Modeling the Relationship between Catchment Attributes and In-stream Water Quality using Artificial Neural Networks

Iman Fatehi

M.Sc. Graduate in Environmental Sciences graduated from University of Tehran

Department of Environmental Science, Faculty of Natural Resources, University of Tehran, Karaj, Iran.

Tel: +98-915-3098326, Email: iman.fatehi@ut.ac.ir

Bahman Jabbarian Amiri*

Assistant Professor

Department of Environmental Science, Faculty of Natural Resources, University of Tehran, Karaj, Iran.

Tel: +98-261-2229721, Fax: +98-261-2229721, Email: jabbarian@ut.ac.ir

* Corresponding author

Afshin Alizadeh

Assistant Professor

Department of Environmental Science, Faculty of Natural Resources, University of Tehran, Karaj, Iran.

Tel: +98-261-2245908, Fax: +98-261-2245908, Email: alizadeh@ut.ac.ir

Jan Adamowski

Assistant Professor

Faculty of Agricultural and Environmental Sciences, McGill University, Ste-Anne-de-Bellevue, Québec H9X 3V9

Tel.: 514-398-7786, Email: jan.adamowski@mcgill.ca

Abstract

The physical attributes of river catchment have a critical influence on chemistry and physical features of in-stream water quality. Therefore, modeling this relationship is crucial to make more punctual management strategies to improve regional water quality. In this paper artificial neural networks (ANN's) are developed to model the relationship between land use/cover, in association with other physical attributes of the catchment, i.e. geological permeability and hydrological soil groups which are used only in few studies in advance, and in-stream water quality parameters (i.e. K^+, Na^+, Mg^{2+}, Ca^{2+}, SO_4^{2-}, Cl^-, HCO_3^-, SAR, pH, EC, TDS) in 88 selected catchments in southern basins of Caspian Sea. To enhance the ANN's architecture, backward elimination multiple linear regressions are developed, which optimize ANN's input nodes by selecting the most correlated variables. A transformation approach is also used to qualify ANN's performance in four quality classes from unsatisfactory to very good. Results showed applying backward method most significant contribution was to TDS model performance, from unsatisfactory to very good. However, pH model performance decreased from very good to satisfactory. Moreover, between all catchment attributes urban areas have the greatest impact on K^+, Na^+, Mg^{2+}, Cl^- and SO_4^{2-}, EC and SAR concentration values. Agricultural areas also had the greatest impact on K^+, TDS and EC. Bare land areas have the greatest impact on Na^+, Ca^{2+} and HCO_3^-. Developed ANN's qualifying approach which is used in this study, showed the most of developed models have "very good" ratings and are reliable to be used practically.

Keywords:

Land cover · Water quality · Soil hydrological groups · Geological permeability · ANN · Linear regression

1- Introduction

Effective integrated management of water resources requires consideration of all the factors that may affect the quantity and quality of water and understanding of the processes involved (Khalil and Adamowski 2013). For instance, in-stream water chemistry is affected by many natural and anthropogenic sources (Amiri et al. 2012), which can be divided according to their spatial extent into point and diffuse sources. Diffuse pollutants are becoming a serious threat to water quality in streams due to land cover and rapid changes in land use (Basnyat et al. 2000).

Recently, rising concern about the condition of water resources has led to an increase in studies of the ecological impact of anthropogenic practices that affect in-stream water. Accordingly, since the 1970s many studies have been conducted on water quality, and particularly on the impacts of land use and land cover (LULC) change, due to the crucial influence of LULC on hydrological processes in catchments (Kalin et al. 2010). For instance, Wilcock (1986) had studied the impact of agricultural runoff as a source of water pollution in New Zealand. Since then, numerous studies have been conducted on the impacts of LULC change on water quality parameters, e.g. Amiri and Nakane 2009a; Amiri et al. 2012; Liden and Arheimer 1988; Miller et al. 2011; Smith and Policy 1993; Tong and Chen 2002; Wan et al. 2014; Williams et al. 2001; Zhou et al. 2012; Obade et al. 2014. The most of these studies concluded that urban and agricultural areas had the greatest negative impacts on water quality.

Other physical aspects of the catchment, such as soil and geological features, can also affect water chemistry. In few studies catchment soil and lithological features are associated with LULC, e.g. Haidary et al. 2013; Hartmann et al. 2014; Pratt and Chang 2012; Reimann et al. 2009; Ryu et al. 2007; Yang and Jin 2010. It is While, fewer studies encompassed both the effects of these physical features and of LULC on the chemistry of in-stream water, e.g. Haidary et al. (2014) or the study of Sangani et al. (2014). Although, numerous features of soil affect runoff potential, e.g. texture, structure, mineral and organic elements, but Hydrological soil groups (HSGs) are one element can be used in determining runoff curve numbers, which is used in this study. HSGs which are A, B, C and D represent the minimum infiltration rate for bare soil after prolonged wetting, while A has the highest runoff potential, and D has the lowest (USDA 1986). As the result of LULC changes, soil profile considerable alters and in this circumstance soil texture of the new surface soil can be used to determine the HSGs, according to table 1. (Brakensiek and Rawls 1983). Geological features of catchment are also used in this study, which are classified according to their permeability. There are three classes of geological permeability i.e. Low, Medium and High, which are related to many attributes of geological formations, such as effective porosity, type and size of cavities and their connection, rock density, pressure gradient and features of the fluid, such as its viscosity. In this paper, soil and geological features of catchments are transformed to HSG and geological permeability classes in association with LULCs, this composition of physical attributes of catchment is applying for the first time in developing hydrological models.

Table 1

2- Methodology

2-1- Backward selection

When there are few candidate covariates (N), one can select a relevant model on the basis of a reasonable criterion e.g. mean squared error (MSE), coefficient of determination (R^2), sum of squared errors (SSE), final prediction error (FPE) or cross-validation error) for all initial subsets of independent variables. However, the greater number of candidate covariates, causes the greater computational capacity of the approach. So it is why step-by-step methods are popular (Noori et al. 2010). Linear regression models can be used to select the most correlated variables. In backward elimination methods as well as enter approach, all independent variables are initially entered into the model, then impact of each variable elimination is evaluated, It is while, stepwise and forward methods involve entering the most correlated independent variables into the model at each step and evaluating the addition impact of each variable using a chosen model comparison criterion until none improves the model. (Efroymson 1960).

2-2- Artificial Neural Networks

The non-linear behavior of ecosystems cannot be effectively modeled by conventional linear methods. ANNs are parametric models, which are generally considered to be lumped (Dawson and Wilby 2001). In this study, ANNs are developed to determine the relationship between water quality parameters and the selected most correlated catchment physical variables. Recently, multi-layer perceptron (MLP) feed forward networks have become a popular ANN architecture (Maier et al. 2010), this three layer ANN models can be described as follows:

If n is the number of input neurons (i.e. catchment physical attributes), h is the number of hidden neurons ($z_1, \ldots, z_h$), and m is the number of output neurons (i.e. water quality parameters, e.g. SAR, $K^+ \ldots, TDS$), which is one for each model in this study. i, j, and k indices represent the input, hidden, and output layers respectively. τ_j is the bias for neuron z_j, ϕ_k is the bias for neuron y_k and w_{ij} is the connection weight between neuron x_i and neuron z_j, and β_{jk} is the connection weight between neuron z_j and y_k. The calculation function of the ANN network is:

$$Y_K = g_{A(\sum_{j=1}^{h} z_j \beta_{jk} + \varphi_k)} \quad (1)$$

$$z_j = f_{A(\sum_{i=1}^{n} x_i w_{ij} + \tau_j)} \quad (2)$$

where g_A and f_A are activation (transfer) functions, that are usually continuous, bounded, and non-decreasing (Amiri et al. 2012).

