Civil Engineering Infrastructures Journal, 51(2): 293 – 310, December 2018
Print ISSN: 2322-2093; Online ISSN: 2423-6691
DOI: 10.7508/ceij.2018.02.004

Application of Machine Learning Approaches in Rainfall-Runoff Modeling (Case Study: Zayandeh_Rood Basin in Iran)

Dastorani, M.T.[1*], Mahjoobi, J.[2], Talebi, A[3] and Fakhar, F.[4]

[1] Professor, Faculty of Natural Resources and Environment, Ferdowsi University of Mashhad, Iran.
[2] M.Sc., Water Recourse Management Company, Yazd Regional Water Authority, Iran.
[3] Professor, Faculty of Natural Resources, Yazd University, Iran.
[4] M.Sc., Graduate in Watershed Management, Yazd University, Iran.

Received: 25 Oct. 2017; Revised: 28 Feb. 2018; Accepted: 13 Mar. 2018

ABSTRACT: Run off resulted from rainfall is the main way of receiving water in most parts of the World. Therefore, prediction of runoff volume resulted from rainfall is getting more and more important in control, harvesting and management of surface water. In this research a number of machine learning and data mining methods including support vector machines, regression trees (CART algorithm), model trees (M5 algorithm) and artificial neural networks have been used to simulate rainfall- runoff process in Zayandeh_rood dam basin in Iran. Data used in this research included 9 years of daily precipitation, minimum temperature, maximum temperature, mean temperature, mean relative humidity of daily times 6:30, 12:30 and 18:30 and run off. A number of 3294 lines of data were totally used, and simulations were carried out in two different conditions: without previous run off data as input vectors (M1 condition), and with previous runoff data as input vectors of the models (M2 condition). Results show that machine learning techniques used in this research are not able to present acceptable predictions of runoff in M1 condition (without previous runoff data). However, predictions are considerably improved when previous runoff data are used as input beside other inputs (M2 condition). Between the models used in this research support vector machines (SVM) presented the most accurate results, as the values of RMSE for results presented by SVM, regression tree, model tree and artificial neural network are 2.4, 6.71, 3.2 and 3.04, respectively.

Keywords: ANN, Cart, Decision Tree, Machine Learning, Rainfall-Runoff, SVM.

INTRODUCTION

Runoff resulted from rainfall and snow melting is one of the main water resources to fulfill agricultural, industrial and domestic requirements. Depending on geological characteristics, land use, vegetation cover, ground slope and watershed form, a considerable part of precipitation generally flows as runoff. Therefore, estimation of the runoff resulted from rainfall events is a very important step in water resources planning to provide enough water for consumers. During recent decades hydrologists have paid specific attention to modeling and prediction of runoff behavior produced by different

* Corresponding author E-mail: dastorani@um.ac.ir

precipitation events. The reasons for this attention are: increase in flood occurrence and as a result more damages, increased demand for hydro-power, marine transport development in the rivers, confident design and construction requirements of hydraulic structures, establishment of flood warning systems, drought damages prevention and water resources management strategies for water treatment. A large number of watersheds in Iran and most of developing countries have no flow measuring system, or the measured data are too short, and not enough for planning purposes. Therefore, development and calibration of suitable runoff prediction methods especially for ungauged sites is necessary to fulfill these requirements.

In recent years, some techniques of artificial intelligence and data mining have been developed and used in hydrology and water resources that are mostly able to model natural conditions. In most of the cases the results produced by these techniques have shown higher level of accuracy. Some of the famous methods in this area are as follows:

- Artificial neural networks
- Fuzzy inference systems
- Adaptive neuro-fuzzy inference systems
- Regression trees
- Model trees
- Support vector machines
- Genetic programming

Several research projects have been carried out in the area of rainfall-runoff using different machine learning and data mining techniques. Aqil et al. (2007) evaluated the efficiency of two machine learning techniques including artificial neural network and neuro-fuzzy systems in prediction of runoff resulted from rainfall in daily and hourly time scales. In this study the effects of input data variation on model efficiency was also evaluated. The study area was Cilalawi river catchment which is a tributary of Citarum river in Indonesia. Finally it was concluded that the results presented by neuro-fuzzy system show higher accuracy in comparison to the ANN technique. Bhadra et al. (2010) simulated rainfall-runoff process using semi-distributed conceptual SCS CN method (in combination with Muskingum routing technique) and compared the results to those presented by a ANN based model. It was concluded that ANN technique, in spite of requiring much less data, predicted daily runoff values more accurately than the SCS CN method. Wu and Chau (2011) used ANN coupled with Singular Spectrum Analysis (SSA) for rainfall-runoff modeling, and concluded that coupling ANN with SSA (as a data preprocessing tool) considerably improves the results of ANN based rainfall-runoff model. Nourani (2016) used a new generation of Artificial Intelligence-based models called Emotional Artificial Neural Network (EANN), for modeling daily rainfall-runoff process. Then he compared the results to those of a conventional feed forward neural network (FFNN), and concluded that the EANN could present results with higher accuracy comparing to FFNN in both training and verification phases. He finally stated that the superiority of EANN over classic ANN refers to its ability to recognize and distinguish dry (rainless days) and wet (rainy days) situations using artificial emotional system hormonal parameters. Machado et al. (2011) evaluated the capacity of artificial neural networks for rainfall-runoff modeling in Jangada river basin of Brazil. They compared the results to those produced by a conceptual model called IPHMEN, and it was reported that the ANN presented the best results. El-shafie et al. (2011) used ANN for prediction of rainfall-runoff relationship in Tanakami region of Japan, and the results were compared to those presented by a classical regression model. Feedforward back propagation ANN was able to describe the behavior of rainfall-runoff process more accurately than the

classical regression model. Yilmaz and Muttil (2014), predicted river flows using various machine learning methods including Feed Forward Neural network (FFNN), Adaptive Neuro Fuzzy Inference System (ANFIS), and Genetic Programming (GP) and also a non-machine learning method (multiple linear regression) in the Euphrates Basin in Turkey. Reconstruction of the missing data in the runoff record of the selected stations in Euphrates Basin was also an objective of this study. The machine learning methods were applied to three main Euphrates sub-basins, namely the Upper, Middle, and Lower Euphrates Basins. ANFIS and FFNN methods were the most successful methods for runoff estimation in the Upper and Lower Euphrates Basins, whereas GP and ANFIS models were the best ones in the Middle Euphrates Basin. The model was able to reconstruct missing flow data successfully in the selected stations. Kamali et al. (2014) used Multi-Objective Fuzzy Optimal models for Automatic Calibration of HEC-HMS hydrological Model. The algorithm employed for this purpose was the Particle Swarm Optimization (PSO) algorithm. Comparison of the results taken from the single and multi-objective scenarios indicated the efficiency of the proposed HMS-PSO simulation-optimization method in the multi-objective calibration of event-based hydrologic models. Karimaee Tabarestani and Zarrati (2015) used artificial neural networks and the empirical models to design stone riprap around the bridge piers. The aim was to develop an approach for sizing stable riprap around the bridge piers based on a large amount of experimental data. To estimate the stable riprap stone size around bridge piers in this research, an empirical equation was developed by multiple regression analysis. Finally, in order to receive a higher accuracy for riprap design, the Artificial Neural Network (ANN) method based on utilizing non-dimensional parameters was used. The results indicated that the ANN model provides about 7% improved prediction for riprap size comparing to the conventional regression formula. Barzegari et al. (2015) compared ANN, Decision Trees (DT) and Sediment Rating Curve (SRC) models for estimating suspended sediment in in ten hydrometric stations of Lorestan province, Iran. The results showed that the accuracy of ANN with Levenberg-Marquardt back propagation algorithm was higher than the two other models, especially in high discharges. Shortridge et al. (2016) used multiple regression and machine learning techniques (generalized additive models, multivariate adaptive regression splines, artificial neural networks, random forests, and M5 cubist models) for simulation of monthly stream flow in five highly seasonal rivers in the highland areas of Ethiopia. Then their performance was compared in terms of predictive accuracy, error structure and bias, model interpretability, and uncertainty when faced with extreme climate conditions. While the relative predictive performance of the models were different across basins, but data-driven approaches were able to achieve reduced errors when compared to physical models developed for the region. Granata et al. (2016) carried out a comparison between a SVM (support vector machine)-based approach and the EPA's Storm Water Management Model (SWMM) abilities for rainfall-runoff modeling. The SVM variant used in this study was Support Vector Regression (SVR). Two different experimental basins located in the north of Italy were considered as case studies. The Root-Mean Square Error (RMSE) and the coefficient of determination were selected as criteria to assess the consistency between the recorded and predicted flow rates. Based on the results, both models showed comparable performance. In particular, both models can properly model the hydrograph shape, the time to peak and the total runoff. The main

