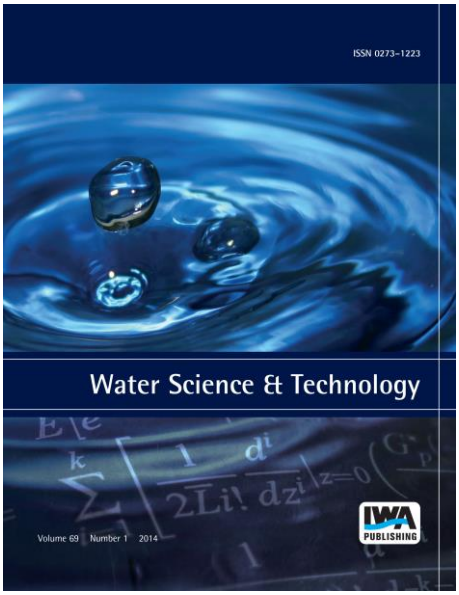

Daily runoff prediction using the linear and non-linear models

Alireza Sharifi, Yagob Dinpashoh and Rasoul Mirabbasi

ABSTRACT

Runoff prediction, as a nonlinear and complex process, is essential for designing canals, water management and planning, flood control and predicting soil erosion. There are a number of techniques for runoff prediction based on the hydro-meteorological and geomorphological variables. In recent years, several soft computing techniques have been developed to predict runoff. There are some challenging issues in runoff modeling including the selection of appropriate inputs and determination of the optimum length of training and testing data sets. In this study, the gamma test (GT), forward selection and factor analysis were used to determine the best input combination. In addition, GT was applied to determine the optimum length of training and testing data sets. Results showed the input combination based on the GT method with five variables has better performance than other combinations. For modeling, among four techniques: artificial neural networks, local linear regression, an adaptive neural-based fuzzy inference system and support vector machine (SVM), results indicated the performance of the SVM model is better than other techniques for runoff prediction in the Amameh watershed.

Key words | data driven techniques, factor analysis, forward selection, gamma test, runoff prediction

Alireza Sharifi (corresponding author)
Rasoul Mirabbasi
Department of Water Engineering, Faculty of Agriculture,
Shahrekord University,
Shahrekord,
Iran
E-mail: *areza_sharifi@yahoo.com*

Yagob Dinpashoh
Department of Water Engineering, Faculty of Agriculture,
University of Tabriz,
Tabriz,
Iran

INTRODUCTION

Rainfall-runoff modeling has a significant role in operational flood management procedures such as design of hydraulic systems and flood prediction. On the other hand, most of the hydrological processes are nonlinear, time varying and spatially distributed. The rainfall-runoff process in a watershed is a nonlinear process that is affected by many factors. Therefore, runoff prediction as a nonlinear and complex process is essential for effective water resources management. So far, many studies have performed runoff modeling with different methods. In recent years, data mining techniques and mathematical methods such as artificial neural networks (ANNs) (Dawson & Wilby 2001; Nayak *et al.* 2005, 2007; Han *et al.* 2007a, 2007b; Aksoy & Dahamsheh 2009), adaptive neural-based fuzzy inference system (ANFIS) (Firat & Güngör 2009; Moghaddamnia *et al.* 2009a, 2009b; Petković *et al.* 2015) and support vector machine (SVM) (Li *et al.* 2013; Wang *et al.* 2014) have been widely used in hydrological modeling. Most of the researchers used effective factors, especially precipitation, on the rainfall-runoff process with a lag time for modeling (Tayfur & Guldal 2006; Remesan *et al.* 2009). But there are still many unsolved issues in hydrological modeling using data driven methods; for example, determination of the best input data for the model and determination of the optimum length of data in the training section.

There are $2^n - 1$ meaningful combinations of n input. Identifying the best input combination can greatly reduce the trial and error steps in the modeling process. For this purpose, various techniques have been proposed and used by researchers; for instance, principal component analysis (Zhang *et al.* 2006; Zhang 2007; Noori *et al.* 2010b; Niu 2013), forward selection (FS) (Chen *et al.* 1989; Wang *et al.* 2006; Noori *et al.* 2010a; Dehghani *et al.* 2014), procrustes analysis (Dinpashoh *et al.* 2004) and gamma test (GT) (Moghaddamnia *et al.* 2008, 2009a, 2009b; Ahmadi *et al.* 2009; Wan Jaafar *et al.* 2011; Kakaei Lafdani *et al.* 2013; Chang *et al.* 2014).

As mentioned above, there are a number of unsolved issues in rainfall-runoff modeling. The main purpose of this study is to investigate and find efficient solutions to solve the two mentioned issues (i.e. to determine the best input

doi: 10.2166/wst.2017.234

combination and to determine the length of data in the training section). Therefore, the following steps were taken in this study. First, the GT, FS and factor analysis (FA) were used to determine the best input combination for the runoff model. Then, the GT was applied in order to determine the appropriate amount of data that was required in the training step.

Thus, four data mining methods: ANNs, ANFIS, SVM and local linear regression (LLR), were selected for estimating the runoff in the Amameh watershed and finally the results of these methods were compared. These methods have been widely applied in rainfall-runoff modeling, more than other data driven techniques in recent years.

METHODOLOGY

Study area and data set

The Amameh watershed is located in the southern area of the central Alborz Mountain, near Tehran (the capital of Iran) and is one of the sub-basins of the Latian dam. It is located between 35° 51′N and 35° 75′N latitudes and between 51° 32′E and 51° 38′E longitudes, with a drainage area of 37.2 km^2 (Figure 1). This watershed is mainly covered by mountainous rangelands, comprising about 80% of the area. The mean annual precipitation of the Amameh watershed is about 840 mm and the mean annual temperature is about 8.6 °C. There are two hydrometric stations in the Amameh watershed, which are located at the outlet (Kamarkhani) and the middle (Amameh) of the watershed on the main stream. Also, there is one climatic station (Amameh) in the middle of the watershed. The wettest and driest months of the watershed are April and September, respectively.

In this study, 9 years (2001–2009) of daily rainfall (from Amameh station) and runoff (from Kamarkhani station) data have been used in order to develop the rainfall-runoff model. The basic statistics of the rainfall (P(t)) and runoff (Q(t)) data set are listed in Table 1. There were no missing data in the data set. Also, the quality of data was checked before the analysis. As a result, we did not find any outlier data.

