Journal of Urban and Environmental Engineering, v.8, n.2, p.198-210

ISSN 1982-3932
doi: 10.4090/juee.2014.v8n2.198210

Journal of Urban and Environmental Engineering
www.journal-uee.org

IMPLICATION OF SOM-ANN BASED CLUSTERING FOR MULTI-STATION RAINFALL-RUNOFF MODELING

Vahid Nourani*, Masoud Mehrvand and Aida Hosseini Baghanam

[1]*Department of Water Resources Engineering,*

Faculty of Civil Engineering, University of Tabriz, Iran

Received 7 April 2014; received in revised form 16 November 2014; accepted 18 November 2014

Abstract: In this study the performance of ANN with feed-forward neural network (FFNN) algorithm evaluated rainfall-runoff modeling in five gauging stations in Florida State. In addition, for investigating the performance of ANN in multi-station discharge prediction, self-organizing map (SOM) clustering tool employed in order to cluster the input data with similar patterns, due to the large amount of records in multiple stations. The main aim of study is to investigate capability and accuracy of ANN based methods in multi-station discharge prediction. In order to consider multiple stations effect on watershed outlet discharge, different combinations for precipitation and discharge data of all stations with antecedent values over the watershed have been taken into account. In this way, application of the representatives from each cluster led to significantly reduction in the numbers of the input variables so that the optimal ANN structure could be proposed. Therefore, ANN as a data-driven model was trained to predict daily runoff for the Peace River basin via recorded values from July 1995 to July 2011. Three scenarios conducted the aim of research; first scenario was an integrated ANN model trained by the data of rainfall and runoff at multiple stations. The second scenario was a sequential ANN model processed with upstream discharge records in addition to rainfall data as inputs and downstream discharge values as target. Finally, third scenario was a SOM-ANN model, in which rainfall and runoff data were clustered according the homogeneity of data via (SOM). The center of each cluster as the dominant component of each cluster was imposed to ANN in order to present an optimal rainfall-runoff model over the watershed. In all scenarios, different data sets at various time lags in both rainfall and stream flow data were applied as inputs in ANN-based model to predict stream flow. Results show that ANN model coupled with SOM is useful tools for forecasting multi-station discharge and precipitation event response in the watershed. Furthermore, the comparison of scenarios leads to select the most efficient and optimal inputs to ANN which subsequently, presents the optimal multi-station rainfall-runoff model over the watershed.

Keywords: Rainfall-runoff modeling; artificial neural networks; self-organizing map; multi-station modeling

* Correspondence to: Vahid Nourani. E-mail:vahidnourani@yahoo.com

INTRODUCTION

In recent decades, especially since the second half of the 20th century, when the effect of the environmental destruction started to become more obvious, great efforts have been made towards the environmental modeling. Although there is a plethora of definitions about the term modeling in the environmental and hydrological literature, all of definitions have one thing in common, the simplifying. A model is a simplified representation of a complex system (Clarke, 1973), so, the definition also includes hydrological models (i.e., models of hydrological systems) such as physical, analogue or mathematical. When the system to be modeled is very complex, it may be adequate for many purposes to adopt some relatively simple form of the system but this simplification should not affect the way the whole system treats. In this way, accurate modeling of the rainfall-runoff process as one of the most considerable elements of environment and hydrological models came into the point of interest in recent years.

Therefore various models have been developed in order to simulate the rainfall-runoff process. Classic time series models such as auto regressive integrated moving average (ARIMA), seasonal ARIMA, ARIMA with exogenous input (ARIMAX) and multiple linear regression (MLR), are widely applied to forecast hydrological time series (e.g. Adamowski *et al.*, 2012; Cleaveland and Stahle, 1989; Graumlich, 1987; Hansen and Nelson, 1997; Nourani *et al.*, 2011a; Pulido-Calvo and Portela, 2007; Salas *et al.*, 1980). These models are basically linear and assume stationary of the dataset. Thus, when it comes to model a complex and non-linear phenomena such rainfall-runoff, some deficiencies are exposed.

In this regard, new generation of hydroinformatic models with capability of nonlinearity modeling which employ new methods and algorithms of forecasting model came to existence. Although linear models may sometimes be inaccurate because of their inability to handle non-stationarity and non-linearity, such conventional methods are still used both in practice because they are simple to use, and because they can be used as 'comparison models' to evaluate newer methods.

Nonlinearity and natural uncertainty of stochastic processes such as rainfall and runoff, the need for long-term historical records, and the complexity of physical-based methods are the reasons that researchers have attempted to develop black box models such as Artificial Neural Network (ANN). ANN has the ability to recognize and identify relationships and patterns from the given data so that solve complex hydrologic problems by handling large amounts of dynamicity, non-linearity of noisy data. This makes black box models well suited to time series modeling problems with a data-driven nature. To be specific, black-box modeling properties, makes it independent to have preliminary perception about the details of the whole process, which in hydrological issues is the physical situation of a watershed. Besides, ANN as a progressive type of black box models can process multiple inputs that are totally differs from each other in characteristics, thus, can represent the time–space variability.

ANN, as a self-learning and self-adaptive approximating function, has great capabilities in modeling and forecasting nonlinear and non-stationary hydrological time series. This modeling have been a topic of interest for many researchers in past two decades and been widely used for hydrological processes modeling such as Rainfall-Runoff, e.g., Tokar & Johnson (1999), Sudheer *et al.* (2000), Kumar *et al.* (2004), Nourani *et al.* (2011a; 2011b). A comprehensive review of ANN application on hydrological models in general and on rainfall-runoff models in particular has been presented by ASCE task Committee (2000) and Abrahart *et al.* (2012). Study on discharge forecasting have shown that ANN is superior to classic regression techniques and time series models including ARIMA (Abrahart & See, 2000) because linear nature of conventional models assume discharge data are stationary, so have a limited ability to capture non-stationarities and non-linearities in discharge data.

With all the improvements achieved by ANN modeling for rainfall-runoff process at the outlet of a watershed in broad range of time scale (i.e., daily, monthly, yearly, etc.), a few studies have addressed the flow estimation at the multiple gauging stations within a watershed. Mutlu *et al.* (2008) employed two different neural network models, the multilayer perceptron and the radial basis neural network to predict stream flow at four gauging stations using antecedents of flow and precipitation in the Eucha watershed in north-west Arkansas and north-east Oklahoma. In their proposed models, the MLP model performed better for forecasting daily flow at multiple gauging stations in the watershed. Turan & Yurdusev (2009) used feed forward back propagation algorithm (FFBP), generalized regression neural networks and fuzzy logic for the estimation of the river flows at one location from the upstream flow records in the case of Birs River in Switzerland. Their work focused on estimating downstream daily mean discharge values at the outlet of watershed from three records of upstream daily mean discharge. The results demonstrated that all the methods considered were capable of yielding satisfactory outputs. However, FFBP algorithm was selected over the other models because of higher performance.

Considering the spatio-temporal distribution of hydrologic processes data, pre-processing of data can improve the efficiency of data-driven methods such as ANN. Clustering is one suggested method to conduct

spatio-temporal pre-processing of data. In the context of ANN-based rainfall-runoff modeling, clustering is usually performed for classification of data, stations or zones into homogeneous classes (Nourani *et al.,* 2013), and/or for optimization of the model structure by selecting dominant and relevant inputs (Bowden *et al.*, 2005). Clustering techniques identify structure in an unlabeled data set by objectively arranging data into homogeneous groups, where the within-group-object dissimilarity is minimized, and the between-group-object dissimilarity is maximized. Conventional clustering methods, such as K-mean, that proceed according to linear characteristics require the number of clusters to be specified in advance (Hsu & Li, 2010).

