Input selection for long-lead precipitation prediction using large-scale climate variables: a case study

Azadeh Ahmadi, Dawei Han, Elham Kakaei Lafdani and Ali Moridi

ABSTRACT

In this study, a precipitation forecasting model is developed based on the sea level pressures (SLP), difference in sea level pressure and sea surface temperature data. For this purpose, the effective variables for precipitation estimation are determined using the Gamma test (GT) and correlation coefficient analysis in two wet and dry seasons. The best combination of selected variables is identified using entropy and GT. The performances of the alternative methods in input variables selection are compared. Then the support vector machine model is developed for dry and wet seasonal precipitations. The results are compared with the benchmark models including naïve, trend, multivariable regression, and support vector machine models. The results show the performance of the support vector machine in precipitation prediction is better than the benchmark models.

Key words | climatic prediction, entropy theory, Gamma test, precipitation prediction, support vector machine

Azadeh Ahmadi (corresponding author)
Department of Civil Engineering,
Isfahan University of Technology,
Isfahan,
Iran
E-mail: *aahmadi@cc.iut.ac.ir*

Dawei Han
Water and Environmental Management Research Centre,
Department of Civil Engineering,
University of Bristol,
UK

Elham Kakaei Lafdani
Department of Watershed Management,
Faculty of Natural Resources and Marine Sciences,
Tarbiat Modares University,
Noor,
Iran

Ali Moridi
Abbaspoor College of Technology,
Shahid Beheshti University,
Tehran,
Iran

INTRODUCTION

Adaptation with climate variability is one of the main challenges in water resources management. Development of accurate precipitation and streamflow forecast models can help for water resources planning and management. Precipitation prediction can be made in the short term such as next hourly prediction for flood management or in the long term such as prediction of the future 6 months. Usually, long-term precipitation forecasts are used for operational purposes: irrigation and water supply management, flood warning and prevention and hydropower planning. In recent decades, much effort has improved development of reliable long-lead forecasting models utilizing large-scale ocean-atmospheric patterns. There are various methods to develop the relationship between the large-scale climatic parameters such as geopotential height and mean sea level pressure (SLP) (predictors) and the local variables (predictands) such as temperature, precipitation and runoff. The most widely used models usually implement general circulation model (GCM) outputs as predictors to predict the precipitation. However, because the GCM data are in coarse resolution, downscaling techniques are employed to model climate outputs into metrological variables appropriate for hydrologic applications.

In recent years, a wide range of rainfall prediction methods has been used to investigate the effects of large-scale climate signals on rainfall variability. Dynamical and empirical methods are two approaches for rainfall prediction. In the dynamical approach, a physical model is utilized to generate and solve a system of equations for precipitation prediction. In the empirical method, the historical atmospheric and oceanic data are analyzed and a simulation model is developed for precipitation prediction of future periods. Recently, a wide range of conceptual models, such as regression, artificial neural networks (ANNs), and *K*-nearest neighbor (KNN) have been used in empirical approaches.

doi: 10.2166/hydro.2014.138

Many efforts have been devoted to the development of prediction models by implementing linear methods, such as simple/multiple linear regression and canonical correlation analysis (Johansson & Chen 2003; Hanssen-Bauer *et al*. 2003; Busuico *et al*. 2006; Crawford *et al*. 2007; Schmidli *et al*. 2007; Hertig & Jacobeit 2008; Hashmi *et al*. 2009), independent component analysis (Moradkhani & Meier 2010), or singular value decomposition (Conway *et al*. 1996; Widmann *et al*. 2003).

When the predictand variable is precipitation, linear regression relationship may not work very well, because the predictor–predictand relationships are often very complex. For this reason, a number of nonlinear regression downscaling techniques, especially ANNs because of their high potential for simulating the complex, nonlinear, and time-varying input–output systems, are employed (e.g., Mpelasoka *et al*. 2001; Cavazos & Hewitson 2005; Haylock *et al*. 2006; Hashmi *et al*. 2009; Najafi *et al*. 2011).

Support vector machines (SVMs) as one of the nonlinear modeling tools are widely used in precipitation prediction. SVMs for regression (SVR), as described by Vapnik (1992), exploit the idea of mapping input data into a high dimensional (often infinite) reproducing kernel Hilbert space where a linear regression is performed. Dibike *et al*. (2001) presented some results showing that radial basis function (RBF) is the best kernel function to be used in SVM models. Liong & Sivapragasam (2002) compared SVM with ANN and concluded that the SVM's inherent properties give it an edge in overcoming some of the major problems in the application of ANN (Han *et al*. 2007). Bray & Han (2004) illustrated the difficulties in SVM identification for flood forecasting problems. Tripathi *et al*. (2006) identified climate variables affecting spatio-temporal variation of precipitation in India. Then, the SVM-based downscaling model is applied to future climate predictions from the second generation coupled global climate model to obtain future projections of precipitation. Ghosh & Mujumdar (2008) developed downscaling models based on sparse Bayesian learning and relevance vector machine to model streamflow at river basin scale for the monsoon period using GCM simulated climatic variables. A decreasing trend is observed for monsoon streamflow of Mahanadi due to high surface warming in the future, with the CCSR/NIES GCM and B2 scenario. Najafi *et al*. (2011) used multilinear regression, SVM, and adaptive-network-based fuzzy inference system.

Other downscaling techniques including KNN (Araghinejad *et al*. 2006) and genetic programming (Hashmi *et al*. 2011) are also utilized. Depending on regions and criteria of comparison, any linear and nonlinear techniques can be employed.

Some other techniques are used for data preprocessing to reduce the dimensionality of the problem, including sensitivity analysis (Nourani & Sayyah Frad 2012), principal component analysis (Schoof & Pryor 2001; Araghinejad & Burn 2005), fuzzy clustering (Ghosh & Mujumdar 2008), wavelet transform (Nourani & Parhizkar 2013) and Gamma test (GT) (Ahmadi *et al*. 2009; Moghaddamnia *et al*. 2009; Ahmadi & Han 2013). The GT is a nonlinear modeling and analysis tool. GT predicts the minimum achievable modeling error before the modeling. GT was first reported by Stefansson *et al*. (1997) and Končar (1997), and later was discussed by many scientists and used to determine the best input combination (Chuzhanova *et al*. 1998; Remesan *et al*. 2008; Jaafar & Han 2011).

A main challenge of developing prediction models is input selection. A basic question is how many input variables should be considered in a model. Although the model performance is higher in model calibration, the systems accuracy in estimation cannot be improved with more inputs. Recent researches have been to investigate the best input variables and data length using the GT (Ahmadi *et al*. 2009; Piri *et al*. 2009; Jaafar & Han 2011).

The aim of this paper is to find the relationship between the large-scale climate parameters provided by the National Centers for Environmental Prediction (NCEP) and the precipitation of the Aharchay watershed in the northwestern part of Iran. This paper deals with the input selection challenge in three parts in identification of: (1) the most effective predictors; (2) the best combination of predictors; and (3) the best simulation method. In this paper, a comparison and assessment have been carried out for selection of the most effective predictors by GT and correlation coefficient analysis. The best combination of selected predictors is identified using entropy theory and GT. The best simulation model is selected among SVM, naïve, trend, and multivariable regression models.

The main novelty of this paper is utilizing the GT in two steps including determining the effective input variables and the best combination of input variables. The results are compared with other methods including correlation coefficient and entropy theory, respectively. The other contribution of this paper is proposing certain steps for achieving a more accurate long-lead precipitation prediction model including determination of effective signals, best combination of input variables, and best simulation models among some alternatives. The paper is organized as follows, with the 'Materials and methods' section including the used models GT, entropy theory, and SVM. This is followed by a case study for implementing the proposed methodology. The obtained results are presented in the next section which is followed by a summary and conclusion.

MATERIALS AND METHODS

The modeling processes in this study including determination of effective signals, the best combination of them, and the best simulation model are presented in Figure 1. First, the time series of effective signals on western Iran's precipitation (suggested by Karamouz *et al.* 2005) are gathered and updated. Then the seasonal precipitation for two seasons of wet (December to May) and dry (June to November) are calculated. In order to select the effective variables out of the total climate signals, two methods including GT and correlation coefficient analysis considering multicollinearity are used. The results of the methods are compared through developing the simulation models.

