World Applied Sciences Journal 13 (6): 1387-1397, 2011
ISSN 1818-4952

Seasonal Meteorological Drought Prediction Using Support Vector Machine

[1]*Ali Reza Nikbakht Shahbazi,* [2]*Banafsheh Zahraie,* [3]*Hossein Sedghi,*
[4]*Mohammad Manshouri and* [5]*Mohsen Nasseri*

[1]Department of Water Science and Engineering,
Science and Research Branch, Islamic Azad University, Tehran, Iran
[2]Member of the Center of Excellence for Infrastructure Engineering and Management,
School of Civil Engineering, University of Tehran, Tehran, Iran
[3]Department of Water Sciences and Engineering,
Science and Research Branch, Islamic Azad University, Tehran, Iran
[4]Department of Water Sciences and Engineering,
Science and Research Branch, Islamic Azad University, Tehran, Iran
[5]School of Civil Engineering, University of Tehran, Tehran, Iran

Abstract: Meteorological drought prediction depends on the ability to forecast seasonal precipitation. In this paper, a well known statistical machine learning method, Support Vector Machine (SVM), is used to predict seasonal variations of the Standardized Precipitation Index (SPI) in four reservoir basins supplying the water demands of Tehran, the capital city of Iran. The historical time series of the meteorological variables including air temperature and geopotential height at the surface, 300, 500, 700 and 850 mbar levels in the geographical zone covering 10° to 60° north latitudes and 0° to 90° east longitudes have been selected as the model predictors. Mutual Information (MI) has been used for feature selection among the aforementioned predictors. The selected predictors in the months of April to August have been used as the SVM model inputs to predict seasonal SPIs in autumn, winter and spring seasons. The results have been compared with those of the Artificial Neural Networks (ANN). The comparison has shown that SVM outperforms ANN in terms of the Normal Mean Squared Error (NMSE), Mean Squared Error (MSE) and the coefficient of determination (R^2). The results have shown that the seasonal SPI values can be predicted by the proposed model with 2 to 5 months lead-time with enough accuracy to be used in long-term water resources planning and management in the study area.

Key words: Meteorological Drought · Standardized Precipitation Index · Support Vector Machine · Mutual Information · Iran

INTRODUCTION

Drought prediction has been a challenge in water resources planning and management for a long time, but in recent years, the severity of the droughts affecting the Middle East and North African (MENA) region has brought into focus the need to improve the techniques for predicting such droughts with some measure of accuracy.

Different studies conducted over the past decades have shown that meteorological drought is never the result of a single cause. A great deal of research has been carried out on the role of interacting systems in recognition of the regional patterns of climatic variability. Two major approaches have been prominent in the search for appropriate meteorological drought prediction techniques. These include the use of teleconnections and development of numerical models. The studies on teleconnections have been mostly focused on recognition of spatial patterns of climate variability in certain regions that are highly affected by some of these well-known teleconnections such as El Nino Southern Oscillation (ENSO) [1]. While these patterns tend to recur periodically with enough frequency to improve our ability for seasonal prediction of dry or wet spells, less work has been done about the regions that are not highly affected by these teleconnections. Iran is not an exception [2].

Corresponding Author: Ali Reza Nikbakht Shahbazi, Department of Water Science and Engineering,
Science and Research Branch, Islamic Azad University, Tehran, Iran. E-mail: nikbakhta@gmail.com.

Nazemossadat [3], Nazemossadat and Cordery [2] and Zahraie and Karamouz [4] investigated the North Atlantic Oscillation (NAO) and ENSO teleconnections to the precipitation and runoff in different areas in Iran. The results of their studies showed that however some correlations between the teleconnections and climate variabilities can be detected but they are not strong enough to offer opportunities for meteorological drought prediction.

In studying statistical properties of the drought events and their correlations with teleconnections, different techniques such as Markov Chain and Log-Linear methods [5-8], Artificial Neural Network (ANN) [9, 10] and pattern clustering [11] have been used. Raziei *et al.* [12] also studied spatial patterns and temporal variability of droughts in Western Iran. They also showed that there is not a clear evidence for a link between hydrological droughts in this region and ENSO events.