In this study the performance of ANN with feed-forward neural network (FFNN) algorithm evaluated rainfall-runoff modeling in five gauging stations in Florida State in order to predict outlet station discharge. Moreover a pre-processing technique was conducted through self-organizing map (SOM) in order to decrease the input variables of ANN and obtain accurate results. In this light, the capability of integrated ANN and SOM-ANN methods in multi-station discharge prediction was compared through different combinations of input variables. In the next section of the book chapter, the concepts of forecasting and clustering methods (i.e., ANN and SOM, respectively) are reviewed. The following section describes the study area and data sources, proposed methodology as well as the evaluation criteria. The obtained results are then presented and discussed, and ultimately followed by some concluding remarks.

MATERIAL AND METHODS

Artificial Neural Network (ANN)

ANN models are often applied in nonlinear data series and results to precise outcomes, especially for complex Phenomenon such as rainfall-runoff that physics of involved variables are not completely comprehended. In this study, different combinations of input data series applied to ANN model in order to evaluate the efficiency of the ANN to predict multi station discharges.

ANN as a data driven method inspired by the way biological nervous systems, such as the brain, process information. The key element for this type of data processing is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons or nodes) working in unison to solve specific problems. ANNs, like people, learn by example. Since ANN is a supervised learning algorithm, it requires target values in order to adopt the way of learning process. This typically implies that a large number of input and output data which considered as examples are required for the iterative process of supervised learning.

ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Among the applied neural networks, the FFNN with a back-propagation (BP) training algorithm is the most common method in solving various engineering problems. In the feed-forward phase, the external input information at the input nodes is propagated forward to compute the output information signal at the output unit, and in the backward phase modifications to the connection strengths are made based on the differences between the computed and observed information signals at the output units (Rumelhart *et al.*, 1986). In any FFNN-based modeling, there are two important points to which attention must be paid: firstly, the architecture, i.e., the number of neurons in the input and hidden layers, and secondly, the training iteration (epoch) number. The Number of neurons in the input layer depends directly to the number of input variables and the number of neurons in the hidden layer is the result of trial and error process starting with a few neurons as initial values and by proceeding the process optimal ANN structure with the number of hidden neurons and epoch number will be determined. Appropriate selection of these two parameters improves model efficiency in both the training and testing steps. Furthermore, a high epoch number and poor quality or quantity of data could cause the network to over fit during the training step. If this occurs, the model cannot adequately generalize new data outside of the training set.

FFNNs allow signals to travel one way only; from input to output. There is no feedback (loops) i.e. the output of any layer does not affect that same layer. FFNNs tend to be straight forward networks that associate inputs with outputs. The term *feed forward* means that a neuron connection only exists from a neuron in the input layer to other neurons in the hidden layer or from a neuron in the hidden layer to neurons in the output layer and the neurons within a layer are not interconnected to each other. The explicit expression for an output value of a three-layered FFNN is given by (Kim & Valdes, 2003):

$$\hat{y}_k = f_o\left[\sum_{j=1}^{M_N} w_{kj} f_h\left(\sum_{i=1}^{N_N} w_{ji} x_i + w_{jo}\right) + w_{ko}\right] \quad (1)$$

where w_{ji} is a weight in the hidden layer connecting the i-th neuron in the input layer and the j-th neuron in the hidden layer, w_{jo} is the bias for the j-th hidden neuron, f_h is the activation function of the hidden neuron, w_{kj} is a weight in the output layer connecting the j-th neuron in the hidden layer and the k-th neuron in the output layer, w_{ko} is the bias for the k-th output neuron, f_o is the activation function for the output neuron, x_i is ith input

variable for input layer and $\hat{Y}_k$ is computed output variable. N_N and M_N are the number of the neurons in the input and hidden layers, respectively. The weights are different in the hidden and output layers, and their values can be changed during the process of the network training.

In order to train a neural network to perform some task, the weights of each unit must be adjusted in such a way that the error between the desired output and the actual output is reduced. This process requires that the neural network compute the error derivative of the weights (EW). In other words, it must calculate how the error changes as each weight is increased or decreased slightly. The BP algorithm is the most widely used method for determining the EW.

The FFNN technique consists of layers of parallel processing elements called neurons, with each layer being fully connected to the preceding layer by interconnection strengths, or weights. It has proved that a FFBP model with three layers is satisfactory for the forecasting and simulating as a general approximator (Hornik et al, 1989). Thus, a three-layer ANN with FFBP algorithm trained by the Levenberg-Marquardt optimization method was chosen for this study (Haykin, 1994).

Initial estimated weight values are progressively corrected during a training process that compares predicted outputs with known outputs. Learning of these ANNs is generally accomplished by BP algorithm. The objective of the BP algorithm is to find the optimal weights, which would generate an output vector $Y= (y_1, y_2, ..., y_p)$, as close as possible to the target values of the output vector $T=(t_1, t_2, ..., t_p)$, with the selected accuracy. The optimal weights are found by minimizing a predetermined error function i.e., E in **Eq. (2)** (ASCE, 2000):

$$\mathrm{E} = \sum_{p} \sum_{p} (y_i - t_i)^2 \qquad (2)$$

where y_i is the component of an ANN output vector *Y*, t_i is the component of a target output vector T, p is the number of output neurons and P is the number of training patterns. The error between the desired and predicted output is propagated backwards through the network and the weights connecting the neurons are updated in the learning phase via a training algorithm. A simple structure is provided in **Fig. 1**.

The Tansig and Purelin functions can be utilized as transfer functions in the hidden and output layers. The Tansig transfer function (hyperbolic Tangent sigmoid) is given as **Eq. (3)** (ASCE, 2000):

$$f(x) = \frac{2}{1+e^{-2x}} - 1 \qquad (3)$$

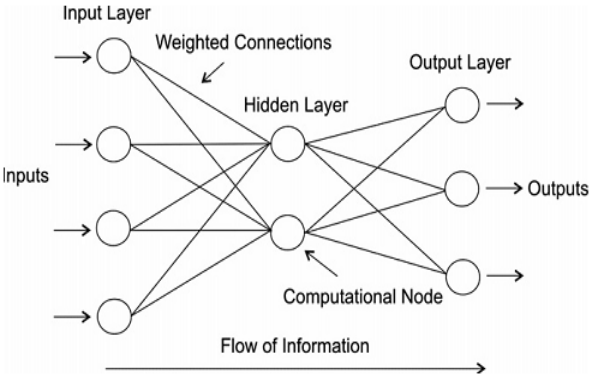

Fig. 1 The basic structure of a FFNN.

Self-Organizing Map (SOM)

The SOM is an effective tool for the visual-based clustering of high-dimensional data. It implements an orderly mapping of a high-dimensional distribution onto a regular low-dimensional grid. Therefore, it is able to convert complex, nonlinear statistical relationships between high-dimensional data items into simple geometric relationships on a low-dimensional display while preserving the topology structure of the data (Kohonen, 1997). SOM reduces dimensions by producing a map of usually 1 or 2 dimensions that plots the similarities of the data by grouping similar data items together.