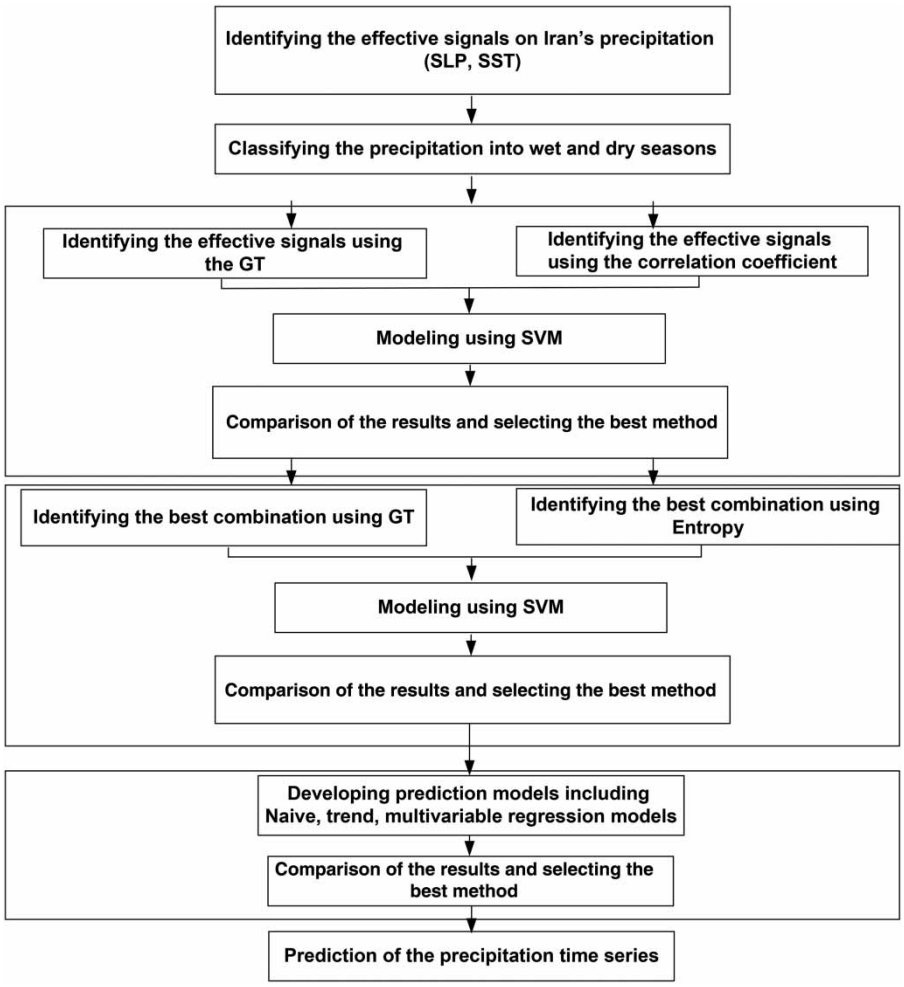

Figure 1 | The flowchart of the study steps.

After identifying the effective signals, in order to reduce the complexity and increase the model's accuracy, the best combination of selected variables using GT and entropy theory is selected. In the third part of modeling, the results of the SVM model are compared with the benchmark models including naïve, trend, and multivariable regression models. In the following sub-sections, brief explanations about the methods are given.

Gamma test

This novel technique enables us to quickly evaluate and estimate the best mean squared error that can be achieved by a smooth model on unseen data for a given selection of inputs, prior to model construction.

GT estimates the minimum mean square error which is achievable in continuous nonlinear models with unseen data. The main idea was somewhat different from previous efforts for the nonlinear analysis. Suppose X_i and X_j are close to each other; therefore, y_j and y_i should also be close to each other. In GT, it is attempted to make this view qualitative through mean distance between the nearest neighbor bounded set of X_i and X_j and mean length between the corresponding output points of y_j and y_i and achieve estimation for error value. Suppose there are a series of observations as the following form:

$$((x_1, \ldots, x_m), y) = (X, y) \tag{1}$$

where $X = (x_1, \ldots, x_m)$ is the input vector at the range of $C \in R^m$, and y is the output vector. The only assumption of this method is that the following equation is established between the systems:

$$y = f(x_1, \ldots, x_m) + r \tag{2}$$

where r is the random variable that illustrates noise of equation and must be determined. Without losing the generality of the function, it can be assumed that mean of this random variable is zero (as any constant bias might be subsumed into an unknown function) and its variance is bounded. GT is based on $N[i, k]$ which includes a set of nearest neighbors from $k(1 \le k \le p)$ for each vector $X_i(1 \le i \le M)$. Delta function calculates the mean square of kth distance from the neighbor:

$$\delta_M(k) = \frac{1}{M}\sum_{i=1}^{M} \left|X_{N[i,k]} - X_i\right|^2 \tag{3}$$

where | | indicates Euclidean distance, corresponding gamma function is as:

$$\gamma_M(k) = \frac{1}{2M}\sum_{i=1}^{M} \left(y_{N[i,k]} - y_i\right)^2 \tag{4}$$

where $y_{N[i,k]}$, y is the corresponding value for the kth neighbor of X_i in Equation (4). In order to calculate Γ, a linear regression line is built from P point on values of $\delta_M(k)$ and $\gamma_M(k)$.

$$\gamma = A\delta + \Gamma \tag{5}$$

Intercept of the vertical axis ($\delta = 0$) is the value of Γ and $\gamma_M(k)$ is equal to variance errors. Drawing the regression line can provide useful data about complexity grade of the model. Vertical intercept of estimated line provides the best obtainable mean square error (Evans & Jones 2002). Furthermore, the gradient of the line provides complexity of the model (high greater complexity models have steeper gradient). Gamma is a conceptual model and its results have nothing to do with used techniques for a model of f function. These results can be standardized by considering the term of V_{ratio} which is defined as follows:

$$V_{\text{ratio}} = \frac{\Gamma}{\sigma^2(y)} \tag{6}$$

where $\sigma^2(y)$ is the y output variance that provides the power of judgment to be formed independent from the output range. When V_{ratio} is close to zero, there would be a higher degree of predictability of the required output of the model. A formal proof for the GT can be found in Evans (2002) and Evans & Jones (2002).

Entropy

Entropy is a tool for quantifying the uncertainty of random processes. It measures the reduction of the uncertainty

using the observation data based on the gained information. Shannon & Weaver (1949) developed the principles of the information theory in terms of 'Entropy'. Singh (1997) reviewed some applications of entropy approaches in hydrology and water resources.

Harmancioglu & Alpaslan (1992), Caselton & Husain (1980), Harmancioglu & Singh (1998), and Husain (1989) used entropy to assess uncertainties of hydrologic variables in water resources systems and to design water quality monitoring and hydrological networks. Also, Krstanovic & Singh (1992), Mogheir & Singh (2002), and Alfonso *et al*. (2010) have used entropy theory in the field of designing groundwater quantity and quality monitoring networks.

Shannon & Weaver (1949) defined the marginal entropy, $H(X)$, of a discrete random variable X as follows:

$$H(X) = -\sum_{i=1}^{N} p(x_i) \log p(x_i) \tag{7}$$

Here, N represents the number of elementary events with probabilities $p(x_i)(i = 1, \ldots, N)$. Transinformation measures the redundant or mutual information between dependent X and Y expressed as follows:

$$T(X, Y) = H(X) + H(Y) - H(X, Y) \tag{8}$$

or as

$$T(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{9}$$

where $H(X|Y) =$ conditional probability density function of X with respect to Y. By definition, mutual information is the reduction in uncertainty with respect to Y due to observation of X (Cover & Thomas 1991).