2-2-1- Model Performance

Five statistical coefficients are used to measure modeling performance, including the coefficient of determination (R^2), bias ratio (R_{BIAS}), Nash-Sutcliffe efficiency (E_{NASH}), normalized mean square error (NMSE) and RMSE-observations standard deviation ratio (RSR). The coefficient of determination used to measure the linear quantitative variables is:

$$R^2 = \left(\frac{n \sum O_i S_i - (\sum O_i)(\sum S_i)}{\sqrt{n(\sum O_i^2) - (\sum O_i)^2} \sqrt{n(\sum S_i^2) - (\sum S_i)^2}} \right)^2 \quad (3)$$

Where n is the number of data points, O is the observed data and S is the simulated outputs. The degree of over- or under-prediction of the model forecast can be measured by the bias ratio as follows:

$$R_{BIAS} = 100\frac{\sum(S_i - O_i)}{\sum O_i} \tag{4}$$

A negative value of R_{BIAS} shows under-prediction and positive values show over-prediction (Salas et al. 2000).

The Nash–Sutcliffe efficiency (E_{NASH}) is a common statistic for assessing the forecasting power of hydrological and environmental models (Nash and Sutcliffe 1970), which is expressed as;

$$E_{NASH} = 1 - \frac{\sum(O_i - S_i)^2}{\sum(O_i - O')^2} \tag{5}$$

Where O' is the mean of observed data. This statistic can be measured from –∞ to 1, where 1 represents a perfect model.

While the MSE statistic is used as a criterion for selecting the optimal model architectures, the NMSE is used as a mean for assessing the model's performance. Contrary to the bias statistic, in the NMSE the deviations are summed, so it can show the most significant differences between models. Using NMSE for each water quality parameter can minimize the effect of sample numbers and the range of measurements (Kalin et al. 2010). NMSE is calculated as follows:

$$NMSE = \sum_{j=1}^{m} \frac{1}{(n_j)^2} \sum_{i=1}^{n_j} \left(\frac{S_{j,i} - O_{j,i}}{\bar{O}_j}\right)^2 \tag{6}$$

Where m represents the total number of catchments, n_j is the number of data in catchment j and $\bar{O}_j$ is the total average of observed values.

RMSE is one of the most common error index statistics (Chu and Shirmohammadi 2004; Vazquez-Amabile and Engel 2005) and lower values show higher model performance. Singh et al. (2004) developed an evaluation statistic, which is based on RMSE and the standard deviation of observed data, called the RMSE-observations standard deviation ratio (RSR). For this statistic, the observation standard deviation is used to standardize the RMSE. RSR can be calculated as follows:

$$RSR = \frac{RMSE}{STDEV_{obs}} = \frac{\left[\sqrt{\sum_{i=1}^{n}(O_i - S_i)^2}\right]}{\left[\sqrt{\sum_{i=1}^{n}(O_i - \bar{O})^2}\right]} \tag{7}$$

Where O_i and S_i are observed and simulated values respectively, and $\bar{O}$ is the mean of observation values. RSR values vary between 0 as the optimal value, which refers to zero RMSE or residual variation, and positive values (Moriasi et al. 2007).

Akaike's information criterion (AIC) and Bayesian information criterion (BIC) information-based statistics are commonly used in the literature to compare ANN architectures, and identify the optimum (Kalin et al. 2010; Qi and Zhang 2001; Ren and Zhao 2002; Zhao et al. 2008). In this study AIC and BIC are used to evaluate the impact of linear regression approach. AIC and BIC are calculated as follows:

$$AIC = \log(\sigma_{MLE}^2) + \frac{2m}{n} \quad if \quad \frac{n}{(m+1)} \geq 40 \tag{8a}$$

$$AIC = \log(\sigma_{MLE}^2) + \frac{2m}{(n-m-1)} \quad if \quad \frac{n}{(m+1)} < 40 \tag{8b}$$

$$BIC = \log(\sigma_{MLE}^2) + \frac{m \log(n)}{n} \tag{9}$$

Where n and m are the number of data and the number of model parameters respectively and (σ_{MLE}^2) is the MSE target and simulated value.

Network architecture has a crucial role in ANN performance, and can be optimized by finding the best network functions and the optimum size of hidden layer nodes. ANN performance is also highly affected by the quality of network training. The fundamental goal of this process is to identify a set of weights and threshold values that minimize the predefined error function by decreasing the gap between the ANN outputs and the target values (Committee 2000).

3- Case Study

3.1- Study area and materials

This study is conducted in catchment of the Caspian Sea in north of Iran, which is about 174 618 square kilometers in area at (49º48' and 54º41') longitude and (35º36' and 37º19') latitude. The majority of the area (65.10%) is covered by forests, while the remainder is covered by rangelands (24.41%), agricultural land (9.41%), urban land (0.88%), water bodies (0.0126%) and bare land (0.186%). First, 108 water quality stations distributed throughout the Caspian Sea catchment are selected and analyzed. Digital elevation models (DEM) at a 30m × 30m resolution obtained from the USGS database is used to delineate the upstream catchment boundaries. User digitizing technique is also used to enhance the boundaries. To consider critical impact of catchment size on hydrological turnover in modeling, macro size catchments i.e. >1000 km^2, which included 18 catchments are eliminated from the process.

Water quality parameters, including K^+, Na^+, Mg^{2+}, Ca^{2+}, SO_4^{2-}, Cl^-, HCO_3^-, Sodium Adsorption Ratio (SAR), pH, electrical conductivity (EC) and total dissolved solids (TDS)) are obtained from the Iran Water Resource Management company (WRMC) (http://www.wrm.ir), which are sampled on a monthly basis. Sampling process and devices are conformed to WRMC Guidelines for Surface Water Quality Monitoring (2009) and EPA-841-B-97-003 standards (Dohner et al. 1997). For statistical analysis five-year means (1998-2002) of water quality data are calculated. Statistical features of water quality data are provided in Table 2. Then data is statistically analyzed to check for normality and outliers, which resulted in outlier records elimination, leaving 88 final stations. The study area is shown in Figure 1.

The used LULC map is created using a 2002 digital LULC map (Scale 1:250 000) obtained from the Forest, Ranges and Watershed Management Organization of Iran (http://frw.org.ir). LULCs categorized in six classes, including; bare land, water body, urban, agriculture, rangeland and forest. Digital geological and soil feature maps (1:250,000) are also obtained from the Geological Survey of Iran (www.gsi.ir). Physical characteristics of selected catchments and their statistical features are represented in Table 3.

