Desalination and Water Treatment
www.deswater.com
doi: 10.1080/19443994.2015.1057535

57 (2016) 12963–12976
June

Taylor & Francis
Taylor & Francis Group

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

Bachir Sakaa[a,b,*], Saifi Merdas[a], Toufik Mostephaoui[a], Hicham Chaffai[b], Azzedine Hani[b], Larbi Djabri[b]

[a]*Scientific and Technical Research Center on Arid Regions CRSTRA, Omar El Bernaoui. BP 1682 RP Biskra 07000, Algeria, Tel. +213 670 16 83 50; email: sakaabachir@yahoo.fr (B. Sakaa), Tel. +213559 09 26 56; emails: saifieco@gmail.com (S. Merdas), mostephaoui66@gmail.com (T. Mostephaoui)*
[b]*Faculty of Earth Sciences, Laboratory of Water Resource and Sustainable Development (REED), Annaba University, BP 12 23000, Algeria, Tel. +213 772 34 65 62; email: hichamchaffai@yahoo.fr (H. Chaffai), Tel. +213 775 17 34 96; email: haniazzedine@yahoo.fr (A. Hani), Tel. +213 661 32 22 02; email: djabri_Larbi@yahoo.fr (L. Djabri)*

Received 1 December 2014; Accepted 22 May 2015

ABSTRACT

With fast social and economic growth, stream water pollution in Saf-Saf river basin must consider appropriate control measures of the pollution sources. Hence, there is a need for a better knowledge and understanding of the pressure variables influencing the total dissolved solids of stream water. Saf-Saf river basin was chosen as the study area, and the data set included data on 9 variables for thirty different municipalities in the Saf-Saf river basin for monitoring year 2012. In this study, the effective variables have been characterized and prioritized using multi-criteria analysis with artificial neural networks (ANNs), and expert opinion and judgment. The selected variables were classified and organized using the multivariate techniques of principal components analysis (PCA) and factor analysis (FA). The results of ANN analysis indicate that domestic wastewater and industrial wastewater are the most pressing pollution sources, which is in contrast with the results of expert opinion in terms of ranking and prioritizing of pressure variables. The PCA/FA grouped the 30 municipalities into four groups based on their similarities, corresponding to municipalities of urban pollution (group I), very low pollution (group II), rural pollution (group III), and industrial pollution (group IV). Therefore, the identification of the main potential pollution sources in different municipalities by this study will help managers make better and more informed decisions about how to improve stream water quality degradation.

Keywords: Saf-Saf river basin; Total dissolved solids; Pressure indicators; Artificial neural networks; Principal component analysis; Factor analysis

1. Introduction

Recently, stream water pollution has been a common problem for many regions and countries [1]. Stream water quality impairment is strongly related to the increasing anthropologic influences in watersheds, such as changing land use pattern, increasing wastewater discharge and fertilizer application [2,3]. Therefore, a major issue in watershed management

*Corresponding author.

studies is the assessment of the linkage between land use and land cover characteristics and surface water quality [4].

Increasingly, water quality degradation from land use conflicts, destruction of wetlands and ecosystems and anthropogenic effects is undermining the sustainable management of water resources and threatening water resource base as part of nature. Anthropogenic effects are caused by local and diffuse sources. The pollution sources are: urban sewage, solid waste, hazardous waste, industrial waste, overuse of fertilizers and pesticides. In addition, over-exploitation of coastal aquifers has already led to many cases of irreversible saltwater intrusion [5]. Toward establishing an efficient watershed management system, therefore, the first step is to distinguish the relationship between the river water pollution and anthropologic influences at a watershed scale [6].

In Saf-Saf river basin, land use patterns and human activities have significant impact on the stream water characteristics. In rural areas, stream water quality is mainly impacted by nutrients from farming systems. However, in urban areas, stream water quality is mainly impacted by nutrient and organic chemical pollutants from household wastewater and industrial sewage [7]. To assist water planners and managers to gain adequate knowledge and understanding of the relationship between pressure indicators and total dissolved solids (TDS) of stream water, there is a need to use a proper methodology to define the effective pressure indicator influencing the increasing values of TDS in different municipalities of Saf-Saf river basin.

The main objectives of the research were to:

(1) Characterize and prioritize the most effective pollution source among pressure indicators and define the municipalities which are under pollution sources;
(2) Establish a modeling relationship between TDS, TDS of stream water and pressure indicators;
(3) Classify municipalities into groups associated with their related pressure indicators using of multivariate statistical techniques (principal component analysis (PCA) and factor analysis (FA)).

In this work, ANNs were employed to relate a set of independent input variables (the pressure indicators) with one dependent output variable (the TDS). In addition, the artificial neural networks (ANNs) and expert opinion are used to characterize and prioritize the most effective variable (indicator). The selected variables have been classified using the PCA and FA.

ANNs have been successfully used to model groundwater, assess quality of water, forecast precipitation, predict stream flow, and support other hydrologic applications. Wang et al. [8] applied ANNs to assess the confined groundwater vulnerability in North of China. Raman and Chandramouli [9] adopted similar ANNs as alternative tools for deriving the general operating policy of reservoirs. Leket et al. [10] applied ANNs to predict the concentration of nitrogen in streams from watershed features. Wen and Lee [11] addressed the multi-objective optimization of water pollution control and river pollution planning, for the Tou-Chen river basin in Taiwan. Rogers and Dowla [12] employed an ANNs trained by a solute transport model, to perform optimization of groundwater remediation.

Diverse multivariate techniques have been used to investigate how environmental indicators are related to explain the dependent variable (indicator), including several methods of ordination, canonical analysis, and univariate or multivariate linear, curvilinear, or logistic regressions [13,14]. Most statistical methods, reviewed by James and McCulloch [15], assume that relationships are smooth, continuous, and either linear or simply polynomials.

In this research, a relation between TDS of stream water and pressure variables in Saf-Saf river basin has been developed based on a cause–effect relationship tackling the life cycle of water resources management. The Driver–Pressure–State–Impact–Response (DPSIR) was selected as a well-established framework to develop the possible pressure variables. The effective variables have been characterized and prioritized using multi-criteria analysis with ANNs and expert opinion and judgment. The selected variables have been classified and organized using multivariate techniques, which are FA and PCA.

2. Materials and methods

2.1. Study area and data description

The Saf-Saf river basin is situated in the North Eastern of Algeria. It is bordered by the Guebli river basin from the West, the upstream of Seybouse river basin from the south, Kebir West river basin from the Eastern, and finally Mediterranean Sea from the north as shown in the Fig. 1. The total area of the Saf-Saf river basin is 1,158 km^2 and contains 30 municipalities (Fig. 1).

