Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 120 (2017) 237–244

Procedia Computer Science

www.elsevier.com/locate/procedia

9th International Conference on Theory and Application of Soft Computing, Computing with Words and Perception, ICSCCW 2017, 24-25 August 2017, Budapest, Hungary

Monthly streamflow estimation using wavelet-artificial neural network model: A case study on Çamlıdere dam basin, Turkey

Gökçen Uysal[a,b]*, Ali Ünal Şorman[c]

[a]Anadolu University, Department of Civil Engineerring, 2 Eylul Campus, Eskisehir, 26555, Turkey
[b] University of Duisburg-Essen, Institute of Hydraulic Engineering and Water Resources Management, Essen, 45141, Germany
[c] Near East University, Department of Civil Engineering, POBOX:99138, Nicosia, North Cyprus, Mersin 10 Turkey

Abstract

Data driven techniques have become well-known application in hydrology in which physical processes are highly nonlinear. They require detailed analyses of different input combinations, selecting the appropriate model structures, assigning the optimization parameters etc. Besides, the model performance are also highly correlated with additional analysis techniques. In this study, the value of using different data sets such as air temperature, precipitation, evaporation and streamflow records, evapotranspiration around the basin are investigated to estimate monthly inflows using a multi-layer perceptron network model. Since the noise always exists in the time-series data, Discrete Wavelet Transform (DWT) is applied for data decomposition. Çamlıdere dam basin, which is one of the vital water supply reservoir of the capital city of Turkey, Ankara, is selected as an application area. The model sets are employed using 1960 – 2016 monthly observed data. The reliability of the modelled flows are verified with: coefficient of determination (R^2), Nash-Sutcliffe model efficiency (NSME), root mean square error (RMSE) and mean absolute error (MAE). According to the results, instead of increasing input vector number, application of data pre-processing have more impact to capture especially high flows. Decomposed discharge data together with meteorological other inputs perform 0.85 – 0.73 both for R^2 and NSME for training and testing periods, respectively.

Peer-review under responsibility of the scientific committee of the 9th International Conference on Theory and application of Soft Computing, Computing with Words and Perception.

Keywords: monthly streamflow estimation; neural netwok; multi-layer perceptron; wavelet transform.

* Corresponding author. Tel.: +90-222-321-35-50/6623; fax: +90-222-323-95-01.
E-mail address: gokcenuysal@gmail.com

1877-0509 © 2018 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the 9th International Conference on Theory and application of Soft Computing, Computing with Words and Perception.
10.1016/j.procs.2017.11.234

1. Introduction

Reservoirs are still one of the important component in water resources systems. They are usually used for multi-purpose demands e.g. water supply, hydropower, irrigation, navigation, therefore optimal operation of them is an important issue. The operational decisions are attributed to the different facets such as operational strategy of the system, resource allocation, environmental and physical constraints etc. Besides, the system performance is highly related with accurate future prediction of hydrological response of the basin even if a perfect decision support system would be utilized. Models to mimic streamflow processes can be classified in different types such as distributed physical models (Refsgaard & Knudsen, 1996), lumped/distributed conceptual models (Lindström et al., 1997), stochastic models (Valipour et al., 2013), statistical models and soft computing methods (Uysal et al., 2016). Data availability is one of the main problem to construct a reliable model. Though physical models are more representative of the rainfall-runoff processes by using parameters which are related directly to the physical characteristics of the catchment (topography, soil, vegetation, geology etc.) and providing spatial variability of physical and meteorological conditions by its distributed framework, they have a predictive capability at model grid scale due to limitations of data availability and uncertainty of model conceptualizations (Refsgaard et al., 2016). When the model becomes more physical representation, the performances would be increased, however more information is required which is not always practical (Karimi et al., 2016).