Nine variables, as input variables, namely lag-1 daily streamflow (Q(t–1)), lag-2 daily streamflow (Q(t–2)), lag-3 daily streamflow (Q(t–3)), and lag-4 daily streamflow (Q(t–4)) as well as daily rainfall (P(t)), lag-1 daily rainfall

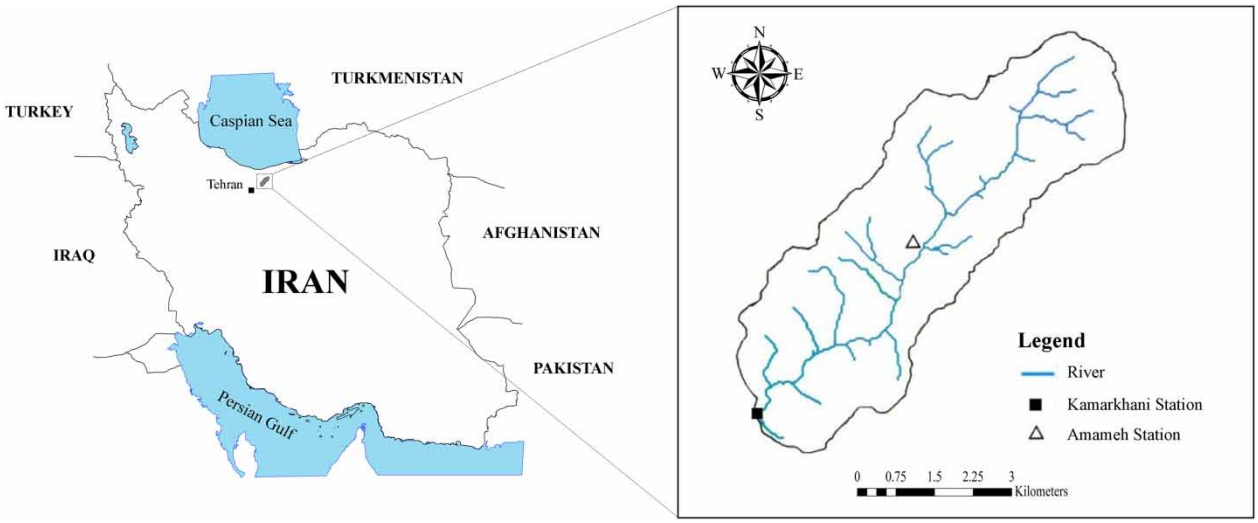

Figure 1 | Location map of the study area, Amameh watershed.

Table 1 | Basic statistics of Amameh watershed rainfall-runoff data

Parameters	Location	Unit	X_{mean}	S_x	Cv	X_{max}	X_{min}
Rainfall (P)	Amameh S.	mm	1.66	5.90	3.55	90.0	0.00
Runoff (Q)	Kamarkhani S.	m^3/s	0.63	0.86	1.36	10.8	0.01

(P(t–1)), lag-2 daily rainfall (P(t–2)), lag-3 daily rainfall (P(t–3)) and lag-4 daily rainfall (P(t–4)) were produced from the data.

All the data were normalized prior to the analysis, by mapping the mean to zero and the standard deviation to 0.5. The training and validation of the data sets were selected by randomizing the input data (Moghaddamnia *et al.* 2009b).

Gamma test

The GT was first reported by Agalbjörn *et al.* (1997), Stefansson *et al.* (1997) and Končar (1997) and later enhanced and discussed in detail by Durrant (2001), Evans & Jones (2002) and Evans (2002).

The GT is based on $N[i, k]$, which are the kth $(1 \leq k \leq p)$ nearest neighbours $\boldsymbol{X}_{N[i,k]}(1 \leq k \leq p)$ for each vector $\boldsymbol{x}_i(1 \leq k \leq M)$. Specifically, the GT is derived from the Delta function of the input vectors:

$$\delta_M(k) = \frac{1}{M}\sum_{i=1}^{M} |\boldsymbol{x}_{N[i,k]} - \boldsymbol{x}_i|^2 (1 \leq k \leq p) \quad (1)$$

where $|\ldots|$ denotes Euclidean distance, and the corresponding Gamma function of the output values is:

$$\gamma_M(k) = \frac{1}{2M}\sum_{i=1}^{M} |y_{N[i,k]} - y_i|^2 (1 \leq k \leq p) \quad (2)$$

where $y_{N[i,k]}$ is the corresponding y-value for the $\boldsymbol{k}$th nearest neighbor of $\boldsymbol{x}_i$ in Equation (3). In order to compute $\boldsymbol{\Gamma}$, a least squares fitted regression line is constructed from the $\boldsymbol{p}$ points $(\delta_M(k), \gamma_M(k))$

$$\gamma = A\delta + \Gamma \quad (3)$$

The intercept on the vertical ($\delta = 0$) axis is the $\boldsymbol{\Gamma}$ value, as can be shown as

$$\gamma_M(k) \to Var(r) \text{ in probability as } \delta_M(k) \to 0 \quad (4)$$

The graphical output of this regression line (Equation (3)) can provide very useful information for hydrological modelers. First, it is remarkable that the vertical intercept $\boldsymbol{\Gamma}$ of $\boldsymbol{y}$ axis offers an estimate of the best MSE achievable utilizing a modeling technique for unknown smooth functions of continuous variables (Evans & Jones 2002). Second, the gradient A offers an indication of the model's complexity (a steeper gradient indicates a model of greater complexity).

We can also determine the reliability of the Gamma statistics by running a series of the GT for increasing M, to establish the size of data set required to produce a stable asymptote. This is known as the M-test. The M-test helps us to decide how much data are required to build a model with a mean squared error that approximates the estimated noise variance. In practice, the GT can be achieved through winGamma™ software implementation (Tsui *et al.* 2002). A formal proof for the GT can be found in Durrant (2001), and Evans (2002).

Forward selection

FS is a data driven model building approach. FS has been widely used for different subjects by many researchers in order to determine the best input combinations and build prediction models (Chen *et al.* 2004; Eksioglu *et al.* 2005; Wang *et al.* 2006; Khan *et al.* 2007; Noori *et al.* 2010a, 2010b). FS is based on a linear regression model.

FS starts with an empty subset. In the first step, variables are ordered according to their correlation with the dependent variable, from the most to the least correlated variable. Then, the first variable is selected by the explanatory variable, which is best correlated with the dependent variable.

After that, at each step, each variable that is not already in the model is tested for inclusion in the model. The most significant of these variables is added to the model, as the second input according to their correlation with the output and the variable that most significantly increases the correlation coefficient (R^2) is selected as the second input. Finally, among N obtained subsets, the subset with optimum R^2 is selected as the model input subset. The optimum R^2 is integral to a set of variables after which adding a new variable does not significantly increase the R^2 value (Noori *et al.* 2010a). In this study, the SPSS software package was used for selecting the best input combination with FS.

Factor analysis

FA is a statistical method, and in this study was used to find the best combination of inputs. This method has frequently applied in different studies (Dinpashoh *et al.* 2004; Malekinezhad *et al.* 2011; Um *et al.* 2011). For more details about FA please refer to Harman (1976), Basilevsky (1994) and Rencher (1995). In this study, the method of principal components and varimax rotation, as one of the most acceptable types of rotation, was used to extract the factors loading matrix (White *et al.* 1991; Um *et al.* 2011).

Artificial neural networks

The first ANN returns to the 1940s, when McCulloch and Pitts introduced it as a mathematical model to create a non-linear relationship between the input and output of a complex system using historical data (McCulloch & Pitts 1943). After that, Rosenblatt (1962) developed the idea of the perceptron. The important phase of a neural network application is the training phase. There are many different learning algorithms for training. Between these algorithms, Levenberg-Marquart (LM), conjugate gradient and quasi-Newton are faster than other algorithms (Lahmiri 2011).

One of the training algorithms based on the quasi-Newton method, which was introduced in 1987 by Fletcher, is BFGS (Fletcher 1987). The BFGS algorithm is performed iteratively using successively improved approximations to the inverse Hessian, instead of the true inverse. The improved approximations are obtained from information generated during the gradient descent process (Jones 2004). ANNs have been widely used for simulating many hydrological processes such as rainfall-runoff simulations (Han *et al.* 2007a). There are a large number of articles and books available on ANN models (Jones 2004; Nayak *et al.* 2005, 2007; Han *et al.* 2007a, 2007b), so no further details are described here. In this study, we used the BFGS algorithm for runoff prediction. WinGamma™ software version 1.97 was used for this purpose.

