

Assessment of climate change impacts on rainfall using large scale climate variables and downscaling models – A case study

AZADEH AHMADI^{1,*}, ALI MORIDI², ELHAM KAKAEI LAFDANI³ and GHASEM KIANPISHEH⁴

¹*Department of Civil Engineering, Isfahan University of Technology, Isfahan, Iran.*

²*Abbaspour College of Technology, Shahid Beheshti University, Tehran, Iran.*

³*Department of Watershed Management Engineering, Faculty of Natural Resources and Marine Sciences, Tarbiat Modares University, Noor, Iran.*

⁴*Department of Civil Engineering, Isfahan University of Technology, Isfahan, Iran.*

**Corresponding author. e-mail: aahmadi@cc.iut.ac.ir*

Many of the applied techniques in water resources management can be directly or indirectly influenced by hydro-climatology predictions. In recent decades, utilizing the large scale climate variables as predictors of hydrological phenomena and downscaling numerical weather ensemble forecasts has revolutionized the long-lead predictions. In this study, two types of rainfall prediction models are developed to predict the rainfall of the Zayandehrood dam basin located in the central part of Iran. The first seasonal model is based on large scale climate signals data around the world. In order to determine the inputs of the seasonal rainfall prediction model, the correlation coefficient analysis and the new Gamma Test (GT) method are utilized. Comparison of modelling results shows that the Gamma test method improves the Nash–Sutcliffe efficiency coefficient of modelling performance as 8% and 10% for dry and wet seasons, respectively. In this study, Support Vector Machine (SVM) model for predicting rainfall in the region has been used and its results are compared with the benchmark models such as K-nearest neighbours (KNN) and Artificial Neural Network (ANN). The results show better performance of the SVM model at testing stage. In the second model, statistical downscaling model (SDSM) as a popular downscaling tool has been used. In this model, using the outputs from GCM, the rainfall of Zayandehrood dam is projected under two climate change scenarios. Most effective variables have been identified among 26 predictor variables. Comparison of the results of the two models shows that the developed SVM model has lesser errors in monthly rainfall estimation. The results show that the rainfall in the future wet periods are more than historical values and it is lower than historical values in the dry periods. The highest monthly uncertainty of future rainfall occurs in March and the lowest in July.

1. Introduction

The rainfall prediction in the next few months or even seasons has many benefits to decision makers of the basin for water allocation to different sectors and proper reservoir operation. Development of the

prediction models leads to more dynamic and flexible decisions about the reservoir operation including water storing/releasing and improving the performance of reservoir operation policies. In recent decades, much effort has been devoted for developing mid-to long-term (monthly and seasonal) rainfall

Keywords. Rainfall prediction; support vector machines; K-nearest neighbourhood; artificial neural network; Gamma test; climate change.

prediction models. There are various methods to develop the relationship between the large scale climatic variables such as geopotential height and mean sea level pressure (predictors) and the local variables (predictands) such as temperature and precipitation. The most widely used models usually implement the general climate model (GCMs) outputs as predictors to predict the rainfall. But because the GCM data are in coarse resolution, the downscaling techniques are employed to change climate model outputs into metrological variables appropriate for hydrologic applications.

Much efforts have been devoted for development of models that are a combination of statistical and conceptual approaches. There have been advances in rainfall prediction by implementing linear methods such as, simple/multiple linear regression (Hanssen-Bauer *et al.* 2003; Johansson and Chen 2003; Dutta *et al.* 2006), canonical correlation analysis (CCA) (Chen and Chen 2003; Xoplaky *et al.* 2004) or singular value decomposition (Conway *et al.* 1996; Widmann and Bretherton 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 artificial neural networks (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; Haylock *et al.* 2006; Bae *et al.* 2007; Nourani *et al.* 2009; Najafi *et al.* 2011). Other downscaling techniques including support vector machine (SVM) (Ghosh and Mujumdar 2008; Najafi *et al.* 2011), K-nearest neighbour Model (KNN) (Araghinejad *et al.* 2006), and Genetic Programming (GP) (Hashmi *et al.* 2011) are also utilized.

Some other techniques used for data preprocessing to reduce the dimensionality of the problem include principle component analysis (e.g., Schoof and Pryor 2001; Araghinejad and Burn 2005), fuzzy clustering (Ghosh and Mujumdar 2008), and Gamma test (GT) (Ahmadi *et al.* 2009; Moghaddamnia *et al.* 2009).

One of the statistical models that is very popular in GCM downscaling named SDSM (statistical downscaling model) is developed by Wilby *et al.* (1999, 2002). SDSM is a hybrid between a multilinear regression method and a stochastic weather generator. The model has been applied in many catchments around the world to predict the daily hydrologic variables and to assess climate change impacts (e.g., Wilby and Dettinger 2000). Haylock *et al.* (2006) compared six statistical models and two dynamic downscale models to predict the seasonal heavy rainfall in northwest and southeast

England. The results showed that the models based on nonlinear ANN were found to be the best at modelling the interannual variability of the indices. Liu and Coulibaly (2011) compared the SDSM and a time lagged feed forward neural network (TLFN), and evolutionary polynomial regression (EPR) in a region in northeastern Canada. The results are more efficient downscaling techniques than SDSM for both the ensemble daily precipitation and temperature. Most of the cases where the SDSM is applied, are located in the region with high rainfall and SDSM is rarely used for rainfall prediction in arid and semi-arid regions.

In this study, three nonlinear methods including ANN, KNN, and SVM are used to find the relationship between the large-scale climate variables provided by the National Centres for Environmental Prediction (NCEP) and the rainfall of the Zayandehrood dam basin. Zayandehrood river is the most important river in the central part of Iran. Climate change may exacerbate the already contentious water supply situation in the basin. In order to select the model inputs, two methods including the correlation coefficient analysis and Gamma test (GT) have been employed and their performances in the model inputs selection are assessed. The results of simulation models are compared and the best one is introduced. Then the statistical downscaling model (SDSM) is utilized for daily rainfall prediction. The results of nonlinear models have been compared to the results of the SDSM in monthly rainfall prediction. Finally, climate change impacts on the Zayandehrood dam basin's rainfall in the future are assessed.

The main contributions of this study are utilizing Gamma test in input selection for developing more accurate downscaling models, comparing monthly and daily downscaling models, and evaluating climate change impacts on the case study. This paper evaluates the impacts of climate change on rainfall of Zayandehrood river basin. The paper is organized as follows: the used models including ANN, KNN, SVM, GT, and SDSM are briefly described in section 2. A case study for implementing the proposed methodology is presented in section 3. The obtained results are presented in section 4.

2. Materials and methods

In this study, firstly the signals that affect the basin's climate are identified out of the effective signals on Iran's climate carried out by Karamouz *et al.* (2005). The relevancies between the case study's rainfall with the climate signals are identified during a period of six months. The statistical period is classified into the wet season (December–May) and the dry season (June–November). The

steps of applying the proposed methodology in this study are presented in figure 1. As shown in this figure, after collecting and categorizing data into two seasons, in order to achieve more precise predictions and reduce the number of input variables between some effective signals, the correlation coefficient analysis and the new Gamma test method are applied. The results of two methods in selecting the effective variables are compared through the assessment criteria.

Then the prediction models are developed using data-driven models including the ANN, KNN, and SVM methods and the results from different models are compared. Also, SDSM is used for predicting

the daily rainfall in the study basin. Finally, the results of monthly rainfall obtained from monthly data driven models and the daily statistical downscaling model are compared.