Water resources are also vulnerable to the fast growing demand of urban and rural populations, demand of economic sectors including agriculture, industry, and public institutions [7]. The population is

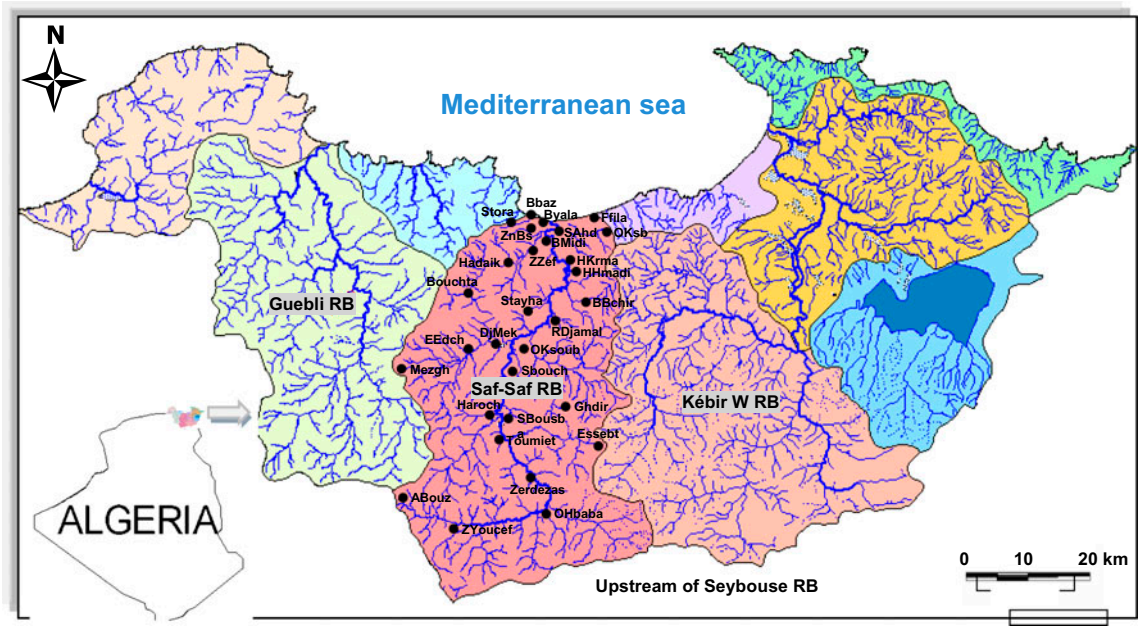

Fig. 1. Geographical projection of Saf-Saf river basin (Sakaa et al. [16]).

estimated at 474,908 capita (in 2012),and domestic water supply ranges from 80 to 170 l per capita per day (l c^{-1} d^{-1}). The industry is concentrated in the downstream of Saf-Saf river basin, which consumes 7.95 hm^3 y^{-1}, and finally, the important agriculture is located along the Valley of Saf-Saf river basin with the consumption of water estimated at 25.15 hm^3 $year^{-1}$.

The water resources balance of 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 $year^{-1}$) which indicates that there is a water deficit. This deficit results in lowering of water resources which face an increase in water demand [16]. The negative balance leads to decreasing the volume of freshwater in the river basin and degradation of water quality.

In this research work, the data of TDS and pressure indicators (variables) were applied to create the ANN model using the software package of STATISTICA 8. The database used is the results of sampling and analysis of stream water from the thirty municipalities (sites of sampling) and the collection of pressure variables from thirty municipalities (statistic data) for the reference year 2012. The pressure variables were the following:

(1) Hazardous wastes (HazWas) refer to generation of domestic, industrial, medical, and agricultural hazardous wastes. They are measured in tons per day (ton d^{-1});
(2) Generation of domestic wastewater (DomWW) represents the liquid waste generated by households, public institutions, schools, hospitals,

Table 1
Estimated water balance of Saf-Saf river basin

Inflows (hm^3 $year^{-1}$)	Min	Max	Outflows (hm^3 $year^{-1}$)	Min	Max
Ground water	29.45	31.38	Municipal mobilization	25.35	26.75
Surface water	22.55	25.75	Agricultural 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			

and public places. It is approximately 80% of the water use. It is measured by million cubic meters per year ($hm^3\ y^{-1}$);

(3) Pesticides (Pesticid) represent all substances used to kill pests, whether the pests are animals or plants. They include insecticides, herbicides, and fungicides. They are measured in tons y^{-1};
(4) Chemical fertilizers (ChemFer) refer to the amounts of chemical fertilizers used in agriculture to promote the plant growth. They include urea, ammonium, nitrate, and sulfates, ammonia, phosphatic fertilizers. They are measured in tons y^{-1};
(5) Organic fertilizers (OrgFer) represent the amounts of organic nitrogen input released by microorganisms in the soil for plants use and growth. They are derived from animal manures and vegetable by-products, composted organic matter and sludges. It is measured in tons per year (ton y^{-1});
(6) Petrol stations (PetrolS) refer to the number of fuel stations that provide the vehicles with fuel. These stations have underground storages which are considered as source for the hydrocarbon contamination;
(7) Industrial wastewater (IndWW) means the volume of liquid waste produced by the industrial facilities both existing in the residential areas and industrial states. It is measured by million cubic meters per year ($hm^3\ y^{-1}$);
(8) Carbon dioxide means the CO_2 content in the air due to the emissions from transport, energy station, fuels, industrial processes, and waste. It is measured in parts per million (ppm);
(9) Total dissolved solids (TDS) reflect the salinity of freshwater and originate from natural sources, sewage, urban, runoff, industrial wastewater, and chemicals. TDS consist mainly of inorganic salts (principally calcium, magnesium, potassium, sodium, carbonates, bicarbonates, chlorides, sulfates, phosphates) and some small amounts of organic matter that are dissolved in water. TDS are measured in milligram per liter ($mg\ l^{-1}$).

The variables representing pressure indicators are considered as the possible input variables, while the target output variable is the TDS. The variables presented in this study have been most frequently selected by the Organization for Economic Cooperation and Development [17]. To develop indicators, a long list of possible indicators will be drawn up based on literature review of indicator systems. Then, a group of different local stakeholders including local governmental officers, beneficiaries, and academics is asked to evaluate each of the proposed indicators. Also, a questionnaire will be administered to ask the participants to rank each indicator according to their judgment on the significance of each indicator to local sustainable development. Many people have participated in the formulation and development of pressure indicators.

2.2. Methods

The tools chosen for this research were ANNs, expert opinion and judgment, basic statistics and multivariate techniques [18–21]. The software selected were the STATISTICA package version 8.0, STATISTICA Neural Networks.