Local linear regression

The LLR is a nonparametric regression method. This technique has been successfully used in many low-dimensional forecasting and smoothing problems. The LLR performs linear regression through the $p_{\max}$ nearest points to a query point to produce a linear model in the locality of that query point (Durrant 2001). Deciding the size of $p_{\max}$, (the number of near neighbours to be included in the LLR modeling) is the tricky part in LLR modeling (Remesan *et al.* 2009). For more information and detail about LLR please refer to Durrant (2001) and Remesan *et al.* (2009).

Adaptive neuro-fuzzy inference system

ANFIS was first introduced by Jang (1993). ANFIS is a network structure consisting of a number of nodes connected through a directional link. Each node is characterized by a node function with fixed or adjustable parameters. The learning or training phase of a neural network is a process to determine parameter values to sufficiently fit the training data. The basic learning rule is the well-known back propagation method, which seeks to minimize some measure of error, usually some of the squared differences between network outputs and the desired outputs. It can be used as a basis for constructing a set of fuzzy 'If–Then' rules with appropriate membership functions in order to generate the preliminary stipulated input–output pairs. The outline of a typical ANFIS is as follows:

Layer 1: Every node in this layer is an adaptive node with a node function that may be a generalized bell-shaped membership function or a Gaussian membership function.

Layer 2: Every node in this layer is a fixed node labeled Π, representing the firing strength of each rule, and is calculated by the fuzzy AND connective of the 'product' of the incoming signals.

Layer 3: Every node in this layer is a fixed node labeled N, representing the normalized firing strength of each rule. The i^{th} node calculates the ratio of the i^{th} rule's firing strength to the sum of two rules' firing strengths.

Layer 4: Every node in this layer is an adaptive node with a node function indicating the contribution of i^{th} rule toward the overall output.

Layer 5: The single node in this layer is a fixed node labelled R, indicating the overall output as the summation of all incoming signals.

For details about ANFIS and the learning algorithm please refer to Moghaddamnia *et al.* (2009a) and Remesan *et al.* (2009).

Support vector machines

In recent years, SVM as a modern tool regarding artificial intelligence is becoming popular in the field of hydrology. SVM was introduced by Vapnik (1995). This method has been successfully used in information categorization and lately in regression. Other models such as ANNs classify the data by a line, a plane or a hyper plane. But SVM classifies the data in a way such that the risk of classification is minimized. SVM can be used in regression problems (Smola 1996; Kecman 2001). A short explanation of SVM is given below. In general, a basic function for the statistical learning process in SVM is

$$y = f(X) = \sum_{i=1}^{M} w_i \varnothing_i(X) = W_{\varnothing}(X) \tag{5}$$

where the output is a linearly weighted sum of M and the

nonlinear transformation is shown by $\varnothing_i(X)$. For using in SVM, the last equation is converted as below:

$$y = f(X) = \left\{ \sum_{i=1}^{N} w_i K(X_i, X) \right\} - b \tag{6}$$

where K is the Kernel function, w_i and b are parameters of the model, N is the number of training data, X_i are vectors for the training process and X is the independent vector. The role of the Kernel function simplifies the learning process by changing the representation of the data in the input space to a linear representation in a higher dimensional space called the output space (Remesan & Mathew 2014). The parameters of models are derived with maximization of the objectives of functions.

SVM models use some of the specific Kernel functions (often standard Kernel) to convert input vector. The standard Kernel functions applied in SVM are linear, polynomial, radial and sigmoidal (Remesan & Mathew 2014).

Suppose the relation between input and output is as below:

$$y = f(x) = \langle w.x \rangle + b \tag{7}$$

where w and b are the parameters of the model. The goal of this linear regression model is to find the linear function that is the best interpolation for the training point. According to the technique, w and b are determined by minimizing the sum of squares obtained data. For w, it is required to minimize the Euclidean norm i.e. $||w||^2$. It can be written as an optimization problem, as below:

$$\text{minimize } \frac{1}{2}||w||^2 \tag{8}$$

$$\text{S.t.} \begin{cases} y_i - \langle w, x_i \rangle - b \le \varepsilon \\ \langle w, x_i \rangle + b - y_i \le \varepsilon \end{cases} \tag{9}$$

This dual formulation can be solved using the Lagrange multiplier. The obtained Lagrangian equation is as below:

$$\begin{aligned} Min\ L = {} & \frac{1}{2}||w||^2 + C\left(\sum_{i}^{l} \xi_i^* + \sum_{i}^{l} \xi_i \right) + \sum_{i}^{l} (\eta_i \xi_i + \eta_i^* \xi_i^*) \\ & - \sum_{i}^{l} \alpha_i (\varepsilon + \xi_i - y_i + \langle w.x \rangle + b) \\ & - \sum_{i}^{l} \alpha_i^* (\varepsilon + \xi_i^* + y_i - \langle w.x \rangle - b) \end{aligned} \tag{10}$$

where $\eta_i,\ \eta_i^*,\ \alpha_i,\ \alpha_i^* \ge 0$ are the parameters of the equation. The partial derivative of the Lagrangian equation compared with w, b, ξ_i and ξ_i^* is as follows:

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{l} (\alpha_i^* - \alpha_i) = 0 \tag{11}$$

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{l} (\alpha_i^* - \alpha_i) x_i = 0 \tag{12}$$

$$\frac{\partial L}{\partial \xi_i^{(*)}} = C - (\alpha_i^{(*)} - \eta_i^{(*)}) = 0 \tag{13}$$

where $\eta_i^{(*)},\ \xi_i^{(*)}$ and $\alpha_i^{(*)}$ correspond with $\eta_i^{(*)},\ \xi_i^{(*)},\ \alpha_i^{(*)}$ and $\eta_i,\ \xi_i,\ \alpha_i$. By replacing the above equation into the Lagrangian equation we have:

$$\begin{aligned} \text{Min } & -\frac{1}{2} \sum_{i,j=1}^{l} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle x_i, x_j \rangle - \sum_{i=1}^{l} (\alpha_i + \alpha_i^*) \\ & + \sum_{i=1}^{l} y_i (\alpha_i - \alpha_i^*) \end{aligned} \tag{14}$$

$$\text{S.t.} \quad \sum_{i=1}^{l} (\alpha_i - \alpha_i^*)$$

$$\alpha_i,\ \alpha_i^* \in [0, C]$$

Equation (12) can be rewritten as follows:

$$w = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) x_i \tag{15}$$

and therefore,

$$f(x) = \sum_{i,j=1}^{l} (\alpha_i - \alpha_i^*) \langle x_i, x_j \rangle + b \tag{16}$$

This developed equation of support vectors is for a linear model which is used for non-linear relationships. It is not proper for many hydrological analyses of linear regression for modeling and therefore, it is proper to convert the Kernel to put data in a space with more dimensions and then using the linear regression. Kernel function $K(x, z)$ is $\langle \varnothing(x), \varnothing(z) \rangle$. In this study, the Radial Basis Kernel Function (RBF) was used. For more detail please refer to Vapnik (1995).

