Hydrol. Earth Syst. Sci., 13, 411–421, 2009
www.hydrol-earth-syst-sci.net/13/411/2009/

Hydrology and Earth System Sciences

Multi-criteria validation of artificial neural network rainfall-runoff modeling

R. Modarres

Faculty of Natural Resources, Isfahan University of Technology, Isfahan, Iran

Received: 19 September 2008 – Published in Hydrol. Earth Syst. Sci. Discuss.: 2 December 2008
Revised: 19 February 2009 – Accepted: 5 March 2009 – Published: 19 March 2009

Abstract. In this study we propose a comprehensive multi-criteria validation test for rainfall-runoff modeling by artificial neural networks. This study applies 17 global statistics and 3 additional non-parametric tests to evaluate the ANNs. The weakness of global statistics for validation of ANN is demonstrated by rainfall-runoff modeling of the Plasjan Basin in the western region of the Zayandehrud watershed, Iran. Although the global statistics showed that the multi layer perceptron with 4 hidden layers (MLP4) is the best ANN for the basin comparing with other MLP networks and empirical regression model, the non-parametric tests illustrate that neither the ANNs nor the regression model are able to reproduce the probability distribution of observed runoff in validation phase. However, the MLP4 network is the best network to reproduce the mean and variance of the observed runoff based on non-parametric tests. The performance of ANNs and empirical model was also demonstrated for low, medium and high flows. Although the MLP4 network gives the best performance among ANNs for low, medium and high flows based on different statistics, the empirical model shows better results. However, none of the models is able to simulate the frequency distribution of low, medium and high flows according to non-parametric tests. This study illustrates that the modelers should select appropriate and relevant evaluation measures from the set of existing metrics based on the particular requirements of each individual applications.

Correspondence to: R. Modarres
(r_m5005@yahoo.com)

1 Introduction

The rainfall-runoff relationship is an important issue in hydrology and a common challenge for hydrologists. Due to the tremendous spatial and temporal variability of watershed characteristics such as snowpack, soil moisture, hydraulic conductivity, watershed slope, seasonal rainfall etc., the rainfall-runoff relationship is usually a nonlinear process. Since the middle of the 19th century, different methods have been applied by hydrologists within rainfall-runoff modeling whereupon many models have attempted to describe the physical processes involved (e.g. mathematical-physical lumped or distributed models).

Over the last decade, there has been a tremendous growth in the interest of application of a class of techniques that operate in a manner analogous to that of biological neurons system, i.e. artificial neural networks (ANNs). While ANNs are capable of capturing non-linearity in the rainfall-runoff process compared with other modeling approaches (Hsu et al., 1995), ANN models have been applied in hydrology and in the context of rainfall-runoff modeling (Smith and Eli, 1995; Dawson and Wilby, 1998; Tokar and Markus, 2000; Zhang and Govindaraju, 2003; Kumar et al., 2005). From these studies, it has been demonstrated that ANN models can be flexible enough to simulate the rainfall-runoff processes successfully.

Various types of neural network models are available for rainfall-runoff modeling. Feedforward artificial neural networks (FFANNs) maintain a high level of research interest due to their ability to map any function to an arbitrary degree of accuracy. This has been demonstrated theoretically for both the radial basis function (RBF) network and the popular multilayer perceptron (MLP) network (Harpham

Published by Copernicus Publications on behalf of the European Geosciences Union.

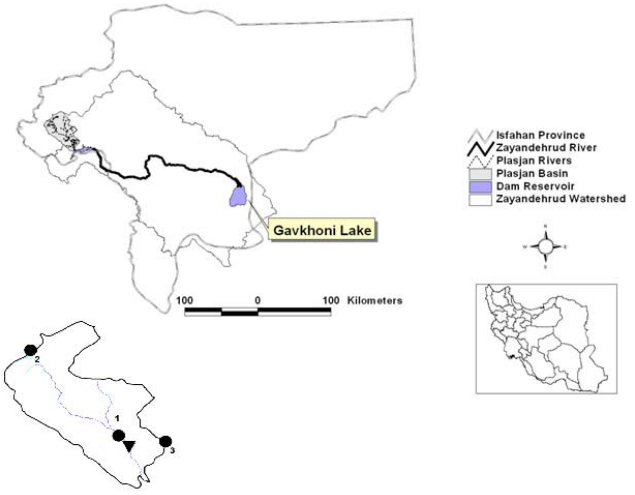

Fig. 1. Location of Zayandehrud watershed in Isfahan Province and the location of Plasjan Basin. The rainfall stations (black circles) and Plasjan hydrometery station (black triangle) are also shown inside the Plasjan Basin.

and Dawson, 2005). The primary goal of ANN modeling is the prediction or forecasting of hydrological variables, e.g. runoff prediction. In this case, a set of variables is divided into two sets prior to the model building: the training set and validation set. The validation set is kept aside to evaluate the accuracy of the model derived from the training test. In the validation phase, the model output is compared with actual outputs using statistical measurements such as root-mean-square error (RMSE) and the coefficient of correlation (CORR).

However, the equality of the probabilistic characteristics of the observed and simulated runoff is usually ignored in validation test. It is important because the simulated runoff should reflect the relevant hydrological characteristics of the observed runoff in terms of both magnitude and frequency. For example, the observations are arranged in order of the magnitude, beginning with 1 for the biggest, when the flow duration curves are depicted. Therefore, the simulated runoff should reproduce the probabilistic behavior of the observed runoff, especially for both upper and lower extreme values.

In this regard, the main objectives of this study are twofold; in the first step, we develop an effective ANN model for studying the rainfall-runoff relationship in the study area and verify the models by the global statistics such as root-mean-square error (RMSE), coefficient of correlation and coefficient of efficiency. In the second step, the non-parametric test for the equality of the mean, variance and probability distribution of the observed and simulated runoff is used to validate rainfall-runoff models and to compare them with global statistics.