There are four steps in the proposed methodology for developing a relationship between TDS and pressure variables (Fig. 2).

Step 1: the first step expresses the creation of the ANN model, the characterization and prioritization of the effective variables, and the establishment of the modeling relationships between pressure indicators and TDS.

Step 2: this step indicates the analysis of the questionnaire undertaken to explore the expert 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 examine the understanding and knowledge of the local experts about the actual baseline conditions of stream water degradation.

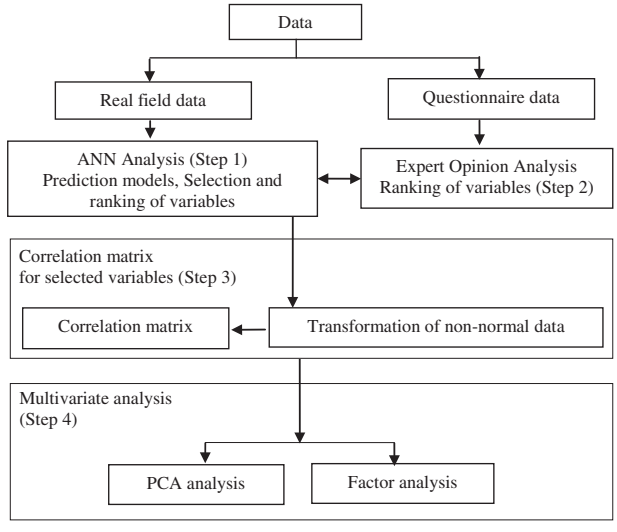

Fig. 2. Steps of data analysis.

Step 3: transformation of data variables that were not normally distributed and calculation of the correlation matrix were carried out in Step 3 for the variables selected from Step 1.

Step 4: two techniques of multivariate analysis were undertaken in Step 4 for the selected variables, to classify them with the relevant municipalities.

2.2.1. Background of ANNs

Neural networks have seen an explosion of interest over the last few years, and are being successfully applied across an extraordinary range of problem domains, in areas as diverse as, engineering, geology, and physics. Neural networks are applicable in virtually every situation in which a relationship between the simulator variables and simulated variables exists, even when that relationship is very complex and not easy to articulate in the usual terms of correlations or differences between groups. The basic idea of an ANN is that the network learns from the input data and the associated output data, which is commonly known as the generalization ability of the ANN.

The application of ANNs and other statistical techniques in this research work have some limitations due to the limited data sets available (30 municipalities). Therefore, the cross-validation was used in ANN as a stopping criteria to determine the optimal number of hidden layer nodes, while avoiding the risk of over training.

A variety of validation criteria that could be used for the evaluation and inter-comparison of different models was proposed by the World Meteorological Organization [22]. They fall into two groups: graphical indicators and numerical performance indicators of the several numerical indicators [22], suitable ones for this study are chosen. These are the root mean square error (RMSE) and the R^2 efficiency [23], given by:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (\mathrm{TDS} - \widehat{\mathrm{TDS}})^2}{N}} \tag{1}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (\mathrm{TDS} - \widehat{\mathrm{TDS}})^2}{\sum_{i=1}^{N} (\mathrm{TDS} - \bar{\mathrm{TDS}})^2} \tag{2}$$

Where TDS is the observed output value, $\widehat{\mathrm{TDS}}$ is the simulated output value, $\bar{\mathrm{TDS}}$ is the mean value of TDS values, and N is the total number of data sets. The RMSE gives a quantitative indication for the network error. It measures the deviation of the estimated values from the corresponding observed values of target output which refers to the estimation accuracy [24,25]. Besides, the RMSE was used to compare the performance of MLP with radial function basic (RBF). 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. A value of R^2 above 90% refers to a very satisfactory model performance. Values range between 80 and 90% indicates unsatisfactory model [26–28]. The ideal value for RMSE is zero and for R^2 is unity.

For this purpose, we developed ANNs using the BFGS (Broyden Fletcher Goldfarb Shanno Quasi-Newton) and Scaled Conjugate Gradient (SCG) back propagation, which is recommended because it is more likely to optimize the simulation performance. The number of neurons in a hidden layer is decided after training and testing. Multi-layered network trained by back propagation is currently the most popular and proven [29]. Training of ANNs consists of showing example inputs and target outputs to the network and iteratively adjusting internal parameters based on performance measures. The MLP is simple, robust, and very powerful in pattern recognition, classification, and mapping. MLP is capable of approximating any measurable function from one finite dimensional space to another within a desired degree of accuracy [30].

In this work, the variables representing the pressure variables were considered as the possible input variables including HazWas, DomWW, Pesticid, ChemFer, OrgFer, PetrolS, IndWW, and CO_2, while the target output variable was the TDS, which is the major parameter in water quality assessment. The MLP network can be represented by the following compact form:

$$\{\mathrm{TDS}\} = \mathrm{ANN}\left[\mathrm{HazWas, DomWW, Pesticid, ChemFer, OrgFer, PetrolS, IndWW, CO_2}\right]$$

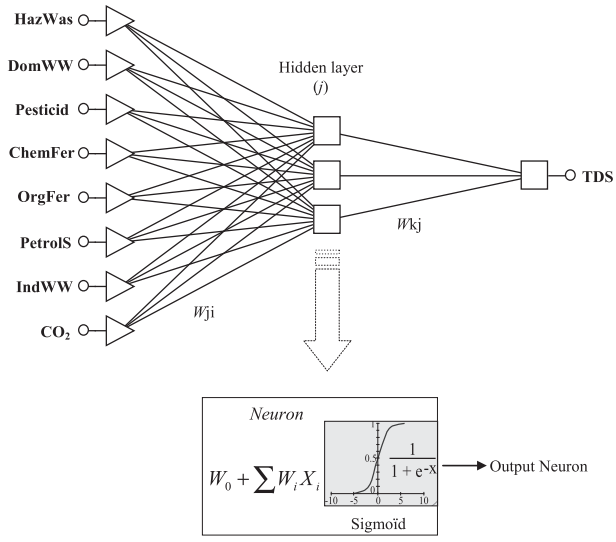

Fig. 3. Schematic diagram of neural networks.

A schematic diagram of neural network is given in Fig. 3. It shows a typical feed forward structure with signals flow from input nodes, forward through hidden nodes, eventually reaching the output node. The input layer is not really neural at all; these nodes simply serve to introduce the standardized values of the input variables to the neighboring hidden layer without any transformation. The hidden and output layer nodes are each connected to all of the nodes in the preceding layer. However, the nodes in each layer are not connected to each other. A numeric weight is associated with each of the internode connections. W_{ij} represents the strength of connections of nodes between input and hidden layer, while Wjk represents the strength of connections of nodes between hidden and output layers.

