



Precipitation forecasting using classification and regression trees (CART) model: a comparative study of different approaches

Bahram Choubin¹ · Gholamreza Zehtabian² · Ali Azareh³ · Elham Rafiei-Sardooi⁴ · Farzaneh Sajedi-Hosseini² · Özgür Kişi⁵

Received: 20 January 2018 / Accepted: 16 April 2018 / Published online: 20 April 2018
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract

Interest in semiarid climate forecasting has prominently grown due to risks associated with above average levels of precipitation amount. Longer-lead forecasts in semiarid watersheds are difficult to make due to short-term extremes and data scarcity. The current research is a new application of classification and regression trees (CART) model, which is rule-based algorithm, for prediction of the precipitation over a highly complex semiarid climate system using climate signals. We also aimed to compare the accuracy of the CART model with two most commonly applied models including time series modeling (ARIMA), and adaptive neuro-fuzzy inference system (ANFIS) for prediction of the precipitation. Various combinations of large-scale climate signals were considered as inputs. The results indicated that the CART model had a better results (with Nash–Sutcliffe efficiency, $NSE > 0.75$) compared to the ANFIS and ARIMA in forecasting precipitation. Also, the results demonstrated that the ANFIS method can predict the precipitation values more accurately than the time series model based on various performance criteria. Further, fall forecasts ranked “very good” for the CART method, while the ANFIS and the time series model approximately indicated “satisfactory” and “unsatisfactory” performances for all stations, respectively. The forecasts from the CART approach can be helpful and critical for decision makers when precipitation forecast heralds a prolonged drought or flash flood.

Keywords Large-scale climate predictors · Classification and regression trees · Machine learning · Semiarid region

Introduction

Precipitation prediction is important for water resources management because of its highly precarious conditions in different climate conditions. Precipitation changes may alter underlying water resources conditions and increases the need for new water management programs and strategies

especially in highly uncertain climate conditions within arid and semiarid regions (Choubin et al. 2017b).

Arid and semiarid regions span approximately one-third of the global land surface (Yatheendradas et al. 2008), one-third of Asia (Lemons 2003) and two-third of Iran. In such regions, rainfall occurs very rapidly and causes severe flash flooding. The resulting extreme events (runoff) are highly localized, heterogeneous and dominated by environmental gradients. These flash floods account for more of all flood-related deaths in Iran causing the highest human and economic impact (e.g., Sharifi et al. 2012).

Perhaps the most effective way to mitigate the risks of flash floods and predict a reliable forecast is through single-event-based hydrologic model. Typically, continuous hydrologic models require a highly reliable precipitation data. Although precipitation data may be available in recent years in semiarid watersheds but continuous prediction approaches require reliable and long-lead data at many climate stations. In Iran, historical precipitation records (daily, monthly and annual) and corresponded climate stations in

✉ Bahram Choubin
Bahram368@gmail.com

¹ Department of Watershed Management, Sari Agricultural Sciences and Natural Resources University, P.O. Box 737, Sari, Iran

² Department of Reclamation of Arid and Mountainous Regions, University of Tehran, Karaj 31585-3314, Iran

³ Department of Geography, University of Jiroft, Kerman, Iran

⁴ Faculty of Natural Resources, University of Jiroft, Kerman, Iran

⁵ Faculty of Natural Sciences and Engineering, Ilia State University, Tbilisi, Georgia

semiarid watersheds have neither long-lead records and nor dense gauges distribution (data scarcity); hence, this approach yields a strong motivation to predict precipitation time series for watershed hydrology.

To effectively predict precipitation characteristics, forecasting models must accurately capture precipitation intensity, magnitude and storm intermittency. Indeed, requirements for such a highly complex system include an advanced model to accurately capture the highly nonlinear processes occurring in the climate regime of a semiarid region (Choubin et al. 2017b).

Many studies have been addressed the impact of large-scale signals on hydroclimatology studies in different climate regions (e.g., Gamiz-Fortis et al. 2010; Fallah-Ghalhary et al. 2010; Gaughan and Waylen 2012; Berg et al. 2013; Choubin et al. 2016a among others). Recent literature applied nonlinear models such as artificial neural networks (ANNs), adaptive neuro-fuzzy inference system (ANFIS) for precipitation prediction in different climate conditions. ANNs have been previously used to predict precipitation and proved promising in prediction accuracy and quantifying precipitation values (e.g., Afshin et al. 2011; Azadi and Sepaskhah 2012; Rezaeian-Zadeh et al. 2012; Sigaroodi et al. 2014; Choubin et al. 2016b). For example, El-Shafie et al. (2011) successfully applied the ANFIS model for precipitation prediction in Klang River, Malaysia. Sanikhani and Kisi (2012) employed two different ANFIS techniques (grid partition and sub-clustering) for the estimation of monthly streamflow in the Tigris-Euphrates Basin of Turkey. Choubin et al. (2014, 2017b) studied the relationship between large-scale signals with drought and precipitation in southwestern Iran and the results indicated that the ANFIS model could predict droughts and precipitation values with significant accuracy and precious. Some studies have been performed about the effect of ENSO on precipitation occurrence in Iran; nevertheless, the influences of climate indices on precipitation occurrence still need further investigation in other regions.

Predictions of climatic information in the arid and semiarid region tend to be a challenging task and mostly uncertain due to heterogeneity and nonlinearity of climate system behaviors. Our hypothesis is that precipitation amount is unconditional and has spatiotemporal errors associated with precipitation forecast and is negotiable at the application sites due to a limited areal extent.

The important overall question we attempt to address in this study is as follows: “how reliable is a precipitation forecast using large-scale climate signals for a semiarid precipitation data?” To this end, this research aims to make a better assessment of arid and semiarid precipitation forecast using classification and regression trees (CART) approach by employing various combinations of large-scale climate predictors.

Already, no study has investigated the performance of the CART model for precipitation forecasting. The current research is a new application of the CART model, as a data-mining algorithm, for forecasting the precipitation over a highly complex semiarid climate system using climate signals. This model is robust in case of outliers and auto-correlated input data (Loh 2011), which its splitting algorithm isolates outliers in individual nodes. Also, CART have no assumption on data distribution. Trees in this model are used for description and prediction of patterns (De'ath and Fabricius 2000; Sutton 2005; Choubin et al. 2018). Therefore, the main objectives of current research are to: (1) develop and apply a CART model for predicting the precipitation; and (2) compare CART model results with the two most commonly used approaches including ANFIS and time series models in predicting precipitation. The modeling approach conducted in this study can be helpful and critical for decision makers when precipitation forecast heralds a prolonged drought or flash flood.