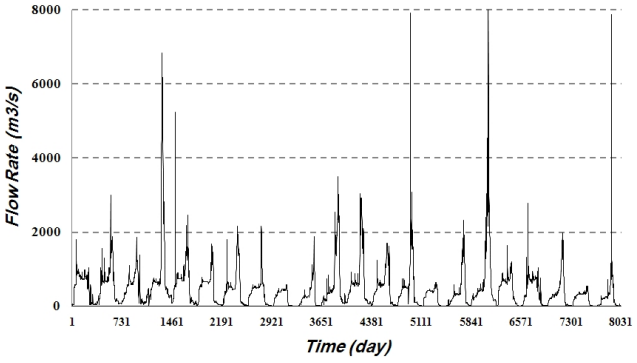

Fig. 2. Daily streamflow of Plasjan River (m^3/s).

2 Study area and data

In this study, the most popular FFANN architecture, i.e. MLP, is used for rainfall-runoff modeling of the main upstream basin of the Zayandehrud watershed in the western region of Isfahan Province in the center of Iran. Zayandehrud watershed has two main basins called Ghaleh Shahrokh and the Plasjan Basin. These two basins connect directly to the Zayandehrud Dam which provides the water supply for Isfahan province. The input and output variables for ANN is the daily rainfall and runoff of the Plasjan basin (Fig. 1). The data set includes Plasjan daily streamflow time series and three daily rainfall time series of the stations within the basin for the period of 1978-2000. The daily streamflow of Plasjan is given in Fig. 2.

3 Multi-layer perceptron

In this study, the multilayer perceptron architecture assumes that the unknown function (rainfall-runoff) is represented by a multilayer feed forward network of sigmoid units. An ANN model with n input neurons ($x_1, \ldots, x_n$), h hidden neurons ($w_1, \ldots, w_h$) and m output neurons ($z_{1,\ldots,Zm}$) is considered in this study. The function that this model calculates is

$$z_k = f\left(\sum_{j=1}^{h} \alpha_{kj} w_j + \varepsilon_k\right) k = 1, \ldots, m \quad (1)$$

$$w_j = g\left(\sum_{i=1}^{n} \beta_{ji} x_i + \tau_j\right) j=1, \ldots, h \quad (2)$$

Where g and f are activation functions, i, j, and k are representing input, hidden and output layers respectively, τ_j is the bias for neuron w_j and ε_k is the bias for neuron z_k, β_{ij} is the weight of the connection from neuron x_i to w_j and α_{jk} is the weight of the connection from neuron w_j to z_k.

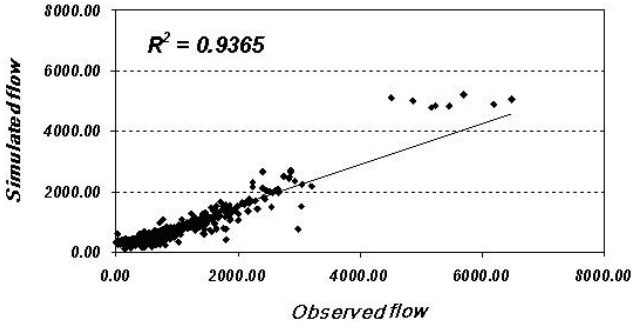

Fig. 3a. Scatter plot of observed versus simulated streamflow (m^3/s) for MLP2 network.

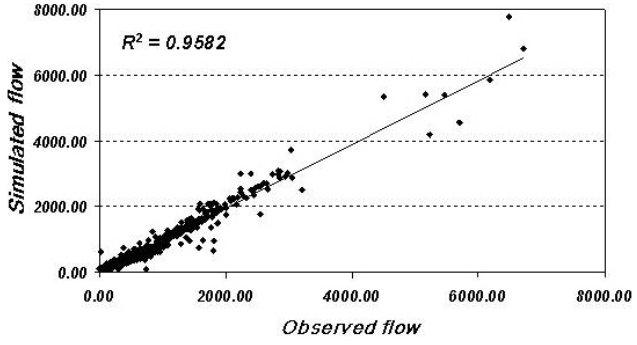

Fig. 3b. Scatter plot of observed versus simulated streamflow (m^3/s) for MLP3 network.

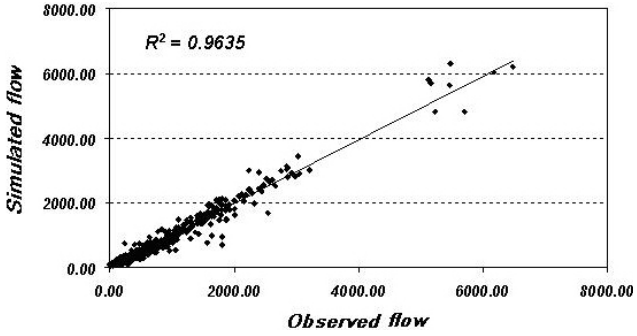

Fig. 3c. Scatter plot of observed versus simulated streamflow (m^3/s) for MLP4 network.

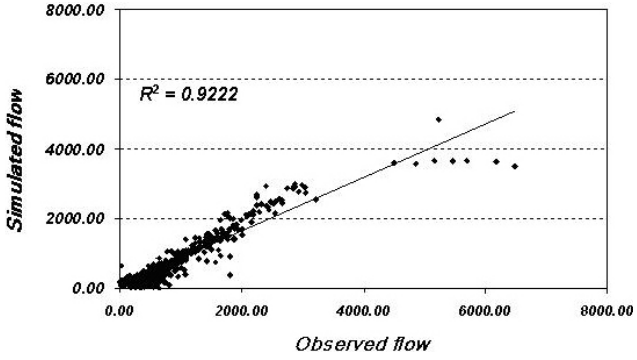

Fig. 3d. Scatter plot of observed versus simulated streamflow (m^3/s) for MLP5 network.

The hyperbolic tangent sigmoid function is used in this study as activation function for the hidden nodes. The function can be written as the following

$$g(s_i) = \frac{e^{s_i} - e^{-s_i}}{e^{s_i} + e^{-s_i}} \tag{3}$$

Where s_i is the weighted sum of all incoming information and is also referred to as the input signal

$$s_j = \sum_{i=1}^{n} \beta_{ji} x_i + \tau_j \tag{4}$$

The major advantage of the MLP is that it is less complex than other artificial neural networks such as Radial Basis Function (RBF), and has the same nonlinear input–output mapping capability (Coulibaly and Evora, 2007). The training of the MLP involves finding an optimal weight vector for the network. The objective function of the training process is:

$$E = \frac{1}{2} \min \sum_{p=1}^{N} \sum_{k=1}^{M} (t_{kp} - z_{kp})^2 \tag{5}$$

Where N is the number of training data pairs, M is the output node number, t_{kp} is the desired value of the *kth* output node for input pattern p, and z_{kp} is the *kth* element of the actual output associated with input p (Antar et al., 2006).