2.1 Gamma test

The Gamma test is used to examine the relationship between inputs and outputs in numerical datasets without a need to construct the prediction model. It is used for estimating the variance of the output before modelling, even though the model is unknown. This error

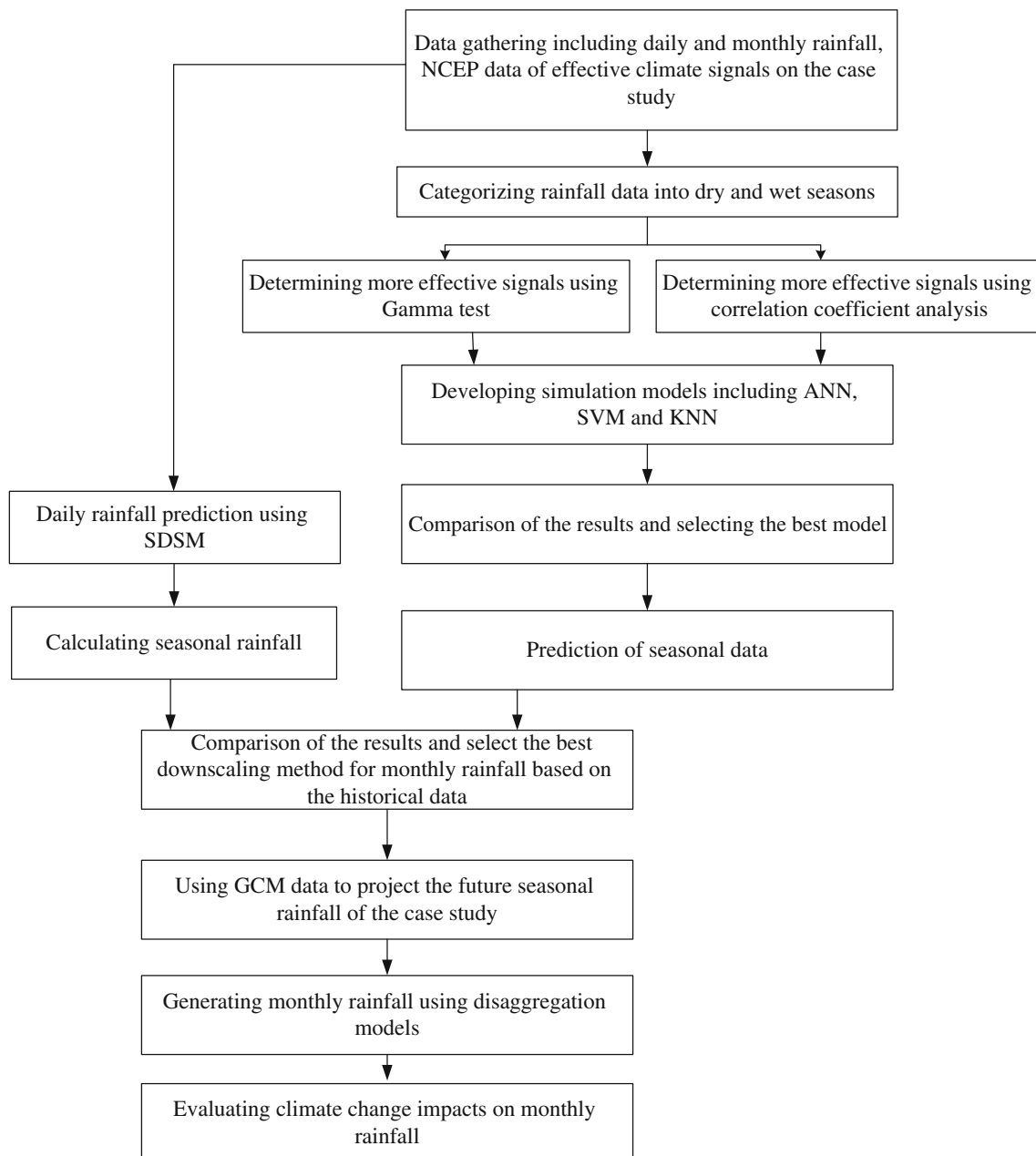


Figure 1. Flowchart of the study.

variance estimate presents a target mean-squared error that any smooth nonlinear function should attain on unseen data. Suppose we have a set of observed data represented by:

$$((x_1, \dots, x_M), y) = (x, y) \quad (1)$$

where the vector $X = (x_1, \dots, x_M)$ is the *input*, confined to a closed bounded set $C \in R^M$ and the scalar y is the corresponding *output*, without loss of generality. The only assumption made is that the relationship of the system is in the following form:

$$y = f(x_1, \dots, x_M) + r \quad (2)$$

where f represents a smooth function and r denotes an indeterminable part, which may be due to real noise lack of functional determination in the assumed input/output relationship. The Gamma test is used to return a data-derived estimate for $\text{Var}(r)$ without knowing the underlying function f , just directly from the data. The estimate of the model's output variance called the Gamma statistic and represented by Γ cannot be accounted for by a smooth data model. The Gamma test is derived from the Delta function of the input vectors:

$$\delta_M(k) = \frac{1}{M} \sum_{i=1}^M |x_{N[i,k]} - x_i|^2 \quad (3)$$

where $x_{N[i,k]}$ denotes the index of the k th nearest neighbour to x_i , and $|\cdot|$ denotes Euclidean distance. Thus $\delta_M(k)$ is the mean square distance to the k th nearest neighbour. The corresponding Gamma function of the output values is:

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

The Gamma test computes the mean-squared k th nearest neighbour distances $\delta(k)$, ($1 \leq k \leq k_{\text{Max}}$) and the corresponding $\gamma(p)^2$. In order to compute Γ the best line is constructed for the p points $(\delta_M(k), \gamma_M(k))$, and the vertical intercept, Γ is returned as the gamma value. The regression line slope is also returned to show the complexity of the model f . The V_{ratio} is the standardized results by considering $\Gamma/\text{Var}(y)$. It returns a scale invariant noise estimate which normally lies between zero and one.

2.2 Support Vector Machines (SVM)

The fundamental of Support Vector Machine (SVM) has been developed by Vapnik (1995, 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 interesting features and promising empirical performance (Choy and Chan 2003; Bray and Han 2004; Yu et al. 2004; Sivapragasam and Liong 2005; Karamouz et al. 2009).

The SVM model is produced by support vectors included in the training data and presents the means of 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 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.

$$y = Wx + b. \quad (5)$$

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

$$\sum_{i=1}^l (y_i - Wx_i - b)^2. \quad (6)$$

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

$$(y_i - Wx_i \pm b) \leq \varepsilon. \quad (7)$$

A band or a tube around the hypothesis function y can be visualized that 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.

$$\text{Min} \frac{1}{2} w^2 + C \sum_{i=1}^l (\xi_i + \xi_i^*) \quad (8)$$

Subject to:

$$\begin{aligned} y_i - (W \cdot x_i + b) &\leq \varepsilon + \xi_i \\ (W \cdot x_i + b) - y_i &\leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* &\geq 0, \quad i = 1, 2, \dots, l \end{aligned} \quad (9)$$

where C is a prespecified and positive constant that determines the degree of penalized loss when a training error occurs, ξ_i and ξ_i^* are slack variables representing 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.

2.3 Artificial Neural Network (ANN)

Artificial neural network is a mathematical structure including the network topology and pattern of interconnected assembly of simple processing nodes and the transformed functions. The unique structure of the ANN and utilizing the nonlinear transfer function corresponding to each hidden and output node have made ANNs a powerful tool to develop the nonlinear relationships without *a priori* assumption. Multilayer perceptrons (MLPs) are the most widely used network architecture. The historical data is used to train the network. The input data flows from the input layer to the output layer. In this study, the GCM predictors are the inputs to the nodes in input layer and the back propagation algorithm is used to train the network.

2.4 *K*-nearest neighbourhood (KNN)

The KNN model is a nonparametric statistical pattern recognition procedure without any assumption of the theoretical or analytical relation between the inputs and the outputs. KNN is used for various hydrological modelling by Karlsson and Yakowitz (1987); Galeati (1990); Kember and Flower (1993); Toddini (2000); Karamouz *et al.* (2010). This model returns the *K*, most similar patterns between the historical data in the time series. A feature vector of past records summarizes the past history in a smaller-dimension vector of observations including most of the information relevant to the forecast.

The distance of each feature vector between the present time (*t*) and historical data (*j* = 1, 2, ..., number of observed data), z_{tj}^d , is calculated based on the Euclidean norm for *d*-dimensional vector as follows:

$$\|z_{tj}^d\| = \left(\sum_{i=1}^d w_i (p_t^i - p_j^i)^2 \right)^{1/2} \quad (10)$$

where w_i is the weight of each predictor and p_t^i is the *i*th predictor at time *t*. The *K* vectors of observations are identified that have the minimum distances. Then the forecast instant *t* is calculated as a weighted average of the *K* nearest neighbours. In

this study, the kernel functions used for estimating the unknown data (R_t):

$$R_t = \frac{\sum_{j=1}^K R_j/j}{\sum_{j=1}^K 1/j} \quad (11)$$

where *j* is the order of the nearest neighbours and R_j is the value of the neighbour *j*. The best values for weights and *K* value are optimized during calibration process using cross-validation.

2.5 Statistical downscaling Model (SDSM)

The SDSM is a multiple linear regression based tool for constructing future ensemble to evaluate the impact of climate change (Wilby *et al.* 2002). It is capable of projecting the unseen ensembles at daily time scale using grid resolution GCM output. Based on the correlation and the partial correlation analysis, and also utilizing scatter plots, and the physical sensitivity between predictors and predictants, the most relevant predictors are selected. Table 1 shows the list of the predictors with their short names as used in SDSM. For downscaling future climate scenarios four sets of GCM output are available: HadCM2, HadCM3, CGCM2, and CSIRO.