Materials and methods

Study area

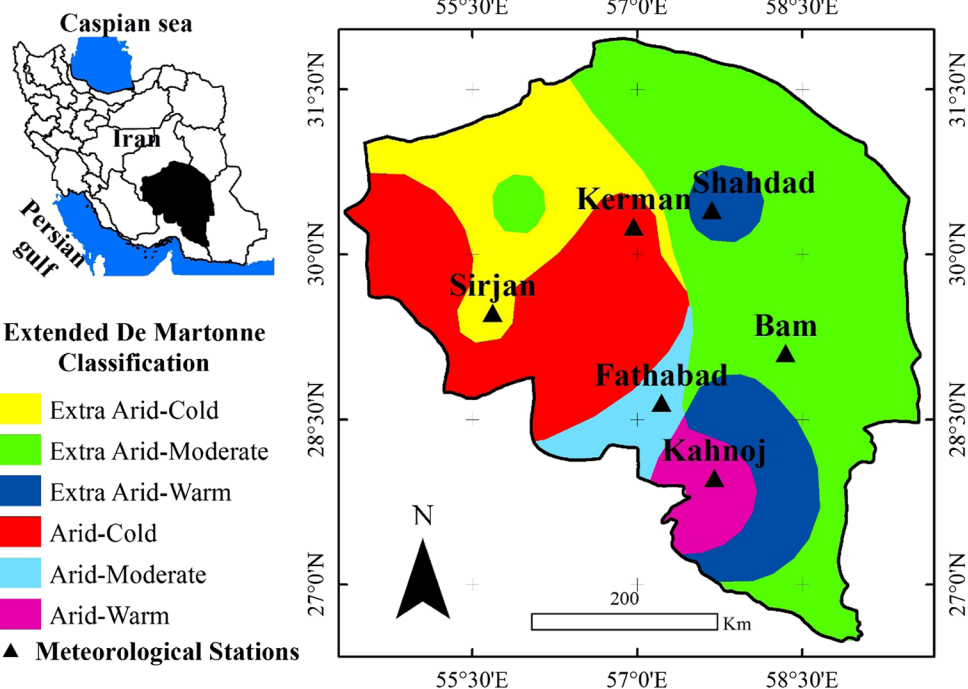
About two-third of Iran is categorized as arid and semiarid regions with a less annual precipitation amount. In most parts of the country, precipitation usually occurs in winter and spring while summer is hot. Kerman Province (54°20' to 59°34' E longitude and 26°29' to 31°57' N latitude) was selected as the study area situated in the south central part of Iran (Fig. 1). The study area is 180,725 km² and the maximum and minimum heights are 4473 and 100 m in the central mountains and north-east of the area, respectively. Based on the statistical center of Iran, the population of Kerman is 2,938,988. So, predicting precipitation is important in this area.

Dataset

Precipitation

To predict precipitation, we collected monthly precipitation data from six stations, namely Bam, Fathabad, Kahnoj, Kerman, Shahdad and Sirjan (Fig. 1). These stations are located in different subclasses of climates in Kerman Province based on extended-De Martonne classification (Khalili 1997). Figure 1 indicates the study area along with extended-De Martonne classification and selected climate stations for each climate subclass of the province. Precipitation data for this research was provided by Iranian Meteorological Service, in September 2015. Then, seasonal precipitation time series and climate indices were calculated for each station during 1985–2014.

Fig. 1 Study area, synoptic stations and extended-De Martonne classification



Climate signals

Climate signals are the global coupled ocean–atmosphere phenomenon to describe the state of a climate system. These signals are created in the various parts of the world and affect the global climate. Climate signals data (sea surface temperature; SST, and sea-level pressure; SLP) were obtained from the National Oceanic and Atmospheric Administration (NOAA) site (<http://www.esrl.noaa.gov/psd>) at the same time as well. Table 1 shows name and coordinates of the points (SST and SLP) used in this study.

To determine the predictor of precipitation (climate signals), the correlation analysis was conducted between precipitation (predictant) and climate signals (predictor). The relationship between precipitation and previous climate signals was determined in seasonal time scale. Thus, effective climate signals were identified for prediction of precipitation.

Adaptive neuro-fuzzy inference system

ANFIS method was developed by Jang (1993), and the most important point in planning a neuro-fuzzy system is the selection of a suitable inference system. Figure 2 exhibits the structure of ANFIS model in which the nodes of the same layers has the same function.

The functioning of ANFIS can be summarized as:

Layer 1: Each node in this layer generates different degrees of membership of an input variable.

Table 1 Name and coordinates of the points (SST and SLP) (Fallah-Ghalhary et al. 2010)

Points	Latitude (°)	Longitude (°)
Aden Gulf (ADE)	12.5	45
Adiabatic Sea (ADI)	43	15
Arabian Sea (ARA)	20	65
Atlantic Ocean (ATL)	30	320
Azores Sea (AZE)	40	330
Black Sea (BS)	45	35
Caspian Sea (CAS)	45	50
East Mediterranean (EM)	35	30
Greenland (GR)	35	30
Indian Ocean (IO)	10	65
Labrador Sea (LS)	60	310
Northern Persian Gulf (NPG)	25	55
Northern Red Sea (NRS)	25	35
Northern Sea (NS)	55	5
Oman Sea (OS)	22.5	65
Soran Network (SN)	20	25
Southern Persian Gulf (SPG)	25	55
Southern Red Sea (SRS)	15	41
West Mediterranean Sea (WM)	40	5

$$OP_i^1 = \mu_{A_i}(x) \leftarrow i = 1, 2 \tag{1}$$

$$OP_i^1 = \mu_{B_{i-2}}(y) \leftarrow i = 3, 4 \tag{2}$$

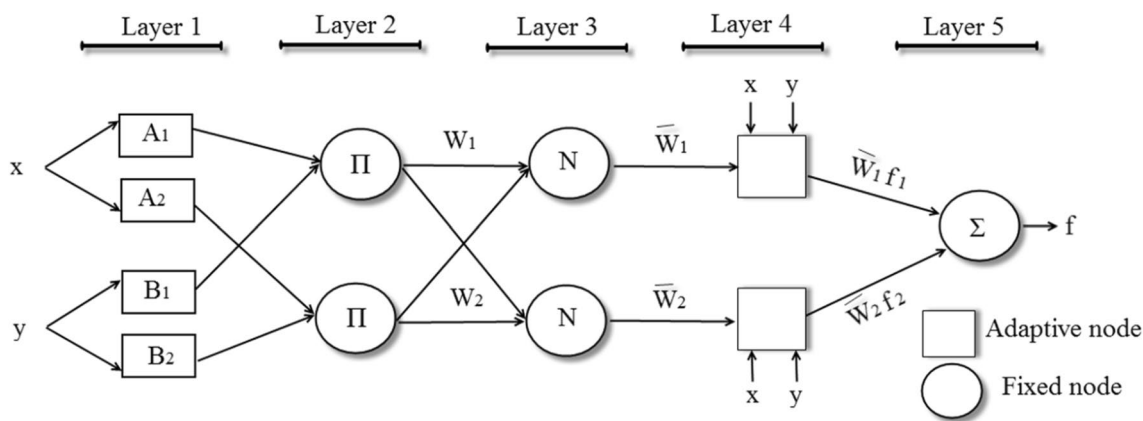


Fig. 2 ANFIS model, where x and y are inputs; A_1, A_2, B_1 and B_2 are fuzzy subsets; Π is the fixed nodes of Layer 2; W_i is the weight of a given fuzzy rule, f_i ; N is the fixed nodes of Layer 3; \bar{w}_i is the

normalized weight; f_i is the fuzzy rule; and f is the final output of the ANFIS model (Kurtulus and Razack 2010)

where x (or y) is the input entered into the selected node and A_i or $(B_i - 2)$ is the fuzzy set associated with this node. For example, Gaussian membership function can be calculated as follows:

$$OP_i^1 = \mu_{A_i}(x) = \frac{1}{1 + \left(\frac{x-c_i}{a_i}\right)^{2b_i}} \tag{3}$$

where $\{a_i, b_i, c_i\}$ are set parameters and the maximum will be 1 and minimum is zero (Kisi et al. 2009).

Layer 2: Each node in this layer multiply the input signals, and the output represents the animated power of the rule.

$$OP_i^2 = w_i = \mu_{A_i}(x)\mu_{B_i}(y) \leftarrow i = 1, 2 \tag{4}$$

Layer 3: The node number i which is named N calculates the normalized starting power:

$$OP_i^3 = \bar{w}_i = \frac{w_i}{w_1 + w_2} \leftarrow i = 1, 2 \tag{5}$$

Layer 4: Node i calculates the contribution of rule i to the output of the model by using the following function.

$$OP_i^4 = \bar{w}_i f_i = \bar{w}_i(p_i x + q_i y + r_i) \tag{6}$$

where w is the output from the previous layer and $\{p_i, q_i, r_i\}$ are set of parameters.