difference is where the SVR algorithm tends to underestimate the peak discharge, while SWMM tends to overestimate it. It can be said that although, SVR shows great potential for applications in the field of urban hydrology, but it also has significant limitations regarding the model calibration. Adamowski and Prasher (2012) studied and also compared two machine learning methods of SVR and wavelet networks (WN) for forecasting daily runoff in the mountainous watershed of Sianji in the Himalayan region of India. The models were based on parameters of runoff, antecedent precipitation index, rainfall, and day of the year that the data have been collected (over the time period from July 1, 2001 and June 30, 2004). It was found that both used methods produced accurate results, with the best WN model slightly outperforming the best SVR model in accuracy. However, is suggested that both the WN and SVR models need to be tested in other mountainous watersheds specially with limited data for further assessment of their suitability in forecasting.

In the present research, rainfall-runoff modeling has been carried out using artificial neural networks, regression trees, model trees and the support vector machine in Zayandeh_rood dam basin of Iran. The purpose of this research is the comparison of the efficiency of these techniques for rainfall runoff modeling. Although many research projects have been carried out using ANN and ANFIS for rainfall-runoff modeling, however quite few investigations are found about other mentioned methods including regression trees, model trees and support vector machines in rainfall-runoff modeling.

MATERIALS AND METHODS

Description of the Selected Models

Several machine learning techniques have been applied in this work including regression trees, model trees, SVM method and ANNs. A short summary of these techniques is presented here.

Regression Trees (Cart Algorithm)

The Classification and Regression Trees (CART) method of Breiman et al. (1984) generates binary decision trees. CART is a non-parametric statistical methodology that has been developed for classification issues analysis either from categorical or continuous dependent variables. The CART tree is constructed by splitting subsets of the data set using all predictor variables to create two child nodes repeatedly, beginning with the entire data set. The best predictor is selected based on using a variety of impurity or diversity measures. The aim is to produce data subsets which have the highest possible homogeneity with respect to the target variable. In CART algorithm for each split, each predictor is evaluated to find the best cut point based on improvement score or reduction in impurity (Breiman et al., 1984). Then after comparison between the predictors, the predictor with the best improvement is chosen for the split. The process repeats recursively until one of the stopping rules is triggered. For controlling to the size of the tree being built, stopping rules are used. The maximum depth of the tree and the minimum number of subjects per parent or child node can be defined.

Regression tree building centers on three main components: 1) a set of questions of the form: is $X \le d$? where X is a variable and d is a constant, 2) Goodness of split criteria for choosing the best split on a variable and 3) the generation of summary statistics for terminal nodes. The least squared deviation (LSD) impurity measure is used for splitting rules and goodness of fit criteria. It is defined as:

$$LSD(t) = \sum_{i=1}^{N_t} \left(y_i(t) - \bar{y}(t) \right)^2 \qquad (1)$$

where y_i: is the value of the target field, and

$\bar{y}(t)$: is the mean of the dependent variable (target field) at node *t*. The *LSD* criterion function for split data set *S* at node *t* is defined as:

$$Q(S,t) = LSD(t) - LSD(t_R) - LSD(t_L) \qquad (2)$$

where $LSD(t_R)$: is the sum of squares of the right child node and $LSD(t_L)$: is the sum of squares of the left child node. The split data set *S* is selected to maximize the value of *Q(S, t)*.

In regression trees, each terminal node's predicted category is the mean of the target values for records in the node ($\bar{y}(t)$).

Model Trees (M5' Algorithm)

Model trees (Quinlan, 1992) are in fact an extension of the regression trees in the sense that they associate leaves with multivariate linear models. Model trees are in fact techniques to deal with continuous class problems that provide a structural representation of the data and a piecewise linear fit of the class. They have a conventional decision tree structure but use linear function at the leaves instead of discrete class labels (Figure 1). M5 Model trees were firstly introduced by Quinlan (1992) and then the idea was reconstructed and improved in a system called M5′ by Wang and Witten (1997). An M5′ Model tree is an effective learning method for predicting real values. M5′ model tree algorithm first constructs a regression tree by recursively splitting the instance space. The splitting criterion is used to minimize the intra-subset variability in the values down from the root through the branch to the node. The variability is measured by the standard deviation of the values that reach that node from the root through the branch with calculating the expected reduction in error as a result of testing each attribute at that node. The attribute that maximizes the expected error reduction is selected. The splitting stops if the values of all instances that reach a node vary slightly or only a few instances remain. The value of standard deviation reduction (*SDR*) is calculated using Eq. (3).

$$SDR = sd(T) - \sum_i \frac{|T_i|}{|T|} \times sd(T_i) \qquad (3)$$

where *T*: is the set of examples that reach the node, T_i: is the sets resulted from splitting the node based on the selected attribute and *sd*: is the standard deviation (Wang and Witten, 1997). After the tree has been grown, M5′ computes a linear multiple regression model for every interior node.

The data associated with that node and only the attributes tested in the sub tree rooted at that node are used in the regression. The attributes will be dropped one by one if they lower the estimated error. Then the tree is pruned from the leaves if those results in a lower expected estimated error. For more information about model trees see Quinlan (1992) and Wang and Witten (1997).

Support Vector Machines

The Support Vector Machines (SVMs) are methods of supervised learning. SVM has been developed by Vapnik (1995) and is gaining popularity because of its attractive features, and promising empirical performance. SVMs have been developed to solve the classification problems, but during the recent years have been also extended to the domain of regression problems. SVM is a tool for empirical risk minimization, a special property of SVMs is that they minimize the empirical classification or regression error and maximize the geometric margin, simultaneously; this is why they are also considered as maximum margin classifiers. A SVM actually constructs a separating hyperplane between the classes in the n-dimensional space of the inputs. This hyperplane maximizes the margin between the two data sets of the two input classes.

Margin refers to the distance between the two parallel hyperplanes, on each side of the separating one, pushed against each of the two datasets. Simply, larger margin indicates better generalization of error of the classifier. In regression case, the only difference is that SVM attempts to fit a curve, with respect to the kernel used in the SVM, on the data points such that the points lie between the two marginal hyperplanes as much as possible, the goal is to minimize the regression error. The formulations and the technical note are explained in more details in Mahjoobi and Adeli Mosabbeb (2009).