Contrary, soft computing approaches have recently been exploited in hydrological modeling (Kentel, 2009). There are some different types like Fuzzy Logic (FL), Evolutionary Computation (EC), Machine Learning (ML) and Probabilistic Reasoning (PR) etc. with the latter subsuming belief networks and parts of learning theory. A neural network is characterized by its architecture that represent the pattern of connection between nodes, its method of determining the connection weights, and the activation function (Fausett, 1994). One of the significant advantage of these models, supervised training (which provides a class of the functions matches the targets such as discharge) with different data sets is possible, even if the data is noisy and contaminated with errors. A network with sufficient parameters can approximate any nonlinear function (target) to any degree of accuracy by flexible nonlinear transfer functions (Kisi & Sanikhani, 2015). Therefore, application of them in hydrology and water resources have become common, since the models can learn, memorize and generalize knowledge from data sets, which makes it potential to solve complex, non-linear problems (Govindaraju, 2000a). Determining the elements of the artificial neural networks issue that affect the forecasting performance of artificial neural networks, and it should be carefully considered. Networks might be within hourly, daily (Uysal et al., 2016) and monthly (Shiri & Kisi, 2010) time intervals depending on different purposes. There are many user defined parts (selection of stopping criteria, normalization techniques, determination of model structure, optimization parameters etc.) in their methodologies, hence it is recommended to try several architectures and select the best algorithm for different data sets. Also, extreme events can create problems in any data analysis and modeling by having the sample mean and standard deviation to be much smaller/higher than the population values.

Recently, hybrid systems which performs better compared to conventional counterparts e.g. the integration of artificial neural networks with conceptual models (Chen & Adams, 2006), wavelet and neuro-fuzzy conjunction model (Shiri & Kisi, 2010), ANFIS (Adaptive Neuro-Fuzzy Inference System) (Tayfur & Brocca, 2015) or hybrid intelligent systems (Bhadra et al., 2010) has been remarked. The wavelet-based seasonal models are more efficient than only Autoregressive models (i.e., ANN and ANFIS) for representing peak values (Nourani et al., 2014). In this study, monthly streamflows into Çamlıdere dam basin, which is the main water supply reservoir of the capital city of Turkey, estimated considering different input data combinations, splitting the training and validation instances. Further, a Discrete Wavelet Transform (DWT) is applied to inputs and improvement of the model performances are compared with pure neural network models.

2. Methodology

The neural network model based streamflows are generated with a feedforward Multi-Layer Perceptrons (MLP) model. These models are capable to represent the input-output relationship by layers and nodes. A node is a processor which is connected to the others by weights, whereas the nodes are generally arranged in layers. The output y that is transmitted to the other nodes is obtained by the following equation:

$$y = f(X.W_j - b_j) \tag{1}$$

where, $X = (x_1, \ldots, x_i, \ldots, x_n)$, $W_j = (w_{1j}, \ldots, w_{ij}, \ldots, w_{nj})$ and X is information from previous nodes, w_{ij} represents the connection weight from the i^{th} node in the preceding layer to this node, where b_j is bias, f is the activation function. The sigmoid function is a bounded, monotonic, nondecreasing function that provides a graded, nonlinear response (Govindaraju, 2000b). Since the relationship in the hydrology is mainly nonlinear, accordingly based on performances of different activation functions, sigmoid activation function is used in this study. Data are transformed to values between [0,1] using minimum and maximum of the data sets.

At the beginning, the initial weights of the nodes are assigned and then these weights are changed/corrected in the training. This is accomplished by a backpropagation learning algorithm that involves two phases: a feedforward phase in which the external input information at the input nodes is propagated forward to compute the output information signal at the output unit, and a backward phase in which modifications to the connection strengths are made based on the differences between the computed and observed information signals at the output units (Eberhard & Dobbins, 1990). The fast, efficient and robust the Levenberg-Marquardt (LM) backpropagation optimization technique (Marquardt, 1963; Levenberg, 1994) is used in the study. The LM algorithm updates the weights as:

$$w_{k+1} = w_k - \left[J^T J + \mu \mathrm{I}\right]^{-1} J^T \varepsilon \tag{2}$$

where w_{k+1} and w_k are weights during $(k+1)^{th}$ & k^{th} epoch, J is the Jacobian matrix that contains first derivatives of the network errors with respect to the weights and biases, μ is learning rate and ε is a vector of network errors.