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 follows:

$$\bar{X}_i = \frac{X_i - X_{\min(i)}}{X_{\max(i)} - X_{\min(i)}} \tag{3}$$

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):

$$\mathrm{Net}_j = \sum_{i=1}^{n} \bar{X}_i W_{ij} + TH_j \tag{4}$$

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

$$Oj = f(\mathrm{Net}_j) = \frac{1}{1 + e^{-\mathrm{Net}_j}} \tag{5}$$

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

$$\mathrm{Net}_k = \sum_{i=1}^{n} O_j W_{jk} + \mathrm{TH}_k \tag{6}$$

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

$$\bar{O}_k = f(\mathrm{Net}_k) = \frac{1}{1 + e^{-\mathrm{Net}_k}} \tag{7}$$

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

$$O_k = \bar{O}_k(O_{\max(k)} - O_{\min(k)}) + O_{\min(k)} \tag{8}$$

Rumelhart et al. [29] explained that the sigmoid function must be continuous, differentiable, and bounded from above and below in the range [0–1]. The calculated error between the observed actual value and the predicted value of the dependent variable is back propagated through the network, and the weights are adjusted. The cyclic process of feed forward and error back propagation is repeated until the validation error is minimal [31].

In case that limited data sets are available, cross-validation can be used as a stopping criteria to determine the optimal number of hidden layer nodes [32]; while avoiding the risk of over training [33]. Cross-validation is a technique used commonly in ANN models and has a significant impact on the division of data [34]. It aims to train the network using one set of data, and to check performance against a validation set not used in training. This examines the ability of the network to generalize properly by observing whether the validation error is reasonably low. The training will be stopped when the validation error starts to increase [27]. The database was divided into training, validation, and testing. For the ANN models described in this study, 50% of the available data was used for training, 25% was used for the validation and 25% to test the validity of network prediction [27,28].

In this study, a sensitivity analysis can be carried out to identify the importance of the input variables; the ranking of variables is based on backward elimination. The network is trained using all the available inputs and the least relevant input or subset of inputs is deleted. Castellano and Fanelli [35] show a fast strategy to adjust the remaining weights after the elimination of input. The sensitivity is presented by the ratio and rank. The ratio denotes the ratio between the error and the baseline error (i.e., the error of the network if all variables are available). The rank simply lists the variables in the order of their importance.

2.2.2. Correlation matrix

Correlation matrix is a table showing inter-correlation among all variables analyzed. It calculates the direction and strength of the relationship between any two variables in the data set. Direction is indicated by positive or negative. Strength is indicated by how close the value of the correlation is to +1 (perfect) in a direct relationship (if one increases then the other increases) and −1 in an inverse relationship (if one increases then the other decreases). 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 [18].

2.2.3. Principal component analysis PCA

The PCA module aims at the reduction in the number of variables to a smaller number of representative and uncorrelated factors and the classification of variables and cases. Two types of analyses are available, depending upon whether the data need to be standardized or centered. In the former case, the analysis is carried out via the correlation matrix, while in the latter, the analysis is carried out via the covariance matrix. The basic method, however, consists of diagonalizing the symmetric matrix (correlation or covariance). The special feature of this module is the graphics that provide visual aid for the classification of variables and cases.

2.2.4. Factor analysis (FA)

FA is a generic term for statistical techniques concerned with the reduction of a set of observable variables into a small number of latent factors and the detection of the structure in the relationships between variables that is to classify variables. This structure is expressed in the pattern of variances and covariances between variables and similarities between observations. The underlying assumption of FA is that there exist a number of unobserved latent factors that account for the correlations within a set of multivariate observations.

3. Results and discussion

3.1. Summary descriptive statistics of pressure variables

Table 2 presents that all pressure variables have positive skewness with different values (right skewed) except TDS which have negative skewness (left skewed). TDS, hazardous wastes, generation of domestic waste water, pesticides, chemical fertilizers, and organic fertilizers have substantial skewness, small spread and normal data distribution. The variables petrol stations, industrial wastewater, and CO_2 have reasonably non-normal distribution of data.

3.2. Artificial neural networks (ANN)

The types of networks considered are MLP with two ways to calculate (Broyden Fletcher Goldfarb

Table 2
Summary descriptive statistics of pressure variables

	N	Mean	Median	Lower. quartile	Upper quartile	Standard deviation	Skewness
TDS	30	1,120	1,177.5	871.00	1,334.40	346.770	−0.352
HazWas	30	13.51	9.00	4.200	17.80	11.754	1.228
DomWW	30	1.32	0.725	0.320	1.650	1.513	1.891
Pesticid	30	1.08	0.719	0.310	1.710	1.005	1.131
ChemFer	30	130.48	100.4	39.500	155.75	124.97	1.565
OrgFer	30	1,244.4	1,122.0	210.00	2,050	1,055.42	0.440
PertolS	30	0.63	0.000	0.000	1.000	0.964	1.324
IndWW	30	9.05	0.004	0.0005	0.005	37.457	4.962
CO_2	30	460.53	377.5	372.00	509.00	132.996	2.263

Shanno Quasi-Newton BFGS and SCG) and RBF. During the analysis, many networks were tested. The best optimal ANNs model found is MLP (BFGS 107) with 09 hidden nodes and a minimal RMSE of 0.0009 compared with the other types of ANNs (Table 3). The model has very good performance in validation with standard deviation of 192.04 and the RMSE for training, validation and testing is small and close, which indicates that the data sub-sets are from the same population (Table 4). In addition, the correlation coefficient is higher than 95% for training, validation, and testing, which shows an excellent agreement between the observed and simulated TDS (Fig. 4).

Table 3
RMSE in various neural networks

ANN	Architecture	RMSE	R^2
RBF	8-10-1	0.0013	0.762
MLP (CG 40)	8-12-1	0.0011	0.780
MLP (BFGS 107)	8-9-1	*0.0009*	*0.840*

Table 4
Regression statistical parameters for the target output (TDS)

	Training	*Validation*	Testing
Data Mean	1,144.65	*1,170.70*	1,008.36
Data SD	358.24	*192.04*	406.21
RMSE	0.0001	*0.0009*	0.0002
Correlation	0.9600	*0.9871*	0.9545

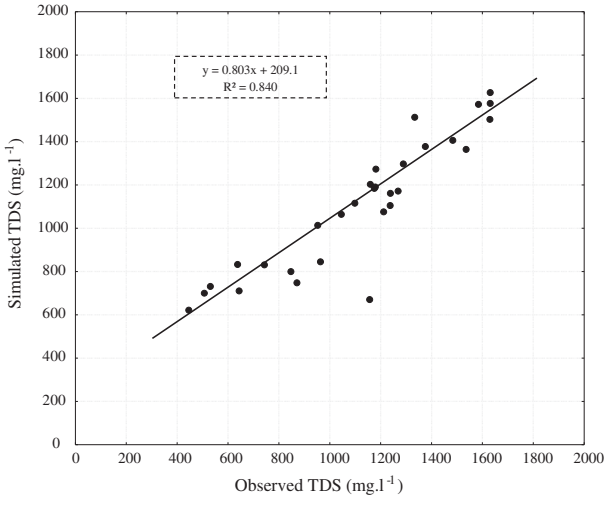

Fig. 4. Simulated TDS vs. observed TDS (in mg l^{-1}).