Layer 5: The only node in this layer calculates the final output of ANFIS by the following equation:

$$OP_i^5 = \text{Final(output)} = \sum_i \bar{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \tag{7}$$

Hybrid learning algorithm that is a combination of least-squares and back propagation gradient descent method has been used in this study.

Time series modeling

In addition of ANFIS model, this study further implemented the ARIMA model (that only uses historical data for prediction) by using the autocorrelation (ACF) and partial autocorrelation (PACF) functions. Modeling parameters were estimated using the method of maximum likelihood. For better verification of the selected models, corrected Akaike information criterion, AICC, (Akaike 1998; Hurvich and Tsai 1989; Choubin and Malekian 2017) was used to compare model predictions and observed values.

Classification and regression trees (CART)

Classification and regression trees (CART) is a recursive algorithm in data mining developed by Breiman et al. (1984). CART uses historical data to construct decision trees. Depending on dependent variable, classification tree (for categorical variable) or regression tree (for continuous variable) can be constructed (Breiman et al. 1984; Singh et al. 2014; Choubin et al. 2018). Constructed tree can then be applied for predicting (regression tree) and classifying (classification tree) the new observations. Classification tree makes classes of dependent variables by user or calculated according to some exogenous rule, while regression trees do not have predefined classes. Instead there is dependent variable which represents the response values for observations in independent variable matrix (Timofeev 2004). CART methodology includes three steps: (1) construction of maximum tree; (2) choice of the optimal tree size (3) classification or production of new data using constructed tree. In this study, regression tree was used

to predict precipitation; hence, we describe steps of CART methodology for regression trees. First step (construction of maximum tree) is the most time-consuming. Splitting and construction of maximum tree in regression trees is based on the squared residuals minimization algorithm. The splitting creates the last observations in learning sample. Maximum tree, especially, in the case of regression tree may be very big, so pruning techniques are necessary for cutting off insignificant nodes (Timofeev 2004). Cross-validation and optimization by number of points in each node are two pruning algorithms which can be used to choice right tree size (second step). In algorithm of optimization by number of points in each node, splitting is stopped when number of observations are less than the minimum number of predefined required observations. Cross-validation procedure, which was used in this study, is based on optimal proportion between misclassification error and the tree complexity. With increasing the size of trees (tree complexity), misclassification error is decreasing. The complexity parameter (cp) is used to select the optimal size of the decision tree. Best cp was identified through trial and error. When regression tree is constructed, it can be used for new data (third step). The output of this step is a certain response value to each of the new observations.

Effective inputs to the CART and ANFIS models were determined using correlation analysis between seasonal precipitation and climate signals. Then, three input combinations based on correlation analysis were considered for precipitation forecasting (Table 2). In the first model, the climate signal that has the highest correlation with precipitation was selected as input. In the second model, two climate signals were selected as inputs based on the high correlations with precipitation. In the third model, all climate signals having significant correlations were used as inputs to the applied model.

Data normalization

Climate data in a semiarid region have sparse and irregular distribution; therefore, the best way to improve the robustness of climate information would be data normalization (Choubin et al. 2017a). Thus, the series was normalized to the range of [0, 1] as follows:

$$X_{\text{norm}} = \frac{X_r - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \tag{8}$$

Table 2 Input and output of the ANFIS and CART models

Model	Input	Output
Model 1	Best climate signal	Precipitation
Model 2	Two of the best climate signals	Precipitation
Model 3	All climate signals with significant correlations	Precipitation

where X_{norm} and X_r are the normalized and the original inputs, and X_{min} and X_{max} are the minimum and maximum ranges of inputs, respectively.

Performance criteria

Some performance criteria were used in this research including; Nash–Sutcliffe efficiency (NSE) coefficient (Nash and Sutcliffe 1970), the ratio of the root mean square error (RMSE) to the standard deviation of measured data (RSR), coefficient of determination (R^2) and BIAS (Eqs. 9, 10, 11 and 12, respectively). The NSE is a normalized statistic that determines the relative magnitude of the residual variance (“noise”) compared to the measured data variance (“information”) (Nash and Sutcliffe 1970). NSE indicates how well the predicted values represent the observed data and it is computed using Eq. 9. Singh et al. (2005) recommended RSR as a model evaluation statistic to evaluate the differences between model and observed data in hydrological subjects. RSR (Eq. 10) is calculated based on RMSE and standard deviation of measured data. Determination coefficient describes the degree of collinearity between predicted and observed data (Moriassi et al. 2007) and it is a linear relationship between measured and simulated data (Eq. 11). Bias calculates the average tendency of the simulated data to be larger or smaller than their measured counterparts. The optimal value of BIAS is 0.0, with low-magnitude values indicating accurate model simulation. Negative values indicate model underestimation bias, and positive values indicate model overestimation bias. BIAS is presented in Eq. 12.

$$\text{NSE} = 1 - \frac{\sum_{i=1}^N (O_i - P_i)^2}{\sum_{i=1}^n (O_i - \bar{O})^2} \tag{9}$$

$$\text{RSR} = \frac{\text{RMSE}}{\text{STDEV}_{\text{obs}}} = \frac{\left[\sqrt{\sum_{i=1}^N (O_i - P_i)^2} \right]}{\left[\sqrt{\sum_{i=1}^N (O_i - \bar{O})^2} \right]} \tag{10}$$

$$R^2 = \left(\frac{\sum_{i=1}^N (O_i - \bar{O})(P_i - \bar{P})}{\sqrt{\sum_{i=1}^N (O_i - \bar{O})^2 \sum_{i=1}^N (P_i - \bar{P})^2}} \right)^2 \tag{11}$$

$$\text{BIAS} = \frac{1}{N} \sum_{i=1}^N (P_i - O_i) \tag{12}$$

where N is the number of data points, O_i and P_i are the observed and predicted i th values, \bar{O} and \bar{P} are the mean of the observed and predicted values.

In this study, data were divided into 80 and 20 percent for training and testing sets, respectively.

Results

Determining the models inputs

Correlations between climate signals and precipitation were conducted to determine the model inputs. So, in each

station, the same model inputs were selected for modeling methods. Table 3 shows the correlation matrix between fall precipitation (t) and summer climate signals ($t - 1$). It must be noted that only fall precipitation has satisfactory significant correlation with climate signals. So, fall results were just presented. In this regard, Nazemosadat and Cordey (2000) demonstrated a strong relationship of Iranian fall precipitation data with the El Niño–Southern Oscillation (ENSO) phenomenon. Therefore, the proposed approach in the current study generates a forecast for the

Table 3 Correlation matrix between fall precipitation (t) and summer climate signals (1 month ago, $t - 1$)