Thus, SOMs accomplish two things: they reduce dimensions and display similarities. The SOM network generally consists of two layers, an input layer and a Kohonen or output layer (**Fig. 2**). The input layer is fully connected to the Kohonen layer, which in most common applications is two-dimensional. The input layer allocates a neuron for each input variable (i.e., precipitation, discharge). Once the size of input layer is determined, an initialization starts to assign weighs for each neuron, then, the training algorithm is employed over the inputs, and finally the clustered datasets are transferred to Kohonen layer.

The SOM is trained iteratively, and initially the weights are randomly assigned. When the *n*-dimensional input vector *x* is sent through the network, the distance between the *w*, weight neurons of SOM and the inputs is computed. The most common criterion to compute the distance is *Euclidean distance*, **Eq. (4)** (Kohonen, 1997):

$$\|x - w\| = \sqrt{\sum_{i=1}^{n} (x_i - w_i)^2} \qquad (4)$$

The weight with the closest match to the presented input pattern is the winner neuron or the best matching unit (BMU). The BMU and its neighboring neurons are

allowed to learn by changing the weights at each training iteration t, to further reduce the distance between the weights and the input vector, **Eq. (5)** (Kohonen, 1997):

$$w(t+1) = w(t) + \alpha(t)\, h_{lm}(x - w(t)) \quad (5)$$

where α is the learning rate, ranging in [0 1], l and m are the positions of the winning neuron and its neighboring output nodes, and h_{lm} is the neighborhood function. The most commonly used neighborhood function is the Gaussian function, **Eq. (6)** (Kohonen, 1997):

$$h_{lm} = \exp\left(-\frac{\|l - m^2\|}{2\sigma(t)^2}\right) \quad (6)$$

where h_{lm} the neighborhood is function of the best matching neuron l at iteration t; and *l-m* is the distance between neurons l and m on the map grid; and σ is the width of the topological neighborhood. The training steps are repeated until convergence. After the SOM network is constructed, the homogeneous regions, or clusters, are defined on the map.

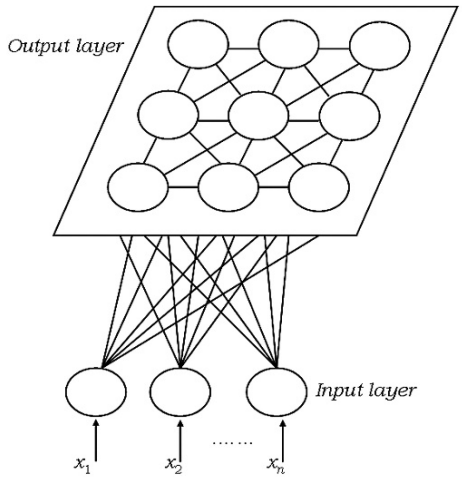

Fig. 2 SOM clustering schematic.

STUDY AREA AND DATA SOURCE

Study area

The Peace-Tampa bay watershed placed in Peace River drainage basin in Florida State was selected for this study (**Fig. 3**). Mentioned watershed connects central Florida to the southwest coast and consists of nine sub-basins covering area of approximately 6,086 km^2. Agricultural land uses encompass about 80% of the

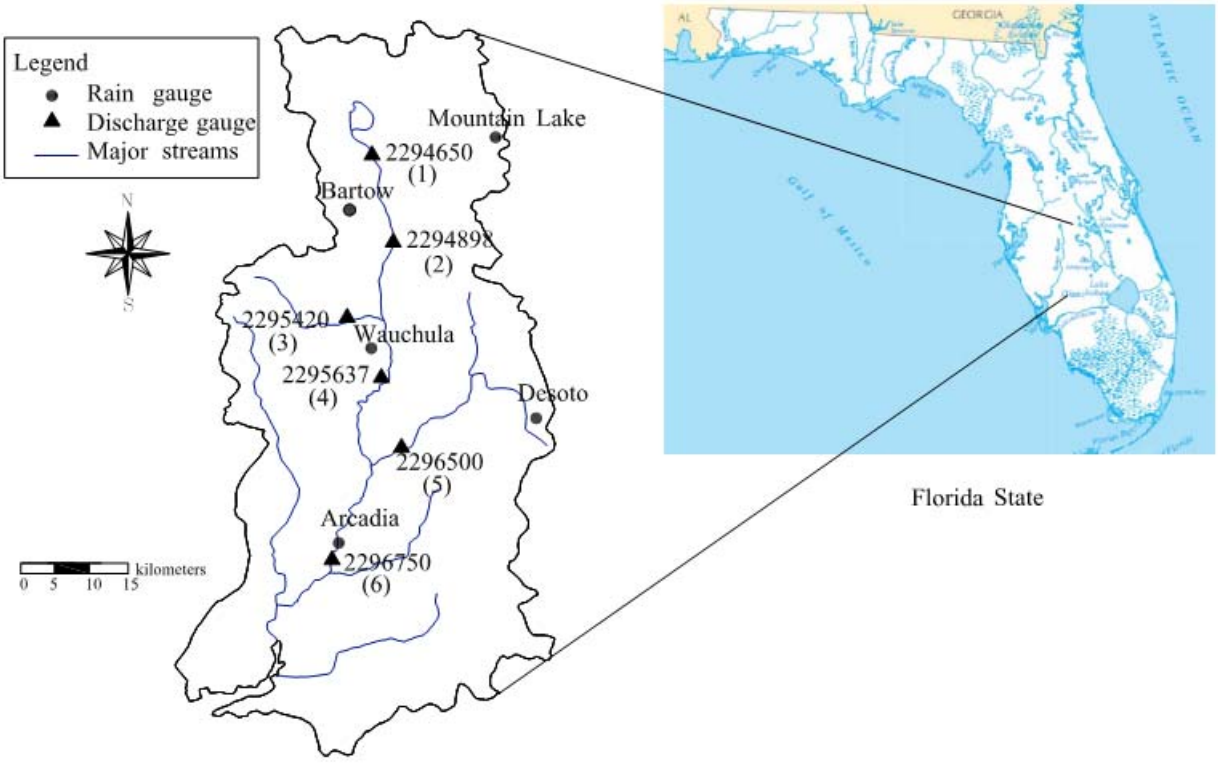

Fig. 3 Peace-Tampa Bay watershed.

basin while urban and mining each cover roughly 10% of the area. The basin contains major portions of three physiographic provinces: the Gulf Coastal Lowlands, the DeSoto Plain, and the Polk Upland. All, or part, of four sand hill ridge provinces are contained within the northern end of the basin. The basin begins in central Polk County, the Polk Uplands, as an internally drained lake region and transitions to a poorly drained upland.

Within the DeSoto Plain of central Hardee and northern DeSoto Counties, the basin becomes a gently sloping plain with well-developed surface drainage features. Downstream of central DeSoto County, the basin enters the Gulf Coastal Lowlands province where elevations are less than 10 meters and the river develops a broad flood plain.

The climate of the area is generally subtropical with an annual average temperature of about 23 degrees in Celsius. Annual rainfall in or near the Peace River drainage basin averages 127 to 142 centimeters.