Support vector machine

The fundamental of SVM was developed by Vapnik (1998). SVM is based on the principle of structural risk minimization from statistical learning theory. The application of SVM has received attention in the field of hydrological engineering and water resources management due to its many interesting features and promising empirical performance (Choy & Chan 2003; Yu *et al*. 2004; Bray & Han 2004; Sivapragasam & Liong 2005; Karamouz *et al*. 2009).

The SVM model is produced by support vectors included in the training data and presents the means of a small subset of training points. The cost function for building the model ignores any training data that are within a threshold ϵ to the model prediction. In the SVM method, the generalization bounds are relied on defining the loss function that ignores errors. In SVM, the problem is to find a linear function that best interpolates a set of training points for the following equation:

$$y = Wx + b \tag{10}$$

The parameters (W, b) should be determined to minimize the sum of the squared deviations of the data utilizing the least squares approach

$$\sum_{i=1}^{l} (y_i - Wx_i - b)^2 \tag{11}$$

Some deviation ε between the eventual targets y_i and the function y is allowed by defining the following constraint:

$$(y_i - Wx_i \pm b) \leq \varepsilon \tag{12}$$

A band or a tube around the hypothesis function y can be visualized, with points outside the tube regarded as training errors, otherwise called slack variables ξ_i. For points inside the tube, the slack variables are zero and increase gradually for points outside the tube. This approach to regression is called ε-SV regression (Vapnik 1998). It can be shown that this regression problem can be expressed as the following convex optimization problem:

$$\text{Min} \quad \frac{1}{2} W^2 + C \sum_{i=1}^{l} (\xi_i + \xi_i^*) \tag{13}$$

Subject to

$$\begin{aligned} &y_i - (W.x_i + b) \leq \varepsilon + \xi_i \\ &(W.x_i + b) - y_i \leq \varepsilon + \xi_i^* \\ &\xi_i, \xi_i^* \geq 0, i = 1, 2, \ldots, l \end{aligned} \tag{14}$$

where C is a pre-specified and positive constant that determines the degree of penalized loss when a training error occurs, ξ_i and ξ_i^* are slack variables that represent the upper and the lower training errors subject to an error tolerance ε.

Then the Lagrange function is constructed from both the objective function and the corresponding constraints to solve the optimization problem. SVMs are characterized by usage of kernel function used to change the representation of the data in the input space to a linear representation in a higher dimensional space called a feature space. Four standard kernels are usually used in classification problems and also used in regression cases: linear, polynomial, radial basis, and sigmoid. The architecture of an SVM algorithm for regression is presented in Figure 2. The input pattern (for which a prediction is to be made) is mapped into feature space. Then the products are computed with the training patterns (support vectors) using kernel functions. Finally, the products are added up using the weights. This, plus the constant term b yields the final prediction output. For more information about SVMs, readers are referred to Vapnik (1992, 2010).

Model evaluation

The criteria of root mean square error, determination coefficient, and Nash–Sutcliffe model efficiency coefficient are used to evaluate the performance of simulation modeling of the historical precipitation. The following formulas are used to calculate them:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{n}} \tag{15}$$

where y_t is the observed value of the historical precipitation, $\hat{y}_t$ is the modeled value of the precipitation, and n is the number of data.

The correlation coefficient indicates the strength and direction of a linear relationship between two variables. The correlation is $+1$ in the case of a perfect increasing linear relationship, and -1 in the case of a decreasing linear relationship

$$R = \frac{\sum_{t=1}^{n} (y_t - \overline{y_t})(\hat{y}_t - \overline{\hat{y}_t})}{\sqrt{\sum_{t=1}^{n} (y_t - \overline{y_t})^2 \cdot \sum_{t=1}^{n} (\hat{y}_t - \overline{\hat{y}_t})^2}} \tag{16}$$

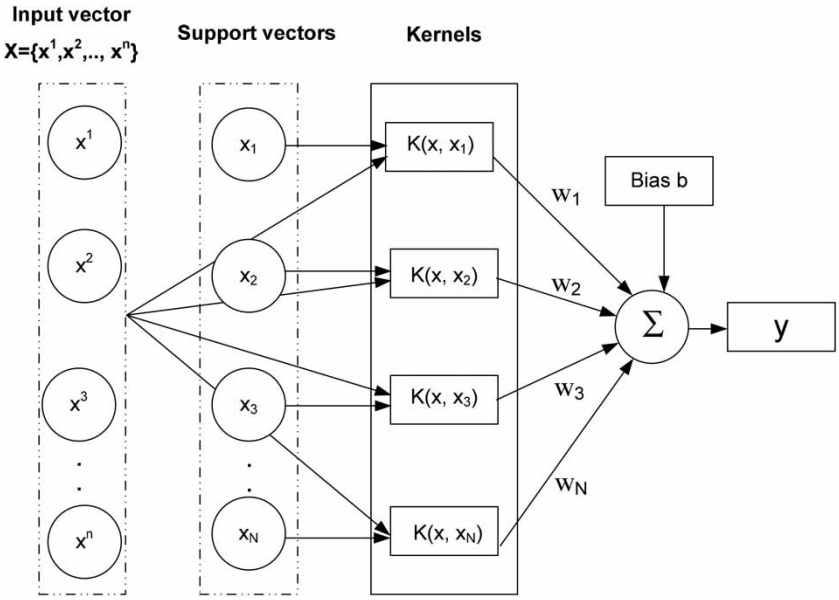

Figure 2 | Architecture of a regression SVM.

where $\overline{y_t}$, $\overline{\hat{y}_t}$ are the mean value of observed and modeled precipitation values. The square of the correlation coefficient (r^2), known as the coefficient of determination, ranges from 0 to 1 which describes how much of the variance between the two variables is described by the linear fit.

The Nash–Sutcliffe model efficiency coefficient is defined as (Nash & Sutcliffe 1970)

$$E = 1 - \frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{\sum_{t=1}^{n} (y_t - \overline{y_t})^2} \quad (17)$$

Nash–Sutcliffe efficiencies can range from $-\infty$ to 1. An efficiency of 1 ($E = 1$) corresponds to a perfect match of modeled precipitation to the observed data. An efficiency of 0 ($E = 0$) indicates that the model predictions are as accurate as the mean of the observed data, whereas an efficiency less than zero ($E < 0$) occurs when the observed mean is a better predictor than the model. Essentially, the closer the model efficiency is to 1, the more accurate the model is.

CASE STUDY

Aharchay river basin in the northwestern part of Iran is located between 47°20′ and 47°30′ east longitude and 38°20′ and 38°45′ north latitude as shown in Figure 3. The precipitations during wet (December to May) and dry (June to November) seasons are about 180 and 112 mm.

The mean annual precipitation, temperature, and inflow at the end of this basin are about 292 mm, 10 °C, and 51 MCM, respectively. About 62% of precipitation occurs in the wet season. The basin is 2,232 km^2 in area and contains the primary tributaries of the Aharchay river which is one of the most important rivers in the Azarbayjan province. Sattarkhan dam, constructed as a multi-purpose dam, is located on the Aharchay River to supply downstream water demands including domestic, industrial, agricultural, and environmental.