Point and data	Bam	Kahnoj	Kerman	Fathabad	Shahdad	Sirjan
Aden Gulf SLP	0.19	0.23	0.13	0.28	0.26	0.37*
Aden Gulf SST	0.09	0.22	0.06	0.36*	0.03	-0.02
Adiabatic Sea SLP	0.24	0.48*	0.06	0.04	0.24	0.49**
Adiabatic Sea SST	0.19	0.02	0.07	0.14	-0.07	-0.01
Arabian Sea SLP	0.18	0.09	0.09	0.23	0.38*	0.30
Arabian Sea SST	-0.22	0.20	-0.34**	-0.10	-0.35*	-0.30
Atlantic Ocean SLP	-0.10	0.08	-0.07	-0.11	-0.16	-0.01
Atlantic Ocean SST	0.21	0.14	-0.02	0.29	0.18	0.10
Azores Sea SLP	-0.20	0.00	-0.01	-0.19	-0.31*	-0.21
Azores Sea SST	-0.21	-0.53**	-0.21	-0.26	-0.09	-0.43*
Black Sea SLP	0.26*	0.45*	0.21	-0.01	0.25	0.43*
Black Sea SST	0.07	-0.32	-0.02	0.09	0.00	-0.21
Caspian Sea SLP	0.13	0.07	-0.00	0.17	0.02	0.14
Caspian Sea SST	0.02	-0.29	-0.05	-0.09	0.01	-0.18
East Mediterranean SLP	0.13	0.49*	0.13	0.20	0.15	0.40*
East Mediterranean SST	0.09	-0.33	-0.03	0.12	0.02	-0.22
Greenland SLP	0.13	0.49*	0.13	0.20	0.15	0.44*
Greenland SST	0.09	-0.33	-0.03	0.12	0.02	-0.22
Indian Ocean SLP	0.26*	0.09	0.11	0.33*	0.37*	0.32
Indian Ocean SST	0.05	-0.09	-0.13	0.21	-0.06	-0.25
Labrador Sea SLP	0.20	0.01	0.07	0.05	0.39*	0.24
Labrador Sea SST	0.01	-0.23	-0.06	0.02	0.02	-0.16
Northern Persian Gulf SLP	0.24	0.49*	0.22	0.19	0.30	0.51**
Northern Persian Gulf SST	-0.02	-0.43*	-0.14	-0.05	0.07	-0.31
Northern Red Sea SLP	0.20	0.31	0.11	0.19	0.26	0.42*
Northern Red Sea SST	-0.01	-0.25	-0.04	0.09	-0.03	-0.26
Northern Sea SLP	0.06	0.01	-0.15	0.11	0.17	0.17
Northern Sea SST	0.13	0.16	-0.01	0.37**	-0.15	0.03
Oman Sea SLP	0.15	0.13	0.06	0.25	0.31*	0.29
Oman Sea SST	-0.18	0.17	-0.32*	-0.09	-0.34*	-0.23
Soran Network SLP	0.13	0.13	0.05	0.13	0.10	0.22
Soran Network SST	-0.17	-0.17	-0.09	-0.01	-0.02	-0.14
Southern Persian Gulf SLP	0.25	0.44*	0.20	0.27	0.42**	0.54**
Southern Persian Gulf SST	-0.05	-0.46*	-0.15	0.01	0.03	-0.40*
Southern Red Sea SLP	0.13	0.27	0.10	0.08	0.20	0.41*
Southern Red Sea SST	-0.02	-0.23	-0.09	0.21	0.02	-0.28
West Mediterranean Sea SLP	0.15	0.41*	0.08	-0.12	0.09	0.38*
West Mediterranean Sea SST	0.22	0.05	-0.07	0.22	0.00	0.09

*Correlation is significant at the 0.05 level; **correlation is significant at the 0.01 level

fall precipitation values based on summer climate signals (previous season).

Results of selected climate signals are represented for each model in Table 4. The best signals with significant correlations are Black Sea SLP (0.264*), Northern Sea SST (0.373**), Arabian Sea SST (−0.340**), Azores Sea SST (−0.527**), Southern Persian Gulf SLP (0.424**) and Southern Persian Gulf SLP (0.538**) for the stations of Bam, Fathabad, Kerman, Kahnoj, Shahdad and Sirjan, respectively. According to Table 4, fall precipitation was predicted by using the input signals of models 1, 2 and 3.

ANFIS results

Various combinations of climate signals represented in Table 4 were used to predict precipitation. In this research, we used grid partition and subtractive fuzzy clustering algorithms to establish the rules based on the relationship between the input and output variables. Grid partition algorithm is used for a few input variables (less than 6, Wei et al. 2007), but in case higher number of input variables, grid partition cannot be used because the fuzzy rules would be too huge (Farokhnia et al. 2011). Therefore, subtractive fuzzy clustering algorithm was used for high number of input variables (higher than 6).

Types and numbers of membership functions were determined by trial-and-error fundamental method. Membership functions of trapezoidal-shaped (trapmf), triangular-shaped (trimf), generalized bell-shaped (gbellmf), gaussian (gaussmf), gaussian 2 (gauss2mf), Π-shaped (pimf), and difference between two sigmoidal functions (dsigmf) were considered for inputs, and linear model was selected as the membership function of output.

ANFIS results for models 1, 2 and 3 are represented in Table 5. Types of membership functions (MF) were determined by trial and error. For training and testing sets, evaluation measures were calculated, and performance rating was represented for the testing set based on Moriasi et al. (2007). The best performance is related to model 2 with good (for Kahnoj station) and satisfactory (for other stations) performance rating (Table 5). As a result, it is clear that the model 2 (two of the best climate signals) provided better results than the models 1 and 3. So, it can be said that the ANFIS model has a satisfactory performance in precipitation forecasting ($0.50 < NSE < 0.65$). Previous studies such as El-Shafie et al. (2011), Choubin et al. (2014, 2017b) also obtained satisfactory results from ANFIS in forecasting precipitation.

Table 4 Selected climate signals for each model based on correlation analysis

Models	Stations	Input: significant climate signals (correlation)
Model 1	Bam	BS_SLP (0.264*)
	Fathabad	NS_SST (0.373**)
	Kerman	ARA_SST (−0.340**)
	Kahnoj	AZE_SST (−0.527**)
	Shahdad	SPG_SLP (0.424**)
	Sirjan	SPG_SLP (0.538**)
Model 2	Bam	BS_SLP (0.264*), IO_SLP (0.263*)
	Fathabad	NS_SST (0.373**), ADE_SST (0.361*)
	Kerman	ARA_SST (−0.340**), OS_SST (−0.318*)
	Kahnoj	AZE_SST (−0.527**), EM_SLP (0.491*)
	Shahdad	SPG_SLP (0.424**), LS_SLP (0.391*)
	Sirjan	SPG_SLP (0.538**), NPG_SLP (0.509**)
Model 3	Bam	Only two signals have significant correlation
	Fathabad	NS_SST (0.373**), ADE_SST (0.361*), IO_SLP (0.330*)
	Kerman	Only two signals have significant correlation
	Kahnoj	AZE_SST (−0.527**), EM_SLP (0.491*), GR_SLP (0.490*), NPG_SLP (0.488*), ADI_SLP (0.484*), SPG_SST (−0.459*), BS_SLP (0.447*), SPG_SLP (0.444*), NPG_SST (−0.430*), WM_SLP (0.409*)
	Shahdad	SPG_SLP (0.424**), LS_SLP (0.391*), ARA_SLP (0.379*), IO_SLP (0.375*), ARA_SST (−0.355*), OS_SST (−0.342*), OS_SLP (0.312*), AZE_SLP (−0.312*),
	Sirjan	SPG_SLP (0.538**), NPG_SLP (0.509**), ADI_SLP (0.492**), EM_SLP (0.440*), GR_SLP (0.440*), AZE_SST (−0.432*), BS_SLP (0.430*), NRS_SLP (0.424*), SRS_SLP (0.407*), SPG_SST (−0.402*), WM_SLP (0.383*), ADE_SLP (0.370*)

*Correlation is significant at the 0.05 level; **correlation is significant at the 0.01 level