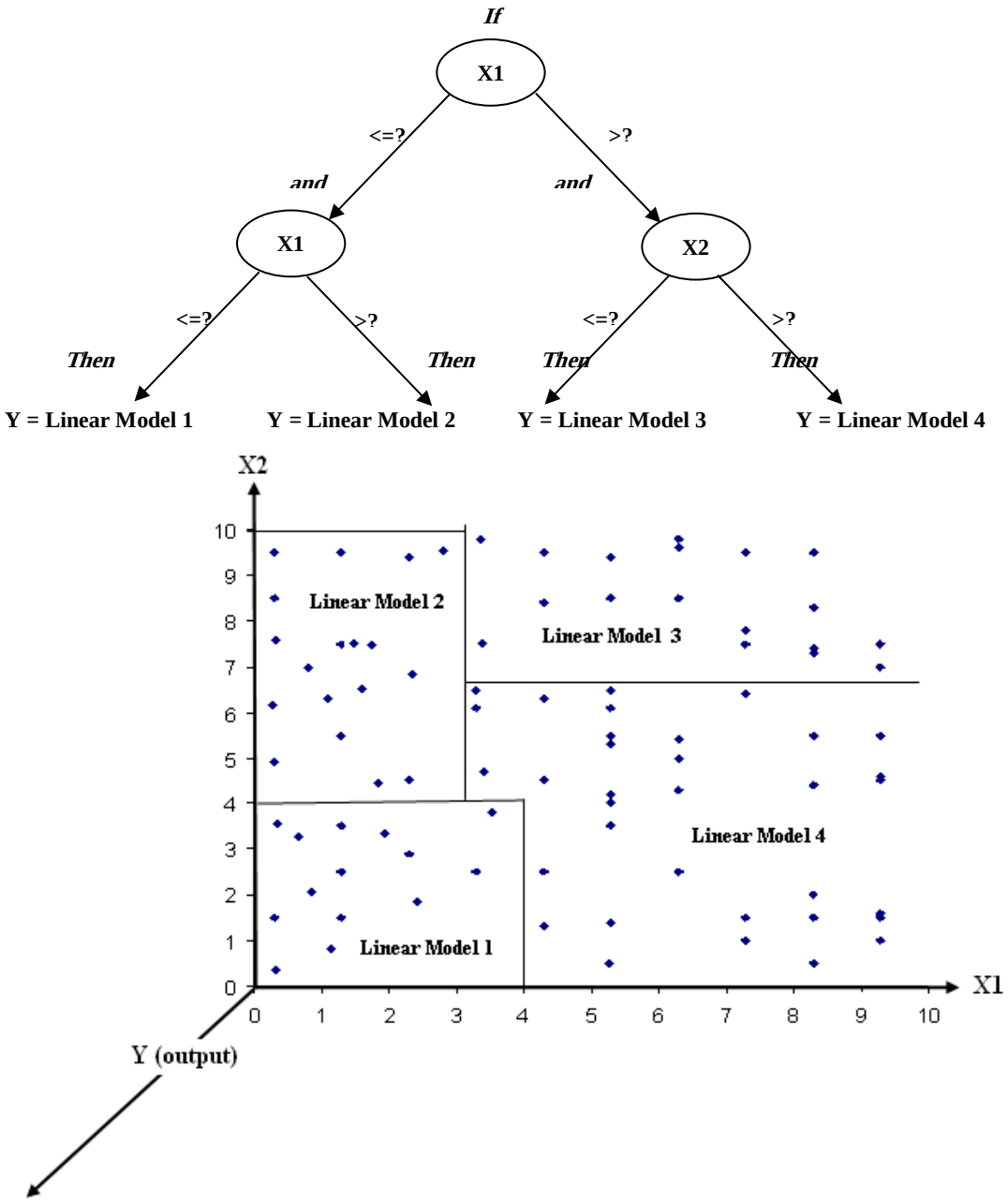


Fig. 1. Splitting the input space X1×X2 by M5 model tree algorithm (Etemad Shahidi and Mahjoobi, 2009)

The SVM parameters used in this study are C = 50 and ε = 0.001. For the optimization process in the regression problem, improved Sequential Minimal

Optimization (SMO) algorithm (Platt, 1999) is used. For the kernel function, Radial Basis Function (RBF) are used. The parameters used for the kernels are $\gamma = 0.01$ for the RBF kernel.

The role of parameter C is to control the tradeoff between SVM errors on training data and margin maximization ($C = \infty$ leads to hard margin SVM). On the other hands, a small value for C will increase the number of training errors, while a large C will lead to a behavior similar to that of a hard-margin SVM. The parameter C must be selected by the user and in this study, we choice empirically on validation data set.

The choices of C and ε control the prediction (regression) model complexity. The problem of optimal parameter selection is quite complicated due to the fact that the complexity of SVM model (and hence its generalization performance) depends on all three parameters (Smola and Schölkopf, 1998). An algorithm for solving the problem of regression with support vector machines was proposed by (Platt, 1999) called Sequential Minimal Optimization (SMO). It puts chunking to the extreme by iteratively selecting subsets only of size 2 and optimizing the target function with respect to them. This algorithm has much simpler background and is easier to implement. The optimization sub-problem could be analytically solved, without the need to use a quadratic optimizer.

A SVM constructs a separating hyperplane between the classes in the n-dimensional space of the inputs. This hyperplane maximizes the margin between the two data sets of the two input classes. This is one of the most advantageous features of SVMs comparing to ANNs. The margin is defined as the distance between the two parallel hyperplanes, on each side of the separating one, pushed against each of the two datasets. Simply, the larger the margin, the better the generalization error of the classifier would be. For the case of regression, the only difference is that SVM attempts to fit a curve, with respect to the kernel used in the SVM, on the data points such that the points lie between the two marginal hyperplanes as much as possible, the aim is to minimize the regression error.

Artificial Neural Networks

An ANN can be defined as an interconnected group of artificial neurons that uses a mathematical model for processing of the information based on a connectionist method for computation. In most cases, ANNs are adaptive systems that change their structures based on external or internal information that flow through the networks. In fact, neural networks are nonlinear statistical data-modeling tools. They are used for modeling complex relationships between inputs and outputs or for finding patterns in data. In many applications, modeling tools have provided better results when used in hydrological time series analysis. Artificial neural networks need to be trained with a group of typical input/output pairs of data which is called training data set. The final weight vectors of a proper trained neural network represents its knowledge about the problem (Dastorani et al., 2010). As different types of neural network deal with the problems in different ways, their ability varies depending on the nature of the problem in hand. Accordingly in this study, A Multi-Layer Perceptron (MLP) with Back Propagation (BP) learning rule was used to train the network. To prevent overfitting during the training of the ANN, the number of nodes of the hidden layer was chosen using expression given by Huang and Foo (2002):

$$M \leq 2Z+1 \tag{4}$$

where M and Z: are the number of nodes in hidden and input layers, respectively.

Study Area and Data

The study area of this research is a part of Zayandeh_rood dam basin. This basin is in fact a part of the larger closed basin located in south west of Isfahan province and its outlet is Zayandeh_rood dam. This basin is divided to three sub basins named Eskandari, Ghale_shahrokh and the central (Makazi) sub basin. The total area is 4265.44 km^2 which is located between 32°, 10' to 32°, 18' N and 50°, 03' to 50°,40' E. The main river of this basin flows from north to west. The climate condition of the study area based on Demartonn method is humid in upper parts and Mediterranean condition near the outlet, but in most parts of the area climate condition is semi humid. Precipitation form especially in upper parts is snow which falls during the winter (it is the main source of river flow especially during the dry seasons) but in lower parts of the basin precipitation form mostly changes from snow to rain. Minimum and maximum altitudes of the basin are1920 and 3900 m above sea level respectively in the outlet and upper mountainous parts. Eskandari sub basin with 1836.95 km^2 area is the largest sub basin. This sub basin is located in the north part and forms the main river with 52.25 km length providing water for Zayandeh_rood dam reservoir. Figure 2 shows the study area of this research. Data measured in Eskandari gauging station which is located in 32°, 49', 20″ N and 50°, 25', 52″ E and 2915 meter above sea level, were used in this research. Daily data of 9 years (21 March 1997 to 21 March 2006) of precipitation, min. temperature, mean temperature, max. Temperature, relative humidity of 6:30, 12:30, 18:30 and runoff have been available to use. The total numbers of data for each parameter are 3294 records. Table 1 shows the minimum, mean and maximum of the measured data for the parameters used in this study.