The neurohydrologist must specify the number of hidden layers and neurons in each hidden layer (Dawson & Wilby, 2001). The input layer number depends on input vector used in the model. The autocorrelation analysis, cross correlation analysis and some physical considerations can be helpful in determining the number of inputs using a trial and error process. On the other hand, more hidden layer increases the element number in the model and does not drastically increases the efficiency of the model, therefore the developed model has three layer by having n input vectors, m neurons with one hidden layers, and one neuron in the output layer (n_m_1_1). The number of neurons are determined by the trial-and-error procedure. Model parameters are effected by stooping criteria. One way of selecting the best network parameters is training the model according to cross-validation by repeating K times, with K as large as possible, and estimating the errors by averaging across the K validation folds (Barrow & Crone, 2013). Similar to this approach, considering two circumstances (randomized initial weights and randomized data partitioning), we run the whole algorithm for several times (each starts with different randomized initial weights, different randomized partitioning and stopping epoch number depending on training and cross-validation instances), and get the average values both for training (including cross-validation) and testing (validation) period. Accuracy performance assessment is accomplished to decide on the number of runs. According to this, running the model for approximately 200 times is found to be satisfactory (after which the errors and relation measures become constant). The codes are generated using MATLAB version 2012a software (License number: 991708).

Data might be decomposed into subsets as a signal process. The transform of a signal might be continuous wavelet transform (CWT) or the discrete wavelet transform (DWT), whereas DWT necessitates not as much of computation time and is relatively simply applied compared to CWT. The formulations have been explained in detailed in literature by many researchers (Kalteh, 2016). In this study, we preferred DWT and 1-D wavelet decomposition is employed to perform a single-level wavelet decomposition of input signals using wavelet family of 'db1'.

The performance of the study is tested with 4 criteria defined as the square of correlation coefficient (R) called as coefficient of determination (R^2), Nash-Sutcliffe Model Efficiency (NSME), Root Mean Square Error (RMSE), Mean Absolute Error (MAE) denoted as:

$$R^2 = \left[\frac{\sum_{t=1}^{n} (Q_m^t - \overline{Q}_m)(Q_o^t - \overline{Q}_o)}{\sqrt{\sum_{t=1}^{n} (Q_m^t - \overline{Q}_m)^2} \sqrt{\sum_{t=1}^{n} (Q_o^t - \overline{Q}_o)^2}} \right]^2 \tag{3}$$

$$NSME = 1 - \frac{\sum_{t=1}^{n} (Q_o^t - Q_m^t)^2}{\sum_{t=1}^{n} (Q_o^t - \overline{Q}_o)^2} \tag{4}$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (Q_m^t - Q_o^t)^2}{n}} \tag{5}$$

$$MAE = \frac{\sum_{t=1}^{n} \left| Q_o^t - Q_m^t \right|}{n} \tag{6}$$

where Q_m^t is modelled flows, Q_o^t is observed flows, $\overline{Q}_m$ is average modelled flows, $\overline{Q}_o$ is average observed flows, n is the number of the data sets.

The coefficient of determination R^2 describes the percentage of total variation explained by the model. To determine systematically over or under-prediction of a model, NSME is popular in evaluation of runoff forecasting and it accounts for model errors in estimating the mean or variance of the observed data sets. ME is sensitive to extreme values thus describes the accuracy of the maximum values and timing of the discharges. Contrary to R^2 and ME, the model's fitness and diagnose the variation in the model errors can be determined by selecting the smallest of RMSE and MAE. While RMSE is a quadratic score, MAE is a linear score of the average magnitude of the errors. RMSE usually results in larger errors that occur in the vicinity of high flows in general; whereas MAE computes all deviations from the original data series and is not weighted towards high values.

