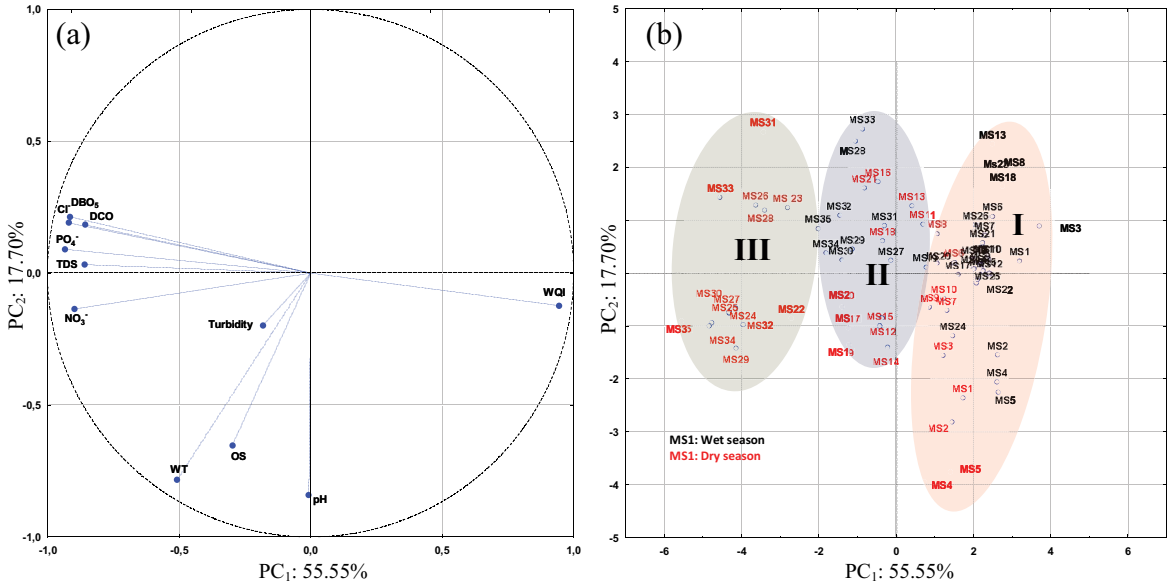

Fig. 7. Projection of variables and sampling sites on the factor-plane (a and b) $PC_1 \times PC_2$.

Table 9
Loadings of water quality parameters (10) and WQI on principal components for the whole datasets (Underlined loadings are >0.70)

PC_s	pH	WT	OS	TDS	Turbidity	NO_3^-	PO_4^{3-}	BOD_5	COD	Cl^-	WQI	Eigenvalue	% variance
PC_1	–0.01	–0.51	–0.30	–0.86	–0.18	–0.90	–0.93	–0.92	–0.86	–0.92	0.94	6.11	55.55
PC_2	–0.84	–0.78	–0.65	0.03	–0.20	–0.13	0.09	0.21	0.18	0.19	–0.12	1.94	17.70

statistical techniques to forecast WQI of surface water in an unmonitored river basin. An MLP neural networks with three different algorithms were trained and tested using datasets (water quality parameters and WQI) measured during wet and dry season in 2015.

The predictive capability of the MLP model is determined using three criteria, namely, RMSE, coefficient of determination (R^2), and the accuracy factor (A_f). The results obtained in this paper show that MLP_{BFGS} neural network demonstrate to be the best ANN structure indicating that BOD_5, Chloride, NO_3^-, DCO, and TDS are the fifth most effective water quality parameters influencing WQI in Saf-Saf river. Selecting and ranking water quality parameters assist decision-makers and water managers to give a priority consideration to these fifth water quality parameters in terms of surface water monitoring.