Statistical criteria for performance evaluation of models

The performance of all models in this study was compared using various statistical criteria. The statistical measures used in this study include the root mean-squared error (RMSE), coefficient of determination (R^2) and Nash Sutcliffe (NS). These statistical terms can be defined as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Q_{pred} - Q_{obs})^2} \quad (17)$$

$$R^2 = \frac{\sum_{i=1}^{N}(Q_{pred} - \overline{Q_{pred}})(Q_{Obs} - \overline{Q_{Obs}})}{\sqrt{\sum_{i=1}^{N}(Q_{pred} - \overline{Q_{pred}})^2 \sum_{i=1}^{N}(Q_{Obs} - \overline{Q_{Obs}})^2}} \quad (18)$$

$$NS = 1 - \left(\frac{\sum_{i=1}^{N}(Q_{Obs} - Q_{pred})^2}{\sum_{i=1}^{N}(Q_{Obs} - \overline{Q_{Obs}})^2}\right) \quad (19)$$

where Q_{obs} and Q_{pred} denotes the observed and predicted runoff by model, respectively, $\hat{Q}_{obs}$ and $\hat{Q}_{mod}$ are the average of the observed and predicted runoff, respectively, and N is the number of data points.

In this study, the GT, FA and FS were used to determine the best input combination of the runoff model. Also, the GT was used for determining the amount of data that were required in the training step. In addition, the ANNs, LLR, ANFIS and SVM methods were used for estimating the runoff of the Amameh watershed (in Kamarkhani station).

RESULTS AND DISCUSSION

First, in this section, we describe the results obtained from the FS, GT and FA to identify the best input combination and length of data for training. Afterwards, the results of modeling using ANNs, LLR, ANFIS and SVM are compared in order to determine the best model for runoff modeling in the Amameh watershed.

Results of model input selection

Forward selection

In this study, the FS method was used as a linear input selection technique in order to select the best input combination of nine input variables. To select the best input combination with the FS method first, the correlation between each input variable and the desired output is evaluated and the variable with the highest correlation (Q(t–1) with $R^2 = 0.809$) is selected as the first and the most important input. Second, the remaining candidates are evaluated and entered into the model one by one based on their correlation coefficient rank. For evaluation of modeling goodness, correlation coefficient (R^2) and Standard Error (SE) were used. This step is repeated several times until the new input variable added to the model does not significantly improve the model performance. Finally, the input variables with the most significant effect on the output are selected and other variables are removed. The result of the FS method is shown in Table 2. From Tables 2 and 7, candidates were selected as input variables according to their importance: Q(t–1), P(t), Q(t–4), Q(t–3), P(t–3), P(t–2) and P(t–4). Also, according to FS, the function between the input and output data is as below:

$$\begin{aligned} Q(t) = {} & 0.682 * \mathrm{Q(t-1)} + 0.197 * \mathrm{P(t)} + 0.144 * \mathrm{Q(t-4)} \\ & + 0.101 * \mathrm{Q(t-3)} - 0.044 * \mathrm{P(t-3)} +, 0.023 * \mathrm{P(t-2)} \\ & - 0.021 * \mathrm{P(t-4)} + 0.001 \end{aligned} \quad (20)$$

Factor analysis

FA, as another method for determination of the best input combination, was also applied in this study. The first six factors, accounting for 96.1% of the total variance, were selected and subjected to Varimax Normalized Rotation in the FA approach. Table 3 shows the value of factor loading for input variables. The larger value shown in bold in the table of the correlation coefficient in each factor was selected as an important variable. Therefore, Q(t–3), P(t), P(t–1), P(t–2), P(t–3) and P(t–4) were determined as important variables for modeling.

Table 2 | Results of FS procedure

Input subset	R^2	SE
Q (t–1)	0.809	0.03510
Q (t–1), P(t)	0.833	0.03284
Q (t–1), P(t), Q(t–4)	0.847	0.03141
Q (t–1), P(t), Q(t–4), Q(t–3)	0.848	0.03132
Q (t–1), P(t), Q(t–4), Q(t–3), P(t–3)	0.849	0.03210
Q (t–1), P(t), Q(t–4), Q(t–3), P(t–3), P(t–2)	0.850	0.03119
Q (t–1), P(t), Q(t–4), Q(t–3), P(t–3), P(t–2), P(t–4)	0.850	0.03117

Table 3 | Rotated factor loading (varimax rotation)

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
Q4	0.935	−0.033	0.008	−0.011	0.155	0.040
Q3	**0.949**	0.008	−0.007	0.144	0.104	0.041
Q2	0.940	0.173	0.038	0.108	0.063	0.025
Q1	0.911	0.141	0.201	0.082	0.012	0.059
P4	0.168	0.049	0.036	0.088	**0.978**	0.020
P3	0.155	0.088	0.050	**0.977**	0.088	0.038
P2	0.129	**0.978**	0.091	0.087	0.049	0.054
P1	0.102	0.089	**0.982**	0.048	0.036	0.097
P	0.075	0.052	0.095	0.037	0.019	**0.990**

Gamma test

For determining the effective variable in the modeling, first the Gamma value was calculated from a combination of all variables (nine input candidates). In the next step, one of the variables was omitted and the Gamma value was calculated for the combination of the remaining variables (eight variables). Then, the omitted variable in the previous stage was returned and another variable was omitted from the original combination and the Gamma value was then calculated for the new combination which again contained eight candidates. This process was repeated for each variable one by one and at each step the Gamma value was computed for an eight variables set. Finally, the variables which are removed increase the Gamma value compared with the original combination with nine variables. The results of GT are shown in Table 4. According to Table 4, P(t) is the most important variable because of having the biggest Gamma value after its omission from the combination. Other important variables are Q(t–1), P(t–1), Q(t–2) and P(t–3), respectively. As a result, these variables were selected as important variables.

The comparison among three combinations selected based on the GT, FS and FA methods indicate two differences among them. First, the number of selected variables and second, the kind of selected variables. For identifying the best input data combination, LLR and ANNs models were used as test models. The results of training and testing of LLR and ANNs models with four different input combinations are given in Table 5. For comparison of modeling results, R^2 and RMSE were used. According to Table 5, although the accuracy of the LLR model with nine input variables is better than other LLR models in the training section, the LLR-GT model has better accuracy in the testing section. In addition, the accuracy of the ANNs-GT model is better than other models in the two sections. Finally, among these eight models, the ANNs-GT model was selected as the best model because it was formed from the lowest number of inputs. Therefore, the combination which was determined by the GT method was selected as the best input data combination for runoff modeling.