Table 5 ANFIS results for models 1, 2 and 3

Models	Stations	Structure ^a	NSE	RSR	R ²	Performance rating
Model 1	Bam	(trimf, 2)	0.22	1.09	0.20	Unsatisfactory
	Fathabad	(pimf, 2)	0.37	0.78	0.37	Unsatisfactory
	Kerman	(trapmf, 2)	0.11	0.93	0.13	Unsatisfactory
	Kahnoj	(trapmf, 3)	0.28	0.83	0.28	Unsatisfactory
	Shahdad	(trapmf, 2)	0.12	0.95	0.12	Unsatisfactory
	Sirjan	(pimf, 2)	0.43	0.74	0.44	Unsatisfactory
Model 2	Bam	(gaussmf, 2)	0.51	0.73	0.52	Satisfactory
	Fathabad	(gaussmf2, 2)	0.59	0.63	0.59	Satisfactory
	Kerman	(trimf, 2)	0.53	0.72	0.53	Satisfactory
	Kahnoj	(pimf, 2)	0.72	0.65	0.79	Good
	Shahdad	(dsigmf, 2)	0.50	0.68	0.50	Satisfactory
	Sirjan	(dsigmf, 2)	0.57	0.64	0.57	Satisfactory
Model 3	Bam ^b	–	–	–	–	–
	Fathabad	(pimf, 2)	0.59	0.63	0.56	Satisfactory
	Kerman ^b	–	–	–	–	–
	Kahnoj	(gaussmf, 2)	0.57	0.66	0.57	Satisfactory
	Shahdad	(gaussmf, 3)	0.30	1.52	0.10	Unsatisfactory
	Sirjan	(gaussmf, 2)	0.56	0.65	0.58	Satisfactory

^aType and number of membership function

^bOnly two signals have significant correlation with rainfall

CART results

Since response variable is continuous or numeric in current research, ANOVA method was used to grow the trees. Then, the optimal tree was determined based on the complexity parameter (cp) for each combination through trial and error. Determining the best cp help to save computing time through cutting off insignificant nodes. Pruning technique reduces overfitting through reducing the decision size and improves the prediction accuracy. CART results for models 1, 2 and 3 are presented in Table 6. Based on the performance ratings (Moriassi et al. 2007), CART results indicate that inputs of the Model 2 have a better performance than the other combinations in all stations (NSE > 0.75; performance rating is very good), while Model 1 is the worst choice of input variable (Table 6).

Time series modeling results

For prediction of precipitation values, this study further implemented the ARIMA model by using the autocorrelation (ACF) and partial autocorrelation (PACF) functions. The best model structures were estimated utilizing the method of maximum likelihood. For example, Fig. 3 shows variation of AICC criterion for determining the best model structure for Kerman station (due to lack of the space, only some of model structures were shown in this figure). As can be seen, ARIMA (0, 1, 1) has the best performance with AICC equal to about 145. Table 7 indicates the best

model for each station which was selected by AICC criterion. Residual analysis based on Fig. 4, the residual autocorrelation function (ACF) and the residual partial autocorrelation function (PACF) of the predicted and observed data, confirm the ARIMA model.

In Table 8, testing results of the ARIMA are represented for each station. Based on Moriassi et al. (2007), performance rating for all stations was found to be unsatisfactory. This implies that the time series modeling is not successful in predicting precipitation.

Models comparison

Seasonal forecasting results of precipitation

Since only fall precipitation, among the other seasons, had significant correlation with climate signals, so, we predicted only fall precipitation data in the study area. Figure 5 indicates the observed and predicted fall precipitation values using CART, ANFIS and ARIMA for Model 2 which has the best performance. As can be seen from Fig. 5, CART has closer fit to the corresponding observed precipitations compared to the ANFIS and ARIMA models. Significantly over- and underestimations of the time series model are obviously seen from the time variation graphs. ARIMA model cannot simulate the variations of precipitation data in the stations.

Table 6 CART results for models 1, 2 and 3

Models	Stations	Structure ^a	NSE	RSR	R ²	Performance rating
Model 1	Bam	(anova, 0.13)	0.41	0.73	0.65	Unsatisfactory
	Fathabad	(anova, 0.44)	0.74	0.48	0.76	Good
	Kerman	(anova, 0.006)	0.74	0.47	0.71	Good
	Kahnoj	(anova, 0.17)	0.51	0.63	0.58	Satisfactory
	Shahdad	(anova, 0.36)	0.73	0.08	0.72	Good
	Sirjan	(anova, 0.008)	0.27	1.03	0.49	Unsatisfactory
Model 2	Bam	(anova, 0.25)	0.83	0.25	0.83	Very good
	Fathabad	(anova, 0.70)	0.81	0.30	0.80	Very good
	Kerman	(anova, 0.13)	0.87	0.34	0.87	Very good
	Kahnoj	(anova, 0.49)	0.76	0.48	0.77	Very good
	Shahdad	(anova, 0.37)	0.84	0.27	0.84	Very good
	Sirjan	(anova, 0.43)	0.78	0.43	0.86	Very good
Model 3	Bam ^b	–	–	–	–	–
	Fathabad	(anova, 0.17)	0.70	0.60	0.73	Good
	Kerman ^b	–	–	–	–	–
	Kahnoj	(anova, 0.06)	0.72	0.64	0.78	Good
	Shahdad	(anova, 0.23)	0.70	0.69	0.73	Good
	Sirjan	(anova, 0.14)	0.74	0.62	0.76	Good

^aThe method used to grow the tree (anova) and complexity parameter (cp)

^bOnly two signals have significant correlation with rainfall

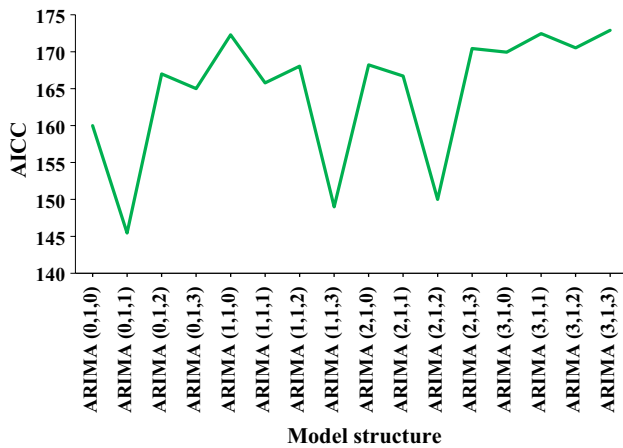


Fig. 3 Variation of AICc criterion for determining the best model structure in Kerman station

Table 7 Best model of time series for climate station

Station	Best model	AICc
Bam	ARIMA (1, 0, 0)	92.53
Fathabad	ARIMA (1, 0, 0)	103.72
Kerman	ARIMA (0, 1, 1)	145.45
Kahnoj	ARIMA (0, 1, 1)	74.87
Shahdad	ARIMA (1, 0, 1)	81.78
Sirjan	ARIMA (0, 0, 1)	75.67

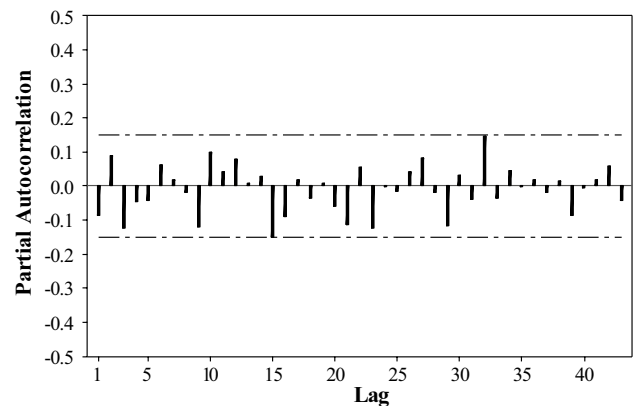
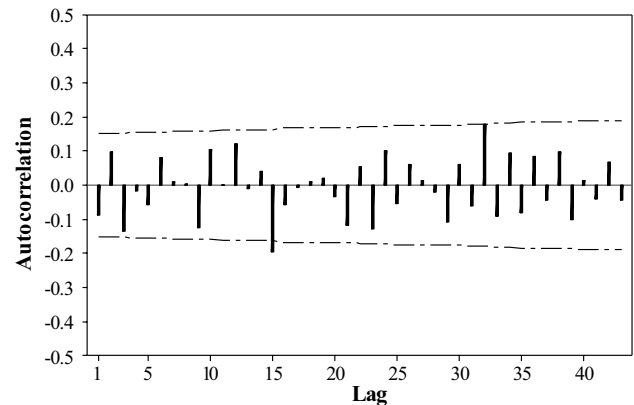