There are a limited number of works on the applications of data mining and statistical learning methods in SPI prediction. Cancelliere *et al.* [13] and Paulo *et al.* [7] utilized Markov Chain models for SPI prediction. The model developed by Cancelliere *et al.* [13] had a 3-month prediction lead time which can be used in the drought warning systems. Mishra and Desai [9] used Autoregressive Integrated Moving Average (ARIMA) and Direct multi-step neural network (DMSNN) models for SPI prediction with 1- to 12-month prediction lead-time. Their results showed DMSNN outperformed ARIMA. Kamban and Elshorbagy [14] also utilized ANN for predicting clustered SPI values. Barros and Bowden [15] tried to extend the lead time of operational drought forecasts. Their research strategy was to explore the predictability of drought severity from space–time varying indices of large-scale climate phenomena relevant to regional hydrometeorology (e.g. ENSO) by integrating linear and non-linear statistical data models. They used MI values to identify and select predictor variables among spatial datasets of precipitation, sea surface temperature anomaly patterns, temporal and spatial gradients of outgoing long-wave radiation and the wind-stress anomaly.

The Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. SVMs have gained grounds in the fields which have traditionally been the ANNs areas of strength. In the recent years, different applications of the SVMs have been reported in the hydrological and climatic studies [16-19]. Dibike *et al.* [20] applied SVM in remotely sensed image classification and rainfall/runoff modeling. They compared performances of SVM, ANN and a conceptual rainfall/runoff model in modeling rainfall-runoff process in three catchments and showed SVM superiority. The flexibility and capabilities of SVMs in detecting data structures make it ideally suited for drought prediction where good generalization performance in capturing non-linear regression relationships between the predictors and the predictand is required.

The aim of this paper is to assess the performance of SVM in recognizing repetitive statistical patterns in variations of meteorological variables in a relatively large region surrounding Iran that might offer opportunities to improve our ability for prediction of seasonal values of SPI as an indicator of meteorological drought severity. The case study of this research includes the catchments of four reservoirs supplying water consumption of Tehran, the capital city of Iran. The major difference between this research and the previous works is in utilizing SVM for seasonal SPI prediction and incorporating a relatively large dataset for choosing the model predictors. This is also the first attempt for developing SPI prediction model for the study area. In the next sections of the paper, a brief introduction to SPI and SVM is presented. Then the proposed approach for feature selection method and the results of the case study are discussed.

MATERIALS AND METHODS

Standard Precipitation Index (SPI): SPI was developed by McKee *et al.* [21] to assess meteorological drought severity and precipitation deficit. Positive (negative) SPI values indicate greater (less) than median precipitation. As a measure of departure from the median, the SPI is a probability indication of the severity of the wetness or aridity. McKee *et al.* [21] selected the Gamma distribution for fitting monthly precipitation data. In calculating SPI, the Gamma distribution is then transformed to a Gaussian distribution. The standardized anomaly is then computed with results having an average of zero and a standard deviation of one. All of the above steps make the SPI independent of both the location and the range of values so that different seasons and climate areas are represented on an equal basis [22,23]. For this purpose, McKee *et al.* [21] divided SPI values to seven linguistic drought classes as shown in Table 1.

Table 1: SPI classes of meteorological drought [21].

Linguistic drought condition	SPI	Acronym
Very wet	more than +2	VW
Wet	1.5 to 1.99	W
Normal Wet	1 to 1.49	NW
Normal	0.99 to -0.99	N
Normal Dry	-1 to -1.49	ND
Dry	-1.5 to -1.99	D
Very Dry	less than -2	VD

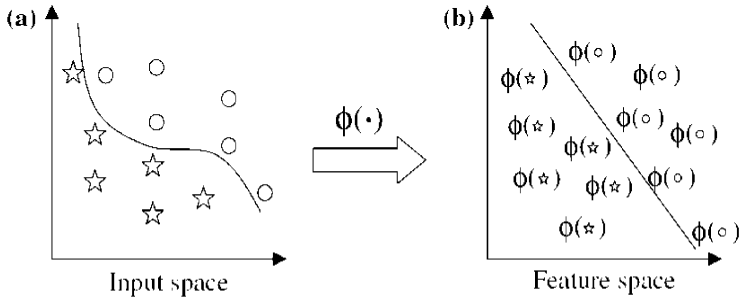

Fig. 1: A nonlinear transformation function $\phi(.)$ defined to convert a non-linear problem in the original lower dimensional input space (a) to linear problem in a higher dimensional feature space (b). (The stars and circles denote the data points) [16].