Land surface elevations in the watershed reach about 61 meters above sea level near the headwaters of the Peace River in Polk County and decline to sea level at Charlotte Harbor. Changes in elevation are most conspicuous along the ridges and scarps. The northwest portions of the Peace River at Bartow basin and the City of Lakeland have an average elevation of 61 meters NGVD (National Geodetic Vertical Datum). Elevations rise to approximately 45 meters NGVD north of Lake Hancock before gradually decreasing again into the Green Swamp. The upland elevations decrease from 50 meters NGVD near Auburndale in the north to 35 meters NGVD near Bartow in the south.

Data sources

Since developed ANN models require discharge and precipitation time series for training and testing periods, data for discharge derived from the records of four USGS discharge gauging stations. Details for each station are shown in **Table 1**. Daily precipitation data derived from five NOAA (National Oceanic and Atmospheric Administration) rain gauges, downloaded from NOAA web site (http://www.noaa.gov). Details for each station are tabulated in **Table 2**.

For first two scenarios, ANN model has been trained and tested with two different input combinations:

1. Q_{t-1}^{i}, Q_{t-2}^{i}, I_{t-1}, I_{t-2}
2. Q_{t-1}^{i}, Q_{t-2}^{i}, Q_{t-3}^{i}, It_{-1}, I_{t-2}, I_{t-3}

Q_t^i is the output variable at station i and $t-1$, $t-2$, $t-3$ refer to three days of antecedent values. It is notable that the selection of three-day lag time was performed according to sensitivity analysis between 1 to 7 days and it is concluded that the effect of lag time greater than 3 days was not prominent enough to enhance the numbers of ANN inputs, thus, such antecedents were ignored. The effect of lag times greater than 3 days relevant to downstream stations can be considered through discharge values of upstream stations, in this way selection of 1, 2 and 3 days lag might include the effects of most important lag patterns in all stations.

Table 1 Discharge Stations' Properties

Station Number	Station Indicator	Station ID	Drainage Area (km²)	Location
1	Q_1	2294650	1,010	Lat 27°54′07″ Long 81°49′03″
2	Q_2	2294898	1,243	Lat 27°45′04″ Long 81°46′56″
3	Q_3	2295420	313	Lat 27°37′13″ Long 81°49′33"
4	Q_4	2295637	2,139	Lat 27°30′15″ Long 81°48′04″
5	Q_5	2296500	855	Lat 27°22′29″ Long 81°47′48″
6	Q_6	2296750	3,540	Lat 27°13′14″ Long 81°52′35″

Table 2 Precipitation Stations' Properties

Rain Gauge Name	Station indicator	Station ID	Location
Mountain Lake	I_1	85973	Lat 27°45′04″ Long 81°46′56″
Bartow	I_2	80478	Lat 27°54′07″ Long 81°49′03″
Wauchula	I_3	89401	Lat 27°30′15″ Long 81°48′04″
Desoto	I_4	82288	Lat 27°22′11″ Long 81°30′49″
Arcadia	I_5	80228	Lat 27°13′14″ Long 81°52′35″

Evaluation criteria of models

In order to train and test ANN it is necessary to have two sets of training data; a calibration set and a validation set. Having trained a network with calibration data the accuracy of the results obtained from that network can be assessed by comparing its responses with the validation set. In this study the network architecture that yielded the best results in terms of determination coefficient (R^2) and root mean square error (*RMSE*) on the training and verifying steps may be determined through trial and error process. For this purpose the data set is divided into two parts: the first 75% of total data were used as

training set and the second 25% are used for verifying purpose. The R^2 and *RMSE* measures of evaluation (**Eqs 8** and **9**) have been used to compare the performance of the different models:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(O_i - C_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2} \quad (8)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - C_i)^2} \quad (9)$$

where n, O_i, C_i and $\bar{O}$ are number of observations, observed data, predicted values and mean of observed data respectively.

The *RMSE* is used to measure forecast accuracy, which produces a positive value by squaring the errors and increases from zero for perfect forecasts through large positive values as the discrepancies between forecasts and observations become increasingly large .Obviously high value for R^2 (up to one) and small value for *RMSE* indicate high efficiency of the model.

RESULTS AND DISCUSSION

In the present study, multi-station modeling of rainfall-runoff have been investigated with three ANN based models. In this way, three scenarios have been proposed. In order to improve the capability of ANN based models for dealing with large amount of multi-station rainfall and runoff data, a preprocessing tool called SOM, have been employed for clustering. In the following, each of the scenarios have been introduced and discussed.

Scenario 1

In the 1st scenario, precipitation and discharge values of both discharge and rain stations 1 to 5 associated to five sub-basins, were imposed to ANN individually, in order to predict one-day-ahead runoff for of each sub-basins. The modeling were performed using the FFNN algorithm by examining 3 to 20 hidden neurons in a single hidden layer using the Levenberg–Marquardt training scheme up to 400 training epochs. The training was terminated at the point where the error in the validation data set began to rise to ensure that the network did not over fit the training data and then fail to generalize the un-seen test data set. No great improvement in modeling performance was found when the number of hidden neurons was increased above a threshold. At this stage, the model efficiency criteria were calculated via observed and computed data of each station to determine the best ANN model. The NNTOOL within the MATLAB software was used for network training purposes (MathWorks, 2010).

Comparison of the best results in terms of evaluation criteria, at **Table 4** indicates accurate performance of ANN-based rainfall-runoff modeling in prediction of discharge.

Although the prediction results are appealing, there are some deficiencies involved in the first scenario which makes it to be under the shadow of doubt. **Table 3** denotes that prediction results for stations number 1 and 6 (i.e., outlet) are not as precise as the results for the middle stations of the watershed such as Q_2, Q_3, Q_4 and Q_5. Since the most important objective of these modeling is to have appropriate outlet prediction, it is of prime importance to have appropriate predicting results for station 6, but the poor evaluation criteria in comparison to other stations do not prove it properly. The deficiency may refer to the fact that the discharge at outlet station is affected by runoff from entire watershed including upstream stations discharge, rather than a sub-basin rainfall-runoff pattern. In order to obtain accurate ANN-based model for outlet discharge station, it is necessary to consider effects between the sub-basins, thus, the 2nd scenario is proposed.

Scenario 2

In the 2nd scenario, the rainfall data of upstream stations are neglected from imposing to ANN directly, assuming that the effect of them is occulted in discharge data of each station. In this way, the discharge data of all the upstream station along with the rainfall data of the nearest station to the outlet of the watershed are imposed to ANN and the prediction pattern are taken out.

The values of R^2 and RSME for the second scenario are presented in **Table 5**. The greatest value for both training and verifying steps for the station number 6 are 0.786 and 0.754, respectively. The results of this 2nd scenario are almost the same as the first scenario in spite of different inputs. The superiority of the 2nd scenario against the first one is that runoff values at the outlet can be predicted via data of upstream without considering data of outlet itself. This may be effective in flood forecasting and subsequently flood alert systems. The probable shortcoming of 2nd scenario can be the imposition of all data without any pre-processing. It means that if some of the discharge stations at upstream involve noises, the noise can be accumulated by considering all the discharge data as inputs. Although ANN assigns the low weights for the noisy data and makes

the outcome less sensitive to inappropriate data, the increase in the number of stations leads to increase data and the noise occulted in high dimensional data, may more greatly be propagated in the long term forecasting. Therefore, in 3rd scenario it is tried to obviate the aforementioned shortcomings and present an effective multi-station discharge forecasting model.