3. Study Area and Data

The developed model is applied to Çamlıdere Dam basin that supplies most of the domestic water to the capital city Turkey (Ankara) where the total supplied water is about 800 000 m^3/day. The location of the basin is shown in Fig. 1. The data used in the study is the monthly records of precipitation (P), air temperature (T), potential evapotranspiration (PET) and streamflow (Q). There is streamflow stations data which was recorded by the State Hydraulic Works (DSI) in Çamlıdere basin for the period 1960 - 2016. The precipitation data were recorded by the State Meteorological Service (DMI) in Esenboga, Ankara, Kizilcahamam stations as shown in Fig.1. The average value of them are used in the modelling. The temperature data is obtained for the same period from Ankara station. PET losses in the basin is simply modeled by temperature values using Thornthwaite's formula and deliberated as an extra input data set.

For input data sets, current month's value as indicated by (n), previous one month values (n-1), previous two month values (n-2) are provided to estimate the current discharge. First 501 months (from January 1960 to September 2001) is used for training while left behind data (180 months) which are not used in any part of the training (from September 2002 to October 2016) is kept for testing. During the training process, a randomly selected 15 % (76 months) are used as cross-validation to determine the stopping epoch.

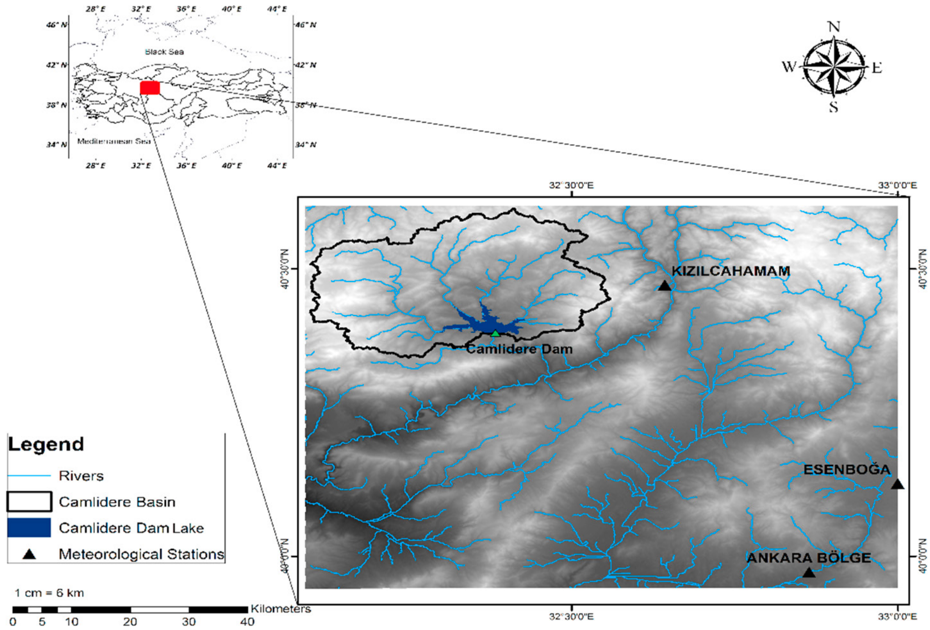

Fig. 1. The study catchment & meteorological stations

There are different method to partition the training and testing data sets. We consider that training and testing periods should be selected by having similarly statistical parameters. Hence, x_{mean}, x_{min}, x_{max}, S_x and C_{sx}, denoting the mean, minimum, maximum, standard deviation and skewness, respectively of all evaluated variables are presented in Table 1. Different periods are tried and abovementioned split is preferred. All periods presents similar discharge characteristics especially in terms of mean and maximum. Last column identifies the linear correlation (R) of each variable with runoff data. These relations are useful to understand the added value of extra inputs in data sets. The highest correlation with discharge values is observed between precipitation data for both training and testing periods.

Table 1. Input data statistics.