This model is assembled and calibrated using the NCEP re-analysis large scale predictors from

Table 1. NCEP reanalysis predictors in the SDSM.

	Predictor	Description
d1	Ncepmslpaz	Mean sea level pressure
d2	Ncepp5_faz	500 hPa airflow strength
d3	Ncepp5_uaz	500 hPa zonal velocity
d4	Ncepp5_vaz	500 hPa meridional velocity
d5	Ncepp5_zaz	500 hPa vorticity
d6	Ncepp5thaz	500 hPa wind direction
d7	Ncepp5zhaz	500 hPa divergence
d8	Ncepp8_faz	850 hPa airflow strength
d9	Ncepp8_uaz	850 hPa zonal velocity
d10	Ncepp8_vaz	850 hPa meridional velocity
d11	Ncepp8_zaz	850 hPa vorticity
d12	Ncepp8thaz	850 hPa wind direction
d13	Ncepp8zhaz	850 hPa divergence
d14	Ncepp500az	500 hPa geopotential height
d15	Ncepp850az	850 hPa geopotential height
d16	Ncepp_faz	Surface airflow strength
d17	Ncepp_uaz	Surface zonal velocity
d18	Ncepp_vaz	Surface meridional velocity
d19	Ncepp_zaz	Surface velocity
d20	Ncepp_thaz	Surface wind direction
d21	Ncepp_zhaz	Surface divergence
d22	Ncepr500az	Relative humidity at 500 hPa
d23	Ncepr850az	Relative humidity at 850 hPa
d24	Nceprhumaz	Near surface relative humidity
d25	Nceprhumaz	Surface specific humidity
d26	Nceptempaz	Mean temperature at 2 m

1961–1991. It regulates the average and variance of downscaled daily rainfall considering factors such as bias, and variance to conform simulated values to observed one. HadCM3 (Hadley Centre Coupled Model, version 3) scenario A2 and B2 data available from year 1961 to 2099, are used for assessing climate change impacts. In this paper, HadCM3 GCM output is used and other sets of GCM output could be applied. In this paper, NCEP and HADCM3 data for BOX 15X and 22Y are utilized.

2.6 Model evaluation

The criteria of mean bias error (MBE), root mean square error (RSME), mean absolute error (MAE) and Nash–Sutcliffe coefficient are used to evaluate the performance of simulation modelling of the historical rainfall. The following formulae are used to calculate them.

$$\text{MBE} = \frac{\sum_{t=1}^n (\hat{y}_t - y_t)}{n} \quad (12)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^n (y_t - \hat{y}_t)^2}{n}} \quad (13)$$

$$\text{MAE} = \frac{\sum_{t=1}^n |\hat{y}_t - y_t|}{n} \quad (14)$$

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

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

$$E = 1 - \frac{\sum_{t=1}^n (y_t - \hat{y}_t)^2}{\sum_{t=1}^n (y_t - \bar{y}_t)^2} \quad (15)$$

where \bar{y}_t is the mean value of observed precipitation values. Nash–Sutcliffe efficiencies can range from $-\infty$ to 1. An efficiency of 1 ($E = 1$) corresponds to a perfect match of modelled 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.

3. Study area and dataset

The study area is the Zayandehrood dam basin in central Iran. It is located within $31^{\circ}15'–33^{\circ}45'$ latitude and $50^{\circ}02'–52^{\circ}45'$ longitude. The basin is 4200 km^2 in area and contains the primary tributaries of the Zayandehrood river. Zayandehrood

river is the most important river in the central part of Iran, which has the semi-arid climate.

The Zayandehrood river basin supplies water demands of about 5 million people of many cities and villages. This area is very important because of Isfahan metropolis, Gavkhunuy wetland ecosystem, about 20% of major national industries, and 8% of national agricultural productions exists in this basin.

The Zayandehrood dam is constructed to cater to 1500 MCM water demands of Isfahan province including domestic, industrial, and agricultural sectors. Therefore, the rainfall prediction for the next month can help decision makers of the basin for water allocation to different sectors and proper reservoir operation. Figure 2 shows the location of the Zayandehrood dam basin. The daily rainfall data of the basin from 1956–2008 is extracted from the data bank of Iran's Meteorological Organization.

In this study, two models are used for long-lead rainfall prediction. The predictors used in the first model are the monthly sea-level pressure (SLP), difference in sea level pressure (DSLPP), and sea surface temperature (SST) of certain points around the world estimated for the National Science Foundation by the National Centre for Atmospheric Research (NCAR). The data is available on the NCEP/NCAR website (<http://dss.ucar.edu/pub/reanalysis/>). In the research by Karamouz *et al.* (2005), the effective points around the world on the Iran's climate, shown in figure 3, are carried out. In this paper, the 16 climate signals addressed by Karamouz *et al.* (2005) are considered as the predictors of the prediction model of Zayandehrood dam basin's rainfall.

4. Results

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 is from December to May. The effective predictors on the seasonal Zayandehrood rainfall are selected between the SLP, DSLPP, and SST of 16 climate variables as presented in table 2. In this stage, the number of predictors is considered to be six.

4.1 Selecting the appropriate predictors

In order to determine the appropriate predictors and select six variables among 16 prediction variables, the correlation coefficient analysis and the new Gamma test method are used. The results of the correlation coefficient analysis show that the effective rainfall predictors of the dry season are the SLP at west of Persian Gulf, SLP at northwest

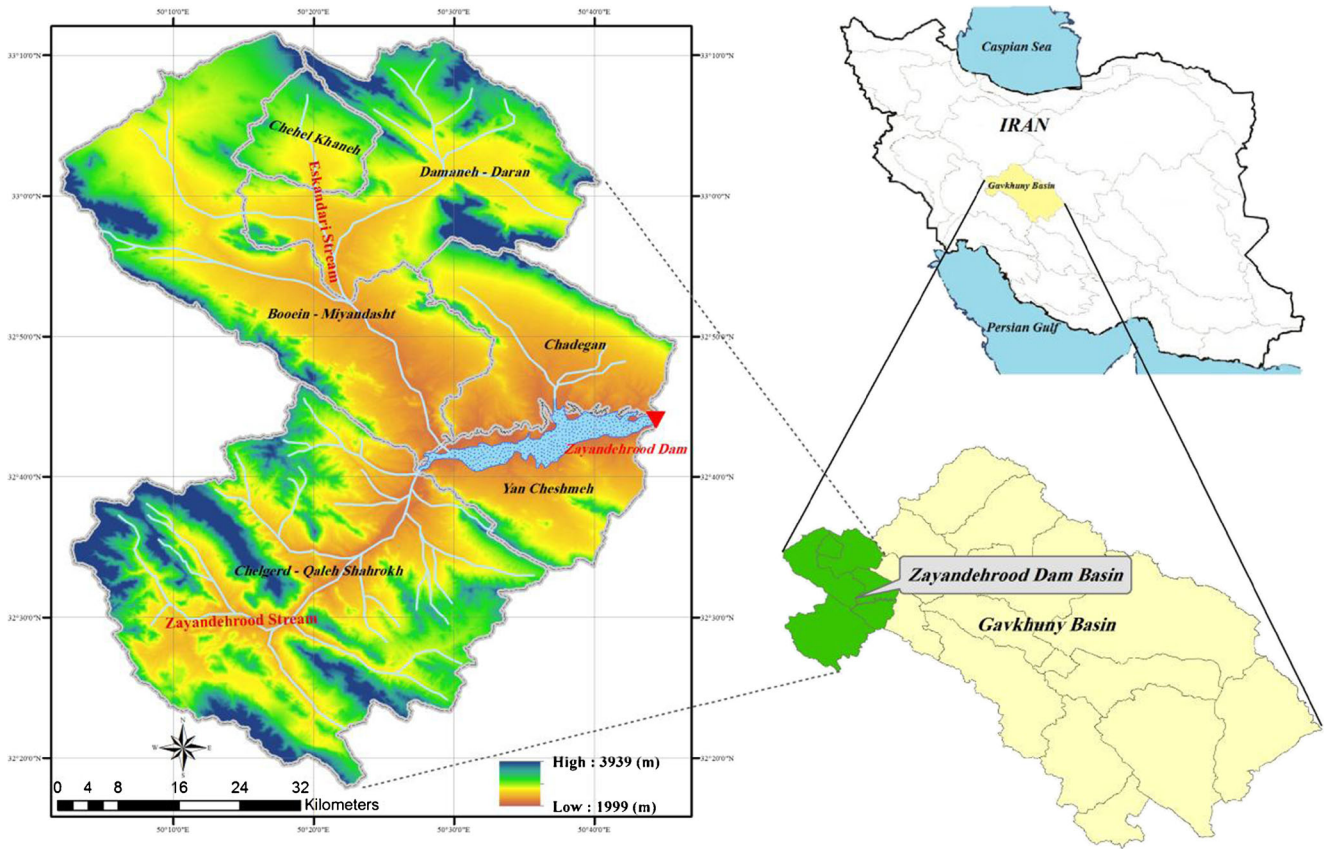


Figure 2. Location of Zayandehrood dam basin.