4 Model development

The total daily observation was divided into training, validation and cross-validation sets prior to the model building. The cross-validation is used to avoid any overfitting during training. In this study, 60, 25 and 15% of data was used for training, validation and cross-validation, respectively.

It is worth noting that the method used to divide the data has significant impact on the results. In other words, the network may use low or high flow samples and give a yield of great precision for training set but fails to simulate outside the range of the training data (Tokar and Johnson, 1999; Shahin et al., 2000). In this study, the rainfall and runoff data were randomized prior to training the network to avoid this problem. The randomization of input data was emphasized by many researchers such as Bras and Rodríguez-Iturbe (1985) and Ochoa-Rivera et al. (2002) for hydrologic variables with large degree of variability and uncertainty. They stated that using only historical data as inputs into ANN may result in a scarcely documented response. However, for the randomization may lead to loosing the historical memory of the basin in cases of the application of ANN for streamflow time series forecasting.

Table 1. The rainfall and runoff variables used to construct neural network with the cross correlation (CCC) and Autocorrelation coefficients (ACC).

Variable	CCC	ACC
$x(1)$: R1(t-1), Daily rainfall of station(1) at lag time 1-day,	0.133	–
$x(2)$: R1(t-2), Daily rainfall of station(1) at lag time 2-days	0.119	–
$x(3)$: R2(t-1), Daily rainfall of station(2) at lag time 1-day	0.076	–
$x(4)$: R3(t-2), Daily rainfall of station(3) at lag time 2-days	0.048	–
$x(5)$: Q(t-1), Daily streamflow at lag time 1-day	–	0.935
$x(6)$: Q(t-2), Daily streamflow at lag time 2-days	–	0.901

In the first step, we select the input data for MLP networks. According to the autocorrelation properties of daily rainfall and runoff time series and the cross correlation between daily rainfall and runoff series, different input variables can be used for ANN. However, due to the possibility of zero rainfall and runoff in the Zayandehrud basin, the initial efforts to construct the ANN showed that data transformation is necessary to reduce the variance of rainfall and runoff time series. In this study, we apply standardized rainfall and runoff time series to construct the ANN. After trial and error, the following standardized variables were selected as input and output data of ANN. The cross-correlation coefficients (CCC) between streamflow and selected rainfall variables and the autocorrelation coefficients (ACC) of streamflow time series at different lags are also given in Table 1. All the coefficients are significant at 1% level.

The output of the model is streamflow discharge of the Plasjan River (Q_t) at the outlet of the basin. We tested different MLP architectures and found that the MLP with 1-hidden layer (i.e. MLP1) is not appropriate while other MLPs (MLP2, MLP3, MLP4 and MLP5) are suitable networks for modeling rainfall-runoff relationship of Plasjan basin. The random order was used for training material and the Levenberg-Marquardt back Propagation algorithm, as the most efficient algorithm (Ramirez-Beltran and Montes, 2002) was used to train neural network and training was stopped at 1000 epochs. The learning rate was set from 0.7 to 0.1 and the learning rule is momentum. Each MLP network contained 7 hidden units positioned in each hidden layer. The performance of these networks is depicted in Fig. 3a–d which shows the network estimated streamflow against observed validation data set.

5 Empirical model

In order to compare ANN with an empirical model, we also develop a multiple linear regression (MLR) model for rainfall-runoff relationship. The discharge of Plasjan River (Q_t) is selected as the dependent variable and the input variables of ANN are selected as independent variables. The best-fit model is estimated using a stepwise procedure and selected based on the highest coefficient of determination (R^2) and residual test for normality. Finally, the following regression model is estimated:

$$Q_t = 0.814x_6 - 0.043x_2 - 0.032x_1 + 0.103x_4 + 0.146x_3 + 0.008x_5$$

The performance of regression model is depicted in Fig. 4 for the validation data set.

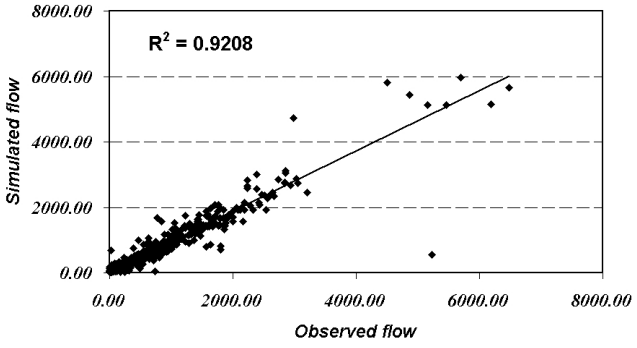

Fig. 4. Scatter plot of observed versus simulated streamflow (m^3/s) with regression model.

6 Comparison of the models: comprehensive multi-criteria analysis

6.1 Global statistics

The performance of hydrologic models is usually evaluated by the comparison of desired and model predicted values. This comparison can be done by graphical or numerical methods. The global statistics (Root Mean Squared Error, Correlation Coefficients, the Coefficient of Efficiency (CE), Index of Agreement (Legates and McCabe, 1999; Harmel and Smith, 2007)) are usually used for model calibration or comparison of different models.

Table 2. Performances indices for MLP and regression models.