Table 10
Factor loadings – water quality parameters and WQI (rotation: Varimax normalized) extraction: principal components (underlined loadings are >0.70)

VF_s	pH	WT	OS	TDS	Turbidity	NO_3^-	PO_4^{3-}	BOD_5	COD	Cl^-	WQI	% variance
VF_1	–0.18	0.32	0.16	0.85	0.11	0.85	0.93	0.94	0.88	0.94	–0.94	59.03
VF_2	0.79	0.86	0.76	0.15	0.07	0.33	0.11	–0.01	0.01	0.02	–0.07	21.07
VF_3	0.23	0.14	–0.25	0.07	0.95	0.01	0.06	–0.02	0.05	–0.04	–0.13	10.62

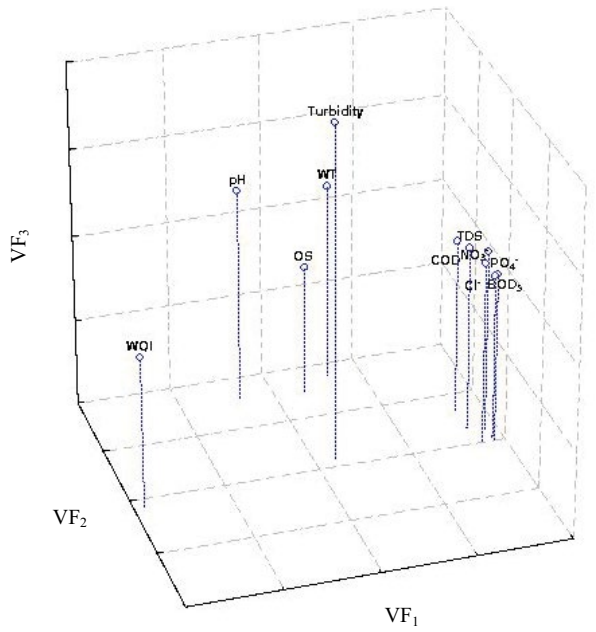

Fig. 8. Factor loadings, VF_1 vs. VF_2 vs. VF_3 – water quality parameters and WQI (rotation: Varimax normalized).

PCA combined with FA were used to assess variations in surface water quality of Saf-Saf river basin, both in time and space. It shows three groups of sampling sites, the group I on the right side of the PC_1 gathers upstream sampling sites that indicate relatively low pollution during two seasons. The group II characterizes sampling sites of Saf-Saf valley and it represents waters with marginal quality in term of WQI. The group III includes downstream sampling sites during dry season; it is characterized by very low WQI and it corresponds to high concentration in BOD_5, Chloride, NO_3^-, DCO, and TDS reflecting very polluted surface water. Therefore, MLP neural network model and multivariate methods enable easy forecasting of surface water quality and allows defining the importance and contribution of water quality parameters to the WQI. In addition, this approach also can be a framework to give reliable and trustful knowledge for decision-makers in improving river basin sustainability and factual strategies.