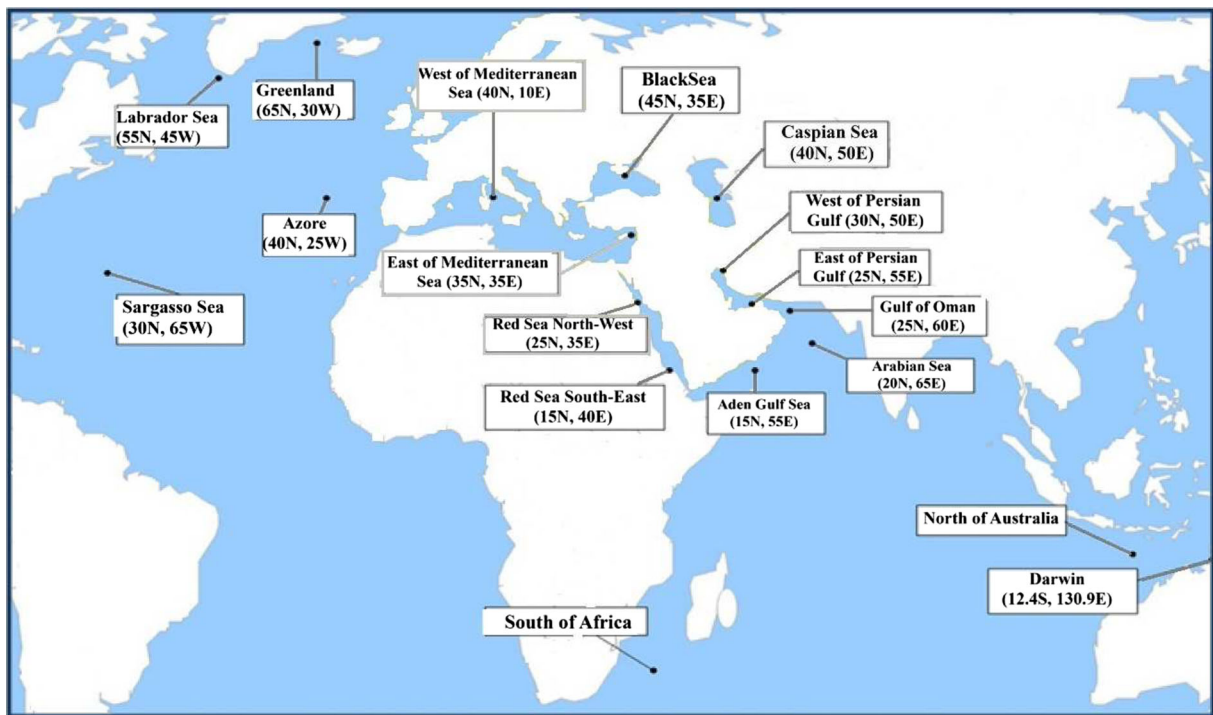


Figure 3. Selected locations for quantifying the effects of climate variables.

of Red Sea, DSLP between Indian Ocean and west of Persian Gulf, DSLP between Indian Ocean and Soudan, DSLP between Indian Ocean and southeast of Red Sea, and DSLP between Indian Ocean and northwest of Red Sea. Also, the results of the correlation coefficient analysis for the wet season show that the effective predictors are the SLP at west of Persian Gulf, SLP at Soudan, SLP at Caspian Sea, DSLP between west of Persian Gulf and west of Mediterranean Sea, DSLP between west and east of Persian Gulf, and DSLP between west of Persian Gulf and Black Sea.

In the Gamma test method, climate signals with the lowest Gamma value are selected as the effective signals. The results of the Gamma test method of the 16 signals with the rainfall are presented in figure 4. According to figure 4, signals with the lowest Gamma values in the dry season are: the SLP at west of Persian Gulf, SLP at northwest of Red sea, DSLP between Indian Ocean and west of Persian Gulf, DSLP between Indian Ocean and southeast of Red sea, DSLP between Indian Ocean and east of Mediterranean Sea, and DSLP between Indian Ocean and northeast of Red sea. Also the Gamma values of the rainfall with the climate variables in the study area in wet season are selected based on the lowest Gamma values including SLP at west of Persian Gulf, SLP at Caspian Sea, DSLP between west of Persian Gulf and west of Mediterranean Sea, DSLP between west and east

of Persian Gulf, DSLP between Caspian Sea and Greenland, and DSLP between Caspian Sea and Black Sea. The results show that sea surface

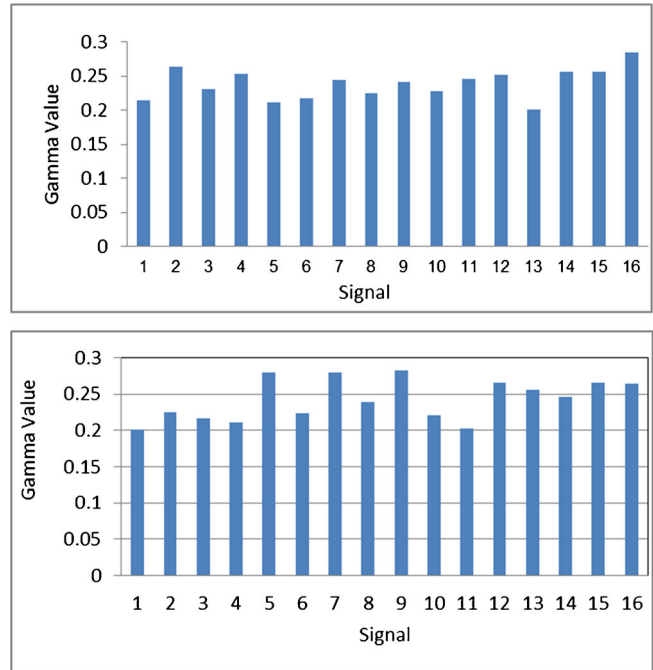


Figure 4. The Gamma values between the rainfall and the climate signals: (a) dry season and (b) wet season.

Table 2. The effective climate variables on central part of Iran for dry and wet seasons.

Dry season (June–November)	Wet season (December–May)
1. SLP at west of Persian Gulf	1. SLP at west of Persian Gulf
2. SLP at southeast of Red Sea	2. SLP at Soudan
3. SLP at east of Persian Gulf	3. SLP at Caspian Sea
4. SLP at Arabian Sea	4. DSLP between west of Persian Gulf and west of Mediterranean Sea
5. SLP at northwest of Red Sea	5. DSLP between west of Persian Gulf and Greenland
6. DSLP between Indian Ocean and west of Persian Gulf	6. DSLP between west of Persian Gulf and east of Persian Gulf
7. DSLP between Indian Ocean and Soudan	7. DSLP between west of Persian Gulf and Black Sea
8. DSLP between Indian Ocean and southeast of Red Sea	8. DSLP between west of Persian Gulf and southeast of Red Sea
9. DSLP between Indian Ocean and Oman Sea	9. DSLP between Caspian Sea and west of Mediterranean Sea
10. DSLP between Indian Ocean and east of Mediterranean Sea	10. DSLP between Caspian Sea and Greenland
11. DSLP between Indian Ocean and east of Arabian Sea	11. DSLP between Caspian Sea and Black Sea
12. DSLP between Indian Ocean and northwest of Red Sea	12. DSLP between Caspian Sea and Arabian Sea
13. DSLP between Indian Ocean and northeast of Red Sea	13. DSLP between west of Mediterranean Sea and Soudan
14. SST at Caspian Sea	14. SST at Greenland
15. SST at east of Mediterranean Sea	15. SST at Azores
16. SST at Greenland	16. SST at Arabian Sea

temperatures (SST) at different locations are not selected as effective inputs by Gamma test and correlation coefficient analysis.

4.2 Selecting the best prediction model

The SVM model is used to predict the rainfall in the study area using the identified effective signals. Therefore, 42 years (1956–1997) and 11 years (1998–2008) are considered for the training

and testing stages, respectively. In order to evaluate the efficiency of the SVM model, the results from the SVM model are compared with the results obtained from the ANN and the KNN models. In ANN modelling, the network consists of three layers containing input, hidden, and output layers with 6, 3, and 1 neurons, respectively. In the SVM modelling, the ν -SVR model and RBF Kernel function are used. The results of different rainfall prediction methods are presented in tables 3 and 4 for both

Table 3. The results of the rainfall prediction models based on the correlation analysis for the dry season.