Support Vector Machine (SVM): Empirical data modeling and structure recognition is a challenge in many engineering realms. To address this challenge, a meta model has to be developed to deduce the system responses that have yet to be observed. Both quantity and quality of the systems observations influence the performance of this constructed empirical model. It must be noted that the performance of the constructed Meta model can be highly influenced by non-uniformity, ambiguity and sparse distribution of the input space in problems that mostly have high dimensions. As a result, the problem may be under ill posed conditions [24] in the sense of Hadamard [25] Performance of the traditional Artificial Intelligence Methods (AIM) has significantly been affected by difficulties in generalization and producing models that can over fit the data.

SVM is a new generation of statistical learning methods which aim to recognize the data structures. The foundations of SVM were developed by Vapnik and Cortes [26]. Its formulation is based on the Structural Risk Minimization (SRM) principle. It has been shown that the application of SRM in SVM leads to a better performances than the application of traditional Empirical Risk Minimization (ERM) principle employed in traditional AIMs. SRM minimizes an upper bound on the expected risk, as opposed to ERM that minimizes the error on the training data. Another SVM feature in detecting the data structure is transformation of original data from input space to a new space (feature space) with new mathematical paradigm entitled Kernel function. For this purpose, a non-linear transformation function $\phi(.)$ is defined to map the input space to a higher dimension feature space, $\Re^{nh}$ (Figure 1). According to Cover's theorem [27] a linear function, $f(.)$, can be formulated in the high dimensional feature space to represent a non-linear relation between the inputs (x_i) and the outputs (y_i) as follows:

$$y_i = f(x_i) = \langle w, \phi(x_i) \rangle + b \quad (1)$$

Where w and b are the model parameters. This mathematical approach has been presented previously by Aizerman *et al.* [28]. Figure 2 shows the schematic structure of a general SVM. SVM can be used for both regression and classification. In this paper, the terms Support Vector Regression (SVR) will be used.

The first generations of SVMs were developed to solve the classification problems, but recently regression based SVMs are developed using more sophisticated error functions [28]. SVM performs regression by using an ε-sensetive loss function $\|y - f(x)\|_\varepsilon = max\{0, \|y - f(x)\| - \varepsilon\}$. This loss function only considers errors bigger than a certain threshold $\varepsilon > 0$. SVM finds the optimal solution of the following primal problem:

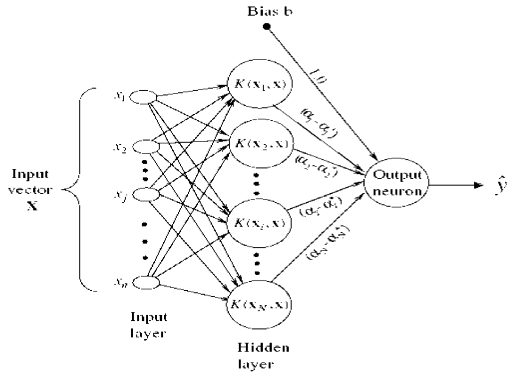

Fig. 2: Schematic architecture of SVM [16].

$$\text{Minimize } \phi(w,\xi,\xi^{(*)})=\frac{1}{2}\|w\|^2+C.\frac{1}{L}\sum_{i=1}^{L}(\xi_i+\xi_i^*), \qquad (2)$$

$$\text{Subject to: } (\langle w,x_i\rangle+b)-y_i\le\varepsilon+\xi_i,$$
$$y_i-(\langle w,x_i\rangle+b)\le\varepsilon+\xi_i^*,$$
$$\xi_i^*\ge 0.$$
$$w\in X.\quad \xi^{(*)}\in\Re^m.\quad b\in\Re\quad i=1,\ldots,L$$

Where:

L: number of data points in the training dataset
C: model parameter
x_i: feature space data points
W: Optimization problem solution
ξ: model residuals $(\xi_i = y_i - f(x_i))$
ξ_i and ξ_i^* are positive slack variables and C is a positive real valued and pre-specified constant. The constant C determines the amount up to which deviations from ε are tolerated. Deviations above ε are denoted by ξ_i, whereas deviations below ε are denoted by ξ_i^*. The large values of C might be a sign of over-fitting.