Results of the training and testing data sets length determination

The result of the M-test analysis is shown in Figure 2 for the best combination selecting by GT. According to Figure 2, we should use about 1,000 data points for training. For validating this result, different scenarios of data partitioning into training were tried in order to determine the optimal length of training data required for modeling with over fitting during training. Table 6 shows different partitioning scenarios and corresponding RMSE, NS and R^2 values

Table 4 | The GT results on the rainfall-runoff data of the Amameh watershed

Input variables	Mask	Gamma (Γ)	Gradient (A)	SE	V_{ratio}
All inputs	111111111	0.0007951	0.0249267	0.0000685	0.123154
All inputs – Q(t–4)	011111111	0.0007573	0.0358710	0.0000842	0.117299
All inputs – Q(t–3)	101111111	0.0007503	0.0394821	0.0000759	0.116216
All inputs – Q(t–2)	110111111	0.0009127	0.0129956	0.0000917	0.141363
All inputs – Q(t–1)	111011111	0.0010774	0.0230902	0.0000796	0.166873
All inputs – P(t–4)	111101111	0.0007300	0.0384730	0.0000504	0.113061
All inputs – P(t–3)	111110111	0.0008647	0.0175060	0.0001163	0.133924
All inputs – P(t–2)	111111011	0.0007224	0.0378475	0.0000771	0.111893
All inputs – P(t–1)	111111101	0.0009514	0.0038840	0.0000943	0.147354
All inputs – P(t)	111111110	0.0012084	−0.0118756	0.0000582	0.187159

Table 5 | Comparing the results of LLR and ANNs models in determining the best input combination

		Training		Testing	
Model	**Number of input variables**	**R^2**	**RMSE**	**R^2**	**RMSE**
LLR	9	0.97	0.015	0.06	0.280
LLR-GT	5	0.97	0.021	0.89	0.033
LLR-FS	7	0.97	0.018	0.31	0.099
LLR-FA	6	0.96	0.020	0.49	0.100
ANNs	9	0.90	0.030	0.85	0.036
ANNs-GT	5	0.94	0.029	0.92	0.028
ANNs-FS	7	0.88	0.032	0.86	0.035
ANNs-FA	6	0.91	0.040	0.85	0.030

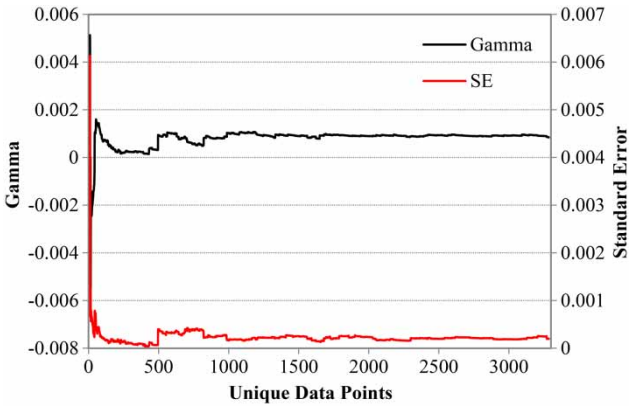

Figure 2 | The variation of gamma statistic and SE with unique data points.

during testing steps using the LLR model. In the training section, the results indicate that after the seventh scenario the values of R^2 and NS are approximately constant. The best values of RMSE, NS, and R^2 in the testing section were obtained for scenario 17, shown in bold in Table 6, with values of 0.03, 0.76 and 0.89, respectively, where 2,400 data points were used in the training section. Therefore, we should select about 2,400 data for the training section. Finally, since the lowest amount of SE and Gamma occurred at point 2,383; we used 2,383 data points for training and the remaining data out of 3,284 data points were used for testing the model.

Results of the ANNs, LLR, ANFIS and SVM techniques

The performance of ANNs, LLR, ANFIS and SVM models were evaluated by error criteria, namely RMSE, NS and R^2. The results of the training and testing for the models are given in Table 7.

Table 6 | Results of different data portioning scenarios for the training and testing periods

		Testing period		
Scenarios	**Training data length**	**RMSE**	**NS**	**R^2**
Case 1	500	0.04	0.65	0.84
Case 2	750	0.05	0.59	0.81
Case 3	1,000	0.05	0.58	0.80
Case 4	1,100	0.04	0.75	0.88
Case 5	1,200	0.05	0.66	0.84
Case 6	1,300	0.05	0.65	0.83
Case 7	1,400	0.05	0.62	0.82
Case 8	1,500	0.05	0.65	0.83
Case 9	1,600	0.04	0.71	0.86
Case 10	1,700	0.04	0.70	0.85
Case 11	1,800	0.05	0.62	0.81
Case 12	1,900	0.05	0.61	0.81
Case 13	2,000	0.04	0.69	0.85
Case 14	2,100	0.04	0.70	0.86
Case 15	2,200	0.04	0.69	0.85
Case 16	2,300	0.04	0.69	0.85
Case 17	**2,400**	**0.03**	**0.76**	**0.89**
Case 18	2,500	0.03	0.74	0.88
Case 19	2,600	0.04	0.69	0.85
Case 20	2,700	0.03	0.70	0.86
Case 21	2,800	0.03	0.75	0.87

Table 7 | Comparison between ANNs, LLR, ANFIS and SVM models for runoff estimation

	Training			Testing		
Models	**RMSE**	**NS**	**R^2**	**RMSE**	**NS**	**R^2**
ANNs	0.03	0.88	0.94	0.03	0.85	0.92
LLR	0.02	0.94	0.97	0.03	0.79	0.89
ANFIS	0.02	0.94	0.97	0.04	0.73	0.88
SVM	0.02	0.93	0.98	0.02	0.92	0.97

The ANNs were trained using the BFGS and the conjugate gradient algorithms. The scatter plots of training and testing results were produced by the ANNs model shown in Figure 3. As can be seen from Table 7, the performance of the ANNs model is not better than other models in the training section, but the accuracy of the ANNs is better than the LLR and ANFIS models in the testing section.

The nonparametric procedure based on LLR models does not require the same training steps as the neural

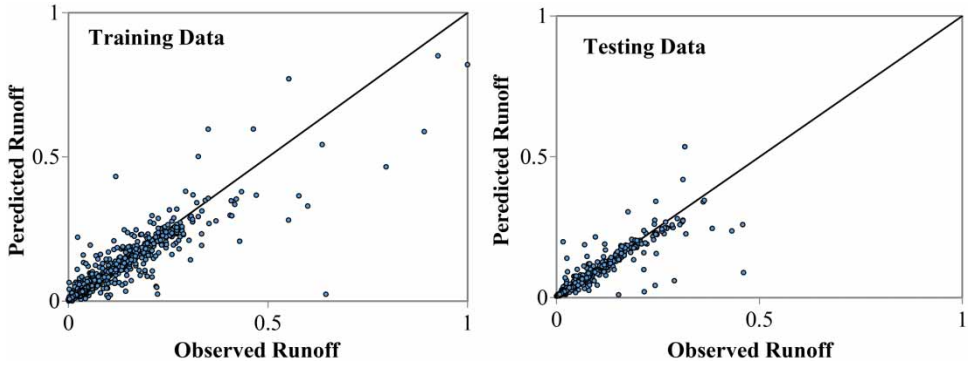

Figure 3 | The scatter plots of training and testing sections produced by ANNs.

network models. The optimal number of nearest neighbors for LLR (principally dependent on the noise level) was determined by a trial and error method and 10 nearest neighbors were implemented. The comparative analysis of this model using some basic statistics has been carried out and is shown in Table 7. In addition, scatter plots of the training and testing sections by the LLR model are shown in Figure 4. It can be seen that the performance of the LLR model is better than the ANNs in the training section, but the accuracy of the ANNs model with R^2 value of 0.92 and NS value of 0.85 is better than the LLR model in the testing section with $R^2 = 0.89$ and NS = 0.79.