The monthly precipitation data of these stations are extracted from the data bank of Iran's Meteorological

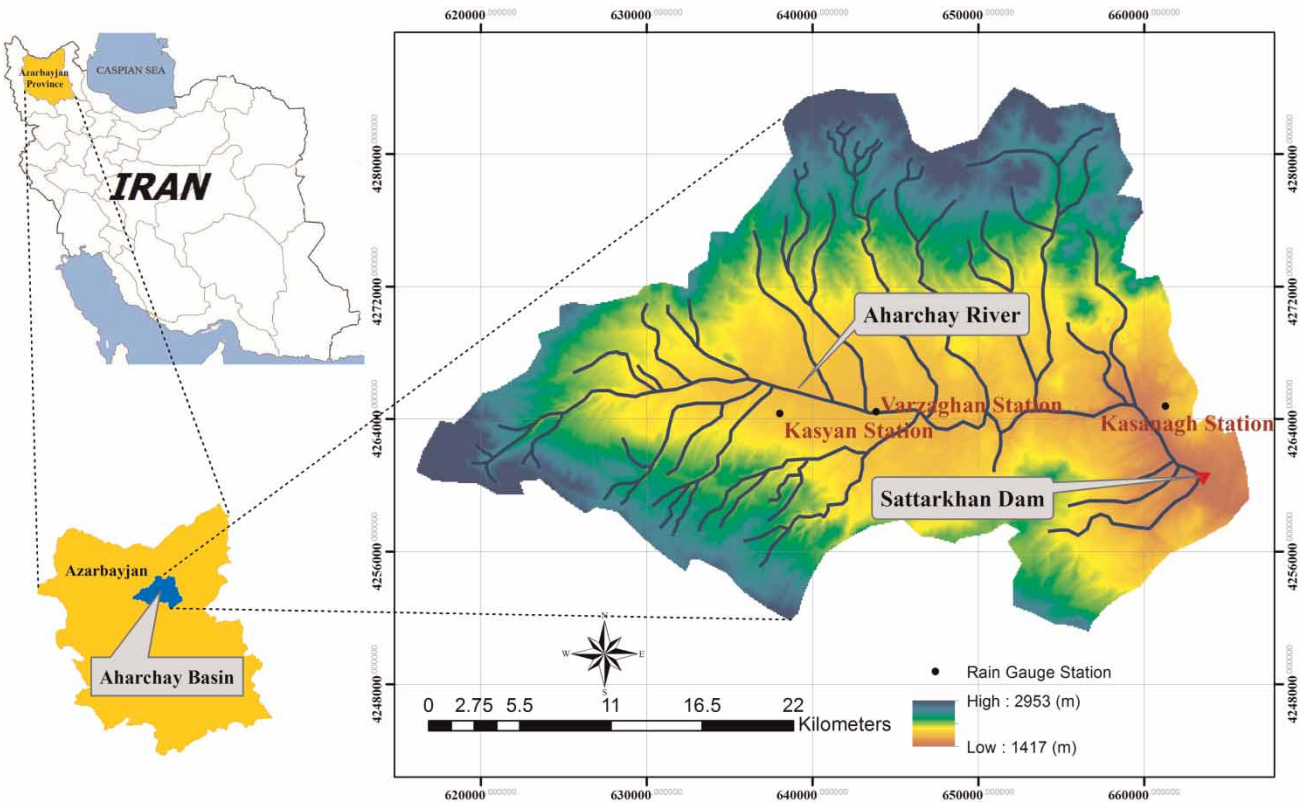

Figure 3 | Rivers and subbasins of the Aharchay river basin.

Organization. The precipitation in the basin has a complete record from 1970 to 2010. The predictors used in this study are the monthly SLP, difference in sea level pressure (DSLP), and sea surface temperature (SST) of the different points around the world which are estimated by the National Center for Atmospheric Research (NCAR). The data come from several sources and are available for free on the NCEP/NCAR internet site (http://dss.ucar.edu/pub/reanalysis/). In research by Karamouz *et al.* (2005), the effects of different points around the world on Iran's climate are carried out, as shown in Figure 4. In this paper, 15 points addressed by Karamouz *et al.* (2005) are considered as the predictors of precipitation of the northwestern part of Iran.

The months of the year are divided into two dry and wet seasons. The dry season is from June to November and the wet season from December to May. In order to use the results of the prediction model for water resources management in a basin, two operational seasons are considered for precipitation prediction. The correct lead time for water resources planning and water allocation is considered to be 6 months. The effective predictors on seasonal Aharchay precipitation are selected between the DSLP and SST of 15 climate variables, as presented in Table 1.

RESULTS

Selecting the appropriate predictors

Correlation coefficient analysis is traditionally used between the input and output variables to identify the effective climate signals. Consider, n is the number of variables, 2^n–1 cases should be examined for calculating transinformation values using the entropy method. Considering mathematical burden, the number of input variables is considered as six. Thus, the number of cases to examine with entropy is reduced to 63. The new GT is also utilized to select six variables among 15 predictor variables. The results of correlation coefficient analysis between the climate

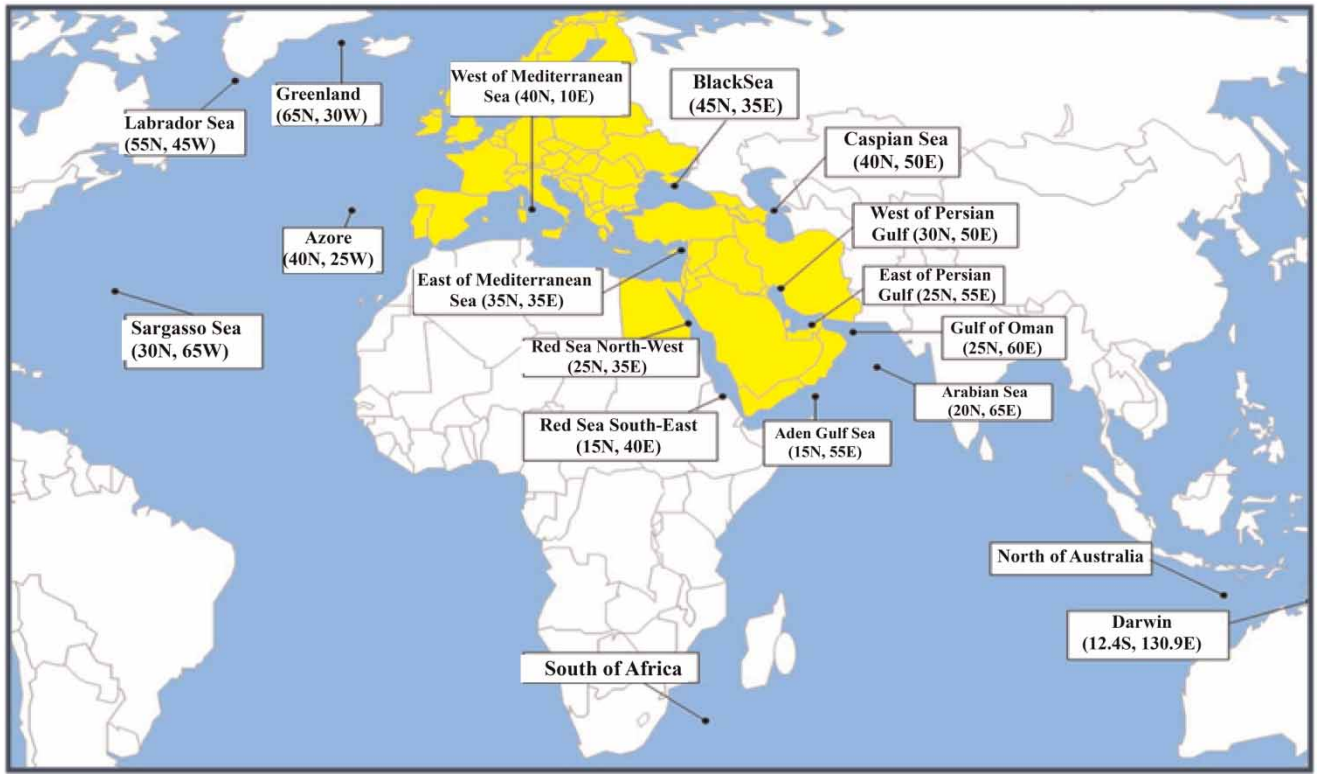

Figure 4 | The selected locations of the climate variables on Iran's climate.

Table 1 | The effective climate signals on climate of northwestern part of Iran

Signal	DSLP between	Signal	SST
1	Greenland and Azores	7	Caspian Sea
2	Greenland and West of Mediterranean Sea	8	Aden Sea
3	Greenland and East of Mediterranean Sea	9	Black Sea
4	Greenland and Black Sea	10	East of Mediterranean Sea
5	Siberia and Sudan	11	Indian Sea
6	Siberia and East of Persian Gulf	12	North West of Red Sea
		13	Arabian Sea
		14	West of Persian Gulf
		15	East of Persian Gulf

variables and the precipitation for the wet and dry seasons are shown in Figure 5.