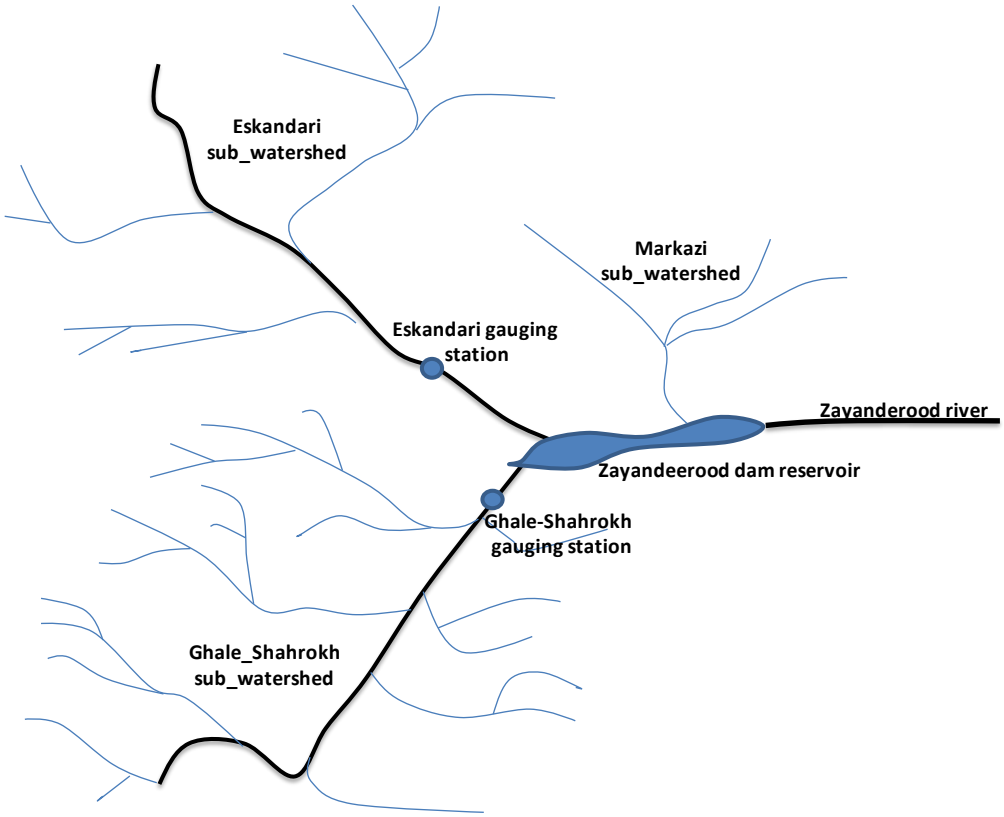


Fig. 2. A schematic view of Zayandeh_rood dam basin

Table 1. The amount of minimum, mean and maximum values of data

Attribute	P (mm)	$RH_{6:30}$	$RH_{12:30}$	$RH_{18:30}$	T_{Min}	T_{Max}	T_{Mean}	Q (m^3/s)
Minimum	0	0	10	21	-43	-14	-12.4	0
Maximum	63.5	89	88	87	6	38	21.1	88.7
Average	1.12	36.6	36.8	66	-8.23	23.18	9.29	2.72

RESULTS AND DISCUSSION

Data was divided in two parts, 70 percent of the total data were used for training of the models (2320 records from 21 March 1997 to 22 July 2003) and the remaining 30 percent of data (947 records from 23 July 2003 to 21 March 2006) were used for evaluation of the efficiency of the models (testing data). For all models, simulations were carried out in two different conditions: M1 condition where P (mm), T min, T mean, T max, RH 6:30, RH 12:30 and RH 18:30 were the inputs of the models and Q as the output. M2 condition where the flow discharge of previous one, two and three days were added to the inputs used in M1 condition.

M1: $Q=f\ (P,\ RH_{6:30},\ RH_{12:30},\ RH_{18:30},\ T_{Min},\ T_{Max},\ T_{Mean})$ (5)

M2: $Q=f\ (P,\ RH_{6:30},\ RH_{12:30},\ RH_{18:30},\ T_{Min},\ T_{Max},\ T_{Mean},\ Q_{t-3},\ Q_{t-2},\ Q_{t-1})$ (6)

For instance, the model tree generated by M5′ algorithm is shown in Figure 3 for M1 condition. As can be seen, 7 rules have been generated. For all four models similar data have been used in training and testing phases.

For statistical comparison of predicted and observed values correlation coefficient (R), root mean square error (RMSE), Scatter Index (SI) and Mean Absolute error (MAE) were used. These statistical measures are defined as:

$$R = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum\left(x_i - \bar{x}\right)^2 \sum\left(y_i - \bar{y}\right)^2}} \quad (7)$$

$$RMSE = \sqrt{\frac{1}{n}\sum (x_i - y_i)^2} \quad (8)$$

$$SI = \frac{RMSE}{\bar{x}} \quad (9)$$

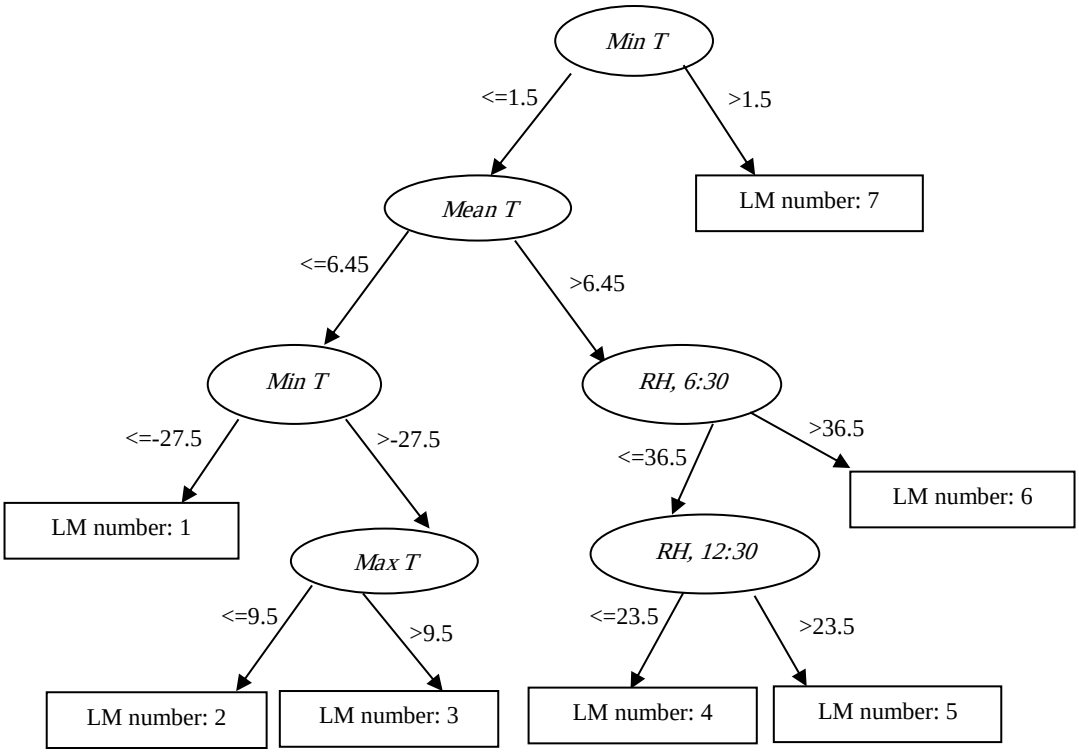
$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - x_i| \quad (10)$$

where x_i: is an observed value, y_i: is a predicted value, n: is the number of observations and $\bar{x}$: is the mean of x and $\bar{y}$: is the mean of y.

Table 2 shows the amount of the statistical measures for the used models in M1 modeling condition. It is seen that quality of the outputs are poor and none of the used models in M1 modeling condition presented acceptable results. Figures 4 to 7 show predictions of the models against the observed values for M1 modeling condition.