For the ANFIS model, the number of membership functions for each input was set to 3 and the input data were scaled between [0,1]. The membership function type was selected among the trimf, trapmf, gbellmf, gaussmf, pimf and dsigmf. The length of training data was also 2,383 data points. Finally, the trapmf membership function was selected as the best membership function. The scatter plots of training and testing sections were produced by the ANFIS model as shown in Figure 5. According to the statistical criteria of this model (Table 7), the performance of the ANFIS model in the training section is approximately similar to the LLR model and better than the ANNs model. But the accuracy of this model is less than the LLR and ANNs models in the testing section.

In this study, the SVM model becomes complex due to the need to consider the distances of all support vectors. Therefore, we used the SVR (SVM for Regression) model and the RBF, which is the standard form of kernel function in SVM. Selection of the kernel function is one of the complex steps in using SVM as well as other parameters such as C and epsilon. Therefore, the parameters of the SVM model were determined by a trial and error method. Figure 6 shows the scatter plots of the training and testing section using the SVM model. According to Table 7, the SVM model shows better results compared with other models in the training section. In this section, the LLR and ANFIS models have acceptable performance as well. But in the testing section, the accuracy of the SVM model is better than other models with RMSE, R^2 and NS values equal to 0.02, 0.97 and 0.92, respectively. Figure 7 shows the curves of observed and predicted runoff by different models in the testing section, and Figure 8 shows the performance of the SVM model at large scale. As can be seen in Figure 8, the SVM

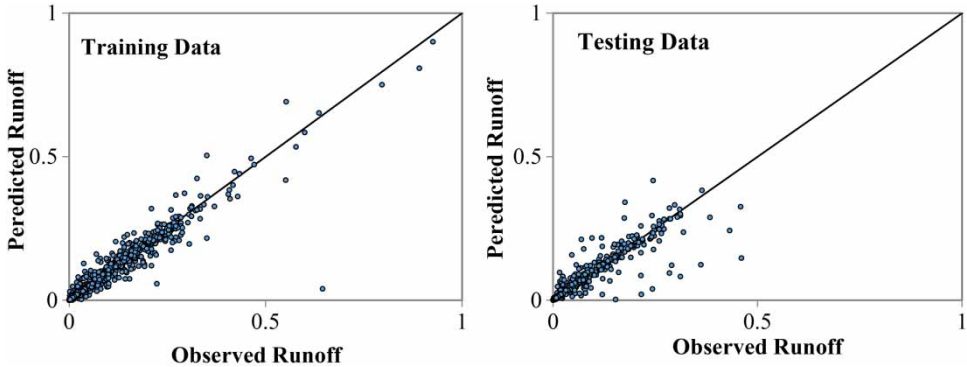

Figure 4 | The scatter plots of training and testing sections produced by LLR.

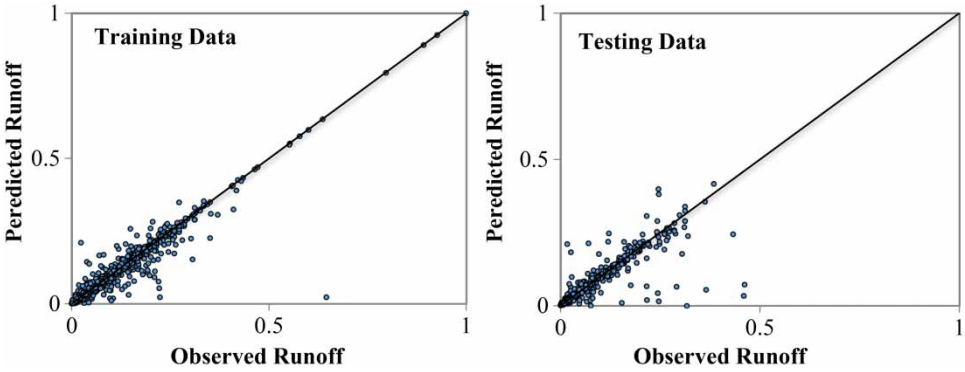

Figure 5 | The scatter plots of training and testing sections produced by ANFIS.

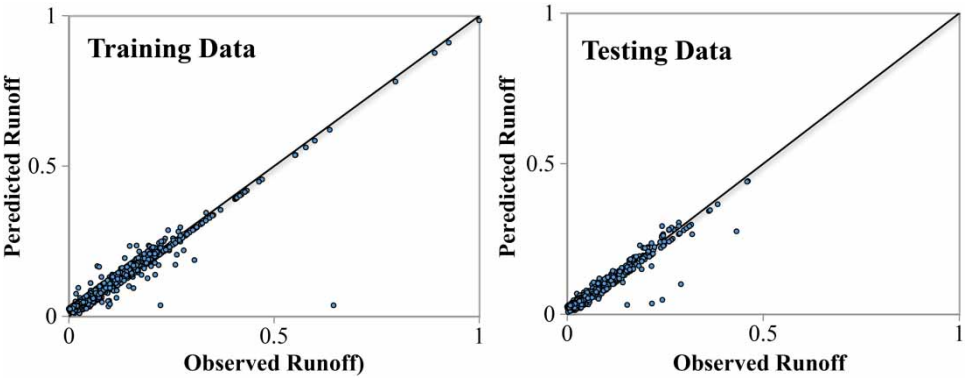

Figure 6 | The scatter plots of training and testing sections produced by SVM.

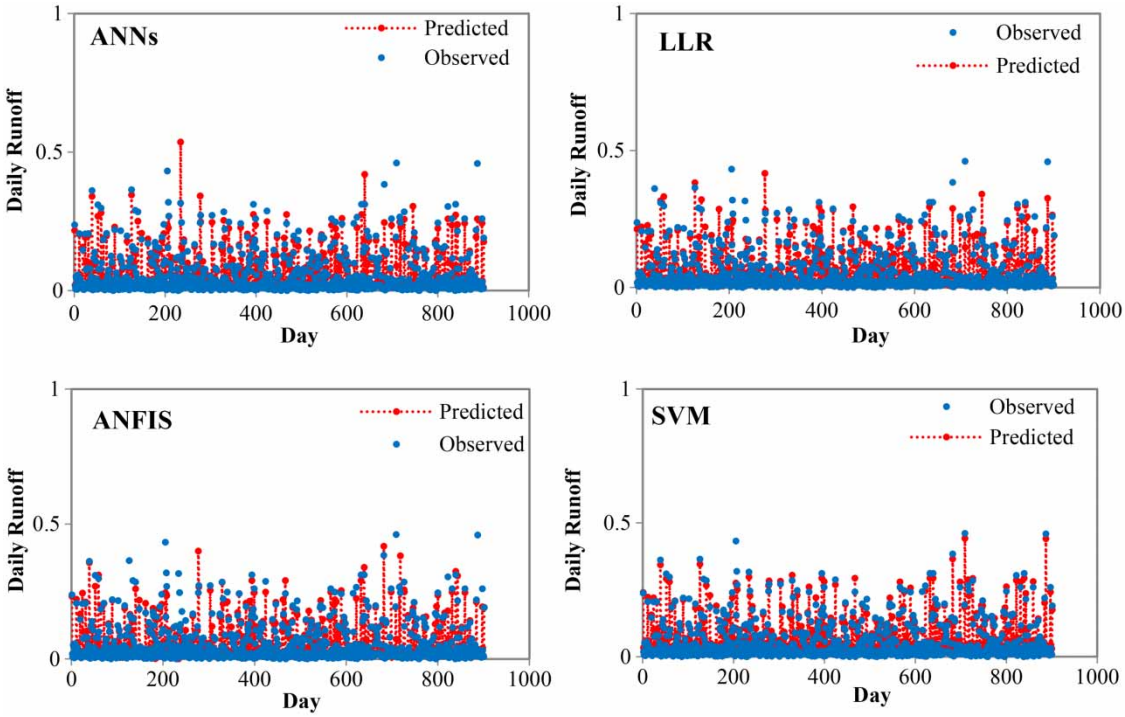

Figure 7 | Predicted and observed curve of daily runoff by ANNs, LLR, ANFIS and SVM models in testing section.