Fig. 4 Autocorrelation and partial autocorrelation functions of the residuals (ACF and PACF) for the Bam station

Table 8 ARIMA results of each station for testing set

Station	NSE	RSR	R ²	Performance rating
Bam	0.09	1.00	0.01	Unsatisfactory
Fathabad	0.08	0.97	0.11	Unsatisfactory
Kerman	0.00	0.96	0.01	Unsatisfactory
Kahnoj	0.04	1.06	0.45	Unsatisfactory
Shahdad	0.16	1.64	0.03	Unsatisfactory
Sirjan	-1.19	1.35	0.02	Unsatisfactory

Evaluation of spatial bias of precipitation

To evaluate the spatial variability of precipitation values, models' bias was used. Figure 6 indicates the spatial biases of ARIMA, ANFIS and CART models for testing set. Also, it is clear that the bias in ARIMA varies between -14.41 and 14.04 mm, whereas for ANFIS, the bias is lower (between -8.07 and 9.82 mm). Lowest bias is related to CART prediction which ranges from -1.10 to 5.50 mm. ARIMA has overestimations in west of region, on the other hand, it has underestimations in the center of the area. Whereas CART and ANFIS reveal approximately same variations where they have overestimations in south and southeast, and have underestimations in north, northeast and northwest of the Kerman Province.

Discussion

The best correlated signals with precipitation in this study were the SPG, BS, NS, ARA, AZE and SPG, respectively, for the stations of Sirjan, Bam, Fathabad, Kerman, Kahnoj and Shahdad, respectively. In this regard, there are some studies such as Fallah-Ghalhary et al. (2010) and Ruigar and Golian (2016) which demonstrated relationships of these climate signals on precipitation of Iran, as well. Fallah-Ghalhary et al. (2010) indicated that sea surface temperature (SST) and sea-level pressure (SLP) of the EM, NPG, SPG, OS, ADE, BS, ARA, IO, ATL and NS have effects on climate of the Middle East and Iran. Ruigar and Golian (2016) confirmed the effects of the above-mentioned signals (and other signals used in this study) on north of Iran.

Results of this study indicated that CART performs better than the ANFIS and ARIMA in forecasting the fall precipitation. Structure of models can be a reason for differences between the models performance. Predictions are related with the computing power of the different algorithms (Loh 2011), which is caused differences in their outcome and response. The main important limitation of ARIMA is that it uses historical data for prediction and requires a long data series (Sen et al. 2016). The ANFIS model is rule-based technique, but it is facing with some limitations. ANFIS not

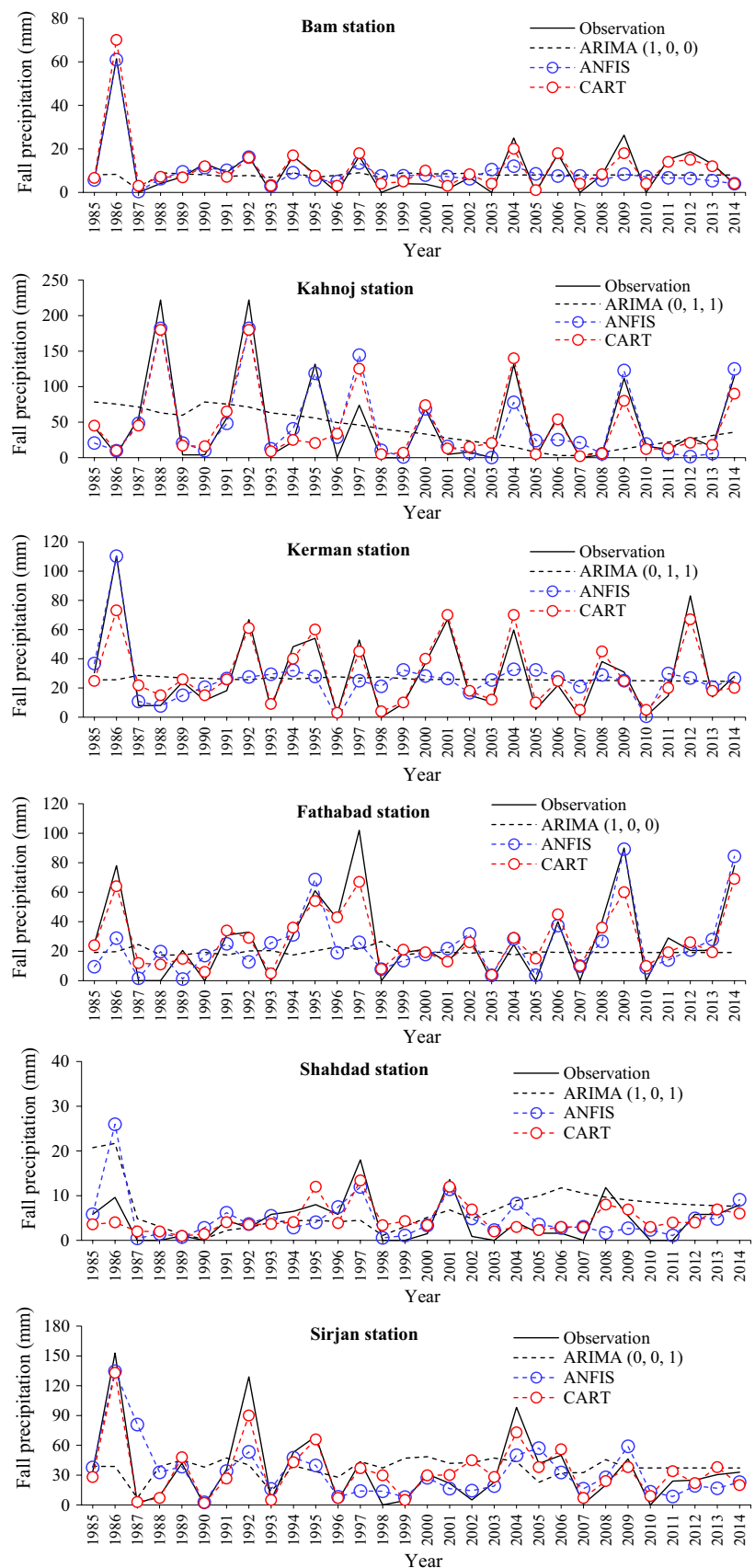
have the capability of producing rules to explain its predictions and it is sensitive to learning datasets (Seera et al. 2012). Also, the ANFIS model can overlearn during training period, which it leads to reducing the performance during testing period (Choubin et al. 2014, 2016a).