Scenario 3

It is concluded from 1st scenario that in order to obtain accurate and reliable prediction, a new method called multi-station modeling, should be employed. Obtained results from the 2nd scenario with the approach of multi-station modeling, demonstrated that plethora of input variables without any preliminary pre-processing may lead to less accurate results. Besides, the 1st and 2nd scenarios do not provide certain solution for determining the dominant inputs and time lags, thus, are dependent on a trial-error procedure. The conventional trial and error procedure to select the most dominant inputs from large datasets is a time consuming due to existence of several input combinations that need to be examined.

The number of trials for a model with *n* input variables is 2^n*-1*. Therefore, if the trial–error method were used in this study to determine the effective and dominant inputs for predicting the outlet watershed, 2^{11}-1combinations (6 discharge stations and 5 precipitation stations) of inputs would need to be examined as the ANN inputs. Since the all precipitation and discharge stations do not have an equal effect on runoff values or do not provide informative input data, the use of only selected inputs into the ANN simplifies the model structure and leads to better results.

Table 4 Performance of ANN model for scenario 1

Discharge Station No.	Combination No.	Input variables	Best structure	Epoch	R2		RMSE(normalized)	
					Calibration	Verification	Calibration	Verification
(1)	1	$*Q_{t\text{-}1}^{1}$, $Q_{t\text{-}2}^{1}$ $**I_{t\text{-}1}^{1}$, $I_{t\text{-}2}^{1}$	*4.7.1*	*130*	*0.728*	*0.706*	*0.036*	*0.034*
	2	$Q_{t\text{-}1}^{1}$, $Q_{t\text{-}2}^{1}$, $Q_{t\text{-}3}^{1}$ $I_{t\text{-}1}^{1}$, $I_{t\text{-}2}^{1}$, $I_{t\text{-}3}^{1}$	*6.9.1*	*310*	*0.752*	*0.738*	*0.035*	*0.033*
(2)	1	$Q_{t\text{-}1}^{2}$, $Q_{t\text{-}2}^{2}$ $I_{t\text{-}1}^{2}$, $I_{t\text{-}2}^{2}$	*4.7.1*	*260*	*0.833*	*0.825*	*0.029*	*0.022*
	2	$Q_{t\text{-}1}^{2}$, $Q_{t\text{-}2}^{2}$, $Q_{t\text{-}3}^{2}$ $I_{t\text{-}1}^{2}$, $I_{t\text{-}2}^{2}$, $I_{t\text{-}3}^{2}$	*4.8.1*	*230*	*0.834*	*0.822*	*0.028*	*0.023*
(3)	1	$Q_{t\text{-}1}^{3}$, $Q_{t\text{-}2}^{3}$ $I_{t\text{-}1}^{3}$, $I_{t\text{-}2}^{3}$	*4.9.1*	*150*	*0.814*	*0.751*	*0.032*	*0.029*
	2	$Q_{t\text{-}1}^{3}$, $Q_{t\text{-}2}^{3}$, $Q_{t\text{-}3}^{3}$ $I_{t\text{-}1}^{3}$, $I_{t\text{-}2}^{3}$, $I_{t\text{-}3}^{3}$	*6.11.1*	*240*	*0.818*	*0.751*	*0.033*	*0.032*
(4)	1	$Q_{t\text{-}1}^{4}$, $Q_{t\text{-}2}^{4}$ $I_{t\text{-}1}^{4}$, $I_{t\text{-}2}^{4}$	*4.8.1*	*140*	*0.832*	*0.815*	*0.028*	*0.026*
	2	$Q_{t\text{-}1}^{4}$, $Q_{t\text{-}2}^{4}$, $Q_{t\text{-}3}^{4}$ $I_{t\text{-}1}^{4}$, $I_{t\text{-}2}^{4}$, $I_{t\text{-}3}^{4}$	*6.6.1*	*220*	*0.834*	*0.811*	*0.028*	*0.025*
(5)	1	$Q_{t\text{-}1}^{5}$, $Q_{t\text{-}2}^{5}$ $I_{t\text{-}1}^{5}$, $I_{t\text{-}2}^{5}$	*4.4.1*	*160*	*0.808*	*0.773*	*0.030*	*0.031*
	2	$Q_{t\text{-}1}^{5}$, $Q_{t\text{-}2}^{5}$, $Q_{t\text{-}3}^{5}$ $I_{t\text{-}1}^{5}$, $I_{t\text{-}2}^{5}$, $I_{t\text{-}3}^{5}$	*6.5.1*	*200*	*0.814*	*0.779*	*0.032*	*0.030*
(6)	1	$Q_{t\text{-}1}^{6}$, $Q_{t\text{-}2}^{6}$ $I_{t\text{-}1}^{5}$, $I_{t\text{-}2}^{5}$	*4.9.1*	*180*	*0.798*	*0.763*	*0.031*	*0.032*
	2	$Q_{t\text{-}1}^{6}$, $Q_{t\text{-}2}^{6}$, $Q_{t\text{-}3}^{6}$ $I_{t\text{-}1}^{5}$, $I_{t\text{-}2}^{5}$, $I_{t\text{-}3}^{5}$	*6.8.1*	*250*	*0.780*	*0.759*	*0.030*	*0.031*

* Superscripts ranging from 1 to 5 denote the discharge station numbers.
** Subscripts *t-1*, *t-2* and *t-3* indicate the lag time with 1, 2 and 3 days.

Table 5 Performance of ANN model for scenario 2

Discharge Station No.	Combination No.	Input variables	Best structure	Epoch	R2		RMSE(normalized)	
					Calibration	Verification	Calibration	Verification
(6)	1	$*Q_{t\text{-}1}^{i}$, $Q_{t\text{-}2}^{i}$ $I_{t\text{-}1}^{5}$, $I_{t\text{-}2}^{5}$	(12,8,1)	290	0.773	0.741	0.032	0.025
	2	$Q_{t\text{-}1}^{i}$, $Q_{t\text{-}2}^{i}$, $Q_{t\text{-}3}^{i}$ $I_{t\text{-}1}^{5}$, $I_{t\text{-}2}^{5}$, $I_{t\text{-}3}^{5}$	(18,9,1)	310	0.786	0.754	0.028	0.025

* *i* ranging from 1 to 5 denotes the discharge stations.

In the 3rd scenario the SOM clustering technique, applied to identify homogenous rainfall and runoff data of all stations. Dominant precipitation and discharge stations time series which are representative of the watershed precipitation and discharge can be identified using a spatial clustering method, such as SOM clustering method. The Euclidean distance criterion (Bowden *et al.*, 2005) was then utilized to select the centroid time series of each cluster, which is the best representation of the discharge and precipitation pattern of each cluster. To apply the SOM on the rainfall and runoff data, the size of the Kohonen layer was determined as a 3-by-3 grid. Since there is no theoretical principle for determining the optimum size of the Kohonen layer, the Kohonen layer should be large enough to ensure that a suitable number of clusters are formed from the training data (Cai *et al.*, 1994). After creation of the 3-by-3 Kohonen layer, the number of rainfall and runoff time series and their position on the SOM was determined by hits map of SOM (**Fig. 4a**). The first cluster involves the discharge records of station 1 (Q_1) and precipitation values of stations 1 and 2 (I_1 and I_2) (**Table 6**).