Feature Selection and Mutual Information (MI) Index: The feature selection process can be considered a problem of global combinatorial optimization in machine learning, which reduces the number of features, removes irrelevant, noisy and redundant data and results in acceptable classification accuracy. Feature selection methods have received much attention in the classification literature. Some useful methods for feature selection such as heuristic optimization, backward and forward sequential approaches and statistical filters such as MI in a number of applications of neural networks in water resources modeling problems can be found in the literature. MI has recently been utilized as a more appropriate statistical measure for feature selection during multi-dimensional model development, since it does not make any additional assumption about the dependency structure of the variables. MI has been found to be robust due to its insensitivity to noisy behavior [30]. MI for two discrete random variables X and Y can be defined as:

$$MI(x,y)=\sum_{y\in Y}\sum_{x\in X}p(x,y)\log\left(\frac{p(x,y)}{p_1(x)p_2(y)}\right) \qquad (3)$$

where $p(x,y)$ is the joint probability distribution function of x and y and $P_1(x)$ and $P_2(x)$ are the marginal probability distribution functions of x and y, respectively. MI is always positive and is also a symmetric function (i.e. $I(x,y) = I(y,x)$). In this study, MI is used for selecting the proper set of predictors of the SVM model which will be trained for seasonal SPI prediction.

The next section of the paper explains the proposed methodology for utilizing SVM, SPI and MI for predicting meteorological drought in the study area.

Methodology: The proposed procedure for meteorological drought prediction consists of the following steps:

Section of the Predictors and the Predictand: In this study, the predictand is the 3-, 6- and 9- month average SPI values. The seasons in which the SPI values are predicted, should be selected in a way to assist in the water resources planning and management decision making processes. The model predictors (meteorological variables) can be selected based on the results of the previous studies on the statistical relationships between the variations of these variables and SPI values. The choice of predictors should create enough prediction lead-time for the water resources managers and the decision makers to benefit from the model predictions.

SPI Estimation: Since the main purpose of this study has been to predict the meteorological drought severity affecting the study area, the SPI values for the selected seasons in step 1 are estimated using areal average precipitation over the basins. In this study, the optimized moving Inverse Distance Weighted (IDW) method presented by Abedini and Nasseri [31] has been utilized. IDW is based on the assumption that the interpolating surface should be influenced mostly by the nearby points and less by the more distant points. The areal average precipitation values for each basin are then transformed to SPI using the procedure explained in section 2.

Feature Selection: Since different combinations of month, meteorological variables and geographical zones can make the number of the candidate predictors very large, using a feature selection method is essential. This process can vary from one region to another. Since there are no general guidelines for selection of the predictors in different parts of the world, site-specific studies should be carried out. In this study, MI index has been used as a feature selection filter to select the suitable subsets of the predictors that provide the best prediction accuracy.

SVM Model Training, Validation and Testing: SVR is used in this study to predict the numerical values of seasonal SPI as a continuous variable. The available dataset should be partitioned into training, validation and testing datasets. Standardization is widely used as a pre-processing step while using algorithms such as SVM and ANN to reduce systematic bias of the datasets. To utilize the SVR model in this study, 50% of the available dataset is selected for training while half of the remaining 50% is used for testing and the rest is used for validation. In the training phase, different Kernel functions are selected and their parameters and also the model parameters including C and ε are calibrated. Mean Square Error (MSE) is used as the statistical criterion for calibrating the model. The next phase is called validation in which the remaining 25% of the available dataset is used to regenerate the SPI values and assess the performance of the trained models. After selection of the best Kernel function based on the validation dataset, the remaining data is used for testing the model performance.