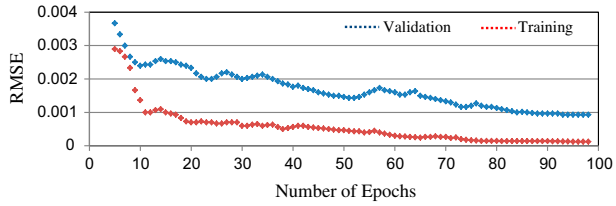

Fig. 5. Training–validation error graph for cases—pressure indicators.

The model training error for the independent cases is shown in Fig. 5. It graphs the RMSE of the network against epochs during iterative training of the back propagation training algorithms. In addition, it plots separate lines for the RMSE on the training and validation sub-sets of the independent cases at the end of the last iterative training run. The graph indicates that the range of RMSE of independent cases for both training and validation is very small [36].

The ANN sensitivity analysis of pressure variables in the validation phases (Table 5) indicates that domestic wastewater is the most pressing pollution source followed by industrial wastewater. The remaining pressure variables according to their ranking in the validation phase are hazardous waste, chemical fertilizers, carbon dioxide, pesticides, organic fertilizers, and petrol stations.

The results of the ANN model and expert opinion (Table 6) are similar only in ranking the third variable which is hazardous waste, while they differ in ranking the remaining variables.

3.3. Correlation matrix

An analysis of the correlation matrix was undertaken to explore the direction, strength, and significance of relationship between any two variables of data set. Transformation of any specified variable that is not normally distributed is a prerequisite, and a transformation to natural logs (base e) worked reasonably well for all intended variables. As an example, the domestic wastewater variable is transformed to ln (domestic wastewater) with an approximately normal distribution.

Table 7 shows a significant and positive linear relationship between ln (domestic wastewater), ln (chemical fertilizer), ln (industrial wastewater), CO_2 and TDS. The increase in domestic and industrial wastewater generation after pretreatment increases the TDS of stream water. The carbon dioxide in the air will be built-up in rain water as carbonic acid which will break up in surface and ground water to

Table 5
Sensitivity analysis of independent input variables

	HazWas	DomWW	Pesticid	ChemFer	OrgFer	PetrolS	IndWW	CO_2
Rang	1	3	7	4	6	8	2	5
Ratio	10.511	5.320	2.866	5.078	3.867	1.088	6.512	4.294
Rang	*3*	*1*	*6*	*4*	*7*	*8*	*2*	*5*
Ratio	*1.897*	*2.213*	*1.059*	*1.771*	*1.017*	*0.181*	*1.996*	*1.067*

Table 6
Ranking of input variables via expert opinion and judgment

	HazWas	DomWW	Pesticid	ChemFer	OrgFer	PetrolS	IndWW	CO_2
Rang	*3*	*2*	*7*	*5*	*8*	*6*	*1*	*4*

Table 7
Correlation matrix – pressure variables

	TDS	ln (HazWas)	ln (DomWW)	Pesticid	ChemFer	OrgFer	PetrolS	ln (IndWW)	CO_2
TDS	1.000								
ln (HazWas)	0.649	1.000							
ln (DomWW)	**0.730**	**0.859**	1.000						
Pesticid	−0.05	0.295	0.224	1.000					
ln (ChemFer)	**0.717**	0.301	0.219	**0.720**	1.000				
ln (OrgFer)	−0.46	0.329	0.232	**0.708**	**0.772**	1.000			
PetrolS	0.361	0.395	0.456	0.199	0.137	−0.032	1.000		
ln (IndWW)	**0.699**	0.116	**0.695**	−0.067	−0.072	−0.017	0.268	1.000	
CO_2	**0.705**	**0.706**	0.263	0.097	0.136	0.059	0.611	0.521	1.000

Note: Correlations are significant at $p < .05000$.

carbonates, thus increasing the TDS content. The use of chemical fertilizers is always associated with irrigation and drainage water increasing the TDS of stream water.

There are significant, positive linear relationships between ln (hazardous waste) and ln (domestic wastewater) and carbon dioxide. The increase in domestic and industrial wastewater increases the production of hazardous waste.

ln (domestic wastewater) has significant and positive linear relationships with ln (industrial wastewater). Domestic wastewater increases with the increase in the industrial wastewater generation since the industrial facilities are connected to the urban wastewater systems immediately after use.

Pesticides have significant and positive linear relationships with organic fertilizers and chemical fertilizers. The use of pesticides is always associated with chemical and organic fertilizers since they are applied for the same agriculture land but with different proportions.

3.4. Principal component analysis (PCA)

The purpose of applying the PCA module was to reduce the number of variables into a smaller number of dimensions (factors) and to classify variables and clusters of observations with similar characteristics with respect to these factors.

Table 8 shows that there are 09 variables in the analysis and thus, the sum of all eigenvalues is equal to 09. The number of factors was chosen in accordance with Kaiser's criterion and Cattell's scree test. It shows that the point where the continuous drop in eigenvalues levels off is at Factor 3. Therefore, three factors were chosen for analysis with a cumulative variance of 77.310%. The remaining eigenvalues each accounts for less than 10% of the total variance.

Table 9 presents variances of factors and their loadings from variables. The first factor corresponds to the largest eigenvalue (2.823) and accounts for approximately 31.373% of the total variance. It is most correlated with the variables: pesticide, chemical, and

Table 8
Eigenvalues of correlation matrix – pressure variables

	Eigenvalue	(%) Total variance	Cumulative eigenvalue	Cumulative (%)
1	2.823	31.373	2.823	31.373
2	2.447	27.188	5.270	58.561
3	1.687	18.749	6.957	77.310
4	0.757	8.412	7.715	85.723
5	0.591	6.560	8.305	92.283
6	0.397	4.409	8.702	96.692
7	0.213	2.370	8.915	99.062
8	0.071	0.791	8.986	99.853
9	0.013	0.146	9.000	100.000

Table 9
Factor–variable correlations (factor loadings), pressure variables (underlined loadings are >.70)

	Factor 1	Factor 2	Factor 3
TDS	0.039	0.595	−0.615
HazWas	0.501	*0.696*	0.482
DomWW	0.345	*0.740*	0.522
Pesticid	*0.883*	−0.315	−0.175
ChemFer	*0.889*	−0.331	−0.146
OrgFer	*0.782*	−0.410	−0.153
PertolS	0.364	0.592	−0.084
IndWW	−0.268	0.285	*−0.692*
CO_2	0.254	0.515	−0.544

organic fertilizers (positive correlations). The second factor corresponding to the second eigenvalue (2.447) accounts for 27.188% of the total variance. It is correlated with hazardous waste and domestic wastewater (positive correlations). The third factor corresponding to the eigenvalue 1.687 accounts for 18.749%. It is significantly correlated with industrial wastewater (negative correlation).