References

[1] S. Lonercan, T. Vansickle, Relationship between water quality and human health: a case study of the Linggi river basin in Malaysia, Soc. Sci. Med., 33 (1991) 937–946.
[2] K.P. Singh, A. Malik, D. Mohan, S. Sinha, Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India)—a case study, Water Res., 38 (2004) 3980–3992.
[3] K.P. Singh, A. Malik, S. Sinha, Water quality assessment and appointment of pollution sources of Gomti River (India) using multivariate statistical techniques: a case study, Anal. Chim. Acta, 538 (2005) 355–374.
[4] R. Koklu, B. Sengorur, B. Topal, Water quality assessment using multivariate statistical methods, a case study: Melen River system, Water Resour. Manage., 24 (2010) 959–978.
[5] P.W. Ramteke, Comparison of standard most probable number method with three alternate tests for detection of bacteriological water quality indicators, Environ. Toxicol. Water Qual., 10 (1995) 173–178.
[6] Canadian Council of Ministers of the Environment CCME, Canadian Water Quality Guide-Lines for the Protection of Aquatic Life: CCME Water Quality Index 1.0, Technical Report, Winnepeg, Canada, 2001.
[7] F.W. Kaurish, T. Younos, Developing a standardized water quality index for evaluating surface water quality, J. Am. Water Resour. Assoc., 43 (2007) 533–545.
[8] N.M. Gazzaz, M.K. Yusoff, A.Z. Aris, H. Juahir, M.F. Ramli, Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors, Mar. Pollut. Bull., 64 (2012b) 2409–2420.
[9] S.F. Pesce, D.A. Wunderlin, Use of water quality indices to verify the impact of Cordóba City (Argentina) on Suquía River, Water Res., 34 (2000) 2915–2926.
[10] C.G. Cude, Oregon water quality index a tool for evaluating water quality management effectiveness, J. Am. Water Resour. Assoc., 37 (2001) 125–137.
[11] A. Sargaonkar, V. Deshpande, Development of an overall index of pollution for surface water based on a general classification scheme in Indian context, Environ. Monit. Assess., 89 (2003) 43–67.
[12] A.A. Bordalo, R. Teixeira, W.J. Wiebe, A water quality index applied to an international Shared River basin: the case of the Douro River, Environ. Manage., 38 (2006) 910–920.
[13] H. Boyacioglu, Development of a water quality Index based on a European classification scheme, Water SA, 33 (2007) 101–106.
[14] T. Song, K. Kim, Development of a water quality loading index based on water quality modeling, J. Environ. Manage., 90 (2009) 1534–1543.
[15] X. Wang, F. Zhang, J. Ding, Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China, Sci. Rep., 7 (2017) 12858.
[16] B. Sakaa, S. Merdas, T. Mostephaoui, H. Chaffai, A. Hani, L. Djabri, The application of ANNs and multivariate statistical techniques to characterize a relationship between total dissolved solids and pressure indicators: a case study of the Saf-Saf river basin, Algeria, Desal. Water Treat., 57 (2015) 12963–12976.
[17] F. Khelfaoui, D. Zouini, L. Tandjir, Quantitative and qualitative diagnosis of water resources in the Saf-Saf river basin (north east of Algeria), Desal. Water Treat., 52 (2014) 2017–2021.
[18] H.R. Maier, G.C. Dandy, Neural networks for the prediction and forecasting of water sources variables: a review of a modeling issues and applications, Environ. Modell. Softw., 15 (2000) 101–124.