Our results are agreement with Choubin et al. (2018), where they found the CART model has better performance compared with ANFIS for estimation of the river-suspended sediment load. The major advantages of the CART model which has better performance compared to others are: (1) it is nonparametric, so does not need specification of any functional form; (2) CART results are fixed to uniform transformations of its independent variables (Timofeev 2004); (3) the splitting algorithm of CART easily isolates the outliers in a separate node (Loh 2011; Timofeev 2004); (4) CART has no assumptions and it is fast in view of computation; (5) it is flexible and has an capability to regulate in time (Timofeev 2004); and (6) CART has the capability of explaining its prediction with rules and it is less sensitive to learning datasets (Seera et al. 2012).

The results of this study demonstrated that the predictions accomplished better performance by two of the best climate indices (Model2) in prediction of fall precipitation at time "t" by using the climate signals at "t-1." During the summer (t-1), large-scale oceanic and atmospheric information can provide important insight into next season climate conditions and fall extreme precipitation events. Specifically, global climate signals were shown to have a good statistical correlation with fall precipitation in the forecasting scheme of next season precipitation values. The CART model showed significant skill at fall precipitation forecast, thus providing crucial advance knowledge of precipitation characteristics which enables efficient water resources planning and management. The outputs of this research can provide useful insights into the skill of longer-lead climate forecasts and also can increase the skill of shorter-lead forecasts that rely on seasonal hydroclimate data. To this end, incorporating large-scale climate predictors into a semiarid forecast model allowed for a shorter-lead time forecast (beginning in September) and also contributed a significant skill during the fall through forecast of early winter extreme events.

In practice, CART model requires more sophisticated data transformation and screening of candidate predictor variables to predict precipitation values precisely. It is recognized that this step in the procedure still assumes a degree of some statistics objectives performance concerning the choice of most appropriate large-scale predictors. Furthermore, in prediction processes, precipitation can be divided into two classifications: dry and wet data, next CART framework could apply on wet class to obtain the precipitation amount. Further efforts are warranted to categorize wet data more finely in technique. Considering days of light, medium and heavy rain by constructing different models for each

Fig. 5 Comparison of observed and simulated (Model 2) fall precipitation data



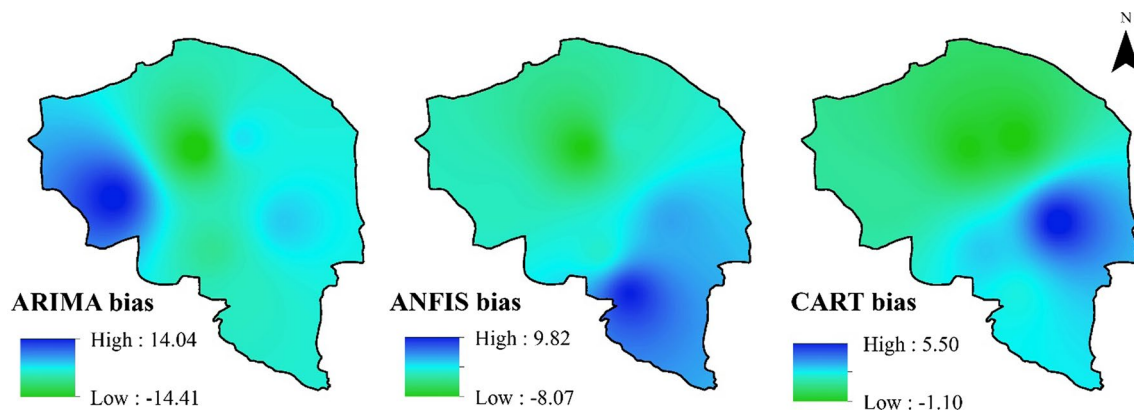


Fig. 6 Spatial bias for testing dataset

class can enhance precipitation forecast of a complex climate system. This approach can effectively reduce the number of invalid values calculated by the model. Moreover, it can reproduce precipitation values with less bias and improve the performance criteria. In order to evaluate whether model indeed decreases the entire projection envelope of seasonal climatic variables, it is necessary to implement an additional study on bias analysis and study additional series (here daily and monthly precipitation datasets), which could provide more insight into the climate system (Samadi et al. 2013). Therefore, future studies can apply other machine learning techniques such as self-organizing maps (SOM) and support vector machine (SVM) to compare the results with CART.

Conclusion

This study focuses on the application and evaluation of CART in prediction of seasonal precipitation. Accuracy of the CART model was compared with two most commonly used models (ANFIS and time series modeling) for prediction of precipitation. The results revealed that the CART produced more accurate fall precipitation values than the other models, and this was also confirmed by spatial bias analysis. The results of the CART, in addition, demonstrated that the predictions accomplished better performance by two of the best climate indices in prediction of fall precipitation at time “ t ” by using the climate signals at “ $t - 1$.”

Findings of this research can be utilized to detect precipitation values in a semiarid region and can be also used as a basis for precipitation forecasting in south and southeast Iran with the same climate conditions. Having a skillful forecast of the upcoming fall precipitation during early summer is a significant importance to water managers and decision makers because fall precipitation usually causes flash floods in semiarid regions. Incorporating this forecast in a decision

support model can be further investigated to provide promising alternatives for arid water resources management.

Acknowledgements This study was partially funded by University of Tehran (grant number 7401001/1/4).

References

- Afshin S, Fahmi H, Alizadeh A, Sedghi H, Kaveh F (2011) Long term rainfall forecasting by integrated artificial neural network-fuzzy logic-wavelet model in Karoon basin. *Sci Res Essays* 6(6):1200–1208. <https://doi.org/10.5897/SRE10.448>
- Akaike H (1998) Information theory and an extension of the maximum likelihood principle. In: Selected papers of Hirotugu Akaike. Springer, New York, pp 199–213. https://doi.org/10.1007/978-1-4612-1694-0_15
- Azadi S, Sepaskhah AR (2012) Annual precipitation forecast for west, southwest, and south provinces of Iran using artificial neural networks. *Theoret Appl Climatol* 109:175–189. <https://doi.org/10.1007/s00704-011-0575-9>
- Berg N, Hall A, Capps SB, Hughes M (2013) El Nino-Southern Oscillation impacts on winter winds over Southern California. *Clim Dyn* 40(1–2):109–121. <https://doi.org/10.1007/s00382-012-1461-6>
- Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth and Brooks/Cole, Monterey
- Choubin B, Malekian A (2017) Combined gamma and M-test-based ANN and ARIMA models for groundwater fluctuation forecasting in semiarid regions. *Environ Earth Sci* 76(15):538. <https://doi.org/10.1007/s12665-017-6870-8>
- Choubin B, Khalighi-Sigaroodi S, Malekian A, Ahmad S, Attarod P (2014) Drought forecasting in a semi-arid watershed using climate signals: a neuro-fuzzy modeling approach. *J Mt Sci* 11(6):1593–1605. <https://doi.org/10.1007/s11629-014-3020-6>
- Choubin B, Khalighi-Sigaroodi S, Malekian A, Kişi Ö (2016a) Multiple linear regression, multi-layer perceptron network and adaptive neuro-fuzzy inference system for forecasting precipitation based on large-scale climate signals. *Hydrol Sci J* 61(6):1001–1009. <https://doi.org/10.1080/02626667.2014.966721>
- Choubin B, Malekian A, Golshan M (2016b) Application of several data-driven techniques to predict a standardized precipitation