Table 2. Amount of statistical measures for the results of the models in M1 modeling condition

Models	RMSE	SI (%)	MAE	R
CART Algorithm	9.37	178.93	4.22	0.12
M5′ Algorithm	9.09	173.6	4.15	0.3
SVM	9.35	178.5	4.12	0.24
ANN	9.07	173.09	4.32	0.20



LM number: 1

Q =-0.0006 * (RH,6:30)- 0.0004 * (RH,12:30)- 0.0002 * Min. T- 0.0029 * Max. T- 0.0016 * Mean T+ 2.6393

LM number: 2

Q =-0.0006 * (RH,6:30)- 0.0004 * (RH,12:30)+ 0.0012 * Min. T- 0.0042 * Max. T- 0.0016 * Mean T+ 1.6773

LM number: 3

Q = -0.0006 * (RH,6:30) - 0.0004 * (RH,12:30) + 0.0012 * Min. T - 0.0033 * Max. T - 0.0016 * Mean T + 2.0504

LM number: 4

Q = 0.0014 * (RH,6:30) + 0.0031 * (RH,12:30) + 0.0019 * Min. T - 0.0028 * Max. T - 0.0016 * Mean T + 1.4258

LM number: 5

Q = 0.0014 * (RH,6:30) + 0.0048 * (RH,12:30) + 0.0019 * Min. T - 0.0028 * Max. T - 0.0016 * Mean T + 3.1786

LM number: 6

Q = 0.0025 * (RH,6:30) + 0.0022 * (RH,12:30)+ 0.0019 * Min. T - 0.0028 * Max. T - 0.0016 * Mean T + 5.1209

Fig. 3. The model tree generated by M5′ algorithm (M1 modeling condition)

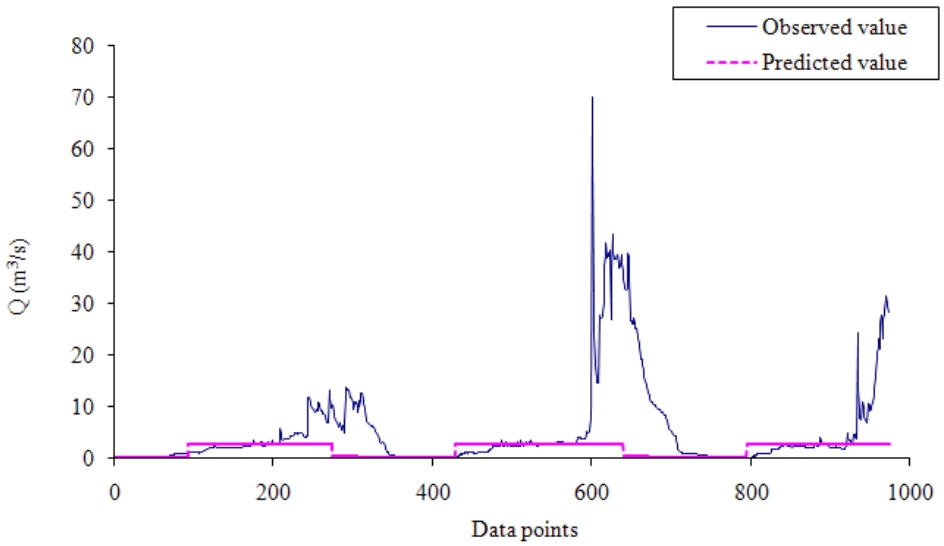


Fig. 4. Comparison between the observed values and predictions of CART algorithm in M1 modeling condition

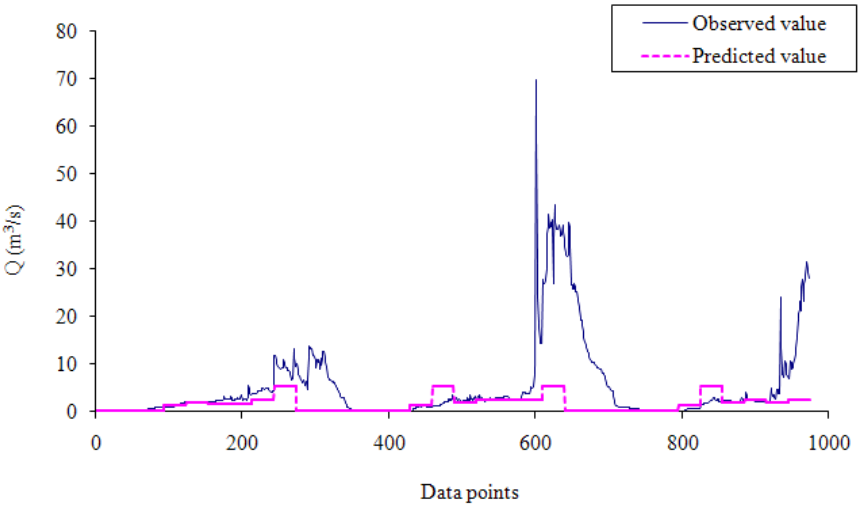


Fig. 5. Comparison between the observed values and predictions of M5' algorithm in M1 modeling condition

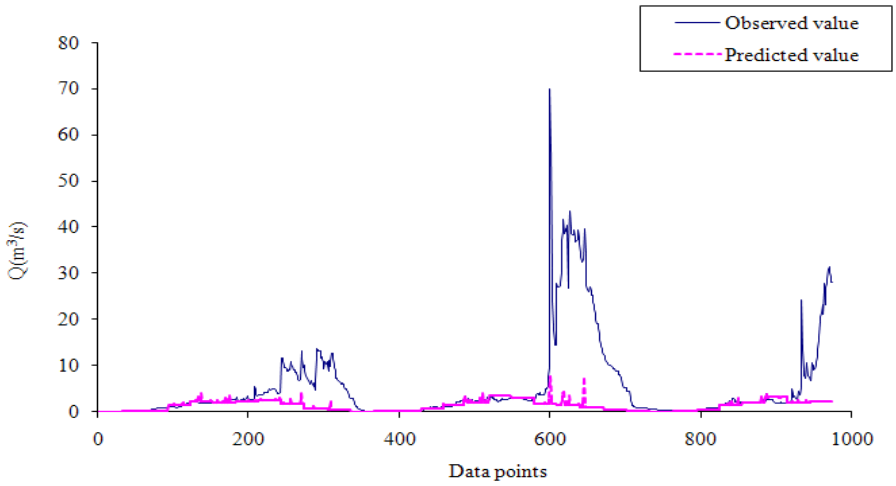


Fig. 6. Comparison between the observed values and predictions of SVM model in M1 modeling condition

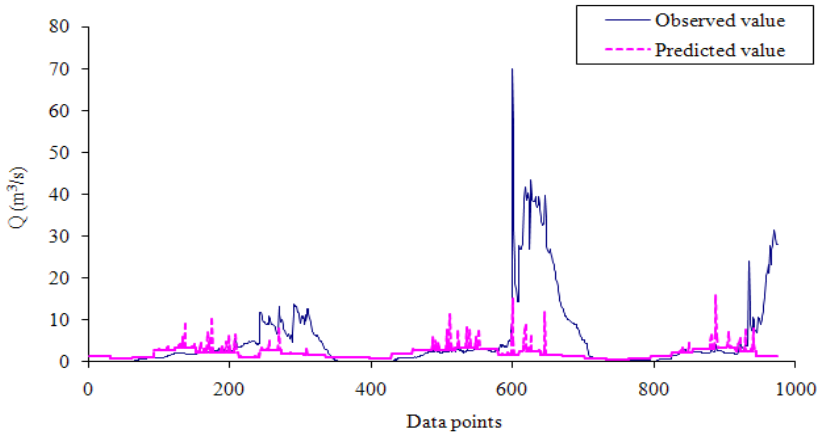
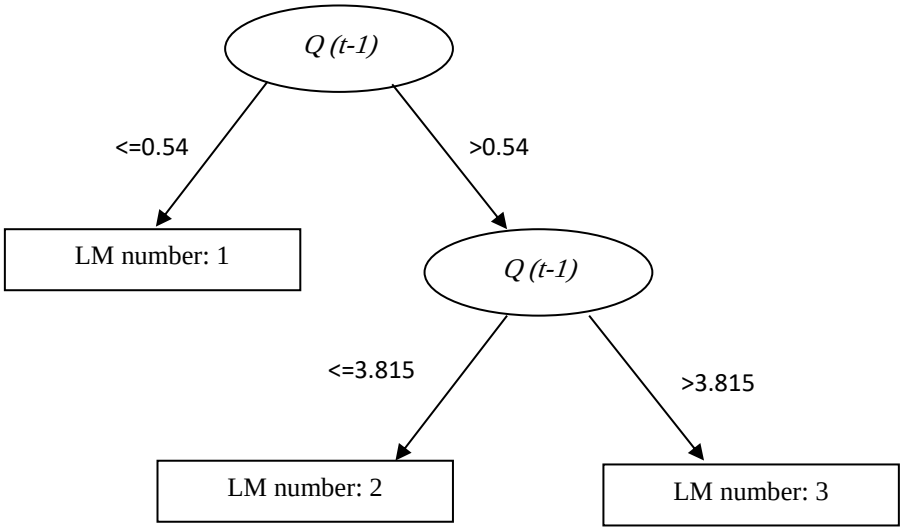