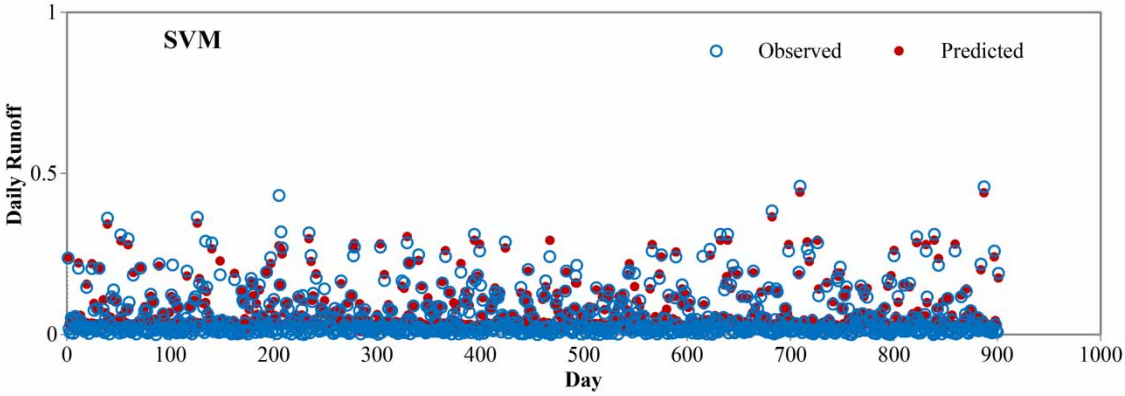

Figure 8 | Predicted and observed curve of daily runoff by SVM model in testing section.

model can predict runoff better than other models, especially at high flows. On the other hand, the comparison of the scatter plots in the testing section shows that the dispersion of points near the bisector line in the SVM model is less than other models. Therefore, the SVM model has the best performance in estimating runoff in the Amameh watershed.

CONCLUSIONS

In this study, the ANNs, LLR, ANFIS and SVM models were used for daily runoff prediction in the Amameh watershed. The daily rainfall-runoff data for the period of 2001–2009 were used for developing the models. To determine the best input combination, GT, FS and FA were applied for runoff modeling. The results showed the GT method had the best performance in determining the best input data combination for modeling compared with the other methods. Based on the GT method, the optimum size of training data was determined to be equal to 2,383 data, and the remaining data were used for testing the models. The results of the modeling showed the accuracy of the SVM model is better than other models in two sections. Therefore, the SVM model is introduced as the best model for runoff estimation in the Amameh watershed.

Determining the best input combination using GT, as the main method in this study, is a less time-consuming procedure than the trial and error method. Moreover, this technique is easy for selection of relevant variables in the construction of nonlinear models for runoff prediction. In addition, GT is quite general, and could be applied to other nonlinear hydrological systems modeling (such as evaporation) and other models because GT is not linked to any specific model.

Generally speaking, increasing the length of data and adding other variables such as temperature, soil humidity etc. caused the results to change. Unfortunately, in the Amameh watershed, other variables affecting the rainfall-runoff process were not measured in the period from 2001–2009. Moreover, daily rainfall and runoff data were not available after 2009.

In the modeling section, four common data driven methods were applied. In recent years, these methods have been widely combined with other methods, and new methods have been developed such as NNARX, which is a combination of ANNs and ARX.

In recent years, these methods have been used in hydrological modeling more than in the past because they are easy and accessible. But there is not a specified relation between input and output, and the results can be varied by changing the length of data and input parameters. As a result, these models are not applied like numerical methods, operationally. On the other hand, the unclear effect of physical factors in the hydrological processes in modeling is one of their other problems. Nevertheless, application of these methods has been expended and development of the new methods are the sign of this progress. Therefore, in order to complete the current study, it is suggested that the result of GT is compared with other input selection techniques, and that the results of the modeling are compared with other methods such as Neuro-wavelet. Finally, we hope this study will persuade more researchers to use and evaluate GT in different catchments.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Agalbjörn, S., Končar, N. & Jones, A. J. 1997 A note on the gamma test. *Neural Comput. Appl.* **5** (3), 131–133.

Ahmadi, A., Han, D., Karamouz, M. & Remesan, R. 2009 Input data selection for solar radiation estimation. *Hydrological Processes* **23**, 2754–2764.

Aksoy, H. & Dahamsheh, A. 2009 Artificial neural network models for forecasting monthly precipitation in Jordan. *Stochastic Environmental Research and Risk Assessment* **23** (7), 917–931.

Basilevsky, A. 1994 *Statistical Factor Analysis and Related Methods: Theory and Applications,* Wiley, New York.

Chang, F., Chen, P., Lu, Y., Huang, E. & Chang, K. 2014 Real-time multi-step-ahead water level forecasting by recurrent neural networks for urban flood control. *Journal of Hydrology* **517**, 836–846.

Chen, S., Billings, S. A. & Luo, W. 1989 Orthogonal least squares methods and their application to nonlinear system identification. *International Journal of Control* **50**, 1873–1896.

Chen, S., Hong, X., Harris, C. J. & Sharkey, P. M. 2004 Sparse modeling using orthogonal forward regression with PRESS statistic and regularization. *IEEE Transactions on Systems, Man and Cybernetics, Part B* **34**, 898–911.

Dawson, C. W. & Wilby, R. L. 2001 Hydrological modeling using artificial neural networks. *Progress in Physical Geography* **25** (1), 80–108.

Dehghani, M., Saghafian, B., Nasiri Saleh, F., Farokhnia, A. & Noori, R. 2014 Uncertainty analysis of streamflow drought forecast using artificial neural networks and Monte-Carlo simulation. *International Journal of Climatology* **34**, 1169–1180.

Dinpashoh, Y., Fakheri-Fard, A., Moghaddam, M., Jahanbakhsh, S. & Mirnia, M. 2004 Selection of variable for the purpose of regionalization of Iran's precipitation climate using multivariate methods. *Journal of Hydrology* **297**, 109–123.

Durrant, P. J. 2001 *Win-Gamma™: A Non-Linear Data Analysis and Modeling Tool With Applications to Flood Prediction*. PhD Thesis, Department of Computer Science, Cardiff University, University of Wales, UK.

Eksioglu, B., Demirer, R. & Capar, I. 2005 Subset selection in multiple linear regression: a new mathematical programming approach. *Computers & Industrial Engineering* **49**, 155–167.

Evans, D. 2002 *Data Derived Estimations of Noise Using Near Neighbour Asymptotic*. PhD Thesis. Cardiff University, University of Wales, UK.

Evans, D. & Jones, A. J. 2002 A proof of the gamma test. *Proceedings of Royal Society. Series A* **458** (2027), 2759–2799.

Feletcher, R. 1987 *Practical Method of Optimization*, 2nd edn. Wiley, New York.