Criteria	ANNs				Regression model
	MLP2	MLP3	MLP4	MLP5	
AME	2246.52	1296.85	**1104.05**	2972.94	4684.42
CE (%)	82	96.5	**97.3**	88.35	92
IoAd	0.930	0.991	**0.993**	0.960	0.970
MAE	190.95	65.77	**53.24**	127.39	56.93
MARE	6.85	1.88	1.25	4.03	**0.66**
MdAPE	41.78	13.68	**11.27**	33.22	12.01
ME	−53.49	−6.11	−3.59	−10.21	**0.003**
MRE	−6.72	−1.79	−1.17	−3.87	**−0.58**
MSRE	406.46	31.87	14.45	141.72	**6.509**
PDIFF	1262.15	−1066.84	**190.30**	1636.13	519.43
PEP	19.47	−15.90	**2.94**	25.24	8.015
PI	0.9119	0.9824	**0.9865**	0.9417	0.9600
R^2	0.9365	0.9582	**0.9635**	0.9222	0.9208
RAE	0.53	0.18	**0.14**	0.35	0.15
R4MS4E	435.28	282.48	**238.26**	555.62	732.54
RMSE	247.17	112.21	**97.15**	201.1	165.77
RVE	−0.121	−0.014	**−0.008**	−0.023	0.009

It is noted that Unal et al. (2004) validated simulation models by using statistical characteristics such as average, standard deviation, skewness coefficient, autocorrelation coefficient, maximum and minimum values, and performance criteria such as relative error, absolute error, frequency of success, ranges of relative and absolute errors. Liu et al. (2003) validated the results of the ANN models with root mean square error and determination coefficient.

Very recently Aksoy and Dahamsheh (2009) used a multi-criteria validation of ANN models developed for Jordan by using graphical and numerical measures including the forecasted and observed time series, scatter diagram, the residual time series between the forecast and observation, mean absolute and relative errors between the forecast and observation, dimensionless mean absolute error and dimensionless mean relative error between the forecast and observation. Additionally following performance measures are adopted: Determination coefficient to quantify the linearity between the forecast and observation, mean square error, mean absolute error; and a and b (the slope and the intercept) in the best-fit linear line of the scatter diagram between the forecast and observation.

As there is no single definite evaluation test, it is important to apply a multi-criteria assessment of ANN skill (Dawson et al., 2002; Kumar et al., 2005). These statistics are summarized in a recent paper by Dawson et al., (2007) and could be calculated automatically on the Hydrotest website available at http://www.hydrotest.org.uk. We apply 17 criteria which are listed in Appendix A. The reader is referred to Dawson et al. (2007) for the mathematical formulation of these criteria.

These error statistics are given for different MLP networks in Table 2. It is evident that the MLP4 network is better than all other networks. Compared with regression model and according to some criteria, i.e. MARE, ME, MRE and MSRE, the regression model performs better than MLP4 network. However, these criteria that are unbounded do not necessarily show the preference of regression model because the low score of these criteria do not necessarily indicate a good model in terms of accurate forecasts, since positive and negative errors will tend to cancel each other out.

6.2 Statistical validation

Although the above error statistics provide relevant information on the overall performance of the models they do not provide specific information about model performances at high or low flows, which are of critical importance in flood or low flow contexts. This study proposes other criteria to evaluate the performance of ANNs, especially for the rainfall-runoff relationship. These criteria are divided into the following graphical and numerical tests:

6.2.1 Graphical tests

In this section we compare the box-plot and probability plot of the observed and computed flows. The probability plot of the observed and simulated streamflow is fitted by Blom's method (Blom, 1958) which is based on the fractional rank of the observation. The parameters of the probability function are estimated by maximum likelihood method.

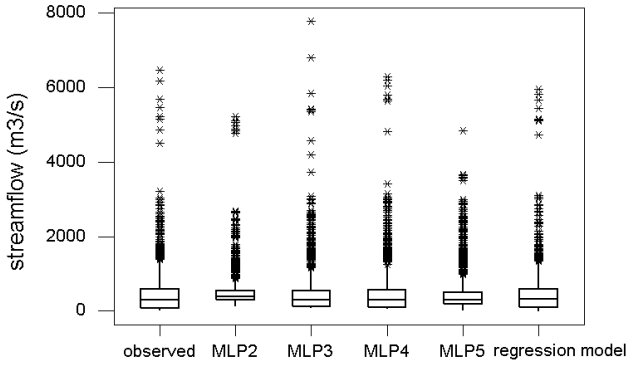

Fig. 5. Comparison of box-plots of observed runoff and simulated runoff by MLP networks.

Table 3. Test results (p-values) of non-parametric methods for the difference between observed and ANN and regression-simulated streamflow data at 95% confidence level.

Nonparametric Method	ANNs				Regression Model
	MLP2	MLP3	MLP4	MLP5	
Wilcoxon	0.003	0.035	0.312	0.008	0.026
Levene	0.002	0.023	0.073	0.005	0.012
K-S	0.001	0.011	0.028	0.004	0.001

These tests are useful for visual comparison of the upper or lower tail of the distribution of the observed and estimated streamflow. The box-plots of observed and estimated streamflow for different MLP networks and regression model are illustrated in Fig. 5 From box-plots, it is clear that the MLP4 network and regression model most closely match the observed streamflow, especially for high flows.

The probability plots for the observed and MLP4 network reveal that the distribution of observed and MLP4-estimated streamflow data are more similar for a normal distribution (Fig. 6) than for a gamma distribution (Fig. 7) because the lower tail of a gamma distribution is very different for observed and estimated streamflow. The gamma distribution for MLP2 and MLP5 networks are also presented in Fig. 8. It is clear that the networks are not able to reproduce the probability distribution of the observed streamflow and there is a significant difference in both upper and lower tails of the quantile distribution of streamflow. The probability plots of estimated streamflow by regression model are also presented in Fig. 9. The normal probability plot (Fig. 9a) is similar to the normal probability plot of observed streamflow and MLP4 network (Fig. 6a and b, respectively). However, the Normal and Gamma probability plots for regression and observed streamflow are different, particularly for lower tail of distribution. These probability plots illustrate that neither the MLP network nor the regression model are able to simulate