Fig. 7. Comparison between the observed values and predictions of ANN model in M1 modeling condition

Predictions produced in M1 modeling condition indicate the fact that M5' and CART algorithms in addition to poor results, are not able even to recognize and follow the trend of runoff variations. In the other word, for the results of these two algorithms, even ascending and descending of predicted time series are not compatible to the measured time series. In contrast, SVM algorithm although has not been able to present acceptable outputs, but its results show a good relevancy with the measured data. As Figure 6 shows, these ascending and descending trends as well as minimum and maximum values presented by SVM are in good agreement with the related measured values.

As mentioned earlier, for increasing the accuracy of the result, flow data of the previous days was added to the input parameters in M2 modeling condition. The model tree generated by M5′ algorithm for M2 condition is shown in Figure 8. As can be seen, three rules have been generated in this stage.

The amounts of statistical measures for the results of different methods in this condition (M2) are shown in Table 3. Figures 9-12 show the predictions of the models against the observed values in M2 modeling condition.



LM number: 1

$$Q = 0.002 * P - 0.0001 * RH,6:30 + 0.0001 * RH,18:30 + 0.0002 * Min.\ T - 0.0002 * Max.\ T - 0.0003 * Mean\ T + 0.0023 * Q_{(t-3)} + 0.0022 * Q_{(t-2)} + 0.0061 * Q_{(t-1)} + 0.1145$$

LM number: 2

$$Q = 0.0349 * P - 0.0022 * RH,6:30 + 0.0001 * RH,18:30 + 0.0034 * Min.\ T - 0.003 * Max.\ T - 0.0063 * Mean\ T + 0.0047 * Q_{(t-3)} - 0.0831 * Qt\text{-}2 + 1.0501 * Qt\text{-}1 + 0.2396$$

LM number: 3

$$Q = 0.3686 * P - 0.0006 * RH,6:30 + 0.0001 * RH,18:30 + 0.0011 * Min.\ T - 0.0002 * Max.\ T + 0.1573 * Mean\ T + 0.1159 * Q_{(t-3)} + 0.1285 * Q_{(t-2)}2 + 0.387 * Q_{(t-1)} + 0.4823$$

Fig. 8. Model tree generated by M5' algorithm in M2 modeling condition (LM is linear model)

Table 3. Values of the statistical measures for the results of the models in M2 modeling condition

Models	RMSE	SI (%)	MAE	R
CART algorithm	6.71	128.08	2.22	0.76
M5´ algorithm	3.2	61.17	1.17	0.97
SVM	2.4	45.8	0.60	0.97
ANN	3.04	58.08	1.33	0.96

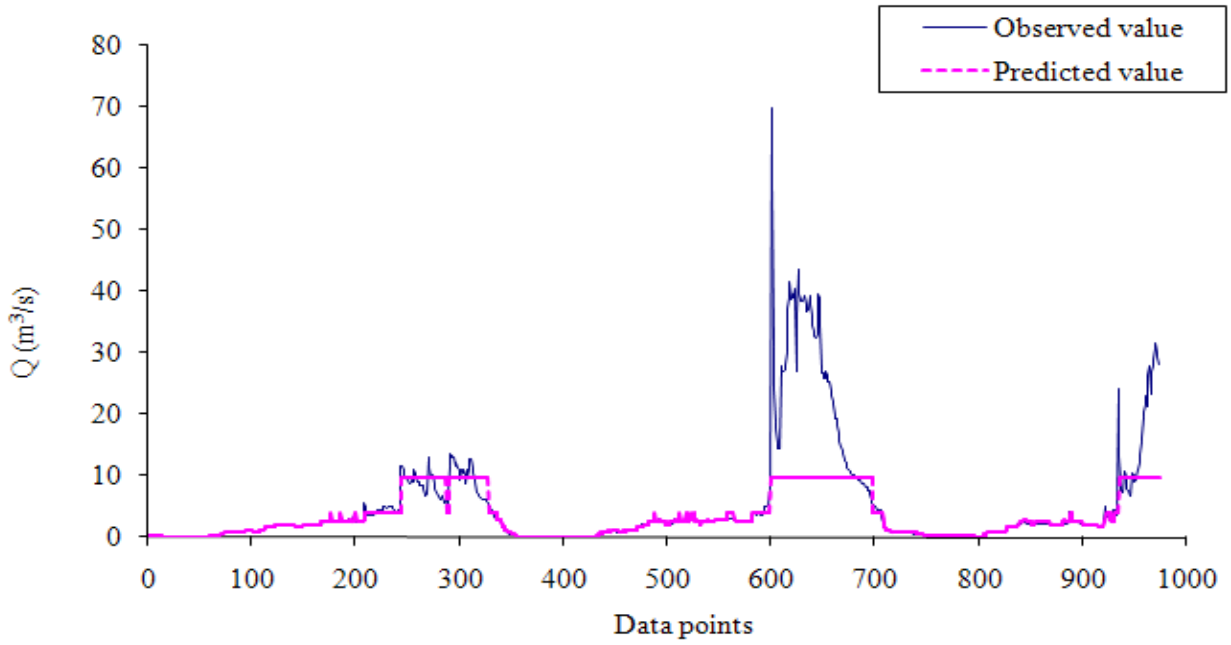


Fig. 9. Comparison between the observed values and predictions of CART algorithm in M2 modeling condition

The values of calculated measures in Table 3 and also comparison of the predicted and observed data in two modeling conditions indicate that addition of previous days flow data to the inputs have considerably increased the accuracy of predictions in all employed models. However, the results presented by SVM method have higher accuracy in comparison to other three methods. Figure 13 shows the distribution of predicted values against the measured values around the exact fit line for SVM output in M2 modeling condition.