Firat, M. & Güngör, M. 2009 Monthly total sediment forecasting using adaptive neuro fuzzy inference system. *Stochastic Environmental Research and Risk Assessment* **24** (2), 259–270.

Han, D., Chan, L. & Zhu, N. 2007a Flood forecasting using support vector machines. *Journal of Hydroinformatics* 9 (4), 267–276.

Han, D., Kwong, T. & Li, S. 2007b Uncertainties in real-time flood forecasting with neural networks. *Hydrological Processes* **21**, 223–228.

Harman, H. H. 1976 *Modern Factor Analysis,* 3rd edn. The University of Chicago Press, Chicago, IL.

Jang, J. 1993 ANFIS: adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems, Man and Cybernetics* **23** (3), 665–685.

Jones, A. J. 2004 New tools in non-linear modelling and prediction. *Computational Management Science* **1**, 109–149.

Kakaei Lafdani, E., Moghaddamnia, A. & Ahmadi, A. 2013 Daily suspended sediment load prediction using artificial neural networks and support vector machines. *Journal of Hydrology* **478**, 50–62.

Kecman, V. 2001 Learning and soft computing: Support Vector Machines. Neural networks and fuzzy logic models. MIT Press, Cambridge, MA, London, UK.

Khan, J. A., Aelst, S. V. & Zamar, R. H. 2007 Building a robust linear model with forward selection and stepwise procedures. *Computational Statistics & Data Analysis* **52**, 239–248.

Končar, N. 1997 Optimisation methodologies for direct inverse neurocontrol. PhD Thesis, Department of Computing. Imperial College, London.

Lahmiri, S. 2011 A comparative study of backpropagation algorithms in financial prediction. *International Journal of Computer Science, Engineering and Applications* **1** (4), 15–21.

Li, W., Yang, M., Liang, Z., Zho, Y., Mao, W., Shi, J. & Chen, Y. 2013 Assessment for surface water quality in Lake Taihu Tiaoxi River Basin China based on support vector machine. *Stochastic Environmental Research and Risk Assessment* **27** (8), 1861–1870.

Malekinezhad, H., Nachtnebel, H. P. & Klik, A. 2011 Regionalization approach for extreme flood analysis using L-moments. *Journal of Agricultural Science and Technology* **13**, 1183–1196.

McCulloch, W. S. & Pitts, W. 1943 A logic calculus of the ideas imminent in nervous activity. *Bull. Math. Biophys.* **5**, 115–133.

Moghaddamnia, A., Ghafari, M., Piri, J. & Han, D. 2008 Evaporation estimation using support vector machines technique. *World Academy of Science, Engineering and Technology* **43**, 14–22.

Moghaddamnia, A., Ghafari, M., Piri, J., Amin, S. & Han, D. 2009a Evaporation estimation using artificial networks and adaptive neuro-fuzzy inference system techniques. *Advances in Water Resources* **32**, 88–97.

Moghaddamnia, A., Remesan, R., Hassanpour Kashani, M., Mohammadi, M., Han, D. & Piri, J. 2009b Comparison of LLR, MLP, Elman, NNARX and ANFIS Models-with a case study in solar radiation estimation. *Journal of Atmospheric and Solar-Terrestrial Physics* **71**, 975–982.

Nayak, P. C., Sudheer, K. P., Rangan, D. M. & Ramasastri, K. S. 2005 Short-term flood forecasting with a neuro fuzzy model. *Water Resources Research* **41**, 1–16.

Nayak, P. C., Sudheer, K. P. & Jain, S. K. 2007 Rainfall-runoff modeling through hybrid intelligent system. *Water Resources Research* **43**, 1–17.

Niu, J. 2013 Precipitation in the Pearl River basin, South China: scaling, regional patterns, and influence of large-scale climate anomalies. *Stochastic Environmental Research and Risk Assessment* **27** (5), 1253–1268.

Noori, R., Hoshyaripour, G., Ashrafi, K. & Nadjar Araabi, B. 2010a Uncertainty analysis of developed ANN and ANFIS models in prediction of carbon monoxide daily concentration. *Atmospheric Environment* **44**, 476–482.

Noori, R., Karbassi, A. & Sabahi, M. S. 2010b Evaluation of PCA and Gamma test techniques on ANN operation for weekly solid waste prediction. *Journal of Environmental Management* **91**, 767–771.

Petković, D., Shamshirband, S., Anuar, N. B., Naji, S., Mat Kiah, M. L. & Gani, A. 2015 Adaptive neuro-fuzzy evaluation of wind farm power production as function of wind speed and direction. *Stochastic Environmental Research and Risk Assessment* **29** (3), 793–802.

Remesan, R. & Mathew, J. 2014 *Hydrological Data Driven Modelling,* Springer, New York.

Remesan, R., Shamim, M. A., Han, D. & Mathew, J. 2009 Runoff prediction using an integrated hybrid modeling scheme. *Journal of Hydrology* 372, 48–60.

Rencher, A. C. 1995 *Methods of Multivariate Analysis,* Wiley, New York.

Rosenblatt, F. 1962 Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan.

Smola, A. 1996 *Regression estimation with Support Vector Learning Machines.* Technics Universitat Munchen, Munich, Germany.

Stefánsson, A., Koncar, N. & Jones, A. J. 1997 A note on the gamma test. *Neural Comput. Appl.* **5** (3), 131–13.

Tayfur, G. & Guldal, V. 2006 Artificial neural networks for estimating daily total suspended sediment in natural streams. *Nordic Hydrology* **37** (1), 69–79.

Tsui, A. P. M., Jones, A. J. & Oliveira, A. G. 2002 The construction of smooth models using irregular embedding determined by the gamma test analysis. *Neural Computing & Applications* **10** (4), 318–329.

Um, M. J., Yun, H., Jeong, C.-S. & Heo, J. H. 2011 Factor analysis and multiple regression between topography and precipitation on Jeju Island, Korea. *Journal of Hydrology* **410**, 189–203.

Vapnik, V. 1995 *The Nature of Statistical Learning Theory.* Springer, New York.

Wan Jaafar, W. Z., Liu, J. & Han, D. 2011 Input variable selection for median flood regionalization. *Water Resources Research* **47**, 1–18.

Wang, X. X., Chen, S., Lowe, D. & Harris, C. J. 2006 Sparse support vector regression based on orthogonal forward selection for the generalised kernel model. *Neurocomputing* **70**, 462–474.

Wang, Y., Guo, S., Chen, H. & Zhou, Y. 2014 Comparative study of monthly inflow prediction methods for the Three Gorges Reservoir. *Stochastic Environmental Research and Risk Assessment* **28** (3), 555–570.

White, D., Richman, M. & Yarnal, B. 1991 Climate regionalization and rotation of principal components. *International Journal of Climatology* **11** 1–25.

Zhang, Y. X. 2007 Artificial neural networks based on principal component analysis input selection for clinical pattern recognition analysis. *Talanta* **73**, 68–75.

Zhang, Y., Li, H., Hou, A. & Havel, J. 2006 Artificial neural networks based on principal component analysis input selection for quantification in overlapped capillary electrophoresis peaks. *Chemometrics and Intelligent Laboratory Systems* **82**, 165–175.

First received 23 September 2016; accepted in revised form 10 April 2017. Available online 28 April 2017