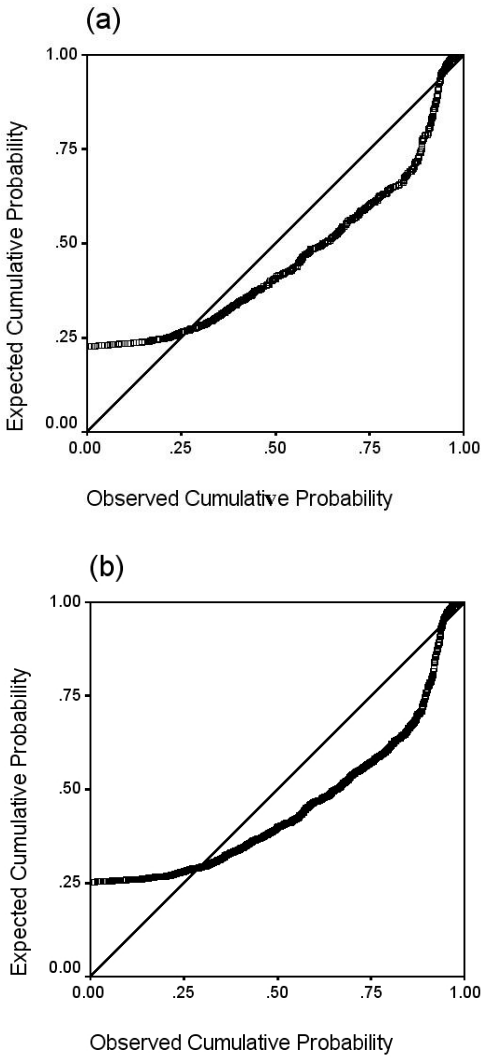

Fig. 6. Normal cumulative probability plots for **(a)** observed and **(b)** MLP4 simulated streamflow.

the probability distribution of the observed streamflow (see also Table 3 and Sect. 6.2.2).

Although the MLP4 network seems to be a better network than other networks and does not achieve very much better results than those of the regression model for rainfall-runoff modeling of the Zayandehrud basin, it would wise to check the validation of the ANN network by statistical measurements presented in the following section.

6.2.2 Statistical tests

In this section, we suggest useful statistical tests to evaluate the performance of the ANNs and to compare these ANNs with each other. These statistical methods include non-parametric tests to compare mean, standard deviation and the cumulative distribution function (CDF) of observed and estimated streamflow. Khan et al. (2006) used these statistics

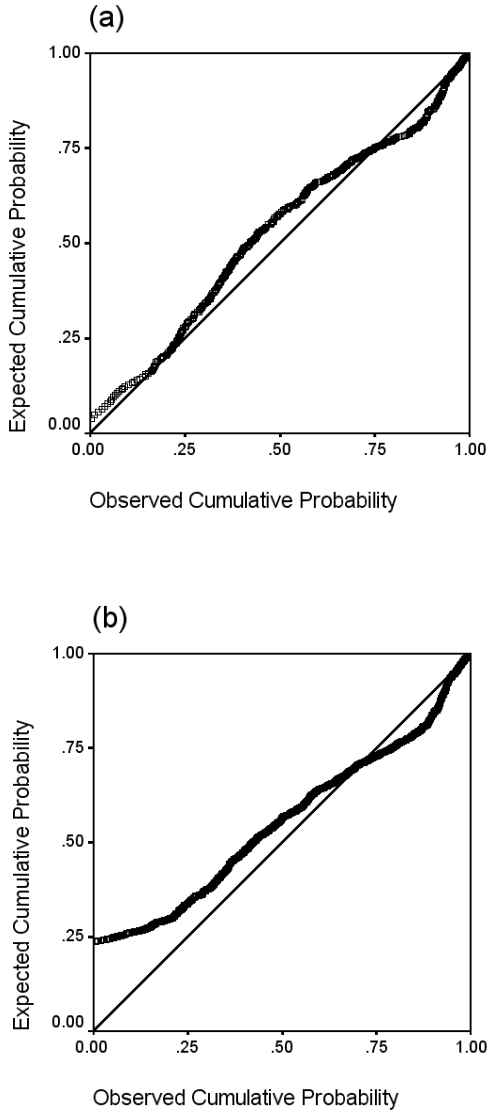

Fig. 7. Gamma cumulative probability plots for **(a)** observed and **(b)** MLP4 simulated streamflow

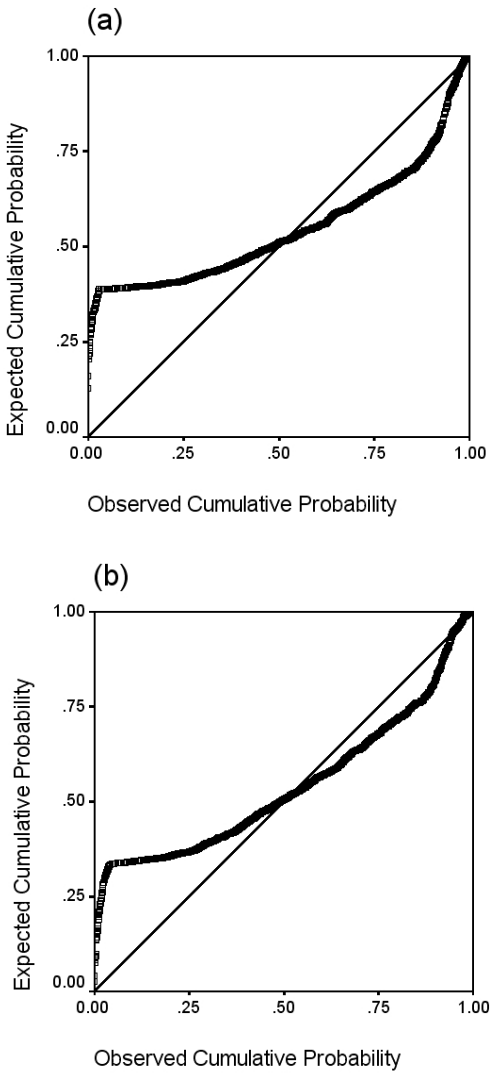

Fig. 8. Gamma cumulative probability plots for **(a)** MLP2 and **(b)** MLP5 simulated streamflow

to compare different precipitation downscaling methods including ANN and Modarres (2007) used a non-parametric method to evaluate drought time series forecasting with ARIMA model for the Plasjan River.