Model	Training dataset			Testing dataset		
	KNN	ANN	SVM	KNN	ANN	SVM
Bias (MBE)	0.287	0.419	-0.821	0.620	0.745	0.587
RMSE	38.2	44.94	46.37	59.85	67.1	53.3
MAE	18.38	23.96	32.7	52.8	64.7	49.9
E	0.88	0.74	0.66	0.55	0.52	0.71

Table 4. The results of the rainfall prediction models based on Gamma test for the dry season.

Model	Training dataset			Testing dataset		
	KNN	ANN	SVM	KNN	ANN	SVM
Bias (MBE)	-0.205	-0.103	-0.542	0.592	0.731	0.490
RMSE	43.06	23.30	36.52	48.69	53.1	40.30
MAE	28.29	9.70	21.49	42.30	47.5	35.36
E	0.67	0.95	0.83	0.64	0.54	0.79

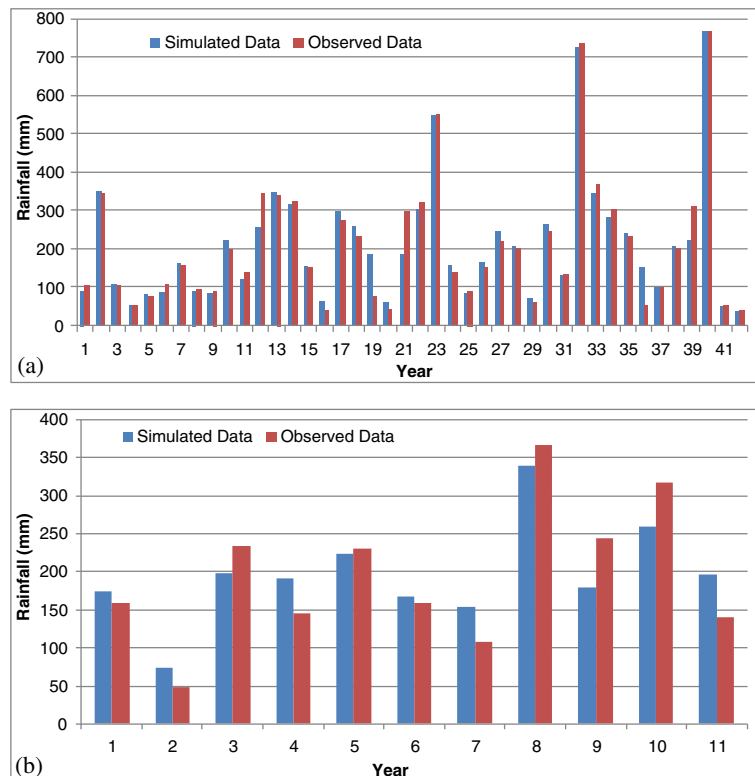


Figure 5. Comparison of simulated and observed rainfall using the SVM model during the dry season for (a) training dataset and (b) testing dataset based on the Gamma test.

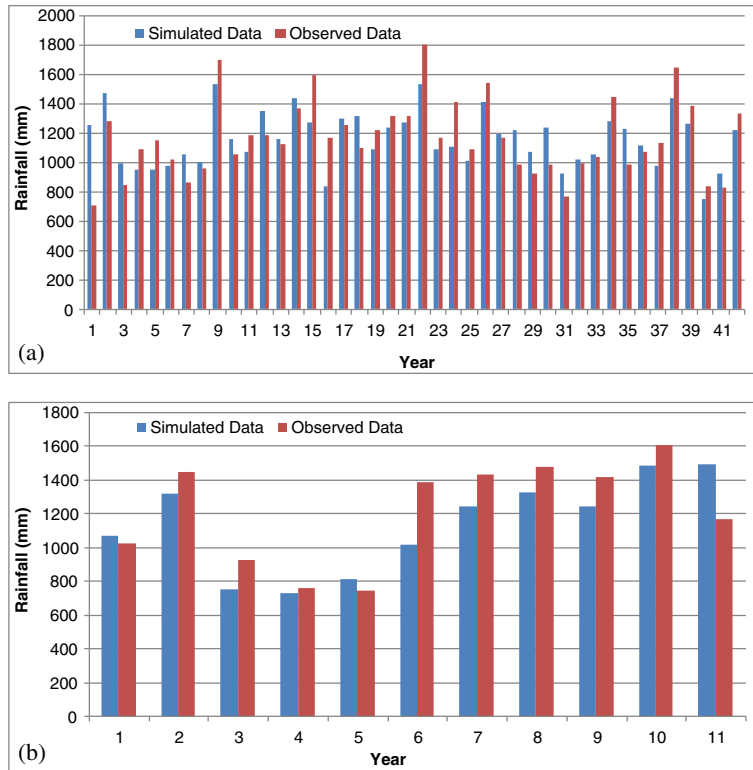


Figure 6. Comparison of simulated and observed rainfall using the SVM model during the wet season for (a) training dataset and (b) testing dataset based on the Gamma test.

Table 5. The results of the rainfall prediction models based on the correlation analysis for the wet season.

Model	Training dataset			Testing dataset		
	KNN	ANN	SVM	KNN	ANN	SVM
Bias (MBE)	-6.78	-5.41	-7.96	23.37	21.7	-14.90
RMSE	93.06	78.30	106.52	219.69	203.1	140.30
MAE	88.29	79.70	92.49	162.30	147.5	135.36
E	0.67	0.75	0.63	0.40	0.34	0.49

Table 6. The results of the rainfall prediction models based on Gamma test for the wet season.

Model	Training dataset			Testing dataset		
	KNN	ANN	SVM	KNN	ANN	SVM
Bias (MBE)	-5.02	3.21	-7.21	16.7	-19.9	-12.1
RMSE	82.37	67.97	92.05	142.4	196.7	117.3
MAE	70.28	69.85	89.80	109.8	122.4	93.2
E	0.75	0.84	0.69	0.5	0.38	0.59

training and testing stages. The simulated and observed rainfall using the SVM model for training dataset and testing dataset based on Gamma test are presented for dry and wet seasons in figures 5 and 6, respectively.

The results show that the ANN model has the least simulation errors in the rainfall prediction during the six months of dry season in the training stage compared with KNN and SVM models. However, in the testing stage, the SVM models, the

KNN models, and the ANN models have the least simulation errors, respectively. Also the results presented in tables 5 and 6 show that the best models at the testing stage are the ANN and the KNN models, respectively. In the wet season, the SVM model has the better performance in the rainfall prediction at the testing stage. Tables 3-6 show that inputs selection using the Gamma test method leads to better performance in the rainfall prediction.

4.3 The results of the SDSM Model

In this study, the Statistical Downscaling Model (SDSM) is used for the daily rainfall prediction. In SDSM, the most effective predictors between 26 variables utilized in SDSM data bank are selected based on their maximum correlation coefficient of their combination with the daily rainfall and minimum P-value between the predictors and rainfall. In this model, the weather generator downscales observed NCEP (National Centre for Environmental Prediction) reanalysis predictors. The downscaling model has been calibrated during 1961–1991 and the results have been validated during 1992–2001.

The selected predictors include the mean sea level pressure, surface zonal velocity, 500 hPa geopotential height, near surface relative humidity, and mean temporal at 2 m. Figures 7 and 8 show the histograms of the monthly values of the observation and modelling results for the calibration and validation periods. Table 7 shows the correlation

and the P-values between climate variables and the rainfall values. In order to compare the developed SVM and SDSM models, their performances in simulating seasonal rainfall of 10 years are presented in table 8. The results show the better performance of SVM model in seasonal rainfall prediction.

4.4 Assessment of climate change impacts

In order to investigate the climate change impacts on Zayandehrood dam basin's rainfall in the future using SVM model, the data of large-scale climate signals under climate change scenarios available from year 2000 to 2099 are received as model inputs. The HadCM3 model outputs are taken from the Hadley Climate Research and Prediction Centre in England and the data of A2 and B2 scenarios were taken from the DCC division through the IPCC website. In these scenarios, greenhouse gas emission and economic and social

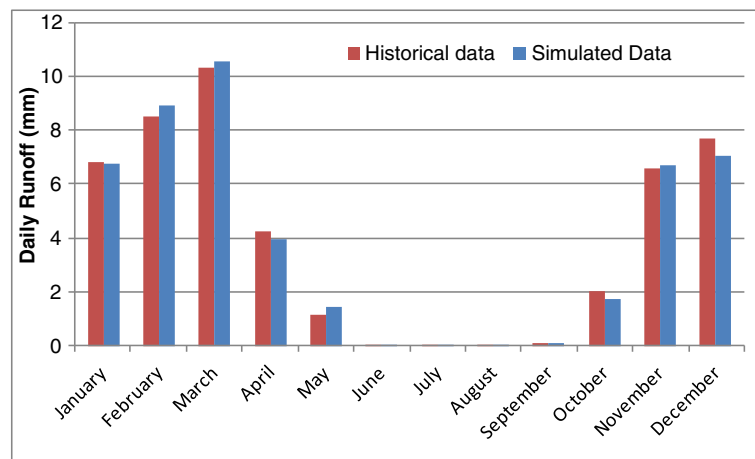


Figure 7. The average monthly observed and simulated rainfall for calibration period (1961–1991) using SDSM.