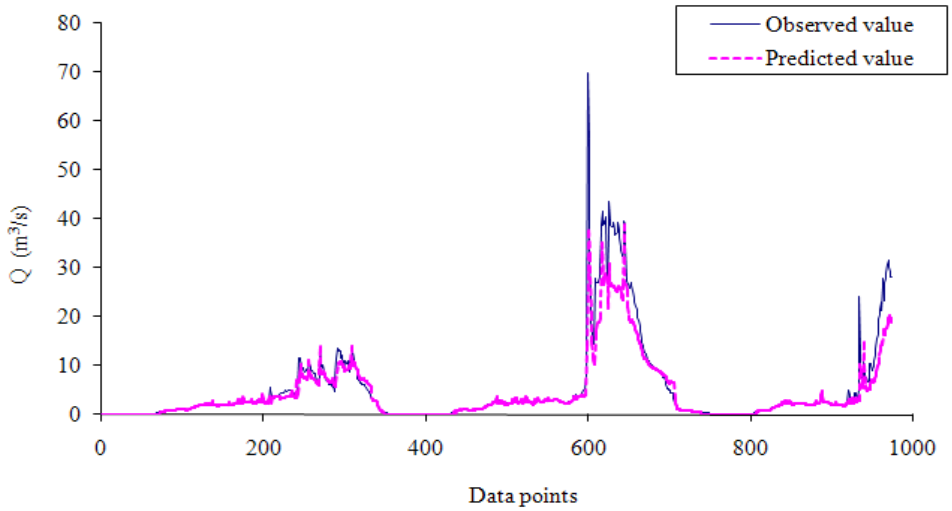


Fig. 10. Comparison between the observed values and predictions of M5 algorithm in M2 modeling condition

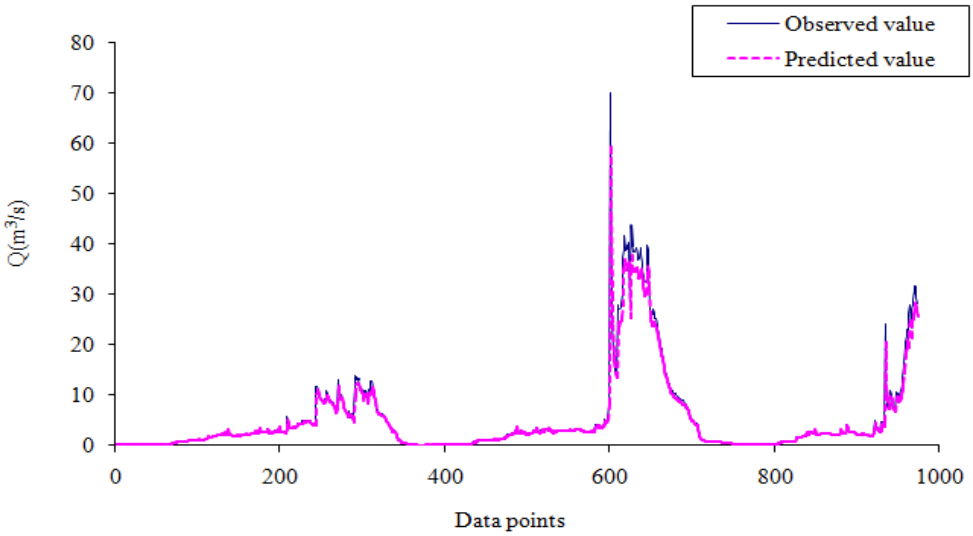


Fig. 11. Comparison between the observed values and predictions of SVM model in M2 modeling condition

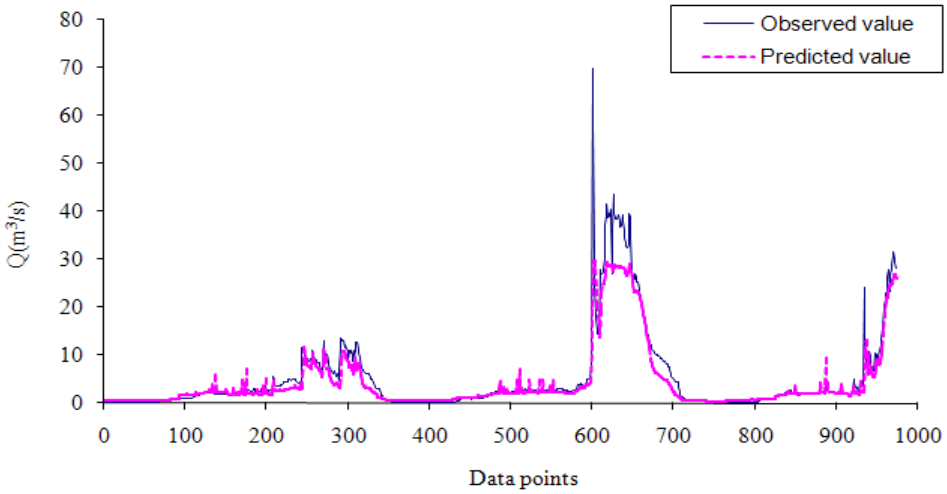


Fig. 12. Comparison between the observed values and predictions of ANN model in M2 modeling condition

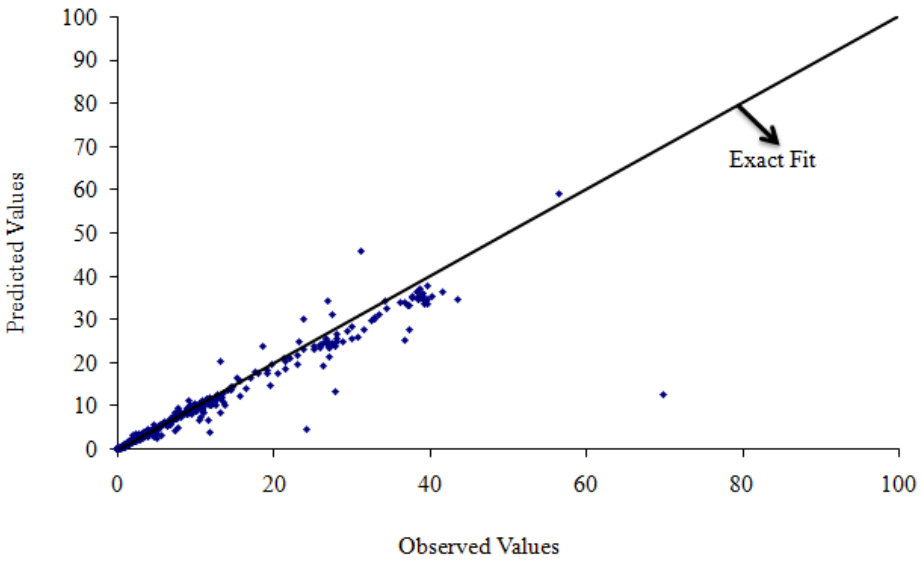


Fig. 13. Scatter plot of the observed data and the predictions produced by SVM method for testing data (M2 modeling condition)

As it is seen from Figure 13, the results of this method have good agreement with the measured values indicating superior ability of SVM in rainfall runoff modeling. As the obtained results of this research show, artificial intelligence models can not present appropriate result in rainfall runoff modeling process until the flow data of previous days are added to the inputs. The relationship between rainfall and runoff in the study area is very complicated, as in most of the cases the amount of runoff resulted from specific rainfall are quite different in different occurrences. In the other word, a specific amount of precipitation creates a considerable volume of runoff in the first time, but for the second time the same precipitation resulted in a relatively small volume of runoff. This complicated relationship between rainfall and runoff causes problem for the models to recognize and learn the process.

Outputs of all models used in this research show that the inputs used in M1 modeling condition cannot provide good training process. The reason for this inability relates to the complicated and unpredictable relationship between rainfall and the resulted runoff. Probably the main reason for this weak correlation between rainfall and runoff in the study area is the diversion of runoff to the farmlands along the river in upper parts. It means that the measured runoff in the gauging station is quite different from what resulted from the rainfall, because a part of it is diverted to the farmlands before reaching the gauging station.

As the amount of runoff diverted from river is quite different from month to month and season to season, it causes large variations in the measured data. Inclusion of the previous runoff data have increased the quality of the outputs in all models, and caused quite acceptable predictions. Previous runoff data show the highest correlation with the predicted runoff, so it plays a positive role in simulation process.

As it is seen from the results, although previous runoff data has considerably increased the quality of outputs but the improvement of accuracy mostly relates to the low and mean amounts and in prediction of high values (peak data) almost all models have problems. As this problem is seen on the results of all models, it cannot be related to the model structure, and probably goes back

to the nature of data. To investigate this problem, total data (the data between minimum and maximum values) were divided to 10 parts based on the values.