Figure 1

Table 2

Table 3

3.2- Methodology implementation

3-2-1- Backward selection

As stated above there are four most common linear regression methods, to evaluate their performance and choose the best linear selection method for this study, linear regressions developed using IBM SPSS Statistics the software. And the backward approach, because of performance criterion value i.e. greater R-square[1] and wider range of variable selection is selected as the pre-processing method. To evaluate the impact of applying linear regression on developed ANNs' performance, ANNs are also developed using all catchment physical. Then performance indexes are used to compare them, results are available in Table 7.

3-2-2- ANN architecture enhancement

The most suitable training function would have the best performance in weight matrix optimization (White 1989). In this study, the most suitable training function for each ANN model is selected between numerous transfer and training functions in MATLAB. Fifteen trials are conducted to identify the best training function for each developed ANN by choosing the function which resulted in minimum model RMS. The general architecture of the ANN used in this study is shown in Figure 2.

Figure 2

ANNs are developed containing two layer feed-forward network with sigmoid hidden neurons and linear output neurons, i.e. a function fitting neural network. The number of hidden layer nodes is one of the most critical aspects of a multilayer feed-forward network, while there is no general rule to determine the optimum number of hidden layer nodes (Committee 2000). However, Hecht-Nielsen (1987) proposed an equation to determine the upper limits of the optimum number of hidden layer nodes:

$$N^H \leq 2N^I + 1 \tag{10}$$

[1] Coefficient of determination

While N^H is the number of hidden layer nodes and N^I is the number of input layer nodes. In this study, upper limit for the size of hidden layer nodes is calculated for each developed ANN based on the Hecht-Nielsen equation. Then optimum hidden layer size for each ANN is evaluated based on modeling performance lowest mean MSE using a trial and error approach over fifteen trials for each size.

Each ANN requires three sample categories; training, validation and testing. In this study, 62 samples (70%) are selected as training samples. These samples are presented to ANN during the training process, and ANNs are adjusted according to their errors. 13 samples (15%) are selected as validation samples, which are used to measure network generalization. Training is completed when the generalization stopped improving. The remaining samples (13 samples; 15%) had no effect on training and hence are used as test samples to provide an independent measure of network performance during and after training (Srivastava et al. 2006).

3.3- Results and discussion

The linear selection approach is used to evaluate the most correlated physical variables of catchments, i.e. LULCs, HSGs and geological permeability classes, for each water quality developed ANN. Results of this method are shown in Table 4. Table 4 also shows parameter sensitivities using NMSE and R^2 statistics. The regression weights for the selected variables are given in descending order.

Table 4

The optimal architecture of developed ANN (i.e. number of input and hidden layer nodes and also training functions) for each water quality parameter is selected based on equation 10 and also NMSE, AIC and BIC values, using a trial - error process. Results are shown in Table 5. The number of hidden layer nodes ranged from four to fourteen, and three training functions are selected for ANNs training. Figure 3 indicates the mean NMSE of developed ANNs for each size hidden layer nodes. The optimal number of nodes is selected according to the lowest model NMSE. Number of epochs varies from 15 to 1000. Greatest number of epochs occurred for SAR using random order incremental training function. With regards to the ANN performance metrics, the NMSE values for EC and TDS are really high, but for other parameters NMSE values ranged from 0.0001 for K^+ to 0.0658 for pH.

Figure 3

Table 5

Results generally show that ANN validation R^2 is above 0.64 for all water quality parameters except Na^+, for which is 0.405. NMSE is a better metric to compare the performance of different models. EC and TDS resulted in very high NMSE values, which is due to wide observed data range values. Of the other parameters, Na^+ had the highest NMSE value (0.0795). The R^2 and NMSE values for each ANN sample categories i.e. training, validation and test are shown in Table 6.

Table 6

In this study, performance of developed ANNs are assessed by calculating statistical indexes i.e. NMSE, R^2, RSR, E_{NASH}, R_{BIAS}, AIC and BIC statistics, and in this paper as only few ones in-advance, a transformation approach is also used to qualify ANN's performance in four quality classes from unsatisfactory to very good, which is shown in Table 7. There is no general established method to qualify and classify hydrological modeling performance (Kalin et al. 2010). But Moriasi et al. (2007) suggested a quality rating method for catchment models with a monthly time scale, using RSR, E_{NASH} and R_{BIAS} performance statistics. Therefore, developed ANNs performance are rated using this approach (Table 8). ANN simulated versus observed values for each water quality parameter are illustrated in Figure 4. As it is obvious in Figure 4, there is "heteroscedasticity" between K^+ observed and simulated values, which shows different variances, therefore, six variance stabilization approaches i.e. log, SQR, positive Poisson, negative Poisson, inverse and binomial are implemented to transform observed data and eliminate the heteroscedasticity, but all resulted in no significant change.

Table 7

Table 8

Figure 4

4- Conclusion

The Objective of current study is to model the impact of catchment physical attributes (i.e. hydrological soil groups and geological permeability classes in association with catchment LULC) on water quality parameters (i.e. SAR, K^+, Na^+, Mg^{2+}, Ca^{2+}, SO_4^{2-}, Cl^-, HCO_3^-, pH, EC and TDS). ANNs are developed using physical attribute data of 88 selected catchments in Caspian Sea basin in the north of Iran. Linear regression backward method is also applied to optimize the size of ANN input layer nodes by selecting the most correlated variables for each. Results showed EC and TDS have high standard deviation values, which is related to wide range of input values. Based on Table 7, applying linear regression approach improved the ANNs performance quality ratings from satisfactory to very good for K^+, from unsatisfactory to very good for Ca^{2+}, from good to very good for EC and from unsatisfactory to very good for TDS. Although, in case of pH, the performance rate decreased from very good to satisfactory.

Results of the linear regression approach demonstrated that forest has no direct association with water pollution, it is while in some cases enhances water quality. It is concurs with the results of Williams et al. (2001) and also Tong and Chen (2002). According to Table 4, between all catchment physical variables, urban and agricultural land uses have the greatest negative impacts on water chemistry, which is concurs with the results of Wilcock (1986), Williams et al. (2001) and also Tong and Chen (2002). According to results, urban land use has the greatest impact on K^+, Na^+, Mg^{2+}, Cl^-, SO_4^{2-}, EC and SAR concentration values. On the other hand, agricultural areas has the greatest impacts on K^+, EC and TDS values. It is while bare land areas has the greatest impacts on Na^+, Ca^{2+} and HCO_3^- concentration values. In this study, according to the results, between all hydrological soil groups, group, A has the greatest impact on water quality parameters, which is contrary to the results of Yang and Jin (2010), D hydrological group was the most effective one. Which can be attributed to high runoff potential also maximum infiltration rate for bare soil after prolonged wetting of A HSG resulting in higher runoff moving minerals and pollutants into water bodies and.