Input variable	Period	x_{min}	x_{mean}	x_{max}	S_x	C_{sx}	R
Runoff	training	0.3	14.0	114.8	19.2	1.92	1.000
[m^3/s]	testing	0.0	13.3	112.1	19.5	2.38	1.000
	whole	0.0	13.8	114.8	19.2	2.04	1.000
Precipitation [P_{ave}, mm]	training	0.0	39.0	145.7	28.4	0.88	0.458
	testing	0.0	39.3	187.8	31.5	1.37	0.431
	whole	0.0	39.1	187.8	29.2	1.05	0.450
Air Temperature [Temp, °C]	training	-4.2	11.8	26.5	8.2	-0.08	-0.294
	testing	-3.9	13.0	28.5	8.6	0.00	-0.322
	whole	-4.2	12.1	28.5	8.3	-0.05	-0.302
Potential Evaporation	training	0.0	5.9	16.9	4.9	0.40	-0.323
[PET, mm]	testing	0.0	5.9	17.6	5.3	0.59	-0.339
	whole	0.0	5.9	17.6	5.0	0.46	-0.327

4. Model results

Alternative models are generated with different input combinations. Due to the high correlation between streamflows, using previous months' observed streamflow data as an input vector is well-known application. It is noted that two month time lag is enough according to the partial auto-correlation function of daily streamflow data. The models inputs are available in Table 2. They are developed from simple to complex, consequently input number increases by the model name id. While only runoff data is considered for MLP_01 and MLP_02 models, the added value of meteorological data into them is analyzed for MLP_03 to MLP_05 models. In case of no runoff data availability, an alternative model is tested by providing only meteorological data into network by MLP_06 model.

Table 2. MLP models according to input vector selection.

Model name id	$Runoff_{(n-1)}$	$Runoff_{(n-2)}$	$Pave_{(n)}$	$Pave_{(n-1)}$	$Pave_{(n-2)}$	$Temp_{(n-1)}$	$PET_{(n-1)}$
MLP_01	X	-	-	-	-	-	-
MLP_02	X	X	-	-	-	-	-
MLP_03	X	-	X	-	-	-	-
MLP_04	X	-	-	X	-	X	-
MLP_05	X	X	-	X	-	X	-
MLP_06	-	-	-	X	X	X	-

4.1. Neural network model results

The performances of aforementioned models are summarized in Table 3. According to that, using previous month discharge data as input vector can explain/model 64 % and 55 % of the observed streamflows for training and testing periods, respectively (MLP_01). There is no drastic improvement adding $Q_{(n-2)}$ discharge into the model input vector (MLP_02). Adding extra inputs on MLP_01 model is not much appreciated according to R^2 and NSME. Considering current P_{ave} data (MLP_03) as an extra input to discharge based model (MLP_01) increases the performances especially for training part. As a substitute, one can model the monthly runoff only with meteorological data (MLP_06) by having quite similar performances to runoff data based models. Though different input combination experiments, the model results have still low performances, and requires to be improved by other techniques.

Table 3. MLP model performances.

model	period	R^2	NSME	RMSE [m^3/s]	MAE [m^3/s]
MLP_01	Training	0.64	0.64	11.5	6.7
	Forecast	0.55	0.55	13.0	7.5
MLP_02	Training	0.63	0.62	11.8	6.8
	Testing	0.56	0.56	12.9	7.6
MLP_03	Training	0.75	0.75	9.7	5.4
	Testing	0.57	0.55	13.0	6.7
MLP_04	Training	0.68	0.67	10.9	6.3
	Testing	0.59	0.58	12.5	7.0
MLP_05	Training	0.67	0.67	11.1	6.3
	Testing	0.60	0.59	13.3	7.4
MLP_06	Training	0.65	0.65	11.33	6.56
	Testing	0.52	0.52	13.48	7.79

4.2. Improvement with Discrete Wavelet Transform (DWT) Decomposition

An improvement is expected by decomposition of runoff/precipitation data. Even if MLP_01 and MLP_02 models have similar performances, pre-processing provide noteworthy improvement (if the one step before runoff data is decomposed into two subsets), then better performance is achieved by MLP_02_DWT (Table 4). DWT on previous month's runoff data (MLP_04_DWT_1) or precipitation data (MLP_04_DWT_2), the improvement is still less than MLP_02_DWT model. MLP_05_DWT improves the testing period performances. MLP_06_DWT, where $P_{(n-1)}$ and $P_{(n-2)}$ are separately decomposed into each two subsets, gives similar performances to MLP_04_DWT_2.