Fig. 6(a) and (b) represents coordinates for the three factors. The graph shows a unit circle with variables that were used to compute the current factor solution. The circle can provide a visual indication of how well each variable is represented by the current set of factors. Based on the magnitudes of the factor coordinates for the variables in the analysis, the first

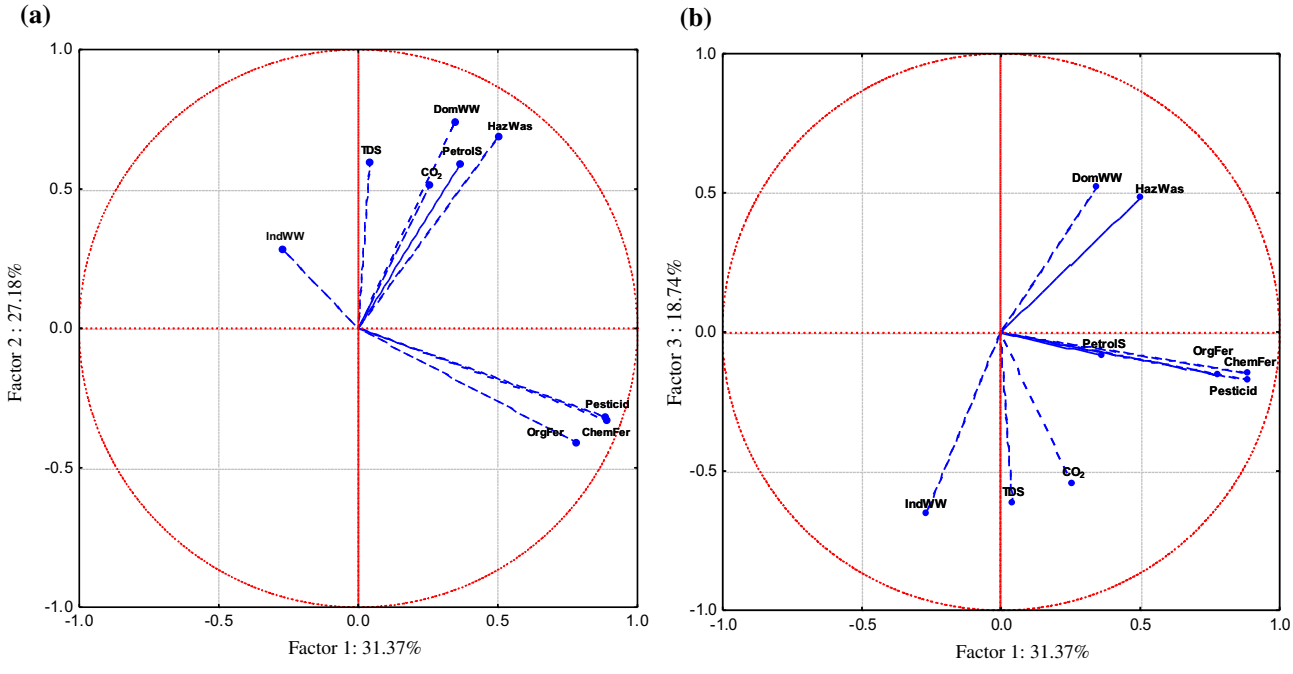

Fig. 6. a, b Projection of the variables on the factor plane, (a) 1 × 2 and (b) 1 × 3.

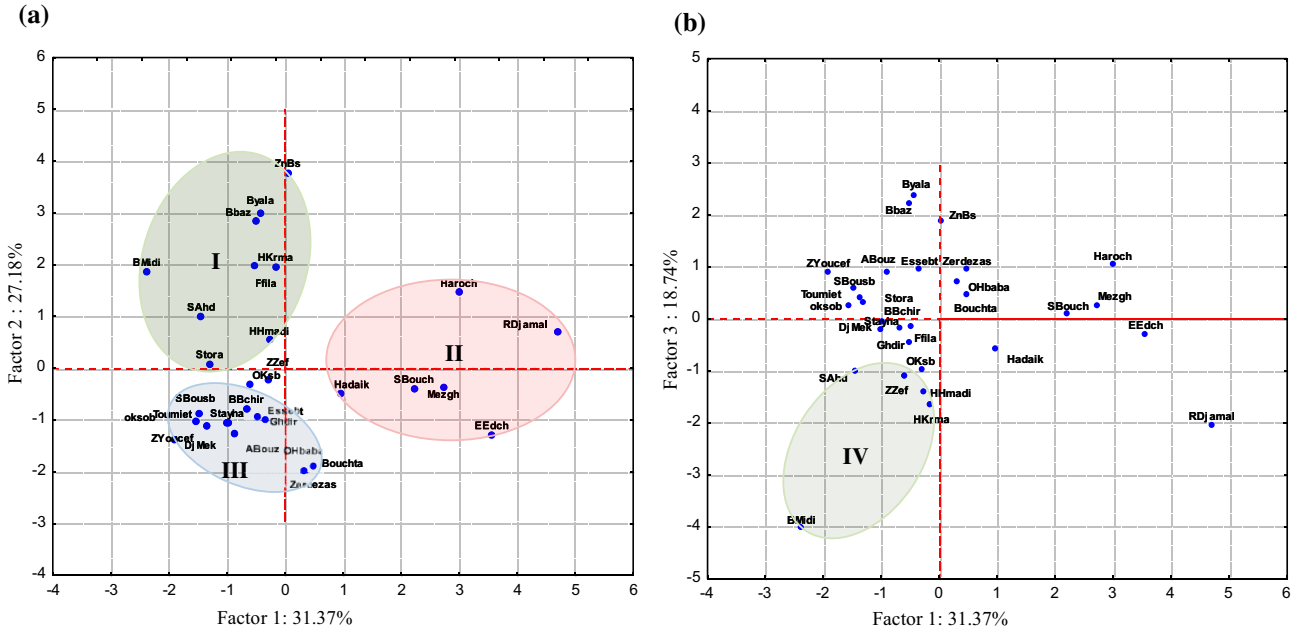

Fig. 7. a, b Projection of the cases on the factor plane, (a) 1 × 2 and (b) 1 × 3.

factor can be labeled as rural pollution. Factor two can be labeled as urban pollution, and factor three can be called as industrial pollution.