To compare the developed ANNs performance, a qualitative rating approach is applied (Table 8). Applying linear regression approach, resulted in significant enhancement in the quality rank of developed ANNs performance for most water quality parameters. It resulted in "very good" quality rank for the majority of water quality parameters. The most significant contribution of applying backward method is to TDS model performance, from unsatisfactory to very good. However, pH model performance decreased from very good to satisfactory (Table 7).

The limitation of this study was in data shortage or unavailability for some biological water quality parameters, which could be used in modeling e.g. NO_3, DO, BOD. To an integrated water quality modeling a wider range of physical variables can also be used, i.e. climate and hydrological parameters; temperature, precipitation and flow discharge. But in this paper the main focus was on the physical features of the catchment. The developed ANNs can be implemented to estimate the water quality in a specific catchment by introducing the catchment physical attributes to the model. Although, it is recommended to use the developed ANNs in catchments < 1000 km^2 and also with similar environmental conditions to Caspian Sea basin in the north of Iran.

Refrences

Amiri B, Nakane K (2009a) Comparative prediction of stream water total nitrogen from land cover using artificial neural network and multiple linear regression approaches Polish Journal of Environmental Studies 18:151-160

Amiri B, Sudheer K, Fohrer N (2012) Linkage between in-stream total phosphorus and land cover in Chugoku district, Japan: an ANN approach Journal of Hydrology and Hydromechanics 60:33-44

Azyana Y, Norulaini NN The Entire Catchment and Site Buffer Radii Landscape Variables, Urban Land Use as Predictors of Water Quality Variation

Basnyat P, Teeter L, Lockaby BG, Flynn KM (2000) RESEARCH: Land Use Characteristics and Water Quality: A Methodology for Valuing of Forested Buffers Environmental Management 26:153-161 doi:10.1007/s002670010078

Brakensiek D, Rawls W (1983) Agricultural Management Effects on Soil-Water Processes. 2. Green and Ampt Parameters for Crusting Soils Transactions of the ASAE 26:1753-1757

Chu T, Shirmohammadi A (2004) Evaluation of the SWAT model's hydrology component in the Piedmont physiographic region of Maryland Transactions of the ASAE 47:1057-1073

Committee AT (2000) Artificial neural networks in hydrology. I: Preliminary concepts Journal of Hydrologic Engineering 5:115-123

Dawson C, Wilby R (2001) Hydrological modelling using artificial neural networks Progress in physical Geography 25:80-108

Dohner E, Markowitz A, Barbour M, Simpson J, Byrne J, Dates G (1997) Volunteer Stream Monitoring: A Methods Manual Environmental Protection Agency: Office of Water (EPA 841-B-97-003)

Efroymson MA (1960) Multiple regression analysis. Mathematical Methods for Digital Computers. Wiley,

Haidary A, Amiri B, Adamowski J, Fohrer N, Nakane K (2013) Assessing the Impacts of Four Land Use Types on the Water Quality of Wetlands in Japan Water Resour Manage 27:2217-2229 doi:10.1007/s11269-013-0284-5

Haidary A, Jabbarian Amiri B, Adamowski J, Fohrer N, Nakane K (2014) Modelling the relationship between catchment attributes and wetland water quality in Japan Ecohydrology:n/a-n/a doi:10.1002/eco.1539

Hartmann J, Moosdorf N, Lauerwald R, Hinderer M, West AJ Global chemical weathering and associated P-release. In: EGU General Assembly Conference Abstracts, 2014. p 15640

Hecht-Nielsen R (1987) Kolmogorov's mapping neural network existence theorem. Paper presented at the Proceedings of the First IEEE International Joint Conference on Neural Networks, San Diego, California,

Kalin L, Isik S, Schoonover JE, Lockaby BG (2010) Predicting Water Quality in Unmonitored Watersheds Using Artificial Neural Networks J Environ Qual 39:1429-1440 doi:10.2134/jeq2009.0441

Khalil B, Adamowski J (2013) Towards a Consistent Approach for the Assessment and Redesign of Surface Water Quality Monitoring Networks Irrigat Drainage Sys Eng 2:e113

Liden GE, Arheimer JS (1988) Effect of agricultural management systems on water quality J Prod ASgric 6:480-486

Maier HR, Dandy GC (2000) Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications Environmental modelling & software 15:101-124

Maier HR, Jain A, Dandy GC, Sudheer KP (2010) Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions Environmental Modelling & Software 25:891-909 doi:http://dx.doi.org/10.1016/j.envsoft.2010.02.003

Miller J, Schoonover J, Williard KJ, Hwang C (2011) Whole Catchment Land Cover Effects on Water Quality in the Lower Kaskaskia River Watershed Water Air Soil Pollut 221:337-350 doi:10.1007/s11270-011-0794-9

Moriasi D, Arnold J, Van Liew M, Bingner R, Harmel R, Veith T (2007) Model evaluation guidelines for systematic quantification of accuracy in watershed simulations Trans ASABE 50:885-900

Muttil N, Chau K-w (2006) Neural network and genetic programming for modelling coastal algal blooms International Journal of Environment and Pollution 28:223-238

Nash J, Sutcliffe J (1970) River flow forecasting through conceptual models part I—A discussion of principles Journal of hydrology 10:282-290

Noori R, Hoshyaripour G, Ashrafi K, Araabi BN (2010) Uncertainty analysis of developed ANN and ANFIS models in prediction of carbon monoxide daily concentration Atmospheric Environment 44:476-482 doi:http://dx.doi.org/10.1016/j.atmosenv.2009.11.005

Obade V, Lal R, Moore R (2014) Assessing the Accuracy of Soil and Water Quality Characterization Using Remote Sensing Water Resour Manag 28:5091–5109 doi:http://dx.doi.org/10.1016/j.atmosenv.2009.11.005

Olawoyin R, Nieto A, Grayson RL, Hardisty F, Oyewole S (2013) Application of artificial neural network (ANN)–self-organizing map (SOM) for the categorization of water, soil and sediment quality in petrochemical regions Expert Systems with Applications 40:3634-3648 doi: 10.1007/s11269-014-0796-7

Pratt B, Chang H (2012) Effects of land cover, topography, and built structure on seasonal water quality at multiple spatial scales Journal of Hazardous Materials 209–210:48-58 doi:http://dx.doi.org/10.1016/j.jhazmat.2011.12.068

Qi M, Zhang GP (2001) An investigation of model selection criteria for neural network time series forecasting European Journal of Operational Research 132:666-680 doi:http://dx.doi.org/10.1016/S0377-2217(00)00171-5

Reimann C, Finne TE, Nordgulen Ø, Sæther OM, Arnoldussen A, Banks D (2009) The influence of geology and land-use on inorganic stream water quality in the Oslo region, Norway Applied Geochemistry 24:1862-1874 doi:http://dx.doi.org/10.1016/j.apgeochem.2009.06.007

Ren L, Zhao Z (2002) An optimal neural network and concrete strength modeling Advances in Engineering Software 33:117-130 doi:http://dx.doi.org/10.1016/S0965-9978(02)00005-4

Ryu J-S, Lee K-S, Chang H-W (2007) Hydrogeochemical and isotopic investigations of the Han River basin, South Korea Journal of Hydrology 345:50-60 doi:http://dx.doi.org/10.1016/j.jhydrol.2007.08.001