As shown in this figure, during the wet season (December to May), variable 5 (DSLP between Siberia and Sudan), variable 6 (DSLP between Siberia and Eastern Persian Gulf), variable 9 (SST in the Black Sea), variable 14 (SST west of the Persian Gulf), variable 10 (SST at the east of the Mediterranean Sea), and variable 13 (SST in the Arabian Sea) have the most correlation with the Aharchay basin precipitation. Variable 3 (DSLP between south Greenland and east Mediterranean Sea), variable 2 (DSLP between south Greenland and west Mediterranean Sea), variable 1 (DSLP between southern Greenland and Azores), variable 4 (DSLP between southern Greenland and the Black Sea), variable 8 (SST in the Aden Sea), and variable 7 (SST in the Caspian Sea) are the most correlated variables with the area's precipitation in the dry season from June to November.

However, it is possible that there are correlations among some of these variables and one may be able to introduce one or more other variables. Multicollinearity occurs when two or more predictor variables in a multiple regression model are highly correlated. In this section, the multicollinearity among predictors in the models resulting from correlation coefficient analysis is explored. For this purpose, the possibility of multicollinearity is assessed by carrying out the correlation matrix for all the variables. The correlation value varies between −1 and +1. The correlation coefficient values are presented in Table 2 and show four pairs of variables having a coefficient value of more than 0.85, which can be classified as highly correlated. In this method, the predictor variables are selected based on the highly correlated with the precipitation and multicollinearity analysis.

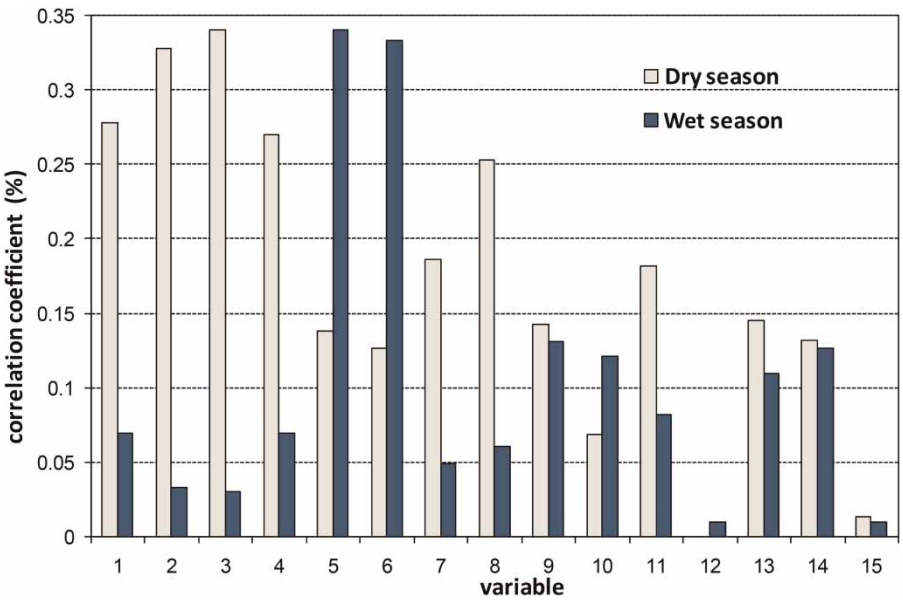

Figure 5 | The correlation coefficient between the climate signals and the precipitation of dry and wet seasons.

Table 2 | The correlation coefficient values between climate variables for wet season

Climate variables		Climate variables														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Climate variables	1	1.00	**0.88**	0.84	**0.85**	−0.21	−0.19	0.13	0.36	−0.14	0.11	0.28	0.11	0.24	0.08	0.09
	2	**0.88**	1.00	**0.91**	**0.91**	−0.27	−0.26	0.10	0.37	−0.12	0.16	0.24	0.19	0.27	0.17	0.18
	3	0.84	**0.91**	1.00	**0.92**	−0.39	−0.32	0.01	0.20	−0.17	0.12	0.02	−0.02	0.12	0.07	0.03
	4	0.85	0.91	**0.92**	1.00	−0.39	−0.35	0.13	0.28	−0.13	0.10	0.08	0.06	0.15	0.02	0.04
	5	−0.21	−0.27	−0.39	−0.39	1.00	**0.93**	0.24	−0.01	0.18	0.18	0.33	0.47	0.24	0.33	0.35
	6	−0.19	−0.26	−0.32	−0.35	**0.93**	1.00	0.24	−0.06	0.14	0.14	0.24	0.39	0.18	0.31	0.27
	7	0.13	0.10	0.01	0.13	0.24	0.24	1.00	0.16	0.70	0.61	0.40	0.52	0.25	0.66	0.62
	8	0.36	0.37	0.20	0.28	−0.01	−0.06	0.16	1.00	0.01	0.24	0.77	0.29	0.78	0.21	0.31
	9	−0.14	−0.12	−0.17	−0.13	0.18	0.14	0.70	0.01	1.00	0.70	0.28	0.49	0.06	0.44	0.36
	10	0.11	0.16	0.12	0.10	0.18	0.14	0.61	0.24	0.70	1.00	0.46	0.71	0.44	0.68	0.64
	11	0.28	0.24	0.02	0.08	0.33	0.24	0.40	0.77	0.28	0.46	1.00	0.56	0.73	0.48	0.53
	12	0.11	0.19	−0.02	0.06	0.47	0.39	0.52	0.29	0.49	0.71	0.56	1.00	0.47	0.61	0.69
	13	0.24	0.27	0.12	0.15	0.24	0.18	0.25	0.78	0.06	0.44	0.73	0.47	1.00	0.48	0.54
	14	0.08	0.17	0.07	0.02	0.33	0.31	0.66	0.21	0.44	0.68	0.48	0.61	0.48	1.00	**0.86**
	15	0.09	0.18	0.03	0.04	0.35	0.27	0.62	0.31	0.36	0.64	0.53	0.69	0.54	**0.86**	1.00

The values greater than 0.85 are shown in bold.

In the wet season from December to May, variable 6 is replaced by variable 11 (SST in the Indian Ocean) in the model input list, since a significant correlation was observed between variables 5 and 6. Therefore, variables 5, 9, 14, 10, 13, and 11 are selected as model inputs using the multicollinearity analysis.

In the dry season (June to November) considering the prediction model inputs covariance matrix, a significant correlation exists between variables 3 and 4 and variables 1 and 2; therefore variables 1, 2, and 4 are replaced by variable 13 (SST at the Arabian Sea), variable 9 (SST at the Black Sea), and variable 5 (DSLP between Siberia and Sudan) in the model inputs list. Therefore, variables 3, 8, 7, 13, 9, and 5 are selected as the prediction model inputs using the multicollinearity method.

The Gamma values between 15 climate signals and the precipitation in the dry and wet seasons are presented in Figure 6. In the wet season, signals with the lowest Gamma value in relation to the precipitation including variables 15, 13, 6, 5, 10, and 11 are selected.

According to the correlation coefficients (CC) between variables 5 and 6, variable 5 is replaced by variable 9 (SST in the Black Sea). Therefore, considering the correlation between the variables, variables 15, 13, 6, 10, 11, and 9 are selected as input variables for the second GT model. In order to evaluate the input variables selection methods, three SVM models have been developed that use 80% historical data at the training stage and 20% in the testing stage. The mean values of wet season precipitation (December to May) for training and testing sets are 194.3, 191.8 (mm) and their standard deviations are 71.0 and 63.9 (mm). The corresponding values of the dry season are 118.3, 109.8, 55.33, and 55.31 (mm), respectively. The results of the developed models are presented in Table 3.

It is observed that input selection using the GT along with replacing the dependent variables is the best model at the training stage since it bears the least average deviation error, square root error, and the most CC between the predicted and historical values. However at the testing stage, the input selection method through GT application without deletion of dependent variables has the best performance indicating the efficiency of the GT in the input variables selection for the prediction model.