Table 4. DWA improvement performances on MLP model.

model	DWA	Period	R^2	NSME	RMSE [m^3/s]	MAE [m^3/s]
MLP_01_DWT	$Q_{(n-1)}$	Training	0.79	0.79	8.8	4.7
		Testing	0.55	0.54	13.2	6.4
MLP_02_DWT	$Q_{(n-1)}$	Training	0.80	0.80	8.65	4.5
		Testing	0.65	0.65	11.5	6.0
MLP_04_DWT_1	$Q_{(n-1)}$	Training	0.80	0.79	8.7	4.6
		Testing	0.59	0.59	12.4	6.3
MLP_04_DWT_2	$Q_{(n-1)}$ & $Pave_{(n-1)}$	Training	0.80	0.80	8.6	4.7
		Testing	0.55	0.54	13.2	6.5
MLP_05_DWT	$Q_{(n-1)}$	Training	0.80	0.80	8.6	4.6
		Testing	0.63	0.63	11.8	6.1
MLP_06_DWT	$Pave_{(n-1)}$ & $Pave_{(n-2)}$	Training	0.73	0.72	10.1	5.73
		Testing	0.54	0.54	13.23	7.75

A final model MLP_05_DWT(n), having $P_{(n)}$ and $T_{(n)}$ together with other inputs, improves the performances up to 0.85, 0.73 for both R^2 and NSME and RMSE is reduced to 7.5 and 10.1 m^3/s and MAE to 4.0 and 5.1 m^3/s for training and testing periods, respectively. Comparison of the two models is presented by a scatter diagram by Fig. 2.

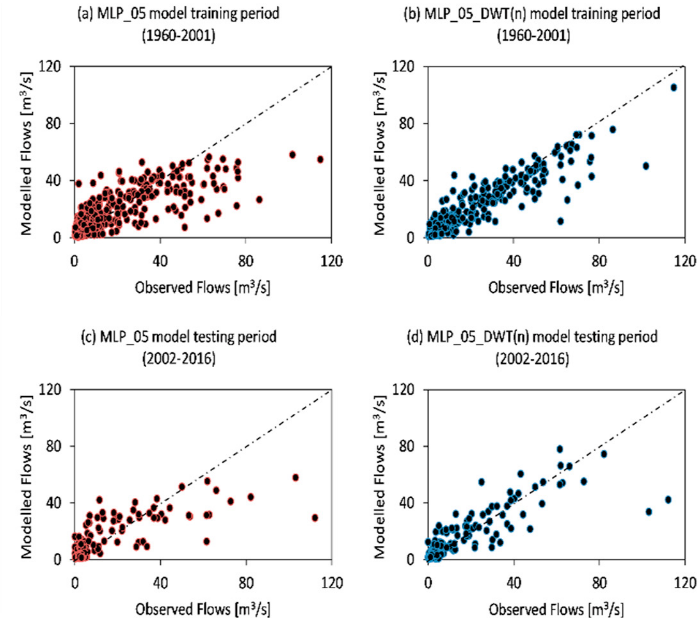

Fig. 2. Scatter diagram of the neural network models (a) MLP_05 model training period, (b) MLP_05_DWT(n) model training period, (c) MLP_05 model testing period, (d) MLP_05_DWT(n) model training period

5. Conclusion and Outlook

In this study, monthly runoff values for Çamlıdere dam basin are estimated by feedforward backpropagation neural network model using multi-layer perceptrons. Previous months' streamflow records has major impact for next month's modelled flow results. Instead of increasing model input number, advanced hybrid techniques like wavelet analysis has much added value to better model the streamflows. Both metrological and discharge data based models provide similar model performances. Moreover, meteorological data should be used to improve discharge based models. Discrete Wavelet Transform of inputs especially on runoff has major impact to improve peak flows, thus correspondingly whole model performances. Since the seasonal pattern is highlighted as the main characteristic of the monthly time series, this can be captured by wavelet analysis in terms of sub-signals. For future studies, Fuzzy theory concept based models might lead to more reliable results instead of new data sets.