Salas JD, Markus M, Tokar AS (2000) Streamflow Forecasting Based on Artificial Neural Networks. In: Govindaraju RS, Rao AR (eds) Artificial Neural Networks in Hydrology, vol 36. Water Science and Technology Library. Springer Netherlands, pp 23-51. doi:10.1007/978-94-015-9341-0_3

Sangani MH, Amiri BJ, Alizadeh AS, Sakieh Y, Ashrafi, S (2015) Modeling relationships between catchment attributes and river water quality in southern catchments of the Caspian Sea Environmental Science and Pollution Research 22:4985–5002 doi:10.1007/s11356-014-3727-5

Singh SK et al. (2004) Identification of human brain tumour initiating cells nature 432:396-401

Smith CM, Policy M (1993) Towards sustainable agriculture: freshwater quality in New Zealand and the influence of agriculture. MAF Policy,

Srivastava P, McNair JN, Johnson TE (2006) Comparison of process based and artificial neural network approaches for streamflow modeling in an agricultural watershed J Am Water Resour Assoc 66:377-393

Tong STY, Chen W (2002) Modeling the relationship between land use and surface water quality Journal of Environmental Management 66:377-393 doi:http://dx.doi.org/10.1006/jema.2002.0593

USDA S (1986) Urban hydrology for small watersheds Technical release 55:2-6

Vazquez-Amabile G, Engel B (2005) Use of SWAT to compute groundwater table depth and streamflow in the Muscatatuck River watershed Transactions of the ASAE 48:991-1003

Wan R, Cai S, Li H, Yang G, Li Z, Nie X (2014) Inferring land use and land cover impact on stream water quality using a Bayesian hierarchical modeling approach in the Xitiaoxi River Watershed, China Journal of Environmental Management 133:1-11 doi:http://dx.doi.org/10.1016/j.jenvman.2013.11.035

Wang XX, Chen S, Lowe D, Harris CJ (2006) Sparse support vector regression based on orthogonal forward selection for the generalised kernel model Neurocomputing 70:462-474 doi:http://dx.doi.org/10.1016/j.neucom.2005.12.129

Water Resources Management Company (WRMC) of Iran (2009) Guidelines for monitoring surface water quality variables. WRMC of Iran, Tehran.
White H (1989) Learning in Artificial Neural Networks: A Statistical Perspective Neural Computation 1:425-464 doi:10.1162/neco.1989.1.4.425
Wilcock, R. 1. 1986. 'Agricultural run-off a source of water pollution in New Zealand?', New Zealand Agricultural Science, 20,98-103.
Williams RA, Onsted CA, Bosch DD, Andrson WP (2001) Soil and water quality. Lewis pub., Boca Ration, FL.
Yang X, Jin W (2010) GIS-based spatial regression and prediction of water quality in river networks: A case study in Iowa Journal of Environmental Management 91:1943-1951 doi:http://dx.doi.org/10.1016/j.jenvman.2010.04.011
Zhao Z, Zhang Y, Liao H (2008) Design of ensemble neural network using the Akaike information criterion Engineering Applications of Artificial Intelligence 21:1182-1188 doi:http://dx.doi.org/10.1016/j.engappai.2008.02.007
Zhou T, Wu J, Peng S (2012) Assessing the effects of landscape pattern on river water quality at multiple scales: A case study of the Dongjiang River watershed, China Ecological Indicators 23:166-175 doi:http://dx.doi.org/10.1016/j.ecolind.2012.03.013

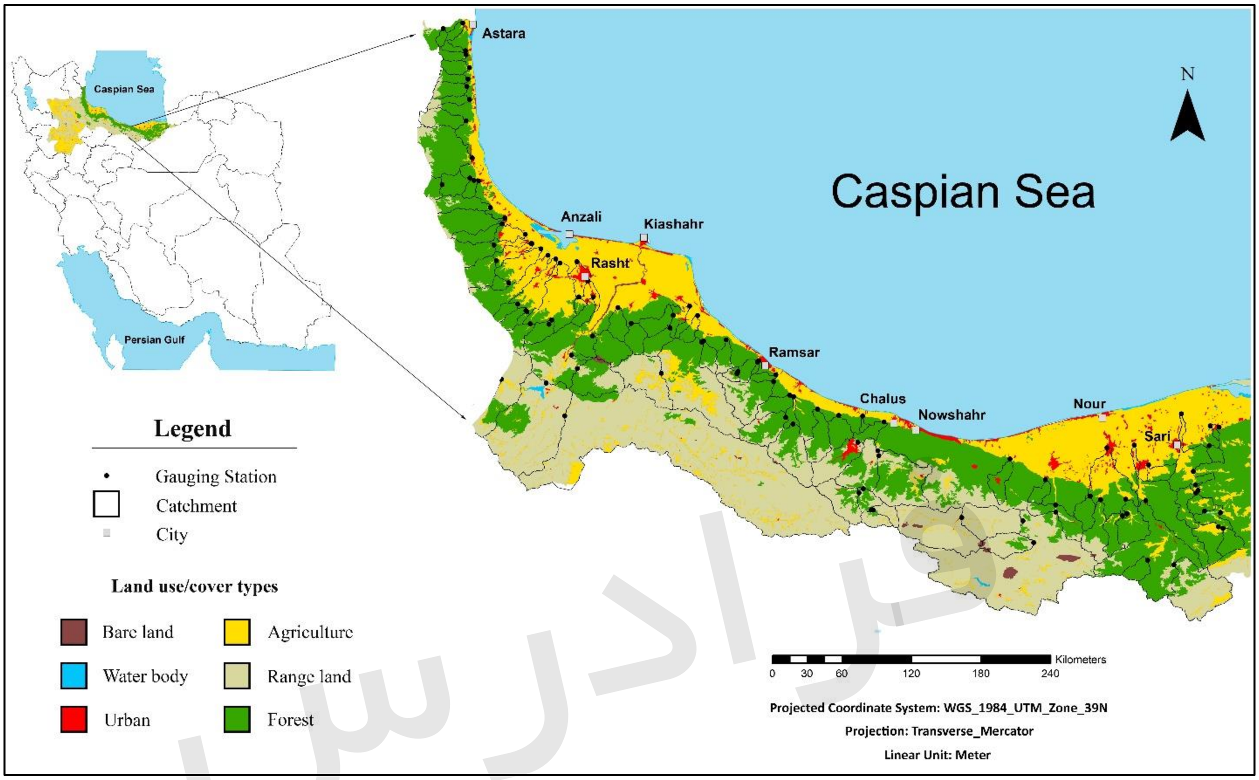

Fig. 1 Study area, which includes eighty-eight selected catchments in southern basin of the Caspian Sea.