[19] B. Sakaa, H. Chaffai, A. Hani, The ANNs approach to identify water demand drivers for Saf-Saf river basin, J. Appl. Water Eng. Res., 8 (2020) 44–54.
[20] H.R. Maier, G.C. Dandy, The use of artificial neural networks for the prediction of water quality parameters, Water Resour. Res., 32 (1996) 1013–1022.
[21] D. Guclu, S. Dursun, Amelioration of carbon removal prediction for an activated sludge process using an artificial neural network (ANN), Clean Soil Air Water, 36 (2008) 781–787.
[22] E. Dogan, B. Sengorur, R. Koklu, Modeling biological oxygen demand of the Melen River in Turkey using an artificial neural network technique, J. Environ. Manage., 90 (2009) 1229–1235.
[23] M. Ay, O. Kisi, Modeling of dissolved oxygen concentration using different neural network techniques in Foundation Creek, El Paso County, Colorado, USA, J. Environ. Eng., 138 (2012) 654–662.
[24] M.M. Hamed, M.G. Khalafallah, E.A. Hassanien, Prediction of wastewater treatment plant performance using artificial neural networks, Environ. Modell. Softw., 19 (2004) 919–928.
[25] S. Heddam, H. Lamda, S. Filali, Predicting effluent biochemical oxygen demand in a wastewater treatment plant using generalized regression neural network based approach: a comparative study, Environ. Process., 3 (2016) 153–165.
[26] T.A. Clair, J.M. Ehrman, Variations in discharge and dissolved organic carbon and nitrogen export from terrestrial basins with changes in climate: a neural network approach, Limnol. Oceanogr., 41 (1996) 921–927.
[27] C. Karul, S. Soyupak, A.F. Çilesiz, N. Akbay, E. Germen, Case studies on the use of neural networks in eutrophication modeling, Ecol. Modell., 134 (2000) 145–152.
[28] Y. Zhao, J. Nan, F.-y. Cui, L. Guo, Water quality forecast through application of BP neural network at Yuquio reservoir, J. Zhejiang Univ. Sci. A, 8 (2007) 1482–1487.
[29] R. Wenning, G.E. Rickson, Interpretation and analysis of complex environmental data using chemometric methods, Trends Anal. Chem., 13 (1994) 446–457.
[30] H. Razmkhah, A. Abrishamchi, A. Torkian, Evaluation of spatial and temporal variation in water quality by pattern recognition techniques: a case study on Jajrood River (Tehran, Iran), J. Environ. Manage., 91 (2010) 852–860.
[31] M. Vega, R. Pardo, E. Barrado, L. Debán, Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis, Water Res., 32 (1998) 3581–3592.
[32] B. Helena, R. Pardo, M. Vega, E. Barrado, J.M. Fernandez, L. Fernandez, Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga river, Spain) by principal component analysis, Water Res., 34 (2000) 807–816.
[33] D.R. Helsel, R.M. Hirsch, Statistical Methods in Water Resources, US Geological Survey, Water Resources Division, Reston, 1992.
[34] A.W. Minns, M.J. Hall, Artificial neural networks as rainfall–runoff models, Hydrol. Sci. J., 41 (1996) 399–417.
[35] D. Patterson, Artificial Neural Networks, Prentice Hall, Singapore, 1996.
[36] S. Haykin, Neural Networks: A Comprehensive Foundation, Prentice Hall, New Jersey, 1999.
[37] M. Caudill, C. Butler, Understanding Neural Networks, Basic Networks, Vol. 1. MIT Press, Cambridge, MA, 1992.
[38] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning Internal Representations by Error Propagation, D.E. Rumelhart, J.L. McClelland, The PDP Research Group, Eds., Paralled Distributed Processing, Explorations in the Microstructure of Cognition, Vol. 1, Foundations, The MIT Press, Cambridge, MA, 1986, pp. 318–362.
[39] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, Oxford, 1995.
[40] J. Liu, H.H.G. Savenije, J. Xu, Forecast of water demand in Weinan City in China using WDF-ANN model, Phys. Chem. Earth, 28 (2003) 219–224.
[41] World Meteorological Organization WMO, Inter-comparison of conceptual models used in operational hydrological forecasting, Technical series, Water Resour. Res., 27 (1975) 2415–2450.
[42] S. Platikanov, X. Puig, J. Martin-Alonso, R. Tauler, Chemometric modeling and prediction of trihalomethane formation in Barcelona's water works plant, Water Res., 41 (2007) 3394–3406.
[43] B. Sakaa, H. Chaffai, A. Hani, The use of artificial neural networks in the modeling of socioeconomic category of integrated water resources management. Case study: Saf-Saf river basin, north east of Algeria, Arabian J. Geosci., 6 (2013) 3969–3978.
[44] M. Kumar, A.L. Ramanathan, M.S. Rao, B. Kumar, Identification and evaluation of hydrogeochemical processes in the groundwater environment of Delhi, India, Environ. Geol., 50 (2006) 1025–1039.
[45] D. Fletcher, E. Goss, Forecasting with neural networks: an application using bankruptcy data, Inf. Manage., 24 (1993) 159–167.
[46] K.P. Singh, A. Basant, A. Malik, G. Jain, Artificial neural network modeling of the river water quality—a case study, Ecol. Modell., 220 (2009) 888–895.
[47] H. Galal-Gorchev, G. Ozolins, X. Bonnefoy, Revision of the WHO guidelines for drinking water quality, Ann. Ist. Super. Sanita, 29 (1993) 335–345.