Non-parametric test for the difference of two population means

The Wilcoxon rank sum method (Conover, 1980) is a robust non-parametric method for constructing a hypothesis test p-value for $\mu_1 - \mu_2$ (difference of two population means). At any significance level greater than the p-value, one rejects the null hypothesis, and at any significance level less than the p-value one accepts the null hypothesis. For example, if p-value is 0.04, one rejects the null hypothesis at a significance level of 0.05, and accepts the null hypothesis at a significance level of 0.01. The null hypothesis of Wilcoxon test can be defined at:

$$H_0 : \mu_1 - \mu_2 = 0 \quad (6)$$

$$H_a : \mu_1 - \mu_2 \neq 0 \quad (7)$$

Non-parametric test for the equality of two population variances

The equality of two population variances can be tested using Levene's test. The hypothesis for the Levene's test can be defined as (Khan et al., 2006):

$$H_0: \sigma_1 = \sigma_2 = ... = \sigma_k \quad (8)$$

$$H_a: \sigma_i \neq \sigma_j \neq ... \neq \sigma_k \text{ for at least one pair } (i, j) \quad (9)$$

In performing Levene's test, a variable X with sample size N is divided into k subgroups, where N_i is the sample size of the ith subgroup, and the Levene test statistic is defined as:

$$W = \frac{(N-K)\sum_{i=1}^{k} N_i(\bar{Z}_i - \bar{Z})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{N_i}(Z_{ij} - \bar{Z}_i)^2} \quad (10)$$

where Z_{ij} is defined as:

$$Z_{ij} = \left|X_{ij} - \bar{X}_i\right| \quad (11)$$

where $\bar{X}_i$ is the median of the ith subgroup, $\bar{Z}_i$ is the group mean of the Z_{ij} and $Z_i$is the overall mean of the Z_{ij}. The Levene's test rejects the hypothesis that the variances are equal if

$$W > F_{(\alpha,k-1,N-k)} \quad (12)$$

where $W > F_{(\alpha,k-1,N-k)}$ is the upper critical value of the F distribution with $k-1$ and $N-k$ degrees of freedom at a significant level of α.

Non-parametric test for equality of CDFs of two populations

Kolmogorov–Smirnov (K-S) non-parametric test (Conover, 1980) is used to compare cumulative distribution function (cdf) of observed and simulated streamflow series. Suppose, $F_1(x)$ and $F_2(x)$ are cdfs of two sample data of a variable x. The null hypothesis and the alternative hypothesis concerning their cdfs are:

H_0:$F_1(x) = F_2(x)$ for all x

H_a:$F_1(x) \neq F_2(x)$ for at least one value of x

and the test statistics, Z is defined as

$$Z = \sup_x |F_1(x) - F_2(x)| \quad (13)$$

which is the maximum vertical distance between the distributions $F_1(x)$ and $F_2(x)$. If the test statistic is greater than the critical value, the null hypothesis is rejected.

To evaluate the performance of MLP networks, we apply the tests in two cases. First, the observed and simulated streamflow time series are compared for the overall validation test. For the second case, the percentiles of observed and simulated streamflow time series are compared in order to check the validation of ANNs for the prediction of high, medium and low streamflows. The streamflow time series are divided into the first 0–25% (P1), the second 25–75% (P2) and the third 75–100% (P3) percentiles.

Table 3 indicates the results of non-parametric tests at 95% significance level for the first case. It is evident that none of the networks can simulate statistical characteristics of the observed streamflow except multi-layer perceptron with 4 hidden layers (MLP4) because all estimated p-values are less than 0.05 except for the MLP4 network.

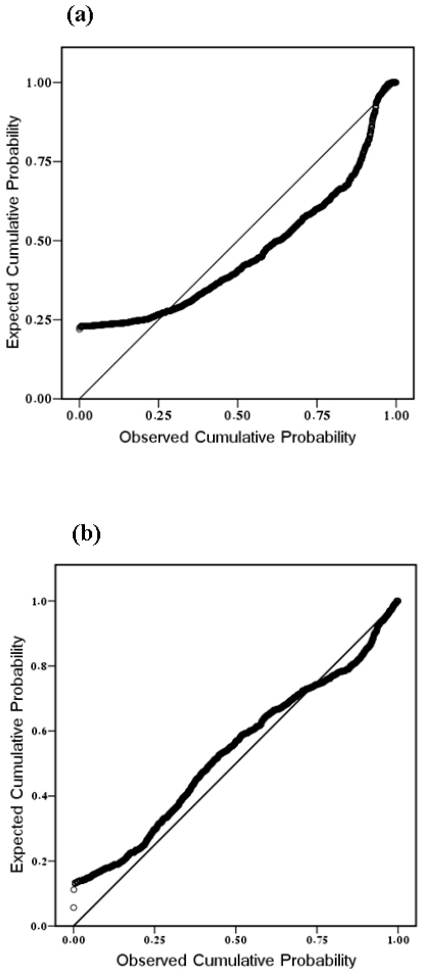

Fig. 9. Normal **(a)** and Gamma **(b)** cumulative probability plots for simulated streamflow by regression model

Although the p-value of the K-S test is close to 0.05 for MLP4 network, the Kolmogorov-Smirnov test does not verify the equality of the CDFs of the observed and ANN simulated streamflow. Table 3 confirms the dissimilarity in the probability plot of the observed and simulated streamflow by different ANN networks and regression model (see also Figs. 7, 8 and 9).

For comparing high, medium and low flow in the second case, the streamflow time series are divided into three percentile groups and the above non-parametric tests are applied for each group. Table 4 represents the global statistics of the networks for each percentile group. Those values highlighted in bold in this table indicate the "best" model out of the five models when assessed using each particular evaluation metric. For example, according to IoAD criterion, the MLP4 network gives the best performance for the third per-

Table 4. Performances indices for MLP and regression models and different percentile groups (P1, P2 and P3).