The SOM clustering led to three classes of data (**Table 6**) and **Fig. 4b** shows the neighbor weight distances, where the dark hexagons represent the neurons. The colors in the regions indicate the distances between neurons, with the darker colors representing larger distances, and the lighter colors representing smaller distances. It is apparent from **Fig. 5b** that the middle dark line divides the rainfall and runoff stations data to three clusters. The cluster numbers of each neuron and each variable dedicated to neurons have been depicted in **Fig. 5a** and **b**.

Table 6 clusters and members distribution

Cluster number	Cluster members	Center of clusters
1	I_1, I_2, Q_1	I_2
2	I_3, I_4, I_5	I_3
3	Q_2, Q_3, Q_4, Q_5	Q_4

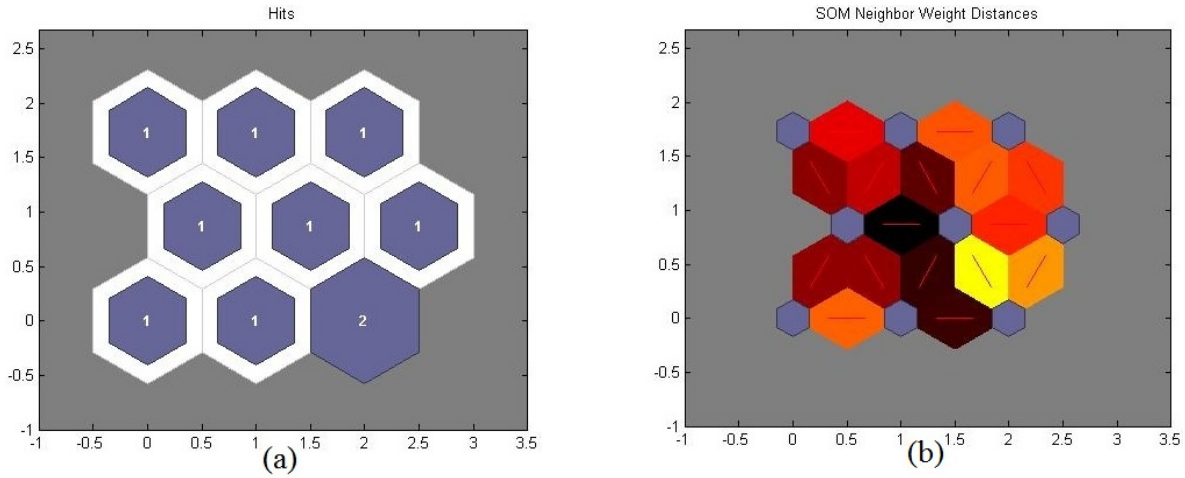

Fig. 4 The 2-Dimensional SOM clustering of rainfall and runoff data **(a)** SOM hits with numbers of members **(b)** SOM neighbor weight distances plan.

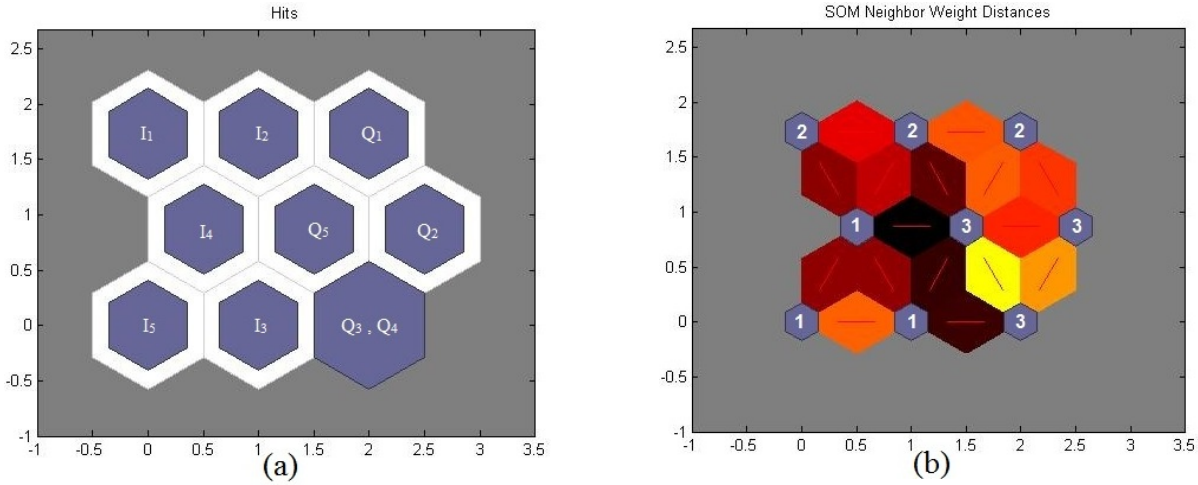

Fig. 5 The 2-Dimensional SOM clustering of rainfall and runoff data **(a)** SOM hits with members **(b)** SOM neighbor weight distances plan including clusters membership numbers.

The second cluster contains the precipitation data of all the remaining rain gauges that have not been clustered in cluster 1(I_3, I_4 and I_5). The 3rd cluster denotes that discharge data at stations 2, 3, 4, 5 have similar values and have similarities in time series that have been clustered in a cluster. SOM clustering results according to topologic property of data adequately corresponded to the topographic position of stations, so that precipitation data of five rain gauges have been clustered in two different clusters; cluster number 1 (I_1, I_2) and cluster number 2 (I_3, I_4 and I_5). As shown in the **Fig. 6** the cluster number 1 includes discharge data of station 1 (Q_1), the first station in the stream network, which may because of geographical position of this station that takes effect only from precipitation values of stations number 1 (I_1) and 2 (I_2), the SOM puts three stations (Q_1, I_1 and I_2) into one cluster. According to the results, cluster 1 encompass one discharge recodes (Q_1) and two rain gauge data (I_1, I_2); since discharge station 1 is located at the top north of the watershed, the only feeding source for the relevant stream is precipitation data of rain gauges 1 and 2 (I_1, I_2), therefore discharge data of station 1 and precipitation data of rain gauges 1, 2 demonstrate the same pattern and clustered in the same cluster (cluster number 1).

Likewise, the second cluster gathered all the remained rain gauges data in a same group, according to the size of watershed at the central part accumulation of rain gauges in one cluster indicates the minimum spatial variability of precipitation data in roughly small area. Ultimately, the 3rd cluster indicates that all the discharge stations which are located on the mainstream of the watershed or near to it are collected in the same cluster. Regarding to the physics of watershed such grouping of the discharge data via SOM greatly coincide the clustering results. Therefore, clustering results of mathematical based and unsupervised modeling of SOM with the physical realities of the watershed arrives to the approach of decreasing variables. Selection of just one member from each cluster which all follow the same pattern and play the same rule in ANN-based forecasting model might be effective from the view point of variable and noise diminishing. Therefore, the dominant member of each cluster is determined via Euclidean distance criterion to select the central time series of each cluster and imposed to ANN in order to model the outlet runoff of the watershed.

Employing the Euclidean distance in order to select the dominant variables from SOM clustering methods led to choose of three dominant time series, which are the representatives for three clusters. As it is apparent from **Table 6**, precipitation data of stations numbers 2 and 3 (I_2 and I_3) were determined as the representatives of clusters 2 and 3. Also, discharge station number 4 (Q_4) was selected as a dominant representative of cluster 3. It should be noted that I_3 and Q_4 are the central stations which located in central part of watershed. This type of dominant determining is inconsistent with the watershed geographical characteristics for example upstream precipitation stations have considerable effects for watershed outlet station (Q_6). As mentioned before, in order to take advantages of multi-station modeling, it is necessary to use multiple stations data but using all the stations data without any pre-processing may cause insufficiency of model such as poor results, consuming time and complexity of model.