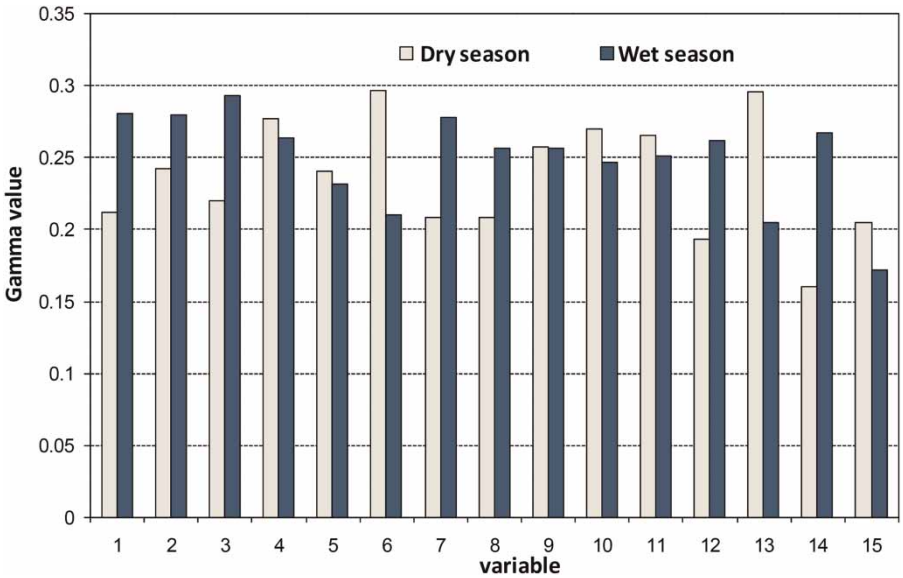

Figure 6 | Gamma values between the precipitation and the climate variables of dry and wet seasons.

Table 3 | Comparison of results obtained from the GT and the CC analysis

		Effective signals	Model	RMSE	R^2	E
Wet season	Training	5, 9, 14, 10, 13, 11	CC	0.22	0.83	0.65
		15, 13, 5, 6, 10, 11	GT 1	0.25	0.81	0.74
		15, 13, 6, 10, 11, 9	GT 2	0.21	0.84	0.77
	Testing	5, 9, 14, 10, 13, 11	CC	0.48	0.51	0.57
		15, 13, 5, 6, 10, 11	GT 1	0.45	0.55	0.60
		15, 13, 6, 10, 11, 9	GT 2	0.73	0.35	0.48
Dry season	Training	3, 8, 7, 13, 9, 5	CC	0.11	0.83	0.74
		14, 12, 15, 7, 8, 1	GT	0.29	0.81	0.73
	Testing	3, 8, 7, 13, 9, 5	CC	0.33	0.64	0.59
		14, 12, 15, 7, 8, 1	GT	0.45	0.51	0.55

Considering the Gamma value presented in Figure 6 for the dry season and the CC between variables, signals with the lowest Gamma value in relation to the precipitation including variables 14, 12, 15, 7, 8, and 1 are chosen. Since a significant correlation is not present between the proposed variables, these variables are selected as the prediction model inputs. In order to evaluate the GT and CC method in input selection, SVM models are developed. The results of their applications are presented in Table 3 and show that determining predictors using the CC method has better performance in both training and testing stages with the least prediction error.

Selecting the best combination of predictors

The aim of selecting the best combination of predictors is to identify and omit the signals whose existence increases the complexity of the model and that have no significant effects on the improvement of the results. In order to choose the best possible combination of predictors, the GT and the entropy method are utilized to find a model with minimum Gamma error and maximum transferred data, respectively. The number of different combinations created by the presence or absence of each one of the six selected signals as the prediction model inputs is $2^6-1=63$ states. In Table 4,

Table 4 | The GT and entropy theory results for the selected combinations of the prediction model inputs

		Combination mask	Gamma	Gradient	S.E.	V ratio	Transinformation
Wet season	1	100000	0.172	1.975	0.040	0.576	0.000
	2	010100	0.175	0.303	0.027	0.640	0.090
	3	001000	0.210	0.411	0.065	0.839	0.085
	4	111100	0.170	0.151	0.025	0.682	0.110
	5	101000	0.235	0.013	0.039	0.939	0.090
	6	111111	0.225	0.019	0.036	0.901	0.133
	7	010001	0.256	0.006	0.053	1.024	0.009
Dry season	1	101010	0.146	0.118	0.035	0.586	0.176
	2	101000	0.220	0.018	0.058	0.881	0.157
	3	101111	0.157	0.068	0.036	0.627	0.209
	4	101011	0.139	0.102	0.040	0.555	0.193
	5	100001	0.181	0.334	0.054	0.724	0.098
	6	111111	0.200	0.017	0.036	0.801	0.256
	7	011000	0.252	0.024	0.040	1.008	0.064

only seven states with the significant transinformation or low Gamma value are presented for the two seasons. The transinformation in this table is calculated based on the entropy theory.

As can be seen in Table 4, the combination of the first four inputs (111100) has the lowest Gamma value in the wet season and (101011) in the dry season. Therefore, the proposed scenarios by the GT are 111100 and 101011 for wet and dry seasons, respectively. According to the amounts of the transferred information present in the last column of Table 4, it can be observed that a combination of all variables (111111) is the scenario suggested by the entropy method in both dry and wet seasons. In the dry season, the Gamma number has the lowest value with the combination of four inputs (101011). The SVM model is developed for evaluating the selected combinations to study the performance of the utilized methods.

The results in Table 5 show that the model with the combination of the first four variables has the better performance in precipitation simulation in the wet season. Therefore, signals affecting precipitation include variables 15, 13, 6, and 5 in the wet season (December to May) and variables 3, 7, 9, and 5 in the dry season (June to November) and these are selected as model inputs, respectively.

Table 5 | Simulation results for different input combinations

		Model	RMSE	R^2	E
Wet season	Training	100000	0.33	0.61	0.38
		010100	0.39	0.53	0.42
		111111	0.25	0.81	0.74
		111100	0.31	0.75	0.68
	Testing	100000	0.51	0.14	0.35
		010100	0.58	0.31	0.40
		111111	0.45	0.55	0.60
		111100	0.4	0.70	0.65
Dry season	Training	101010	0.07	0.68	0.60
		111111	0.11	0.83	0.74
		101111	0.12	0.79	0.72
		101011	0.13	0.78	0.68
	Testing	101010	0.31	0.53	0.56
		111111	0.33	0.64	0.59
		101111	0.43	0.38	0.38
		101011	0.27	0.75	0.62

Table 6 | Modeling results of different methods in the wet season

		Model	RMSE	R^2	E
Wet season	Training	Naïve	0.44	0.56	0.43
		Trend	0.7	0.52	0.28
		Regr	0.43	0.66	0.46
		SVM	0.31	0.75	0.68
	Testing	Naïve	0.47	0.43	0.33
		Trend	0.73	0.35	0.21
		Regr	0.5	0.49	0.35
		SVM	0.4	0.7	0.65
Dry season	Training	Naïve	0.61	0.26	0.18
		Trend	1.12	0.24	0.22
		Regr	0.2	0.59	0.34
		SVM	0.13	0.78	0.68
	Testing	Naïve	0.74	0.05	0.15
		Trend	1.32	0.08	0.16
		Regr	0.46	0.46	0.31
		SVM	0.27	0.75	0.62

Selecting the best model

The SVM technique is used for precipitation prediction in wet and dry periods. The results obtained from SVM in the wet and dry seasons are compared with the basic models, such as the naïve, the trend, and the multivariable regression models as shown in Table 6. The data are divided into training and testing data by the ratios of 80% and 20%, respectively.

The training phase of the learning machine involves adjusting the parameters considering a training sample of 32 patterns in four-dimensional space ($N = 32$ and $n = 4$ in Figure 2). The input vector in the SVM model includes variables 15, 13, 6, and 5 in the wet season (December to May) and variables 3, 7, 9, and 5 in the dry season (June to November), respectively. Seasonal precipitation (predictand) constitutes the output from the model. In SVM modeling processes, after carrying out the sensitivity analysis, the υ–SVR model with the kernel RBF function is developed.