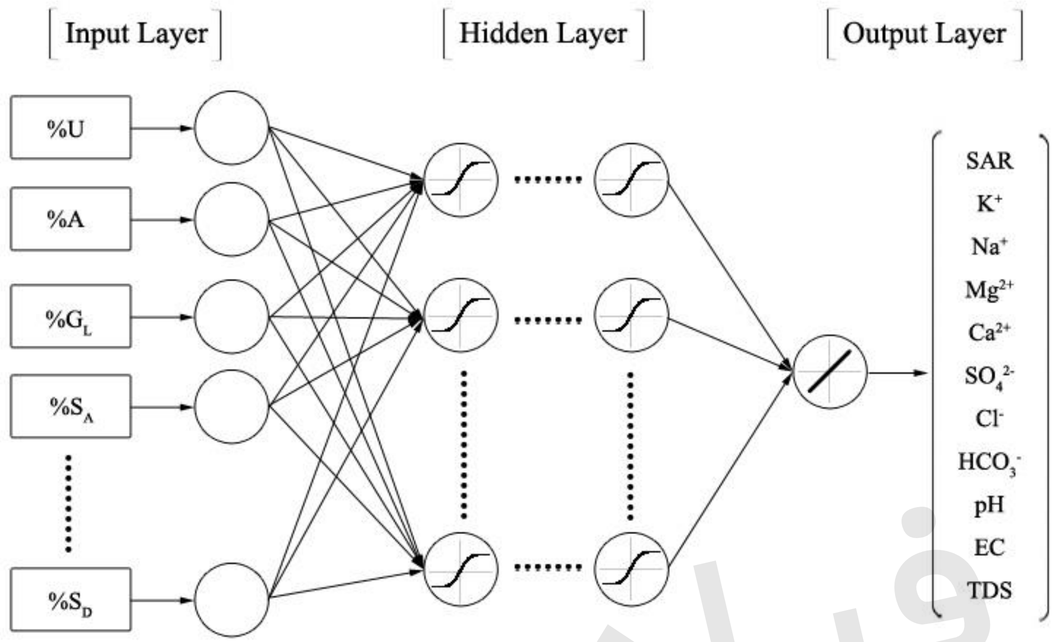

Fig. 2 General structure of an ANN model. See Table 2 for the full list of input variables.

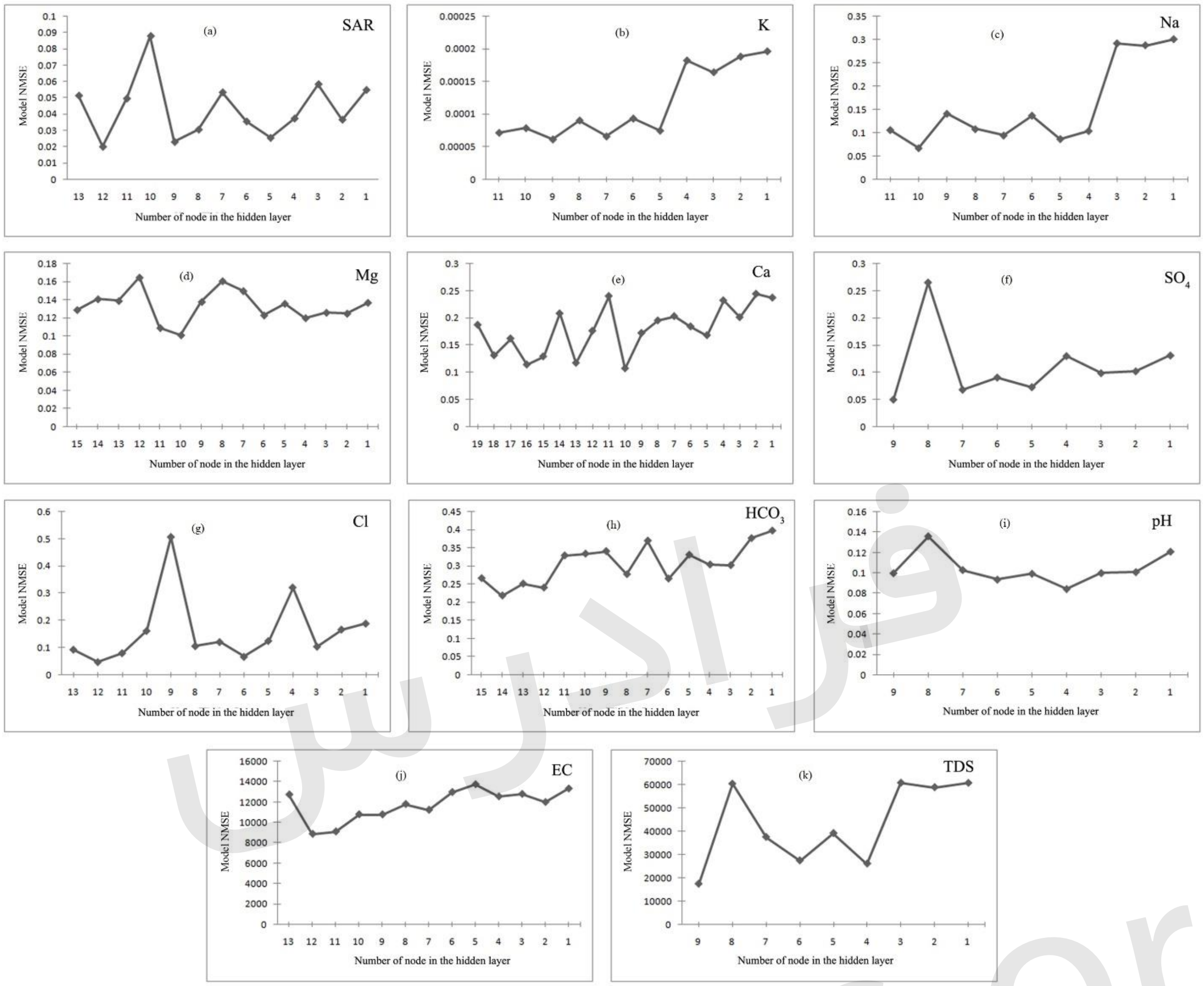

Fig. 3 The mean NMSE associated with using various numbers of nodes in ANNs hidden layer. A lower NMSE means less error in the model.