So, by applying the SOM and extracting dominant data sets, ineffective data is prevented from entering into the model. It can be concluded that, in this study, precipitation data of I_2 and I_3 and discharge data of Q_4 are the best representatives of inputs. Employing clustering and selection of dominant time series resulted in selecting Q_4 among the other discharge stations in cluster 3 because considering the situation of discharge stations located on the main stream and other branches, Q_4 is the better choice than Q_3 and Q_5. In other words, Q_4 which is located on the main stream and central part of watershed may lead to accurate results than other stations. Accompanying the Q_4 with I_2 and I_3 consider the precipitation and upstream effect which is suitable for forecasting purposes. The results tabulated in **Table 7** denote the high performance of SOM-ANN multiple stations model in predicting outlet discharge.

Third scenario provides superior solution in order to predict discharge by reducing the input variables into 3 variables and the structure of ANN simplified with 5 neurons in hidden layer in comparison to second scenario and finally the R^2 values for both training and verification steps increased to 0.829 and 0.818, respectively. **Figure 7** shows the scatterplots of the computed versus observed discharge values of outlet station (Q_6) related to the coupled SOM and ANN (scenario 3) modeling.

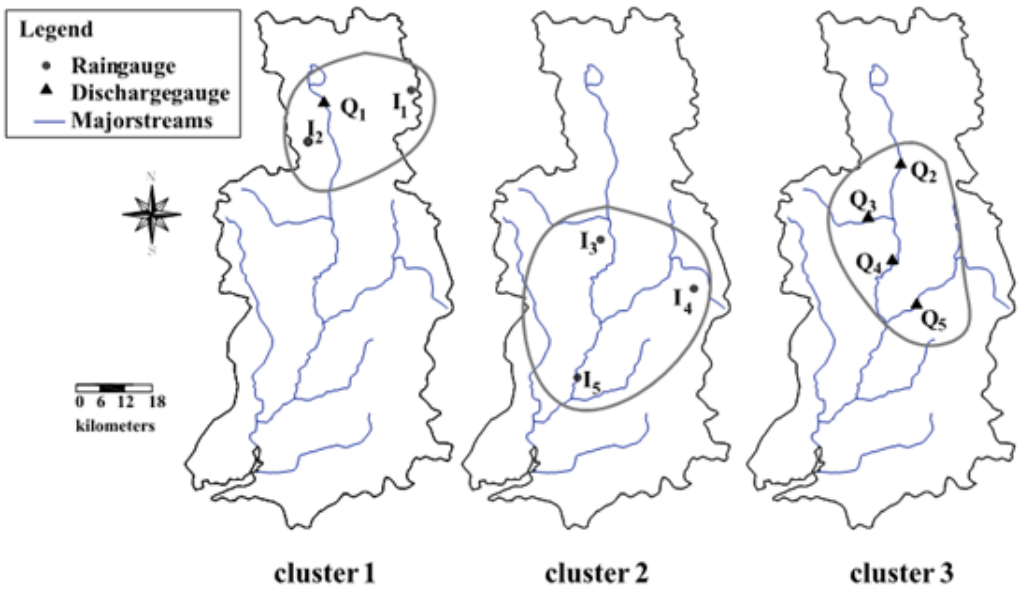

Fig. 6 Members of three clusters on watershed.

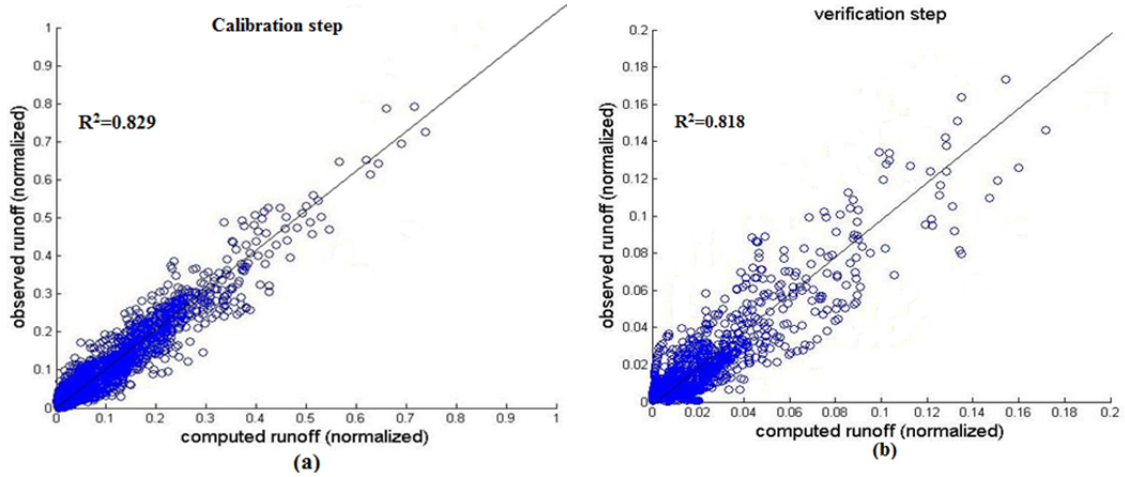

Fig.7 Scatterplots for the station 6 employing the scenario number 3 **(a)** calibration step **(b)** verification step.

Table 7 Performance of SOM-ANN model for scenario 3

Discharge Station No.	Input variables	Best structure	Epoch	R^2		RMSE(normalized)	
				Calibration	Verification	Calibration	Verification
(6)	I_2, I_3, Q_4	(4,5,1)	210	0.829	0.818	0.021	0.015

In the proposed coupled SOM-ANN model, application of the SOM, forced the model to capture similar patterns of data in various stations according to an un-supervised technique which determined the dominant time series that best represented the rainfall and runoff pattern over the watershed.

The selection of dominant time series among a number of time series reduced time and labor in modeling, as the dimensionality of input dataset was decreased as well as the number of trial–error procedures required to optimize the model. Furthermore, the positions of the clustered stations were compatible with the geographical characteristics of the watershed. Following spatial pre-processing, the ANN rainfall–runoff model was constructed to find the non-linear relationship between the selected precipitation data and runoff.

In the proposed coupled SOM-ANN model, application of the SOM, forced the model to capture similar patterns of data in various stations according

The selection of dominant time series among a number of time series reduced time and labor in modeling, as the dimensionality of input dataset was decreased as well as the number of trial–error procedures required to optimize the model. Furthermore, the positions of the clustered stations were compatible with the geographical characteristics of the watershed. Following spatial pre-processing, the ANN

rainfall–runoff model was constructed to find the non-linear relationship between the selected precipitation data and runoff.

CONCLUDING REMARKS

The objective of this study was to predict discharge at outlet station via multiple stations approach in Peace-Tampa Bay watershed at Florida State. For this purpose the ANN with FFNN and BP training algorithm was utilized under three scenarios to have appropriate rainfall-runoff model in predicting outlet discharge over the watershed by using data of discharge and precipitation in several stations. The comparison of scenarios revealed that the 3rd scenario which utilized SOM clustering method, performed effectively in terms of R^2 and RMSE criteria.