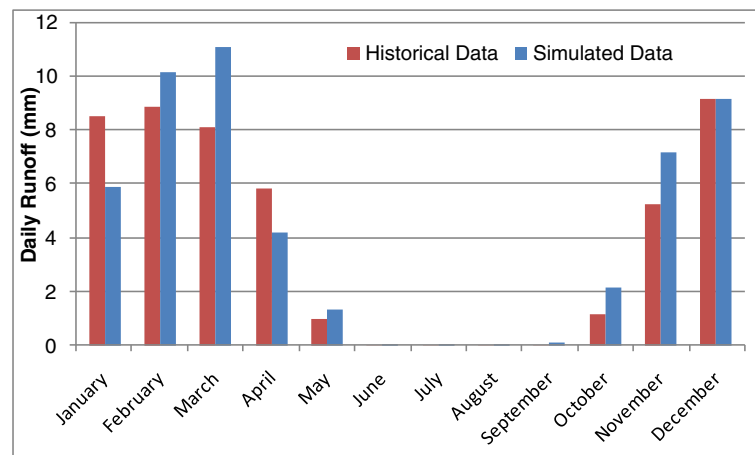


Figure 8. The average monthly observed and simulated rainfall for verification period (1992–2001) using SDSM.

Table 7. The correlation coefficients between climate variables and the rainfall.

Aerologic variables	Description	R^2	P_v (P value)	P_r (Partial r)
Ncepmslpz	Mean sea level pressure	0.225	0.005	0.017
Ncepp_uaz	Surface zonal velocity	0.159	0.024	0.019
Ncepp500az	500 hPa geopotential height	-0.219	0.021	0.033
Nceprhumaz	Near surface relative humidity	0.220	0.005	0.043
Nceptempaz	Mean temperature at 2 m	-0.249	0.022	0.002

Table 8. The performance criteria for SVM and SDSM (1992–2001).

Model	Dry season		Wet season	
	SDSM	SVM	SDSM	SVM
Bias (MBE)	4.32	0.49	19.05	-12.1
RMSE	86.46	40.30	204.009	117.3
MAE	118.06	35.36	158.9	93.2
E	0.48	0.79	0.64	0.59

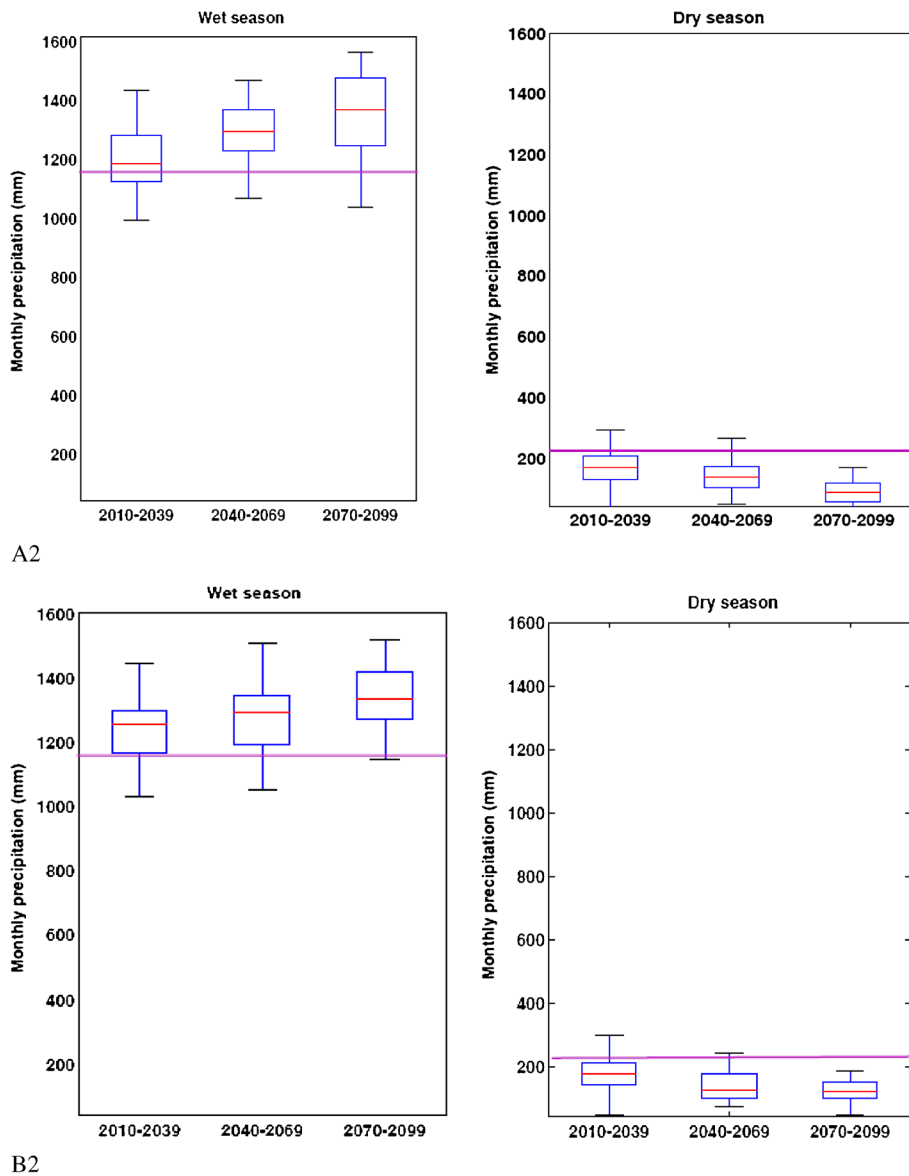


Figure 9. Uncertainty range of seasonal downscaled model under A2 and B2 scenarios with mean observed value shown by straight solid line.

development effects on the future climate are considered. The A2 scenario describes a very heterogeneous region. The underlying theme is self reliance and preservation of local identities. Fertility patterns across regions converge very slowly, which results in a continuously increasing population. Economic development is primarily regionally oriented and per capita economic growth changes more fragmented and slower than other scenarios. The B2 scenario describes a region in which the emphasis is on local solutions for economic, social, and environmental sustainability. The calibrated models were then used to downscale all the GCM data from 1961–2099 including B2 and A2 scenarios.

The seasonal mean rainfall over the periods of 2010–2039, 2040–2069, and 2070–2099 are shown as box plots in figure 9. As shown in this figure, the mean values increase in wet season and decrease in dry season rather than the historical values. The lengths of the box plots show the uncertainties of different years in each time period. The observed mean values for the period of 1956–2008 for dry and wet seasons are shown as the straight line. The

results show that wet season prediction has more uncertainty for all periods whereas dry season has lower uncertainties. The period of 2070–2099 is the most uncertain period. In wet season the uncertainty of scenario A2 is more than scenario B2 in the period of 2070–2099 and it is *vice versa* in dry season.

4.5 The statistical disaggregation model

In order to preserve the statistical properties of a time series at more than one level (spatial/temporal), the disaggregation models are used. Disaggregation facilitates the use of long-term persistence levels, particularly in multisites, multiseasons time series modelling. Salas *et al.* (1980) classified disaggregation models into temporal and spatial types. In this paper, the seasonal prediction results are disaggregated in the sub-series of monthly data. To this purpose, the basic disaggregation model proposed is used. The linear dependence model is as follows:

$$Y = AX + B\varepsilon \tag{16}$$

Table 9. Matrix B in the statistical disaggregation model for wet and dry seasons.