Criteria	MLP2			MLP3			MLP4			MLP5			Regression Model		
	P1	P2	P3	P1	P2	P3	P1	P2	P3	P1	P2	P3	P1	P2	P3
AME	580.95	418.24	2246.52	607.66	361.82	1943.9	625.06	976.76	1429.71	620.9	976.76	2972.94	**23.19**	**28.91**	**1390.61**
CE (%)	22	14	81	94	**89**	93	**95**	85	94	75	47	83	22	**89**	**98**
IoAd	0.1	0.69	0.95	0.31	0.97	0.98	0.39	0.96	**0.99**	0.17	0.81	0.95	**0.78**	**0.99**	**0.99**
MAE	273.89	118.44	251.33	73.35	37.75	118.29	48.23	32.07	105.73	158.69	32.07	179.95	**20.21**	**8.81**	**44.5**
MARE	25.41	0.72	0.32	6.99	0.17	0.13	4.64	0.13	0.10	14.90	0.13	0.44	**1.81**	**0.05**	**0.02**
MdAPE	1330.22	35.15	22.83	356.66	9.42	7.17	227.11	6.90	6.27	765.54	6.90	11.20	**97.22**	**2.16**	**2.05**
ME	−273.89	−89.59	−243.08	−73.35	4.94	−36.03	−48.23	**3.70**	**−20.99**	−158.67	3.70	−149.39	**−20.21**	−4.51	43.11
MRE	−25.41	−0.66	−0.32	−6.99	−0.07	−0.09	−4.64	−0.04	−0.05	−14.90	−0.26	−0.42	−1.81	−0.04	**0.02**
MSRE	1600.33	1.25	0.25	125.53	0.068	0.13	58.72	−0.04	0.03	558.01	0.05	11.79	**7.81**	**0.007**	**0.001**
PDIFF	−532.95	−86.41	−1262.15	−559.66	−151.82	1296.85	−577.06	−872.76	**−190.3**	−572.9	−872.76	−1636.13	**−19.86**	**8.12**	657.19
PEP	−761.24	−14.54	−24.18	−799.4	−25.55	16.67	−824.25	−146.92	**−3.02**	−818.31	−146.92	−33.77	**−28.36**	**1.36**	10.14
PI	−42 6337.79	−11 251.92	−0.84	−3421.35	−1390.95	0.50	−18 043.25	−1917.65	**0.52**	−14 6902.42	−1917.65	−0.72	−2312.78	−61.3	−0.80
R^2	0.40	0.72	0.96	0.42	**0.95**	0.97	0.42	0.92	**0.98**	0.32	0.69	0.94	**0.93**	**0.95**	**0.98**
RAE	18.10	0.89	0.61	4.84	0.28	0.20	3.18	0.24	0.18	10.48	0.24	0.34	**1.34**	**0.06**	**0.08**
R4MS4E	277.04	172.25	607.73	135.60	82.92	490.81	137.01	185.33	407.15	177.23	185.33	785.21	**20.30**	**13.2**	**310.77**
RMSE	274.6	61.3	410	78.1	41.2	230	79.6	50.5	209	162.1	61.3	368	**20.25**	**10.6**	**100.91**
RVE	−10.89	−0.28	−0.27	−2.91	0.016	−0.03	−1.91	**0.01**	**−0.01**	−6.31	−0.05	−0.15	**−0.80**	**−0.01**	0.038

Table 5. Test results (p-values) of non-parametric methods for the difference between observed and ANN and regression-simulated streamflow percentile groups at 95% confidence level.

Nonparametric Method	MLP2			MLP3			MLP4			MLP5			Regression Model		
	P1	P2	P3	P1	P2	P3	P1	P2	P3	P1	P2	P3	P1	P2	P3
Wilcoxon	0.001	0.001	0.001	0.001	**0.466**	**0.534**	0.001	**0.607**	**0.714**	0.001	0.028	0.004	0.001	**0.37**	**0.43**
Levene	0.001	0.001	0.001	0.028	0.01	0.004	**0.545**	**0.301**	**0.44**	0.016	0.001	0.045	0.001	**0.063**	**0.47**
K-S	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.021	0.011

centile compared with other networks while the regression model gives the best results for all percentiles. It can also be seen that the MLP4 network is the best model according to PDIFF for the third percentile while the regression model gives better results for the first and second percentile. As an another example, according to the correlation coefficient, the MLP3 network and regression model have the highest R^2 for the second percentile while for the third percentile the MLP4 network and regression model give the same results. Therefore, it is difficult to select the best model based on one single criterion.

For the first percentile, or the low flows, MLP4 network performs better than other networks based on most of the criteria. However, for some criteria such as AME, PDIFF and PEP, the MLP2 is better than other MLP networks. These criteria illustrate the error of the highest output between the modeled and the observed dataset which is not suitable for low flow error measurement. For the second percentile, the same results can be seen for MLP4 and MLP3 networks. However, for the third or the upper percentile which shows the efficiency of the model for estimating high flows, the MLP4 is the best network. Jain and Srinivasulu (2004) also mentioned that the high flows can be effectively modeled by MLP networks. However, they concluded that for medium and low flow simulation by ANNs, the use of genetic algorithm (GA) may be advantage because the watershed condition is much more complex and dynamic for low flows than high flows.

On the other hand, the regression model seems to be more effective than MLP networks for rainfall-runoff modeling according to almost all criteria and different percentiles. The regression model scores well in terms of most of the metrics. However, the MLP is still better than regression model in terms of PDIFF and PEP. In other words, the MLP4 networks estimate high flows more accurate than regression model while the regression model performs better than MLP4 for medium and low flows. The results of the total data (Table 2) also indicated the better performance of MLP4 network over regression model for high flows.

Table 5 presents the results of non-parametric tests for three percentile groups. It is found that MLP2 is still an insignificant model for rainfall-runoff relationship modeling for the Plasjan River because all p-values are below 0.05.

The MLP3 network can reproduce the mean of observed streamflow for the second and third percentiles but the network is weak in simulating standard deviation and the probability distribution of the observed streamflow because the p-values are below 0.05. The MLP4 network indicates the best simulation results for the mean and standard deviation of the observed streamflow similar to the MLP5, it also fails to reproduce the mean and standard deviation of the observed streamflow. On the other hand, the regression model is similar to the MLP4 network.

However, the Kolmogorov-Smiornov test demonstrates that neither the ANNs nor the regression model can reproduce the probability distribution of streamflow in the validation phase of the modeling. Although the MLP4 network and regression model are able to simulate the mean and standard deviation of the observed streamflow but they could not reproduce the probability distribution of the observed streamflow.

7 Conclusion and summary

Artificial neural networks are powerful tool for modeling nonlinear relationships in hydrology such as rainfall-runoff relationship. The validation phase of the neural network modeling plays an important role in the efficiency testing of the modeling. The global statistics are common methods used in this phase. However, the findings reported in this paper show that the global statistics broadly reflect the accuracy of the model but are insufficient indicators of the best ANN because they do not capture the mean, standard deviation and probability distribution of the observed streamflow. This paper also illustrate s the dangers of relying on one metric alone to evaluate and select different models.