The ANN modeling which employed in the first scenario predicted the discharge with acceptable range of R^2 and RMSE criteria but mentioned methodology in first scenario could not reveal efficient performance in all discharge stations within the watershed. Through the first scenario the discharge for one particular station takes effect not only from the precipitation of relevant sub-basin but also the whole watershed as well as precipitation of other sub-basins, thus first scenario could not be able to predict the outlet discharge very effectively, the evaluation results in terms of R^2 could demonstrate the inefficiency of the proposed model (i.e., 0.78 and 0.759 for training and verification steps, respectively). The 2nd scenario considers multiple discharge stations of the watershed (i.e., stations 1 to 5) beside the precipitation data of outlet rain gauge as inputs to ANN model. The rainfall data of upstream stations are neglected from inputs, assuming that the effect of precipitation would appear in downstream discharge stations. Imposition of the aforementioned data increase the dimension of input and may cause the noise propagation, moreover the ANN structure becomes more complex as the hidden neurons number increase to 8 or 9 and epoch numbers ranges from 290 to 310 comparison to 1st scenario.

Although the 1st and the 2nd scenarios had acceptable results, the third scenario performed more reliable than 1st and 2nd scenarios in multi-station discharge prediction with more simple ANN structure. Application of SOM classified the discharge and precipitation data with similar patterns in a same group, which was compatible to physical properties of watershed. Finally, selection of dominant member of each cluster and imposition of 3 representatives among all 10 discharge and precipitation data as the inputs variables of ANN led to the most effective multi-station prediction of discharge at this study area.

The methodology presented herein not only is applicable in other regions with different climatic regimes, but also it could be utilized in other hydrological processes such as sediment. As a suggestion for future studies, and to improve the model results, in addition to the spatial pre-processing of the stations, a temporal data pre-processing (using wavelet transform, Nourani *et al.*, 2009, 2011a,b, 2013) may also be applied on the discharge and precipitation time series before any ANN training.

REFERENCES

Abrahart, R.J., Anctil, F., Coulibaly, P., Dawson, Ch., Mount, N.J., See, L., Shamseldin, A., Solomatine, D., Toth, E. & Wilby, L.R. (2012) Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting. *Prog. Phys. Geog.* **36**, 480-513

Abrahart, R.J. & See, L. (2000) Comparing neural network (NN) and auto regressive moving average (ARMA) techniques for the provision of continuous river flow forecasts in two contrasting catchments. *Hydrol. Process* **14**, 2157–2172

Adamowski, J., Chan, H.F., Prasher, S.O., Ozga-Zielinski, B., Sliusarieva1, A. (2012) Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. *Water Resour. Res.* **48**, doi:10.1029/2010WR009945.

ASCE task Committee on Application of Artificial Neural Networks in hydrology (2000) Artificial Neural Networks in hydrology II: hydrology application. *Hydro. Eng.* **5**, 124–137

Bowden, G.J., Dandy, G.C., & Maier, H.R. (2005) Input determination for neural network models in water resources applications. Part 1-background and methodology. *J. Hydrol.* **301**, 75-92

Cai, S., Toral, H., Qiu, J., Archer, J.S. (1994) Neural network based objective flow regime identification in air–water two phase flow. *Can. J. Chem. Eng.* **72**, 440–445

Clarke, R.T. (1973) A review of some mathematical models used in hydrology with observation on their calibration and use. *J. Hydrol.* **19**, 1-20

Cleaveland, M.K. & Stahle, D.W. (1989) Tree ring analysis of surplus and deficit runoff in the White River, Arkansas. *Water Resour. Res.* **25**(6), 1391-1401

Graumlich, L.J. (1987) Precipitation variation in the Pacific Northwest (1675-1975) as reconstructed from tree rings. *Ann. Assoc. Am. Geogr.* **77**(1), 19-29

Hansen, J.V. & Nelson, R.D. (1997) Neural networks and traditional time series methods: a synergistic combination in state economic forecasts. *IEEE T. Neural Networ.* **8**(4), 863-873.

Haykin, S. (1994) *Neural Networks: a comprehensive foundation.* Macmillan, New York, USA.

Hornik, K., Stichcombe, M. & White, H. (1989) Multi-layer feed forward networks are universal approximators. *Neural Networks* **2**, 359–366

Hsu, K. & Li, S. (2010) Clustering spatial–temporal precipitation data using wavelet transform and self-organizing map neural network. *Adv. Water Resour.* **33**, 190-200

Kim, T., Valdes, J.B. (2003) Nonlinear model for drought forecasting based on a conjunction of wavelet transforms and neural networks. *J. Hydrol. Eng.* **8**(6), 319–328

Kohonen, T. (1997) *Self-organizing maps*. Springer-Verlag, Heidelberg-Berlin, Germany.

Kumar, A.R.S., Sudheer, K.P., Jain, S.K. & Agarwal, P.K. (2004) Rainfall–runoff modeling using artificial neural network: comparison of networks types. *Hydrol. Process* 19:1277–1291

Legates, D.R. & McCabe, Jr. GJ. (1999) Evaluating the use of "goodness-of-fit" measures in hydrologic and hydro-climatic model validation. *Water Resour. Res.* **35**(1), 233-241

Math Works Inc. (2010) MATLAB: Neural Networks Toolbox (NNTOOL) User's Guide, Version 7, *The Math works, Inc*., Natick

Mutlu, E., Chaubey, I., Hexmoor, H. & Bajwa, S.G. (2008) Comparison of artificial neural network models for hydrologic predictions at multiple gauging stations in an agricultural watershed. *Hydrol. Process* **22**, 5097–5106

Nourani, V., Kisi, Ö., Komasi, M., (2011a) Two hybrid Artificial Intelligence approaches for modeling rainfall–runoff process. *J. Hydrol*. **402**, 41–59

Nourani, V., Ejlali, R.G. & Alami, M.T. (2011b) Spatiotemporal groundwater level forecasting in coastal aquifers by hybrid artificial neural network-geostatistics model: A case study. *Environ. Eng. Sci*. **28**, 217–228

Nourani, V., Baghanam, A, H., Adamowski, J., Gebremicheal, M. (2013) Using self-organizing maps and wavelet transforms for space-time pre-processing of satellite precipitation and runoff data in neural network based rainfall-runoff modeling. *J. Hydrol.* **476**, 228-243

Pulido-Calvo, I. & Portela, M.M. (2007) Application of neural approaches to one-step daily flow forecasting in Portuguese watersheds. *J. Hydrol.* **332**, 1-15

Rumelhart, D.E., Hinton, G. E. & Williams, R. J. (1986) *Learning internal representations by error propagation. In: Parallel distributed processing*. MIT Press, Cambridge, p318–362

Salas, J.D., Delleur, J.W., Yevjevich, V., Lane, W.L. (1980) *Applied Modeling of Hydrological Time Series*. Water Resources Publications, Littleton, Colorado, USA.

Sudheer, K.P., Gosain, A.K., Ramasastri, K.S. (2000) A data-driven algorithm for constructing artificial neural network rainfall–runoff models. *Hydrol. Process* **16**, 1325–1330

Tokar, A.S. & Johnson, P.A. (1999) Rainfall-runoff modeling using artificial neural network. *J. Hydrol. Eng.* **4**, 232-239

Turan ME & Yurdusev MA (2009) River flow estimation from upstream flow records by artificial intelligence methods. *J. Hydrol*. **369**, 71.