Evaluation of the Model Predictions: Various error estimation indices such as NMSE, MSE and or R^2 can be used for assessing the accuracy of SVR predictions. They can be estimated using the following equations:

$$MSE = \frac{\sum_{1}^{n}(X_p - X_o)^2}{n} \quad (4)$$

$$NMSE = \frac{\sum_{1}^{n}(X_p - X_o)^2}{n.S^2} \quad (5)$$

$$R^2 = \frac{(n\sum X_p.X_o - \sum X_p \sum X_o)^2}{(n\sum X_p - (\sum X_p)^2)(n\sum X_o - (\sum X_o)^2)} \quad (6)$$

Where X_p is predicted SPI and X_o is observed SPI and n is the number of the data points. It should be noted that accuracy of the predictions is usually correlated with the lead-time of the predictions. The longer the lead-time is, the less accuracy is expected however the predictions might be more useful in terms of guiding the multi-seasonal water resources planning decision making processes.

Comparison: It is usually suggested to compare the predictions of the proposed models with other techniques which have been previously used by different researchers for the same purpose. In this study, the results of the SVR models are compared with those of ANN models (Multilayer Perceptron Models), respectively. Readers could find comprehensive introduction and interpretation of ANN in Haykin [31].

Case Study: The study area is located between 34°-36.5° North latitudes and 50°-53° East longitudes. Over 70 percent of about one billion cubic meters of water consumption in Tehran, the capital city of Iran and its suburbs is supplied by five dams, namely Latian, Karaj, Taleghan, Mamloo and Lar out of which the basins of the first four dams are considered as the case study of this research. The long-term average seasonal rainfall over these basins is shown in Table 3. The observed time series of the precipitation data in 31 rain gauges have been used in the period of 1976-2007 for estimating areal average precipitation over the basins using the optimized IDW method. Figure 3 shows the location of these rain

Table 2: Kernel functions and corresponding parameters

Parameters	Kernel type	Kernel function
-	Linear	$K(\mathbf{x},\mathbf{z}) = \langle \mathbf{x},\mathbf{z} \rangle$
0<c<5, 0<p<5	Polynomial	$K(\mathbf{x},\mathbf{z}) = (\langle \mathbf{x},\mathbf{z} \rangle + c)^p$
-5<b<5, -5<c<5	Sigmoid	$K(\mathbf{x},\mathbf{z}) = \tanh(b\langle \mathbf{x},\mathbf{z} \rangle - c)$
-5 < γ < 5	Radial Basis Function (RBF)	$K(\mathbf{x},\mathbf{z}) = \exp(-\frac{\|\mathbf{x}-\mathbf{z}\|^2}{\gamma})$

Table 3: Scenarios and long term average precipitation of the four basins of the case study (cm)

		Long term Average Annual Rainfall (cm)			
Scenario	Seasons	Taleghan	Karaj	Latian	Mamloo
SPI1	Spring	22.5	22.9	20.2	21.1
SPI2	Autumn	18.5	17.4	16.5	20
SPI3	Winter	16.3	16.5	14.8	16.2
SPI4	Autumn+Winter	34.8	33.9	31.3	36.2
SPI5	Winter+Spring	38.8	39.4	35	37.3
SPI6	Autumn+Winter+Spring	57.3	62.3	55.2	58.4