By checking the results of the testing stage, the overfitting problem is controlled. The naïve model is a model in which it is supposed that the next period prediction is like

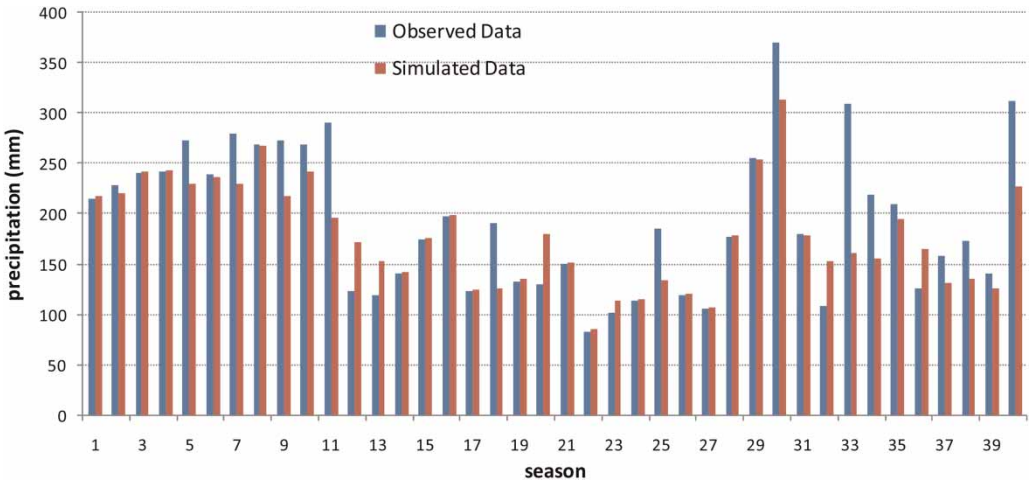

Figure 7 | Seasonal predicted precipitation and observed precipitation for wet seasons.

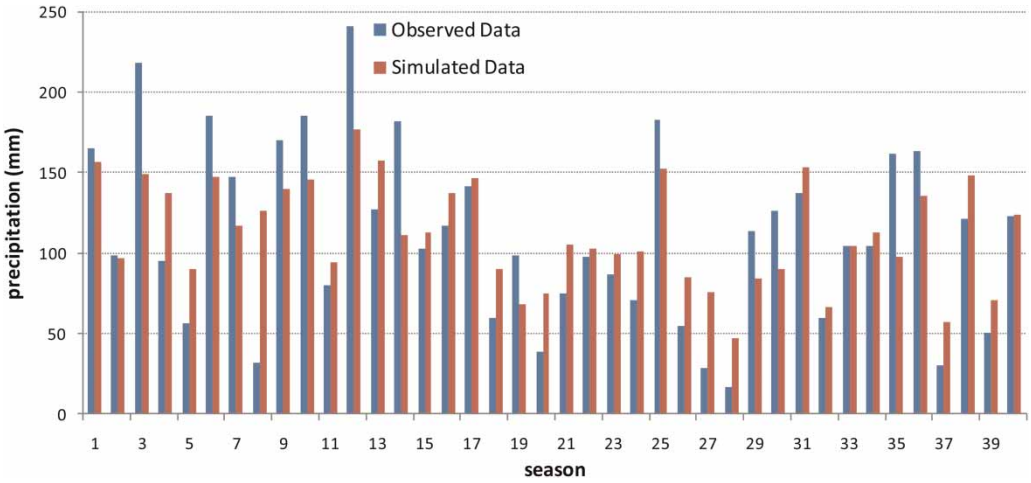

Figure 8 | Seasonal predicted and observed precipitation for dry seasons.

the current period. In the trend model, it is supposed that the next period prediction is based on the linear trend of the two previous periods. As can be gathered from the results presented in Table 6, in both wet and dry seasons, the SVM model gives better results than the naïve and trend models at the testing stage. Also, this model has less modeling error compared to a multivariable regression model which indicates better performance in nonlinear modeling. Figures 7 and 8 show the seasonal predicted precipitation and observed precipitation for wet and dry periods.

SUMMARY AND CONCLUSION

In this study, the relationships between variations of SST, SLP, and DSLP of some certain points and the precipitation of Aharchay basin in two wet and dry seasons are explored. The GT and correlation coefficient analysis are used to select more effective variables between the climatic signals. The results of the SVM model for evaluating the methods show better performance of GT in input selection. In the second part of this paper, two techniques, GT and entropy are used for the best combination selection. The results show the entropy method selects the best model with more input variables which may be the best model in the training stage without this guarantee at the testing stage. The GT selects the model with the best inputs combination which has the better performance in comparison with the entropy method. In the third part of the paper, the SVM model is used for precipitation prediction and its performance is compared with the results of precipitation modeling utilizing the naïve, trend, and multivariable regression models as the benchmark models. The results show better performance of the SVM model at the testing stage.

REFERENCES

Ahmadi, A. & Han, D. 2013 Identification of dominant sources of sea level pressure for precipitation forecasting over Wales. *J. Hydroinform.* **15** (3), 1002–1021.

Ahmadi, A., Han, D., Karamouz, M. & Remesan, R. 2009 Input data selection for solar radiation estimation. *Hydrol. Process.* **23** (19), 2754–2764.

Alfonso, L., Lobbrecht, A. & Price, R. 2010 Optimization of water level monitoring network in polder systems using information theory. *Water Resour. Res.* **46**, W12553, 13.

Araghinejad, S. & Burn, D. H. 2005 Probabilistic forecasting of hydrological events using geostatistical analysis. *Hydrol. Sci. J.* **50** (5), 838–856.

Araghinejad, S., Burn, D. H. & Karamouz, M. 2006 Long-lead probabilistic forecasting of streamflow using ocean-atmospheric and hydrological predictors. *Water Resour. Res.* **42**, W03431, 11.

Bray, M. & Han, D. 2004 Identification of support vector machines for runoff modeling. *J. Hydroinform.* **6** (4), 265–280.

Busuico, A., Giorgi, F., Bi, X. & Ionita, M. 2006 Comparison of regional climate model and statistical downscaling simulation of different winter precipitation change scenarios over Romania. *Theor. Appl. Climatol.* **86**, 101–123.

Caselton, W. F. & Husain, T. 1980 Hydrologic networks: information transinformation. *J. Water Res. Plan. Manage.* **106**, 503–529.

Cavazos, T. & Hewitson, B. C. 2005 Performance of NCEP-NCAR reanalysis variables in statistical downscaling of daily precipitation. *Clim. Res.* **28**, 95–107.

Choy, K. Y. & Chan, C. W. 2003 Modelling of river discharges and rainfall using radial basis function networks based on support vector regression. *Int. J. Syst. Sci.* **34** (14–15), 763–773.

Chuzhanova, N. A., Jones, A. J. & Margetts, S. 1998 Feature selection for genetic sequence classification. *Bioinformatics* **14** (2), 139–143.

Conway, H., Gades, A. & Raymond, C. F. 1996 Albedo of dirty snow during conditions of melt. *Water Resour. Res.* **32** (6), 1713–1718.

Cover, T. M. & Thomas, J. A. 1991 Elements of information theory (D. L. Schilling, ed.) *Wiley Series in Telecommunications*, John Wiley & Sons, New York.

Crawford, T., Betts, N. L. & Favis-Mortlock, D. 2007 GCM grid-box choice and predictor selection associated with statistical downscaling of daily precipitation over Northern Ireland. *Clim. Res.* **34** (2), 145.

Dibike, Y. B., Velickov, S., Solomatine, D. & Abbott, M. B. 2001 Model induction with support vector machines: introduction and applications. *J. Comput. Civil Engng.* **15** (3), 208–216.

Evans, D. 2002 Data Derived Estimates of Noise Using Near Neighbour Asymptotics. PhD Thesis, Department of Computer Science, University of Cardiff, UK.

Evans, D. & Jones, A. J. 2002 A proof of the gamma test. *Proc. Roy. Soc. Lond. Series A* **458** (2027), 2759–2799.