J/I	December–May						June–November					
	1	2	3	4	5	6	1	2	3	4	5	6
1	35.1	0	0	0	0	0	0.3	0	0	0	0	0
2	3.2	38.2	0	0	0	0	0.0	0.2	0	0	0	0
3	-7.6	-1.6	37.2	0	0	0	0.0	0.0	0.3	0	0	0
4	-20.7	-25.7	-18.0	42.7	0	0	-0.1	-0.2	-0.2	0.5	0	0
5	1.4	-3.9	-11.0	-17.0	18.3	0	-0.4	-1.0	-4.5	-1.1	11.9	0
6	-2.5	-7.7	-4.2	-11.0	-3.1	15.6	0.1	0.9	4.3	0.5	-11.9	4.0

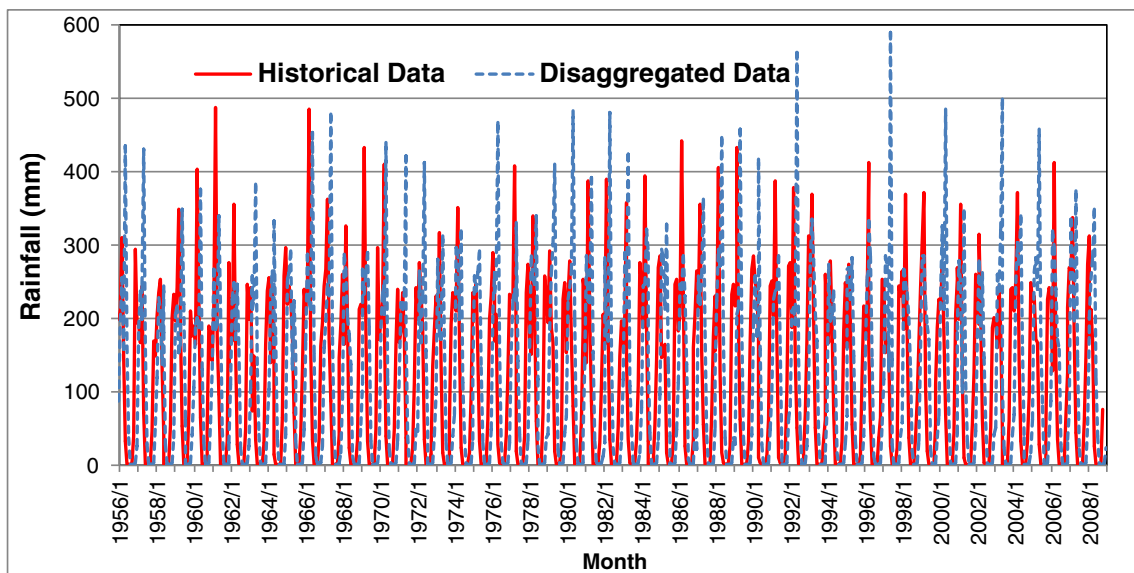


Figure 10. The predicted monthly rainfall.

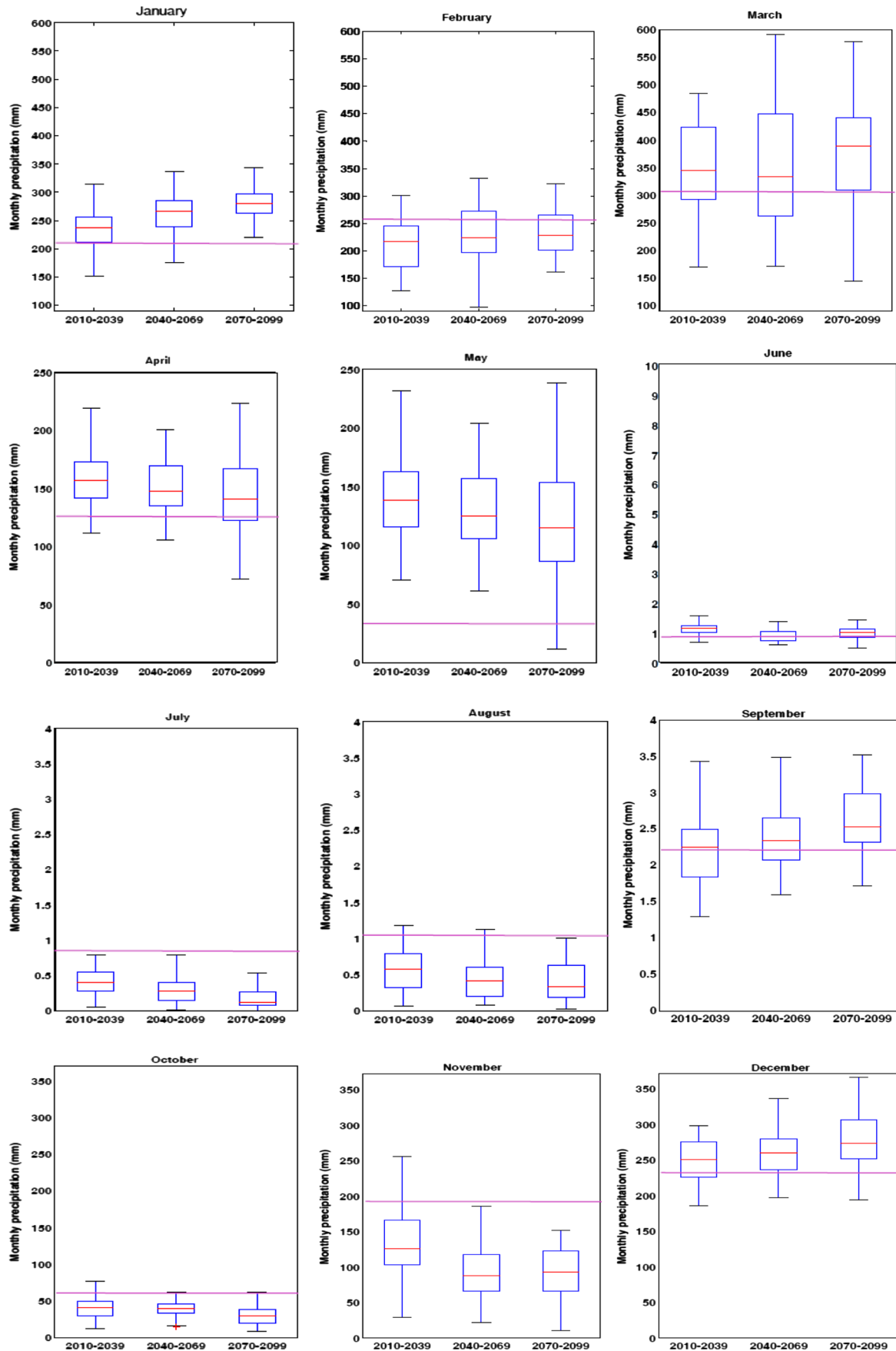


Figure 11. Uncertainty range of monthly downscaled model under scenario B2 with mean observed value shown by straight solid line.

$$\hat{A} = S_{YX}S_{XX}^{-1} \quad (17)$$

$$\hat{B}\hat{B}^T = S_{YY} - \hat{A} \cdot S_{XY} \quad (18)$$

where X_t is the seasonal rainfall at period t , Y_t is the dependent monthly rainfall of the series being generated, ε is the independent random variable, A , B are the parameters matrix, and S_{XY} is the covariance matrix between series.

According to the Zayandehrood basin historical data, components of matrix A for the wet season for December–May are 0.19, 0.18, 0.17, 0.27, 0.15, and 0.04, respectively. These components are 0.0, 0.01, 0.01, 0.02, 0.19, and 0.77 for the period June–November. Components of matrix B are presented in table 9. Utilizing parameter matrices A and B , the monthly rainfall time series are generated based on the predicted seasonal time series. The disaggregated monthly rainfall is shown in figure 10. The correlation coefficient between the observed and simulated for the historical time series is about 68%. The monthly prediction with six months lead time could result in more flexibility in water resources operation.

Therefore, the future GCM data was used to project the monthly and seasonal data utilizing the selected predictors along with the SVM and disaggregation models and associated calibrated parameters. The results of the monthly downscaling under B2 scenario over the periods of 2010–2039, 2040–2069, and 2070–2099 as box plots are shown in figure 11 along with the observed data. The analysis of the length of the box plots show that the highest monthly uncertainty occurs in March for all periods, and the lowest is in July.

As seen in this figure, the mean values increase in wet season and decrease in dry season rather than the historical values. It can be deduced that during rainfall downscaling in wet months, models tend to diverge and increase the uncertainty. It can be concluded that climate change in this area in the beginning of the next century may increase in the rainfall in wet season and it may decrease in dry season.

5. Summary and conclusion

In this study, the effects of large scale climate variables have been studied on the rainfall in the Zayandehrood dam basin area located in the central plateau of Iran. For this purpose, the effective climate signals on Iran's climate are considered. In order to select six signals, two methods including correlation coefficient analysis and Gamma test method are used based on high correlation coefficient and low Gamma test between the climate signals and the rainfall, respectively. The performances of two methods in input selection are

compared through nonlinear model applications with the suggestive signals. The results show the better performance of Gamma test method in climate signal selection.