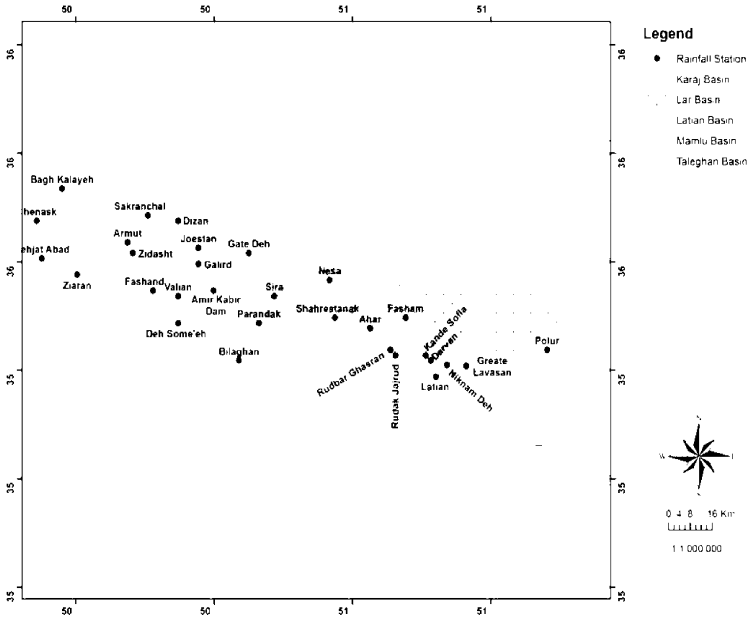

Fig. 3: Locations of the rain gauges the basins

gauges. The four time series of the monthly average precipitation have then been converted to SPI time series based on the methodology presented in Section 2 of the paper.

The major decisions regarding water allocation from the Tehran reservoirs are made during the summer and at the first three months of the water year based on the available water at the end of the summer and the precipitation and runoff predictions of the following reservoir refill seasons. Therefore, the SPI predictions issued in summer and autumn seasons showing aridity or wetness of the reservoirs refill seasons can be helpful in this decision making process. The seasons in which the SPI values is predicted in this study, are shown in Table 3. As can be seen in this Table, the six scenarios (seasons) include 3-, 6- and 9-month average SPIs starting at the beginning of October. Accurate predictions of SPIs for the selected seasons provide valuable information for the water managers responsible for making decisions regarding supplying the water demands of Tehran. The goodness of fit of Gamma distribution to the average precipitation estimated for the study area basins is investigated and the results have shown a proper fit.

In the previous studies carried out by Karamouz *et al.* [10] and Zahraie and Karamouz [4], statistical relations between the SLP and the seasonal precipitation variations in certain locations in geographical zone shown in Figure 4 have been studied. While they also considered some locations in the North and Central parts of the Atlantic Ocean (which are not shown in Figure 4), less attention was paid to the statistical relations that might exist between variations of the precipitation over Iran and the meteorological variables in certain locations in east of the region shown in Figure 4 (e.g. North of India and China). In this study, the time series of the meteorological variables including air temperature and geopotential heights at the surface, 300, 500, 700 and 850 mbar levels in the geographical zones shown in Figure 4 in the period of 1976-2007 are considered as the candidate predictors for the SPI prediction model. These time series are obtained from NCEP/NCAR reanalysis dataset [33].

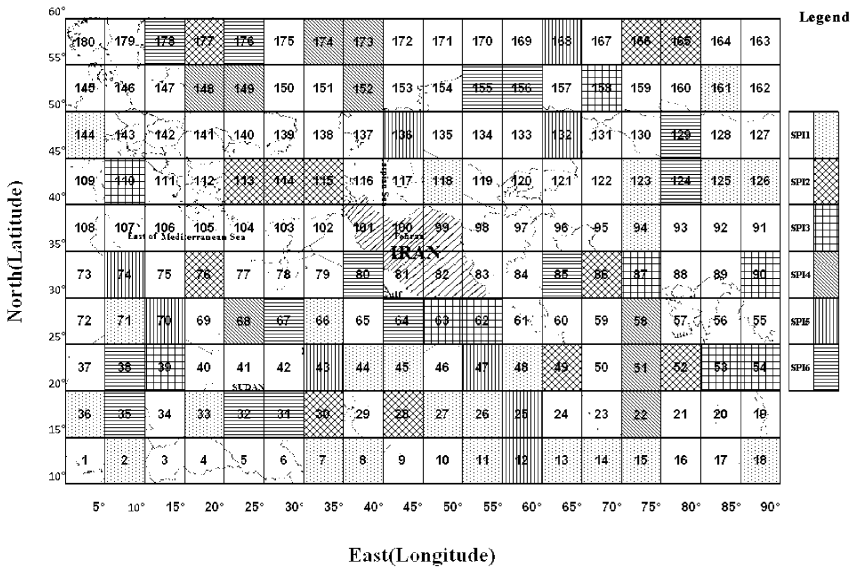

Fig. 4: Geographical zones for estimating predictors and with highest MI values

As shown in Figure 4, In order to cover the selected geographical region, it is divided to 180 square zones with the dimensions of 5° ×5°. In order to achieve a suitable prediction lead-time, the time series of the meteorological variables in the months of April through August are utilized to predict SPI values in the first three seasons of the following water year (seasons are shown in Table 3). The set of predictors in this study includes 180 (square zones) × 9 (predictors described in the before paragraph) × 5 (months of April through August)=8100 time series which must be filtered to find the best subset that provides the highest accuracy of the predictions.