The results show that about 90 percent of data belongs to first two decimals and the last 4 decimals (containing the maximum peak values) contain less than 1 percent of the total records. In the other word, these four decimals contain only 9 records. Even if all of these 9 records are used for training of the models, it cannot be expected to have an optimized training for the models to be able to deal with the problem. Therefore, probably this is the main reason to have higher error between predicted and measured values for peak data. Comparison of the results produced by different methods in this research indicates that SVM has higher efficiency compared to other employed techniques.

In most of the research projects such as Aqil et al. (2007), Nourani (2017), Machado et al. (2011) and El-shafie et al. (2011), although the results presented by artificial intelligence models have been satisfactory but some of the models presented results with higher accuracy than others. Findings of the present research are in general similar to the mentioned researches. All four models used in this research are able to present acceptable results (when the flow data of previous days are also used as inputs), but SVM and CART algorithm respectively presented the results with highest and lowest accuracy. The difference between the qualities of the results becomes highlighted in peak flow data.

Accurate prediction of peak flow is the most important factor in many water-related projects such as flood management measures. Superiority of the new artificial intelligence models over the related traditional methods has been clarified and almost proved in many publications. However, the challenge is the selection of the most relevant type of artificial intelligence model for the problem in hand, which needs more investigations.

The advantages of using SVM in comparing to ANN are the fact that: SVM can create a more reliable model with better generalization error, independent from the variations of the training data, and also SVMs do not over fit, while ANNs may face such a problem and need to deal with it. SVMs need way fewer parameters, comparing to ANNs. Also, SVMs required less computational time comparing to ANN. Model trees, in contrast to ANNs, lead to the division of the input space into a number of subspaces for each of which a separate specialized model is built. They build a piecewise linear model, whereas ANNs build nonlinear models. Furthermore, the model trees represent understandable rules, which is of a great interest.

Also in neural networks we need to find the best topology, both the number of the hidden layers and the number of neurons in each hidden layer. The process of finding these parameters could be performed via trial and error, which is a time-consuming sequence of actions. On the contrary, model trees are non-parametric and therefore are more convenient to use. Besides, they need lower run-time and are automatic. At last but not the least important, the advantage of model trees is that, it represents understandable and simple rules which is in contrast to ANNs.

CONCLUSIONS

The efficiency of different artificial intelligence methods for rainfall runoff modeling in Zayandeh_rood dam basin was evaluated in this research. Employed methods are artificial neural networks, regression trees, model trees and the support vector machines. 3294 records of daily precipitation, min. temperature, max. Temperature, relative humidity of 6:30, 12:30, 18:30 and runoff collected in Eskandari gauging station were used for modeling. 70 percent of data were

used for training and the remaining 30 percent for testing purpose to evaluate the applicability of the models.

Simulations were carried out in two different conditions: one without previous runoff data as input (M1 condition), and the other one with previous runoff data as input of the models beside other inputs (M2 condition). In both conditions prediction of runoff discharge was the output of the models. In M1 condition none of the methods was able to present acceptable prediction. However, in M2 condition (using runoff data of previous days) all of the models presented results with quite good accuracy. Between the four models used in this research, support vector machine presented the results with the highest accuracy.

REFERENCES

Adamowski, J. and Prasher, S.O. (2012). "Comparison of machine learning methods for runoff forecasting in mountainous watersheds with limited data", *Journal of Water and Land Development*, 17(1), 89-97.

Aqil, M., Kita, I., Yano, A. and Nishiyama, S. (2007). "A comparative study of artificial neural networks and neuro-fuzzy in continuous modeling of the daily and hourly behaviour of runoff", *Journal of Hydrology*, 337(1-2), 22-34.

Barzegari, F., Yosefi, M. and Talebi, A. (2015). "Estimating suspended sediment by Artificial Neural Network (ANN), Decision Trees (DT) and Sediment Rating Curve (SRC) models (Case study: Lorestan Province, Iran)", *Civil Engineering Infrastructures Journal*, 48(2), 373-380.

Bhadra, A., Bandyopadhyay, A., Singh, R. and Raghuwanshi, N.S. (2010). "Rainfall-runoff modeling: Comparison of two approaches with different data requirements, water resources management", *Water Resources Management*, 24(1), 37-62.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone C.J. (1984). *Classification and regression trees*, Belmont, Wadsworth Statistical Press.

Dastorani, M.T., Moghadamnia, A.R., Piri, J. and Rico-Ramirez, M. (2010). "Application of ANN and ANFIS models for reconstructing missing flow data", *Environmental Monitoring and Assessment*, 166(1-4), 421-434.

Etemad-Shahidi, A. and Mahjoobi, J. (2009). "Comparison between M5' model tree and neural networks for prediction of significant wave height in Lake Superior", *Ocean Engineering*, 36(15-16), 1175-1181.

El-shafie, A., Mukhlisin, M., Najah, A.A. and Taha, M.R. (2011). "Performance of artificial neural network and regression techniques for rainfall-runoff prediction", *International Journal of the Physical Sciences*, 6(8), 1997-2003.

Granata, F., Gargano, R. and Marinis, G. (2016). "Support vector regression for Rainfall-Runoff modeling in urban drainage: A comparison with the EPA's storm water management model", *Water*, 8(3), 1-13.

Huang, W. and Foo, S. (2002). "Neural network modelling of salinity variation in Apalachicola river", *Water Research*, 36(1), 356-362.

Kamali, B. and Mousavi, S.J. (2014). "Automatic calibration of HEC-HMS model using Multi-Objective fuzzy optimal models", *Civil Engineering Infrastructures Journal*, 47(1), 1-12.

Karimaee Tabarestani, M. and Zarrati, A.R. (2015). "Design of riprap stone around bridge piers using empirical and neural network method", *Civil Engineering Infrastructures Journal*, 48(1), 175-188.

Machado, F., Mine, M., Kaviski, E. and Fill, H. (2011). "Monthly rainfall-runoff modelling using artificial neural networks", *Hydrological Sciences Journal*, 56(3), 349-361.

Mahjoobi, J. and Adeli Mosabbeb, E. (2009). "Prediction of significant wave height using regressive support vector machines", *Ocean Engineering*, 36(5), 339-347.

Nourani, V. (2017). "An emotional ANN (EANN) approach to modeling rainfall-runoff process", *Journal of Hydrology* , 544, 267-277

Platt, J. (1999). "Fast training of support vector machines using sequential minimal optimization", *Advances in Kernel Methods*, *Support Vector Learning*, Schölkopf_Burges, C.J.C. and Smola, A.J., (eds.), Cambridge, MA, MIT Press, 185-208.

Quinlan, J.R. (1992). "Learning with continuous classes", *Proceedings of the Fifth Australian Joint Conference on Artificial Intelligence*, World Scientific, Singapore, 343-348.

Shortridge, J.E., Guikema, S.D. and Zaitchik, B.F. (2016). "Machine learning methods for empirical streamflow simulation: A comparison of model accuracy, interpretability and uncertainty in seasonal watersheds", *Hydrological Earth System Sciences*, 20, 2611-2628.

Smola, A.J. and Schölkopf, B. (1988). "A tutorial on support vector regression", Royal Holloway College, London, U.K., NeuroCOLT Technology Report, TR 1998-030.

Vapnik, V. (1995). *The nature of statistical learning tTheory*, Springer, N.Y.

Wang, Y. and Witten, I.H. (1997). "Induction of model trees for predicting continuous lasses", *Proceedings of the Poster Papers of the European Conference on Machine Learning*, University of Economics, Faculty of Informatics and Statistics, Prague.

Wu, C.L. and Chau, K.W. (2011). "Rainfall-runoff modeling using artificial neural network coupled with singular spectrum analysis", *Journal of Hydrology*, 399(3-4), 394-409.

Yilmaz, A. and Muttil, N. (2014). "Runoff estimation by machine learning methods and application to the Euphrates Basin in Turkey", *Journal of Hydrologic Engineering*, 19(5), 1015-1025.