Acknowledgements

The authors wish to thank the General Directorate of Meteorology (MGM) and the State Hydraulic Works (DSI) for data contribution.

References

Barrow, D. K., Crone, S. F., 2013. Crogging (cross-validation aggregation) for forecasting—A novel algorithm of neural network ensembles on time series subsamples. In Neural Networks (IJCNN), The 2013 International Joint Conference on (pp. 1-8). IEEE.

Bhadra, A., Bandyopadhyay, A., Singh, R., Raghuwanshi, N. S., 2010. Rainfall-runoff modeling: comparison of two approaches with different data requirements. Water Resour. Manag., 24(1), 37-62.

Chen, J., Adams, B. J., 2006. Integration of artificial neural networks with conceptual models in rainfall-runoff modeling. J. Hydrol., 318(1), 232-249.

Dawson, C.W., Wilby, R.L., 2001. Hydrological modelling using artificial neural networks. Prog. Phys. Geogr. 25 (1), 80–108.

Eberhard, R. C., Dobbins, R. W., 1990. Neural network PC tools: a practical guide. Academic Press.

Fausett, L. V., 1994. Fundamentals of neural networks. Prentice-Hall.

Govindaraju, R. S., 2000a. Artificial neural networks in hydrology. II: hydrologic applications. J. Hydrol. Eng., 5(2), 124-137.

Govindaraju, R. S., 2000b. Artificial neural networks in hydrology. I: Preliminary concepts. J. Hydrol. Eng., 5(2), 115-123.

Kalteh, A.M., 2016. Improving forecasting accuracy of streamflow time series using least squares support vector machine coupled with data-preprocessing techniques. Water Resour. Manage 30 (2), 747–766.

Karimi, H., Zeinivand, H., Tahmasebipour, N., Haghizadeh, A., & Miryaghoubzadeh, M., 2016. Comparison of SRM and WetSpa models efficiency for snowmelt runoff simulation. Environmental Earth Sciences, 75(8), 1-16.

Kentel, E., 2009. Estimation of river flow by artificial neural networks and identification of input vectors susceptible to producing unreliable flow estimates. J. Hydrol., 375(3), 481-488.

Kisi, O., Sanikhani, H., 2015. Prediction of long‐term monthly precipitation using several soft computing methods without climatic data. Int. J. Climatol.., 35(14), 4139-4150.

Levenberg, K., 1944. A method for the solution of certain non-linear problems in least squares. Q. Appl. Mater., 2(2), 164-168.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., & Bergström, S., 1997. Development and test of the distributed HBV-96 hydrological model. J. Hydrol., 201(1-4), 272-288.

Marquardt, D. W., 1963. An algorithm for least-squares estimation of nonlinear parameters. J. Soc. Ind. Appl. Math., 11(2), 431-441.

Nourani, V., Baghanam, A. H., Adamowski, J., Kisi, O., 2014. Applications of hybrid wavelet–Artificial Intelligence models in hydrology: A review. J. Hydrol., 514, 358-377.

Refsgaard, J. C., Højberg, A. L., He, X., Hansen, A. L., Rasmussen, S. H., & Stisen, S., 2016. Where are the limits of model predictive capabilities?. Hydrol. Process., 30(26), 4956-4965.

Refsgaard, J. C., Knudsen, J., 1996. Operational validation and intercomparison of different types of hydrological models. Water Resour. Res., 32(7), 2189-2202.

Shiri, J., Kisi, O., 2010. Short-term and long-term streamflow forecasting using a wavelet and neuro-fuzzy conjunction model. J. Hydrol., 394(3), 486-493.

Valipour, M., Banihabib, M. E., Behbahani, S. M. R., 2013. Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir. J. Hydrol., 476, 433-441.

Tayfur, G., Brocca, L., 2015. Fuzzy logic for rainfall-runoff modelling considering soil moisture. Water Resour. Manag., 29(10), 3519-3533.

Uysal, G., Şensoy, A., Şorman, A. A., 2016. Improving daily streamflow forecasts in mountainous Upper Euphrates basin by multi-layer perceptron model with satellite snow products. J. Hydrol., 543, 630-650.