Then the results of three models including ANN, KNN, and SVM are compared in rainfall simulation. The results show the better performance of ANN model in training stage and SVM model in testing stage. The daily rainfall prediction is also carried out utilizing SDSM model. The effective climate signals are the mean sea level pressure, surface zonal velocity, 500 hPa geopotential height, near surface relative humidity, and mean temporal at 2 m.

The monthly rainfall prediction using SVM and SDSM models are compared. The results show that SVM model has the better performance in monthly rainfall prediction. Then the data of large-scale climate signals under climate change scenarios are used to project the future wet and dry seasons under A2 and B2 scenarios.

The seasonal results show that the mean rainfall increases in future wet season and decreases in future dry season rather than the historical values. Also the wet season prediction has more uncertainty for all periods whereas dry season has lower uncertainties. In wet season, the uncertainty of scenario A2 is more than scenario B2 and it is *vice versa* in dry season. The monthly results show that the highest monthly uncertainty occurs in March for all periods, and the lowest is in July. During rainfall downscaling in wet months, models tend to diverge and increase the uncertainty. However, future studies are needed to carry out the future drought and flood frequencies.

References

- Ahmadi A, Han D, Karamouz M and Remesan R 2009 Input data selection for solar radiation estimation; *Hydrol. Process.* **23**(19) 2754–2764, doi: [10.1002/hyp.7372](https://doi.org/10.1002/hyp.7372).
- Araghinejad S and Burn D H 2005 Probabilistic forecasting of hydrological events using gestatistical analysis; *Hydrol. Sci. J.* **50**(5) 838–856.
- Araghinejad S, Burn D H and Karamouz M 2006 Long-lead probabilistic forecasting of streamflow using ocean-atmospheric and hydrological predictors; *Water Resour. Res.* **42** W03431.
- Bae D, Jeong D M and Kim G 2007 Monthly dam inflow forecasts using weather forecasting information and neuro-fuzzy technique; *Hydrol. Sci. J.* **52**(1) 99–113.
- Bray M and Han D 2004 Identification of support vector machines for runoff modelling; *J. Hydroinformatics* **6**(4) 265–280.
- Chen D and Chen Y 2003 Association between winter temperature in China and upper air circulation over East Asia revealed by canonical correlation analysis; *Global Planet. Change* **37** 315–325.
- Choy K Y and 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.

- Conway H, Gades A and Raymond C F 1996 Albedo of dirty snow during conditions of melt; *Water Resour. Res.* **32(6)**, doi: [10.1029/96WR00712](https://doi.org/10.1029/96WR00712).
- Dutta S C, Ritchie J W, Freebairn D M and Yahya Abawi G 2006 Rainfall and streamflow response to El Niño Southern Oscillation: A case study in a semi-arid catchment, Australia; *Hydrol. Sci. J.* **51(6)** 1006–1020.
- Galeati G 1990 A comparison of parametric and non-parametric methods for runoff forecasting; *Hydrol. Sci. J.* **35(1)** 79–94.
- Ghosh S and Mujumdar P P 2008 Statistical downscaling of GCM simulations to streamflow using relevance vector machine; *Adv. Water Resour.* **31** 132–146.
- Hanssen-Bauer I, Forland E J, Haugen J E and Tveito O E 2003 Temperature and precipitation scenarios for Norway: Comparison of results from dynamical and empirical downscaling; *Clim. Res.* **25** 15–27.
- Hashmi M Z, Shamseldin A Y and Melville B W 2011 Comparison of SDSM and LARS-WG for simulation and downscaling of extreme precipitation events in a watershed; *Stochastic Environmental Research and Risk Assessment* **25(4)** 475–484, doi: [10.1007/s00477-010-0416-x](https://doi.org/10.1007/s00477-010-0416-x).
- Haylock M R et al. 2006 Trends in total and extreme south American rainfall in 1960–2000 and links with sea surface temperature; *J. Climate* **19** 1490–1512.
- Johansson B and Chen D 2003 The influence of wind and topography on precipitation distribution in Sweden: Statistical analysis and modelling; *Int. J. Climatol.* **23** 1523–1535.
- Karamouz M, Zahraie B, Fatahi E, Mirzaie E, Remezani F and Hashemi R 2005 Predictors for long-lead precipitation forecasting in western Iran, The First Iran–Korea Joint Workshop on Climate Modelling, Nov. 16–17, Mashhad, Iran.
- Karamouz M, Ahmadi A and Moridi A 2009 Probabilistic reservoir operation using Bayesian Stochastic Model and Support Vector Machine; *Adv. Water Resour.* **32** 1588–1600, doi: [10.1016/j.advwatres.2009.08.003](https://doi.org/10.1016/j.advwatres.2009.08.003).
- Karamouz M, Mojahedi A and Ahmadi A 2010 Inter-basin water transfer: An economic-water quality based model; *J. Irrig. Drain. Eng., ASCE* **136(2)** 90–98, doi: [10.1061/\(ASCE\)IR.1943-4774.0000140](https://doi.org/10.1061/(ASCE)IR.1943-4774.0000140).
- Karlsson M and Yakowitz S 1987 Rainfall-runoff forecasting methods, old and new; *Stoch. Hydrol. Hydraul.* **1** 303–318.
- Kember G and Flower A C 1993 Forecasting river flow using nonlinear dynamics; *Stoch. Hydrol. Hydraul.* **7** 205–212.
- Liu X and Coulibaly P 2011 Downscaling ensemble weather predictions for improved week-2 hydrologic forecasting; *J. Hydrometeorol.* **12** 1564–1580, doi: [10.1175/2011JHM1366.1](https://doi.org/10.1175/2011JHM1366.1).
- Moghaddamnia A, Ghafari M, Piri J and Han D 2009 Evaporation estimation using support vector machines technique; *Int. J. Eng. Applied Sci.* **5(7)** 415–423.
- Mpelasoka F, Mullan A B and Heerdegen R G 2001 New Zealand climate change information derived by multivariate statistical and artificial neural network approaches; *Int. J. Climatol.* **21** 1415–1433.
- Najafi M, Moradkhani H and Wherry S 2011 Statistical downscaling of precipitation using machine learning with optimal predictor selection; *J. Hydrol. Eng.* **16(8)** 650–664, doi: [10.1061/\(ASCE\)HE.1943-5584.0000355](https://doi.org/10.1061/(ASCE)HE.1943-5584.0000355).
- Nash J E and Sutcliffe J V 1970 River flow forecasting through conceptual models. Part 1 – A discussion of principles; *J. Hydrol.* **10(3)** 282–290.
- Nourani V, Alami M T and Aminfar M H 2009 A combined neural-wavelet model for prediction of Ligvanchai watershed precipitation; *Eng. Appl. Artificial Intell.* **22(3)** 466–472.
- Salas J D, Deulleur J W, Yevjevich V and Lane W L 1980 Applied modelling of hydrologic time series, *Water Resources Publications*, Littleton, Colorado, USA.
- Schoof J T and Pryor S C 2001 Downscaling temperature and precipitation: A comparison of regression-based methods and artificial neural networks; *Int. J. Climatol.* **21** 773–790.
- Sivapragasam C and Liang S Y 2005 Flow categorization model for improving forecasting; *Nordic Hydrol.* **36(1)** 37–48.
- Toddini E 2000 Real-time flood forecasting: Operational experience and recent advances; In: *Flood Issues in Contemporary Water Management* (eds Marsalek J, Watt E, Zeman E and Sieker F, Kluwer Academic Publishers, pp. 261–270.
- Vapnik V 1995 *The Nature of Statistical Learning Theory*, Springer, New York.
- Vapnik V 1998 *Statistical Learning Theory*, Wiley, New York.
- Widmann M and Bretherton C S 2003 Statistical precipitation downscaling over the northwestern United States using numerically simulated precipitation as a predictor; *J. Climate* **16(5)** 799–816.
- Wilby R L and Tomlinson O J 2000 The ‘Sunday Effect’ and weekly cycles of winter weather in the UK; *Weather* **55** 214–222.
- Wilby R L, Hay L E and Leavesley G H 1999 A comparison of downscaled and raw GCM output: Implication for climate change scenarios in the San Juan River Basin, Colorado; *J. Hydrol.* **225** 67–91.
- Wilby R L, Dawson C W and Barrow E M 2002 A decision support tool for the assessment of regional climate change impacts; *Environmental and Modelling Softwares* **17** 145–157.
- Xoplaky E, Gonzales-Rouco J F, Luterbacher J and Wanner M 2004 Wet season Mediterranean precipitation variability: Influence of large-scale dynamics and trends; *Clim. Dyn.* **23** 63–78.
- Yu X Y, Liang S Y and Babovic V 2004 EC-SVM approach for real-time hydrologic forecasting; *J. Hydroinform.* **6(3)** 209–223.