The projection of the cases on the factor plane (1 × 2) and (1 × 3) demonstrates four groups of municipalities (Fig. 7(a) and (b)). The group I gathers the urban municipalities which are typical by the high hazardous waste and domestic wastewater (urban pollution), the group II (mountainous municipalities) which are characterized by low values of hazardous waste, domestic wastewater, and CO_2, the group III includes municipalities for agricultural purposes distinguished with pesticides, chemical, and organic fertilizers (rural pollution), the group IV gathers municipalities located in the industrial area which are characterized by the high values of industrial wastewater and high concentrations of CO_2 (industrial pollution).

3.5. Factor analysis (FA)

FA was used for the comparison with PCA results. It reduces the number of observed variables to a smaller number of unobserved latent factors which are uncorrelated with each other and classifies variables within these factors. Varimax normalized rotation was adopted to maximize the variance of factors on the

Table 10
Factor loadings–pressure variables (varimax normalized) extraction: principal components (underlined loadings are >.700000)

	Factor 1	Factor 2	Factor 3
TDS	−0.066	0.134	*0.844*
HazWas	0.084	*0.972*	0.007
DomWW	−0.084	*0.960*	−0.001
Pesticid	*0.952*	0.063	0.011
ChemFer	*0.956*	0.069	−0.021
OrgFer	*0.892*	−0.039	−0.064
PertolS	0.116	0.550	0.417
IndWW	−0.216	−0.246	*0.697*
CO_2	0.143	0.205	*0.750*
Proportion of the total variance	0.301	0.257	0.214

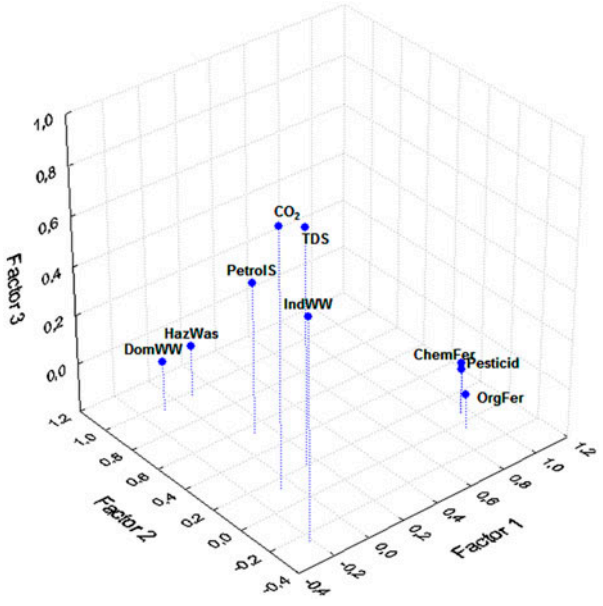

Fig. 8. Factor loadings, Factor 1 vs. Factor 2 vs. Factor 3—pressure variable (rotation: varimax normalized).

new axes and to obtain a pattern of variable loadings on each factor.

Table 10 and Fig. 8 present the three-factor rotated solution with the cross-loadings of their classified variables. The first factor represents 31.37% of the total variance. It contains inter-correlated observed variables which are pesticides, chemical, and organic fertilizers. This underlying factor explains the determinants of rural pollution due to the intensive agriculture.

The second factor represents 27.18% of total variance and has two inter-correlated variables which are domestic wastewater and hazardous waste. This latent factor represents the urban activities as a driver to improve the urban pollution.

The third factor has 18.74% of the total variance and contains three variables which are TDS, industrial wastewater and CO_2. This underlying factor represents the emissions from transport sector and other industrial facilities as a source of global warming and climate change.

Comparing the results of the FA with the PCA results for the pressure variables (Table 9), two variables were introduced to factor 3 with factor loadings greater than 0.7, the variables are TDS and CO_2.

In comparison with the PCA results for the pressure variables (Table 9), the FA introduced a new important determinant of pollution sources which are TDS and CO_2.

4. Conclusion

This study introduces a novel methodology to develop a relationship between pressure indicators and TDS of stream water based on cause–effect relationship to define in the first time, the most effective pressing pollution source on stream water quality and monitoring besides the geographical areas under pollution stresses on objective scientific basis.

Given the differences between the results of the ANN model and the expert opinion about the significance and priority of pressure variables, the research output assists water decision-makers and planners to gain better knowledge and understanding of the actual baseline conditions that ensure the success of undertaking management response measures.

Defining and prioritizing the pollution determinants of stream water quality degradation assist water managers to devise proactive and proper water pollution control measures with the objective of protecting stream water. This strengthens the preventive approach and mainstream environmental sustainability into groundwater management.

The selection of the optimal model configurations for a pressure indicators model using different ANNs was investigated. The results obtained in this study indicate that MLP network (BFGS 107) proved to be the best ANN structure showing that domestic wastewater is the most pressing pollution source on stream water quality followed by industrial wastewater. The selected and prioritized variables assist water managers and planners to introduce cheap proactive- and preventive-based water management policy measures in place of the existing expensive engineering-based water protection actions. Focus should be given to domestic wastewater and industrial wastewater.

The PCA supported by FA helped extract and identifies the different latent pollution sources responsible for variation in stream water quality at three different groups. The result of the PCA\FA indicated that the parameters responsible for stream water quality variation were mainly related to industrial pollution, urban, and rural pollution. Since ANN models and multivariate techniques are easily applied to water quality modeling, using them can be a practical approach to environmental impact assessment.