- index. *Atmósfera* 29(2):121–128. <https://doi.org/10.20937/ATM.2016.29.02.02>
- Choubin B, Solaimani K, Habibnejad Roshan M, Malekian A (2017a) Watershed classification by remote sensing indices: a fuzzy c-means clustering approach. *J Mt Sci*. <https://doi.org/10.1007/s11629-017-4357-4>
- Choubin B, Malekian A, Samadi S, Khalighi-Sigaroodi S, Sajedi-Hosseini F (2017b) An ensemble forecast of semi-arid rainfall using large-scale climate predictors. *Meteorol Appl*. <https://doi.org/10.1002/met.1635>
- Choubin B, Darabi H, Rahmati O, Sajedi-Hosseini F, Kløve B (2018) River suspended sediment modelling using the CART model: a comparative study of machine learning techniques. *Sci Total Environ* 615:272–281. <https://doi.org/10.1016/j.scitotenv.2017.09.293>
- De'ath AG, Fabricius KE (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology* 81:3178–3192
- El-Shafie A, Jaafer O, Seyed A (2011) Adaptive neuro-fuzzy inference system based model for rainfall forecasting in Klang River, Malaysia. *Int J Phys Sci* 6(12):2875–2888. <https://doi.org/10.5897/IJPSI1.515>
- Fallah-Ghalhary GA, Habibi-Nokhandan M, Mousavi-Baygi M, Khoshhal J, Shaemi Barzoki A (2010) Spring rainfall prediction based on remote linkage controlling using adaptive neuro-fuzzy inference system (ANFIS). *Theoret Appl Climatol* 101:217–233. <https://doi.org/10.1007/s00704-009-0194-x>
- Farokhnia A, Morid S, Byun HR (2011) Application of global SST and SLP data for drought forecasting on Tehran plain using data mining and ANFIS techniques. *Theoret Appl Climatol* 104:71–81. <https://doi.org/10.1007/s00704-010-0317-4>
- Gamiz-Fortis SR, Esteban-Parra MJ, Trigo RM, Castro-Diez Y (2010) Potential predictability of an Iberian river flow based on its relationship with previous winter global SST. *J Hydrol* 385:143–149. <https://doi.org/10.1016/j.jhydrol.2010.02.010>
- Gaughan AE, Waylen PR (2012) Spatial and temporal precipitation variability in the Okavangoe–Kwandoe–Zambezi catchment, southern Africa. *J Arid Environ* 82:19–30. <https://doi.org/10.1016/j.jaridenv.2012.02.007>
- Hurvich CM, Tsai C-L (1989) Regression and time series model selection in small samples. *Biometrika* 76:297–307. <https://doi.org/10.1093/biomet/76.2.297>
- Jang JSR (1993) ANFIS: adaptive network-based fuzzy inference systems. *IEEE Trans Syst Man Cybern* 23(3):665–685. <https://doi.org/10.1109/21.256541>
- Khalili A (1997) Integrated water plan of Iran. Vol. 4: meteorological studies, ministry of energy, Iran. Lecha, L. and P. Shackelford, 1997. Climate services for tourism and recreation. WMO Bull 46:46–47
- Kisi O, Haktanir T, Ardiclioglu M, Ozturk O, Yalcin E, Uludag S (2009) Adaptive neuro-fuzzy computing technique for suspended sediment estimation. *Adv Eng Softw* 40(6):438–444. <https://doi.org/10.1016/j.advengsoft.2008.06.004>
- Kurtulus B, Razack M (2010) Modeling daily discharge responses of a large karstic aquifer using soft computing methods: artificial neural network and neuro-fuzzy. *J Hydrol* 381(1–2):101–111. <https://doi.org/10.1016/j.jhydrol.2009.11.029>
- Lemons J (2003) Conserving biodiversity in arid regions: best practices in developing nations. Springer, New York
- Loh WY (2011) Classification and regression trees. *Wiley Interdiscip Rev Data Min Knowl Discov* 1:14–23. <https://doi.org/10.1002/widm.8>
- Moriasi DN, Arnold JG, Van Liew MW, Binger RL, Harmel RD, Veith T (2007) Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans ASABE* 50(3):885–900. <https://doi.org/10.13031/2013.23153>
- Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models: part 1-A discussion of principles. *J Hydrol* 10(3):282–290. [https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6)
- Nazemosadat MJ, Cordey I (2000) On the relationship between ENSO and autumn rainfall in Iran. *Int J Climatol* 20(1):47–61. [https://doi.org/10.1002/\(sici\)1097-0088\(200001\)20:1<47::aid-joc461>3.0.co;2-p](https://doi.org/10.1002/(sici)1097-0088(200001)20:1<47::aid-joc461>3.0.co;2-p)
- Rezaeian-Zadeh M, Tabari H, Abghari H (2012) Prediction of monthly discharge volume by different artificial neural network algorithms in semi-arid regions. *Arab J Geosci* 6(7):2529–2537. <https://doi.org/10.1007/s12517-011-0517-y>
- Ruigar H, Golian S (2016) Prediction of precipitation in Golestan dam watershed using climate signals. *Theoret Appl Climatol* 123(3–4):671–682
- Samadi S, Wilson CA, Moradkhani H (2013) Uncertainty analysis of statistical downscaling models using Hadley Centre Coupled Model. *Theoret Appl Climatol*. <https://doi.org/10.1007/s00704-013-0844-x>
- Sanikhani H, Kisi O (2012) River flow estimation and forecasting by using two different adaptive neuro-fuzzy approaches. *Water Resour Manag* 26(6):1715–1729. <https://doi.org/10.1007/s11269-012-9982-7>
- Seera M, Lim CP, Ishak D, Singh H (2012) Fault detection and diagnosis of induction motors using motor current signature analysis and a hybrid FMM–CART model. *IEEE Trans Neural Netw Learn Syst* 23(1):97–108
- Sen P, Roy M, Pal P (2016) Application of ARIMA for forecasting energy consumption and GHG emission: a case study of an Indian pig iron manufacturing organization. *Energy* 116:1031–1038
- Sharifi F, Samadi SZ, Wilson CAME (2012) Causes and consequences of recent floods in the Golestan catchments and Caprecipitation-Sea regions of Iran. *Nat Hazards* 61(2):533–550. <https://doi.org/10.1007/s11069-011-9934-1>
- Sigaroodi SK, Chen Q, Ebrahimi S, Nazari A, Choobin B (2014) Long-term precipitation forecast for drought relief using atmospheric circulation factors: a study on the Maharloo Basin in Iran. *Hydrol Earth Syst Sci* 18(5):1995–2006. <https://doi.org/10.5194/hess-18-1995-2014>
- Singh J, Knapp HV, Arnold JG, Demissie M (2005) Hydrologic modeling of the Iroquois River watershed using HSPF and SWAT. *J Am Water Resour Assoc* 41(2):343–360. <https://doi.org/10.1111/j.1752-1688.2005.tb03740.x>
- Singh R, Wagener T, Crane R, Mann ME, Ning L (2014) A vulnerability driven approach to identify adverse climate and land use change combinations for critical hydrologic indicator thresholds: application to a watershed in Pennsylvania, USA. *Water Resour Res* 50:3409–3427. <https://doi.org/10.1002/2013WR014988>
- Sutton CD (2005) Classification and regression trees, bagging, and boosting. *Handb Stat* 24:303–329. [https://doi.org/10.1016/S0169-7161\(04\)24011-1](https://doi.org/10.1016/S0169-7161(04)24011-1)
- Timofeev R (2004) Classification and regression trees (CART) theory and applications. In: Master Thesis. Center of Applied Statistics and Economics, Humboldt University, Berlin
- Wei M, Bai B, Sung AH, Liu Q, Wang J, Cather ME (2007) Predicting injection profiles using ANFIS. *Inf Sci* 177(20):4445–4461. <https://doi.org/10.1016/j.ins.2007.03.021>
- Yatheendradas S, Wagener T, Gupta H, Unkrich C, Goodrich D, Schaffner M, Stewart A (2008) Understanding uncertainty in distributed flash flood forecasting for semiarid regions. *Water Resour Res* 44(5):W05S19. <https://doi.org/10.1029/2007wr005940>