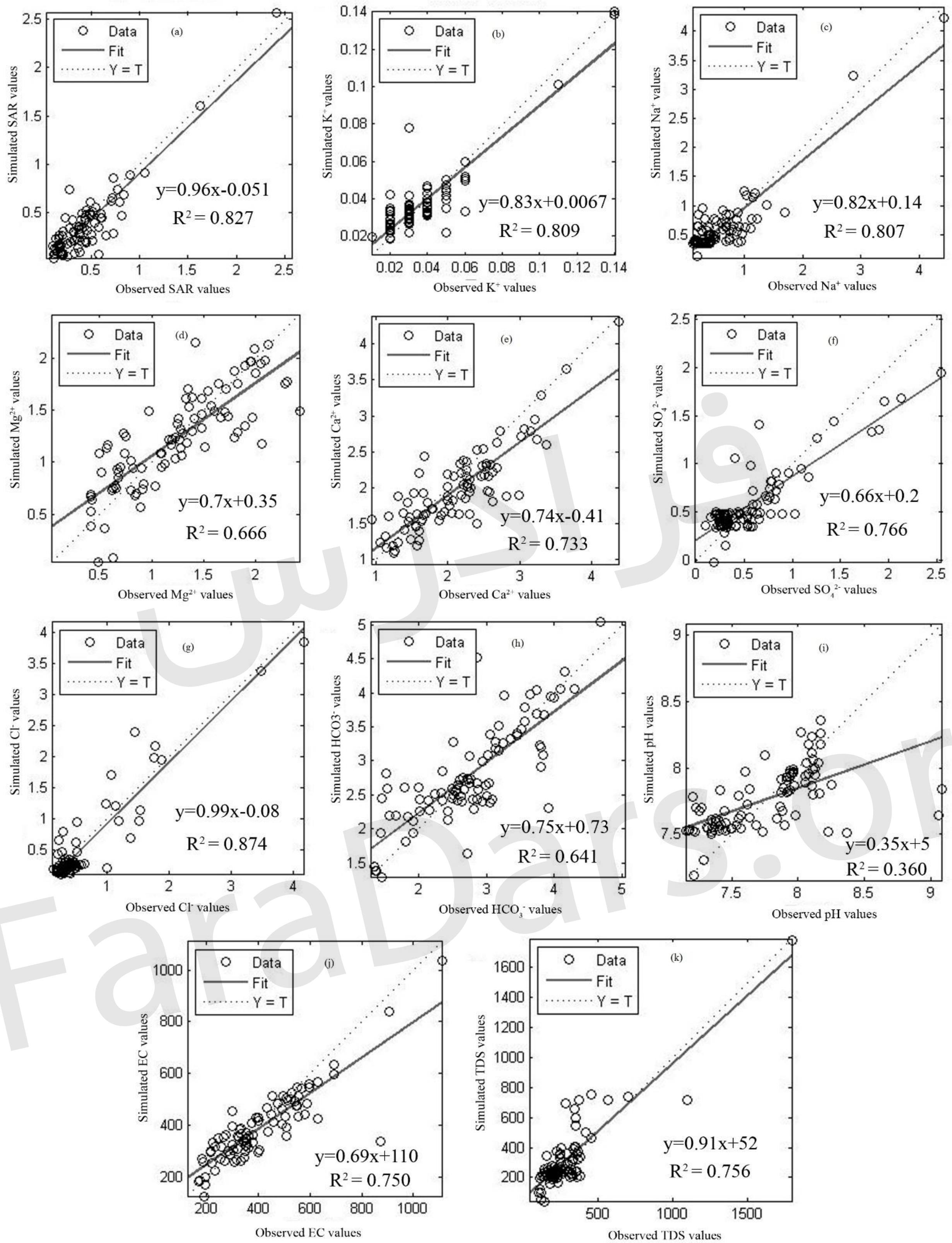

Fig. 4 Observed values versus the predicted values which are simulated using developed ANNs.

Table 1 Hydrological soil groups (HSGs) according to the surface soil texture

HSG	Soil texture
A	Sand, loamy sand, or sandy loam
B	Silt loam or loam
C	Sandy clay loam
D	Clay loam, silty clay loam, sandy clay, silty clay, or clay

Table 2 Statistical features of water quality parameters in selected stations.

	SAR	K^+	Na^+	Mg^{2+}	Ca^{2+}	SO_4^{2-}	Cl^-	HCO_3^-	pH	EC	TDS
Min.	0.1000	0.0100	0.1000	0.4100	0.9300	0.1000	0.1300	1.3500	7.1500	172.0500	108.8900
Max.	2.4200	0.1400	4.4200	2.4200	4.3800	2.5500	4.1600	4.6800	9.0900	1108.9600	1793.3300
Mean	0.4224	0.0369	0.5870	1.2364	2.1105	0.5786	0.5603	2.7802	7.7503	407.4257	287.7130
Median	0.3700	0.0300	0.4100	1.2550	2.1150	0.4150	0.3500	2.7550	7.8500	358.7650	240.6900
SD	0.3270	0.0210	0.5870	0.5340	0.6490	0.4600	0.6360	0.7710	0.3930	168.6700	210.9690
CV	0.7741	0.5691	1.0000	0.4319	0.3075	0.7950	1.1351	0.2773	0.0507	0.4140	0.7332

Min., Minimum value; Max., Maximum value; SD, Standard Deviation; CV, Coefficient of Variation.

Table 3 Statistical features of input data in selected catchments.

	B	F	R	U	W	A	G_L	G_A	G_H	S_A	S_B	S_C	S_D
Min.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Max.	3.33	100.00	94.01	20.21	.01	84.39	100.00	99.47	90.90	79.37	100.00	100.00	100.00
Mean	0.15	71.56	18.90	1.03	0.00	8.34	61.09	28.63	10.25	4.63	20.50	58.16	16.71
Median	0	81.84	6.45	0	0	0.75	64.74	24.62	2.78	0	1.79	63.77	0
SD	0.63	28.65	25.78	3.20	0.00	17.07	27.54	24.79	16.61	12.92	29.51	38.86	26.83
CV	4.14	0.40	1.36	3.10	8	2.04	0.45	0.86	1.62	2.79	1.44	0.67	1.60

B, Bare Land; F, Forest; R, Rangeland; U, Urban; W, Water Body; A, Agriculture; G_L, Low Geological Permeability; G_A, Average Geological Permeability; G_H, High Geological Permeability; S_A, Hydrological Class A; S_B, Hydrological Class B; S_C, Hydrological Class C; S_D, Hydrological Class D; Min., Minimum value; Max., Maximum value; SD, Standard Deviation; CV, Coefficient of Variation.

Table 4 Results of linear regression backward elimination method.

Water quality parameter	Selected variables	R^2	Std. Error of the Estimate	Std. Deviation of predicted values	ANOVA	
					F	Sig.
SAR	U, A, S_A, G_H, R, G_A	0.6178	0.2094	0.2528	21.860	0.000
K^+	A, U, G_H, S_A, S_B	0.5791	0.0140	0.0158	22.582	0.000
Na^+	U, B, A, G_H, G_A	0.6432	0.3611	0.4663	29.525	0.000
Mg^{2+}	U, A, S_A, G_H, S_B, R, S_D	0.4186	0.4247	0.3386	8.238	0.000
Ca^{2+}	B, U, A, S_A, R, S_D, G_A, S_B, G_H	0.5145	0.4781	0.4559	9.166	0.000
SO_4^{2-}	U, A, R, S_D	0.4277	0.3564	0.2942	15.492	0.000
Cl^-	U, A, S_A, G_H, G_A, S_B	0.7709	0.3154	0.5538	45.469	0.000
HCO_3^-	B, A, S_A, G_H, R, S_D, G_A	0.3733	0.6368	0.4608	6.794	0.000
pH	S_B, R, G_H, S_A	0.2970	0.3378	0.2035	8.789	0.000
EC	U, A, S_A, G_H, R, G_A	0.4830	125.6358	115.9407	12.635	0.000
TDS	A, S_A, G_H, R	0.3306	176.7819	115.8908	10.226	0.000

B, Bare Land; F, Forest; R, Rangeland; U, Urban; W, Water Body; A, Agriculture; G_L, Low Geological Permeability; G_A, Average Geological Permeability; G_H, High Geological Permeability; S_A, Hydrological Class A; S_B, Hydrological Class B; S_C, Hydrological Class C; S_D, Hydrological Class D; ANOVA, Analysis of variance; Selected variables, selected variables by linear backward regression approach to enter each ANN.