Abbreviation

TDS—Total dissolved solids measured by mg l^{-1}

The pressure indicators are as follows:

(1) HazWas: Hazardous wastes measured in ton - day^{-1}

(2) DomWW: Generation of domestic wastewater measured by $hm^3\ year^{-1}$
(3) Pesticid: Pesticides measured in tons.year-1
(4) ChemFer: Chemical fertilizers measured in tons $year^{-1}$
(5) OrgFer: Organic fertilizers measured in ton $year^{-1}$
(6) PetrolS: Petrol stations in numbers
(7) IndWW: Industrial wastewater measured by $hm^3\ year^{-1}$
(8) CO_2: Carbon dioxide measured in ppm

References

[1] D.J. Chen, J. Lu, S.F. Yuan, S.Q. Jin, Y.N. Shen, Spatial and temporal variations of water quality in Cao-E River of eastern China, J. Environ. Sci. 18 (2006) 680–688.
[2] Y.F. Li, G.H. Song, Y.G. Wu, W.F. Wan, M.S. Zhang, Y.J. Xu, Evaluation of water quality and protection strategies of water resources in arid-semiarid climates: A case study in the Yuxi River Valley of Northern Shaanxi province, China, Environ. Geol. 57(8) (2009) 1933–1938.
[3] H. Xu, L.Z. Yang, G.M. Zhao, J.G. Jiao, S.X. Yin, Z.P. Liu, Anthropogenic impact on surface water quality in Taihu lake region, China, Pedosphere 19(6) (2009) 765–778.
[4] R.S. King, M.E. Baker, D.F. Whigham, D.E. Weller, T.E. Jordan, P.F. Kazyak, M.K. Hurd, Spatial considerations for linking watershed land cover to ecological indicators in streams, Ecol. Appl. 15(1) (2005) 137–153.
[5] Blue Plan, Results of the Fiuggi Forum on "Advances of water demand management in the Mediterranean", Findings and recommendations. Blue Plan, Sophia Antipolis, 2003, p. 30.
[6] B.M. Dowd, D. Press, M. Los Huertos, Agricultural non-point source water pollution policy: The case of California's Central Coast, Agric. Ecosyst. Environ. 128 (2008) 151–161.
[7] F. Khelfaoui, D. Zouini, L. Tandjir, Quantitative and qualitative diagnosis of water resources in the Saf-Saf river basin (north east of Algeria), Desalin. Water Treat. 52 (2014) 2017–2021.
[8] Z. Wang, Q. Wu, Y. Zhang, J. Cheng, Confined groundwater pollution mechanism and vulnerability assessment in oilfields, North China, Environ. Earth Sci. 64 (2011) 1547–1553.
[9] H. Raman, V. Chandramouli, Deriving a general operating policy for reservoirs using neural network, J. Water Resour. Plann. Manage. 122(5) (1996) 342–347.
[10] M. Leket, Guiresse, J.L. Giraudel, Predicting stream nitrogen concentration from watershed features using neural networks, Water Res. 33 (16) (1999) 3469–3478.
[11] C.W. Wen, C.S. Lee, A neural network approach to multiobjective optimization for water quality management in a river basin, Water Resour. Res. 34(3) (1998) 427–436.
[12] L.L. Rogers, F.U. Dowla, Optimization of groundwater remediation using artificial neural networks with parallel solute transport modeling, Water Resour. Res. 30(2) (1994) 457–481.
[13] C.L. Adamus, M.J. Bergman, Estimating nonpoint source pollution loads with a GIS screening model, J. Am. Water Resour. Assoc. 31(4) (1995) 647–655.
[14] R.C. Smith, S.E. Stammerjohn, K.S. Baker, Surface air temperature variations in the western Antarctic Peninsula region. in: E.E. Hofmann, R.M. Ross, L.B. Quetin (Eds.), Foundations for Ecological Research West of the Antarctic Peninsula, AGU, Washington, 1996, pp. 105–121.
[15] F.C. James, C.E. McCulloch, Multivariate analysis in ecology and systematics: Panacea or pandora's box?, Annu. Rev. Ecol. Syst. 21 (1990) 129–166.
[16] 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.
[17] Environmental Indicators: Development Measurement and Use, Organization for Economic Co-operation and Development, Paris, 2003.
[18] D.R. Helsel, R.M. Hirsch, Statistical Methods in Water Resources US Geological Survey, Science Publishing Company, Reston, 1992, pp. 17–59.
[19] A.W. Minns, M.J. Hall, Artificial neural networks as rainfall–Runoff models, Hydrol. Sci. J. 41(3) (1996) 399–417.
[20] D. Patterson, Artificial Neural Networks: Theory and Applications, Prentice Hall, Singapore, 1996, pp. 141–179.
[21] H.R. Maier, G.C. Dandy, Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications, Environ. Modell. Software 15(1) (2000) 101–124.
[22] World Meteorological Organization, Inter-comparison of conceptual models used in operational hydrological forecasting, WMO, Technical series, Water Resour. Res. 27(9) (1975) 2415–2450.
[23] J.E. Nash, J.V. Sutcliffe, River flow forecasting through conceptual models part I—A discussion of principles, J. Hydrol. 10 (1970) 282–290.
[24] A. Hani, S. Lallahem, J. Mania, L. Djabri, On the use of finite difference and neural network models to evaluate the impact of underground water overexploitation, Hydrol. Processes 20 (2006) 4381–4390.
[25] S. Lallahem, J. Mania, A. Hani, Y. Najjar, On the use of neural networks to evaluate groundwater levels in fractured media, J. Hydrol. 307 (2005) 92–111.
[26] S. Riad, J. Mania, L. Bouchaou, Y. Najjar, Predicting catchment flow in a semi-arid region via an artificial neural network technique, Hydrol. Processes 18 (2004) 2387–2393.
[27] S. Lallahem, J. Mania, A nonlinear rainfall–runoff model using neural network technique: Example in fractured porous media, Math. Comput. Modell. 37 (2003a) 1047–1061.
[28] S. Lallahem, J. Mania, Evaluation and forecasting of daily groundwater outflow in a small chalky watershed, Hydrol. Processes 17(8) (2003b) 1561–1577.

[29] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning internal representations by error propagation. in: Rumelhart DE, McClelland J.L., The PDP Research Group (Eds.), Paralled Distributed Processing. Explorations in the Microstructure of Cognition. 1: foundations, The MIT Press, Cambridge 1986, pp. 318–362.
[30] K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators, Neural Networks 2 (1989) 359–366.
[31] 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.
[32] J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P.Y. Glorennec, H. Hjalmarsson, A. Juditsky, Nonlinear black-box modeling in system identification: A unified overview, Automatica 31(12) (1995) 1691–1724.
[33] R.R. Rongrui Xiao, V. Chandrasekar, Development of a neural network based algorithm for rainfall estimation from radar observations, IEEE Trans. Geosci. Remote Sens. 35 (1997) 160–171.
[34] F.R. Burden, F.R. Burden, R.G. Brereton, Cross-validatory selection of test and validation sets in multivariate calibration and neural networks as applied to spectroscopy, Analyst 122(10) (1997) 1015–1022.
[35] G. Castellano, A.M. Fanelli, Variable selection using neural-network models, Neurocomputing, 31(1–4) (2000) 1–13.
[36] S. Jalala, A. Hani, I. Shahrour, Characterizing the socio-economic driving forces of groundwater abstraction with artificial neural networks and multivariate techniques, Water Resour. Manage. 25 (2011) 2147–2175.