Ghosh, S. & Mujumdar, P. P. 2008 Statistical downscaling of GCM simulations to streamflow using relevance vector machine. *Adv. Water Res.* **31**, 132–146.

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

Hanssen-Bauer, I., Forland, E. J., Haugen, J. E. & Tveito, O. E. 2003 Temperature and precipitation scenarios for Norway:

Comparison of results from dynamical and empirical downscaling. *Clim. Res.* **25**, 15–27.

Harmancioglu, N. B. & Alpaslan, N. 1992 Water quality monitoring network design a problem of multi-objective decision making. *Water Res. Bull.* **28**, 179–192.

Harmancioglu, N. B. & Singh, V. P. 1998 Entropy in environmental and water resources. In: *Encyclopedia of Hydrology and Water Resources* (R. W. Herschy & R. W. Fairbridge, eds). Kluwer Academic Publishers, Boston, MA, pp. 225–241.

Hashmi, M. Z., Shamseldin, A. Y. & Melville, B. W. 2009 Statistical downscaling of precipitation: State-of-the-art and application of Bayesian multi-model approach for uncertainty assessment. *Hydrol. Earth Syst. Sci. Discuss.* **6**, 6535–6579.

Hashmi, M. Z., Shamseldin, A. Y. & Melville, B. W. 2011 Comparison of SDSM and LARS-WG for simulation and downscaling of extreme precipitation events in a watershed. *Stoch. Environ. Res. Risk A* **25** (4), 475–484.

Haylock, M. R., Peterson, T. C., Alves, L. M., Ambrizzi, T., Anunciação, Y. M. T., Baez, J., Barros, V. R., Berlato, M. A., Bidegain, M., Coronel, G., Corradi, V., Garcia, V. J., Grimm, A. M., Karoly, D., Marengo, J. A., Marino, M. B., Moncunill, D. F., Nechet, D., Quintana, J., Rebello, E., Rusticucci, M., Santos, J. L., Trebejo, I. & Vincent, L. A. 2006 Trends in total and extreme south American rainfall in 1960–2000 and links with sea surface temperature. *J. Climate* **19**, 1490–1512.

Hertig, E. & Jacobeit, J. 2008 Assessments of Mediterranean precipitation changes for the 21st century using statistical downscaling techniques. *Int. J. Climatol.* **28** (8), 1025–1045.

Husain, T. 1989 Hydrologic uncertainty measure and network design. *Water Resour. Bull.* **25**, 527–534.

Jaafar, W. Z. W. & Han, D. 2011 Calibration catchment selection for flood regionalisation modeling. *J. Am. Water Resour. Assoc.* **48** (4), 698–706.

Johansson, B. & Chen, D. 2003 The influence of wind and topography on precipitation distribution in Sweden: Statistical analysis and modeling. *Int. J. Climatol.* **23**, 1523–1535.

Karamouz, M., Zahraie, B., Fatahi, E., Mirzaie, E., Remezani, F. & Hashemi, R. 2005 Predictors for long-lead precipitation forecasting in western Iran. The first Iran-Korea Joint Workshop on Climate Modelling, November 16–17, Mashahd, Iran.

Karamouz, M., Ahmadi, A. & Moridi, A. 2009 Probabilistic reservoir operation using Bayesian stochastic model and support vector machine. *Adv. Water Res.* **32**, 1588–1600.

Končar, N. 1997 Optimisation Methodologies for Direct Inverse Neurocontrol. PhD Thesis, Department of Computing, Imperial College of Science, Technology and Medicine, University of London, London, UK.

Krstanovic, P. F. & Singh, V. P. 1992 Evaluation of rainfall networks using entropy: I. Theorical development. *Water Resour. Manage.* **6** (4), 279–293.

Liong, S. Y. & Sivapragasam, C. 2002 Flood stage forecasting with support vector machines. *J. Am. Water Res. Assoc.* **38**, 173–186. doi: 10.1111/j.1752-1688.2002.tb01544.

Moghaddamnia, A., Ghafari, M., Piri, J. & Han, D. 2009 Evaporation estimation using support vector machines technique. *Int. J. Math.* **3** (3), 134–142.

Mogheir, Y. & Singh, V. P. 2002 Application of information theory to groundwater quality monitoring networks. *Water Resour. Manage.* **16** (1), 37–49.

Moradkhani, H. & Meier, M. 2010 Long-lead water supply forecast using large-scale climate predictors and independent component analysis. *J. Hydrol. Eng.* **15** (10), 744–762.

Mpelasoka, F., Mullan, A. B. & Heerdegen, R. G. 2001 New Zealand climate change information drived by multivariate statistical and artificial neural network approaches. *Int. J. Climatol.* **21**, 1415–1433.

Najafi, M., Moradkhani, H. & Wherry, S. 2011 Statistical downscaling of precipitation using machine learning with optimal predictor selection. *J. Hydrol. Eng.* **16** (8), 650–664.

Nash, J. E. & Sutcliffe, J. V. 1970 River flow forecasting through conceptual models, Part 1- A discussion of principles. *J. Hydrol.* **10** (3), 282–290.

Nourani, V. & Parhizkar, M. 2013 Conjunction of SOM-based feature extraction method and hybrid wavelet-ANN approach for rainfall-runoff modeling. *J. Hydroinform.* **15** (3), 829–848.

Nourani, V. & Sayyah Frad, M. 2012 Sensitivity analysis of the artificial neural network outputs in simulation of the evaporation process at different climatologic regimes. *Adv. Eng. Softw.* **47** (1), 127–146.

Piri, J., Amin, S., Moghaddamnia, A., Keshavarz, A., Han, D. & Remesan, R. 2009 Daily pan evaporation modelling in a hot and dry climate. *J. Hydrol. Eng.* **14** (8), 803–811.

Remesan, R., Shamim, M. A. & Han, D. 2008 Model input data selection using gamma test for daily solar radiation estimation. *Hydrol. Process.* **22**, 4301–4309.

Schmidli, J., Goodess, C. M., Feri, C., Haylock, M. R., Hundecha, Y., Ribalaygua, J. & Schmith, T. 2007 Statistical and dynamical downscaling of precipitation: an evaluation and comparison of scenarios for the European Alps. *J. Geophys. Res.* **112**, D04105.

Schoof, J. T. & Pryor, S. C. 2001 Downscaling temperature and precipitation: a comparison of regression-based methods and artificial neural networks. *Int. J. Climatol.* **21**, 773–790.

Shannon, C. E. & Weaver, W. 1949 *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, IL.

Singh, V. P. 1997 The use of entropy in hydrology and water resources. *Hydrol. Process.* **11**, 587–626.

Sivapragasam, C. & Liong, S. Y. 2005 Flow categorization model for improving forecasting. *Nord. Hydrol.* **36** (1), 37–48.

Stefansson, A., Koncar, N. & Jones, A. J. 1997 A note on the gamma test. *Neural Comput. Appl.* **5**, 131–133.

Tripathi, S., Srinivas, V. V. & Nanjundiah, R. S. 2006 Downscaling of precipitation for climate change scenarios: A support vector machine approach. *J. Hydrol.* **330** (3–4), 621–640.

Vapnik, V. 1998 *Statistical Learning Theory*. Wiley, New York.

Vapnik, V. N. 1992 Principles of risk minimization for learning theory. In: *Advances in Neural Information Processing Systems* (J. Moody, S. Hanson & R. Lippmann, eds). Elsevier, New York, p. 4.

Vapnik, V. N. 2010 *The Nature of Statistical Learning Theory*, 2nd edn. Springer, New York.

Widmann, M., Bretherton, C. S. & Salathé, E. P. 2003 Statistical precipitation downscaling over the northwestern United States using numerically simulated precipitation as a predictor. *J. Climate* **16** (5), 799–816.

Yu, X. Y., Liong, S. Y. & Babovic, V. 2004 EC-SVM approach for realtime hydrologic forecasting. *J. Hydroinform.* **6** (3), 209–223.

First received 5 December 2013; accepted in revised form 6 July 2014. Available online 11 August 2014