Table 5 Developed ANNs architecture attributes.

W.Q. Param.	Number of neurons		Epoch	Tr. Function ¥	Best performance		
	Input layer	Hidden layer			NMSE	AIC	BIC
SAR	6	12	1000	trainr	0.0040	-2.172	-2.188
K^+	5	9	23	trainlm	0.0001	-4.770	-4.781
Na^+	5	10	36	trainbfg	0.0129	-1.791	-1.803
Mg^{2+}	7	10	25	trainlm	0.0254	-1.412	-1.433
Ca^{2+}	9	10	18	trainlm	0.0443	-1.093	-1.125
SO_4^{2-}	4	9	31	trainlm	0.0243	-1.519	-1.528
Cl^-	6	12	26	trainlm	0.0062	-1.984	-2.000
HCO_3^-	7	14	23	trainlm	0.0384	-1.245	-1.266
pH	4	4	53	trainbfg	0.0658	-1.083	-1.091
EC	6	12	15	trainlm	2849.9028	3.621	3.605
TDS	4	9	19	trainlm	1132.1117	3.112	3.104

¥ Trainr, Random order incremental training with learning functions; trainlm, Levenberg-Marquardt backpropagation; trainbfg; BFGS, quasi-Newton backpropagation; NMSE, normalized mean square error, AIC, Akaike information criterion; BIC, Bayesian information criterion.

Table 6 Statistical features for water quality parameters in training, validation and test processes of developed ANNs.

W.Q. Param.	Training		Validation		Test	
	R^2	NMSE	R^2	NMSE	R^2	NMSE
SAR	0.876	0.0035	0.931	0.0081	0.245	0.0209
K^+	0.896	0.0000	0.830	0.0000	0.500	0.0000
Na^+	0.834	0.0134	0.405	0.0795	0.591	0.0429
Mg^{2+}	0.723	0.0293	0.677	0.0713	0.643	0.0103
Ca^{2+}	0.815	0.0361	0.654	0.0432	0.512	0.1738
SO_4^{2-}	0.721	0.0196	0.864	0.0338	0.919	0.0358
Cl^-	0.909	0.0056	0.815	0.0036	0.671	0.0120
HCO_3^-	0.727	0.0464	0.642	0.0243	0.403	0.0395
pH	0.362	0.0800	0.787	0.0136	0.752	0.0338
EC	0.764	2574.8869	0.619	5882.8905	0.703	4125.3716
TDS	0.867	853.3733	0.671	18217.9913	0.322	12153.8461

NMSE, normalized mean square error

Table 7 Quantitative and qualitative performance features of the developed ANN models.

Water Quality Parameter	ANN input variables	ANNs best performance							
		Quantitative							Qualitative
		NMSE	R2	RSR	E_{NASH}	R_{BIAS}	AIC	BIC	Performance Rate
SAR	All variables	0.0002	0.8904	0.0419	0.998	0.972	-3.379	-3.443	VG
	B.S. Selected variables	0.0040	0.8275	0.2125	0.955	-16.074	-2.172	-2.188	VG
K^+	All variables	0.0002	0.2460	0.6633	0.560	-3.723	-3.364	-3.429	S
	B.S. Selected variables	0.0001	0.8096	0.1712	0.970	1.141	-4.770	-4.781	VG
Na^+	All variables	0.0024	0.7232	0.0806	0.993	8.006	-2.304	-2.368	VG
	B.S. Selected variables	0.0129	0.8069	0.1894	0.964	5.848	-1.791	-1.803	VG
Mg^{2+}	All variables	0.0625	0.5180	0.4925	0.757	-8.603	-0.813	-0.877	VG
	B.S. Selected variables	0.0254	0.6659	0.3026	0.908	-1.691	-1.412	-1.433	VG
Ca^{2+}	All variables	0.2623	0.3951	0.9347	0.126	-28.003	-0.087	-0.151	U
	B.S. Selected variables	0.0443	0.7327	0.3373	0.886	-6.573	-1.093	-1.125	VG
SO_4^{2-}	All variables	0.0047	0.2731	0.1501	0.977	-0.310	-1.975	-2.039	VG
	B.S. Selected variables	0.0243	0.7665	0.3401	0.884	0.564	-1.519	-1.528	VG
Cl^-	All variables	0.0049	0.8044	0.1102	0.988	0.778	-1.962	-2.026	VG
	B.S. Selected variables	0.0062	0.8737	0.1357	0.981	-15.27	-1.984	-2.000	VG
HCO_3^-	All variables	0.1258	0.5368	0.4608	0.787	0.759	-0.552	-0.616	VG
	B.S. Selected variables	0.0384	0.6411	0.2541	0.935	1.257	-1.245	-1.266	VG
pH	All variables	0.0136	0.3660	0.2965	0.912	0.805	-1.519	-1.583	VG
	B.S. Selected variables	0.0658	0.3606	0.6571	0.568	-0.487	-1.083	-1.091	S
EC	All variables	6858.6887	0.6013	0.5407	0.707	-16.610	4.266	4.202	G
	B.S. Selected variables	2849.9028	0.7506	0.3249	0.894	-4.001	3.621	3.605	VG
TDS	All variables	15862.624	0.1524	0.8321	0.307	-47.941	4.835	4.771	U
	B.S. Selected variables	1132.1117	0.7561	0.1528	0.976	9.073	3.112	3.104	VG

B.S., Backward selection; VG, Very good, G, Good; S, Satisfactory; U, Unsatisfactory; NMSE, normalized mean square error, RSR, RMSE-observations standard deviation ratio; E_{NASH}, Nash-Sutcliffe efficiency, R_{BIAS}, Bias ratio; AIC, Akaike information criterion; BIC, Bayesian information criterion.

Table 8 General hydrological model performance rating for monthly time scales.

	RSR	E_{NASH}	R_{BIAS} (%)	
			Sediment	Nutrient
Very good	$0.00 \leq RSR \leq 0.50$	$0.75 < NSE \leq 1.00$	$\lvert R_{BIAS} \rvert < 15$	$\lvert R_{BIAS} \rvert < 25$
Good	$0.50 < RSR \leq 0.60$	$0.65 < NSE \leq 0.75$	$15 < \lvert R_{BIAS} \rvert < 30$	$25 < \lvert R_{BIAS} \rvert < 40$
Satisfactory	$0.60 < RSR \leq 0.70$	$0.50 < NSE \leq 0.65$	$30 < \lvert R_{BIAS} \rvert < 55$	$40 < \lvert R_{BIAS} \rvert < 70$
Unsatisfactory	$RSR > 0.70$	$NSE \leq 0.50$	$\lvert R_{BIAS} \rvert > 55$	$\lvert R_{BIAS} \rvert > 70$

RSR, RMSE-observations standard deviation ratio; E_{NASH}, Nash-Sutcliffe efficiency, R_{BIAS}, Bias ratio.