Although the multi layer perceptron with four hidden layers was selected as the best neural network based on the global statistics, it failed to reproduce the probability distribution of observed streamflow. The MLP4 network also gives better results than regression model for entire testing data set.

However, it is important to reproduce streamflow statistics such as the mean, standard deviation and probability distribution for high, medium and low flows. According to the objectives of the ANN, i.e. flood or low flow simulation or forecasting, it is very important to check the accuracy of the ANN output separately in future studies. For example, the best ANN in this study, MLP4, gives better estimation for high flows than for low flows. But the MLP4 network is not able to reproduce the probability functions of different percentiles according to the Kolmogorov-Smirnov test. Although the regression model is better than ANNs based on different criteria, it is also inadequate to reproduce probability distribution of the observed streamflow.

In general, the findings of this study conclude that, for validation phase of ANN, the common global statistics are not sufficient and relying on one measurement is not relevant. A multi-criteria assessment based on different global and non-parametric tests is essential for verifying and selecting an optimum ANN. One should use a range of methods to evaluate the methods. This study also shows the advantage of the application of empirical, physical or conceptual models together with ANN because some of these models may give better results with more simple modeling procedure than ANNs.

Edited by: J. Liu

Appendix A

Abbreviations for global criteria used in this study	
AME:	Absolute maximum error
CE:	Coefficient of efficiency
IoAd:	Index of agreement
MAE:	Mean absolute error
MARE:	mean absolute relative error or
RME	Relative mean error
MdAPE:	Medium absolute percentage error
ME:	Mean error
MRE:	mean relative error
MSRE:	mean squared relative error
PDIFF:	Peak difference
PEP:	Error in peak
PI:	coefficient of persistence
R^2:	Correlation of determination
RAE:	Relative absolute error
R4MS4E:	Fourth root mean quadrupled Error
RMSE:	Root Mean squared error
RVE:	Relative volume error

References

Aksoy, H. and Dahamsheh, A.: Artificial neural network models for forecasting monthly precipitation in Jordan, Stoch. Environ. Res. Risk Assess., doi:10.1007/s00477-008-0267-x (in press and online available), 2009.

Antar, M. A., Elassiouti, I., and Allam, M. N.: Rainfall-runoff modelling using artificial neural networks technique: a Blue Nile catchment case study, Hydrol. Proc., 20, 1201–1216, 2006.

Blom, G.: Statistical estimates and transformed beta variables, John Wiley and Sons, NY, USA, 174 pp., 1958.

Conover, W. J.: Practical Nonparametric Statistics, 3rd, Wiley, NY, USA, 592 pp., 1999

Coulibaly, P. and Evora, N. D.: Comparison of neural network methods for infilling missing daily weather records, J. Hydrol., 341, 27–41, 2007.

Dawson, C. W., Abrahart, R. J., and See, L. M.: Hydrotest: A web-based toolbox of evaluation metrics for the standardized assessment of hydrological forecasts, Environ. Model. Softw., 22, 1034–1052, 2007.

Dawson, C. W., Wilby, R. L., and Chen, Y.: Evaluation of artificial neural network techniques for flow forecasting in River Yangtze, China, Hydrol. Earth Syst. Sci., 6, 619–626, 2002, http://www.hydrol-earth-syst-sci.net/6/619/2002/.

Dawson, C. W. and Wilby, R.: An artificial neural network approach for rainfall-runoff modeling, Hydrol. Sci. J., 43, 47–66, 1998.

Jain, A. and Srinivasulu, S.: Development of effective and efficient rainfall-runoff models using integration of deterministic, real coded genetic algorithms and artificial neural network techniques, Water Resour. Res., 40, W04302, doi:10.1029/2003WR002355, 2004.

Harmel, R. D. and Smith, P. K.: Consideration of measurement uncertainty in the evaluation of goodness-of-fit in hydrologic and water quality modeling, J. Hydrol., 337, 326–336, 2007.

Harpham, C. and Dawson, C. W.: The effect of different basis functions on a radial basis function network for time series prediction: A comparative study, Neurocomputing, 69, 2161–2170, 2005.

Hsu, K., Gupta, H. V., and Sorooshian, S.: Artificial neural network modeling of the rainfall–runoff process, Water Resour.

Res., 31(4), 2517–1530, 1995.
Khan, M. S, Coulibaly, P., and Dibike, P.: Uncertainty analysis of statistical downscaling methods, J. Hydro., 319, 357-382, 2006
Kumar, A. R. S., Sudheer, K. P., Jain, S. K., and Agarwal, P. K.: Rainfall-runoff modeling using artificial neural networks: comparison of network types, Hydrol. Proc., 19, 1277–1291. 2005.
Legates, D. R. and McCabe Jr., G. J.: Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation, Water Resour. Res. 35(1), 233–241, 1999.
Liu, J., Savenije, H. H. G., and Xu, J.: Forecast of water demand in Weinan City in China using artificial neural networks, Phys. Chem. Earth, 28(4–5), 219–224, 2003.
Modarres, R.: Streamflow drought time series forecasting, Stoch. Env. Res. Risk A., 21, 223–233. 2007.
Ramirez-Beltran, N. and Montes, J. A.: Neural networks to model dynamic systems with time delays, IIE T., 34, 313–327, 2002.
Tokar, A. S., and Markus, M.: Precipitation-runoff modeling using artificial neural networks and conceptual models. J. Hydrol. Eng., 5, 156-161, 2000.
Tokar, A. S. and Johnson, P. A.: Rainfall-runoff modeling using artificial neural networks, J. Hydrol. Eng., 4, 232–239, 1999.
Smith, J. and Eli, R. B.: neural Network models of rainfall-runoff process, J. Water Res., 4, 232–239, 1995.
Shahin, M. A, Maier, H. R., and Jaksa, M. B.: Evolutionary data division methods for developing artificial neural network models in geotechnical engineering. Research Report No. R 171., Department of Civil and Environmental Engineering, The University of Adelaide, Australia, 2000.
Unal, N, E., Aksoy, H., and Akar, T.: Annual and monthly rainfall data generation schemes, Stoch. Environ. Res. Risk Assess., 18, 245–257, 2004.
Zhang, B. and Govindaraju, R. S.: Geomorphology-based artificial neural networks for estimation of direct runoff over watersheds, J. Hydrol., 273, 18–34, 2003.