RESULTS AND DISCUSSION

As it was explained in the previous section of this paper, 8100 candidate predictors should be analyzed to find the best set of predictors for each SPI scenario. MI which is used as a static filter for selecting the best set of predictors has been estimated for all combinations of six SPI scenarios and 8100 series of predictors. Figure 4 shows highest MIs for different SPI scenarios. This figure shows that the predictors in the following zones have the highest mutual dependence with the SPI variations:

SPI1 : Mostly over Red Sea and Arabian Sea
SPI2 : Scattered over Black Sea, Arabian Sea, Indian Ocean, North of India and South of Russia
SPI3 : Mostly Southeast of Iran and Oman Sea and North of Indian Ocean
SPI4 : Eastern part of India, Eastern Europe and west of Russia
SPI5 : Mostly over Arabian Sea and Red Sea
SPI6 : Mostly over Saudi Arabia and Sudan

Based on these results, the predictors with highest MI values have been chosen for each SPI scenario (Table 4). In Table 4, only the set of the predictors which have resulted in higher accuracies of predictions are presented. The LIBSVM toolbox [32] has been used for model training, validating and testing. Table 5 shows the selected Kernel functions and optimal values of the parameters obtained for each scenario-basin. The results of this study have shown that overall the linear Kernel function represents the best results however in three scenarios shown in Table 5, Polynomial and RBF Kernel functions have shown better performances. The calibrated hyper-parameters C and ε and the kernel function parameters, γ, c and p, are also shown in Table 5.

According to derived MSE, NMSE and R^2 indices in the model training and validation, the best model validation results based on NMSE have been obtained for SPI6, SPI1/SPI2, SPI6 and SPI1 for Karaj, Latian-Mamloo, Latian and Taleghan Basins, respectively.

The characteristics of the ANN models are set to two types of back-propagation training paradigm with one and two hidden layers and sigmoid transfer function. Since the selection of the number of hidden neurons and network configuration in ANNs is dependant to the problem and difficult, it is determined by trial and error.

Table 4: Selected SVM predictors for SPI scenarios

Seasonal Scenario	Reservoir Basin	Selected Predictors Using MI Values
SPI1	Latian	AT1000*,AT300,AT500,AT700,AT850,GH850**
	Mamloo	AT1000,AT300,AT500,AT850,GH300,GH500
	Taleghan	GH500,GH700
SPI2	Latian	AT300,AT700,AT850,GH300,GH500,GH700,GH850
	Mamloo	AT300,GH500,GH850
SPI3	Karaj	AT500,GH500,GH850
SPI4	Taleghan	AT1000,AT500,AT700,AT850,GH300,GH700,GH850
SPI5	Latian	AT300,AT500,AT500,GH500,GH700
	Taleghan	GH500,GH700,GH850
SPI6	Karaj	AT300,AT500,AT850,GH700,GH850
	Latian	AT300,AT700,GH500,GH700,GH850

*AirTemprature at 1000 mbar

**Geopotential Height at 850 mbar

Table 5: The best Kernel functions and the calibrated values of their parameters and the model hyper-parameters in each scenario.

Scenario	Basin	$C\times1000$	ε	Kernel Type*	Kernel Parameters*		MSE
SPI1	Latian	58	0.9	Linear	-	-	0.06
	Mamloo	650	0.3	Linear	-	-	0.08
	Taleghan	345	0.7	Linear	-	-	0.14
SPI2	Latian	581	0.9	Linear	-	-	0.34
	Mamloo	68	0.1	RBF	γ=-0.91	-	0.4
SPI3	Karaj	12	0.5	Polynomial	c= 4	p= 4.72	0.34
SPI4	Taleghan	154	0.7	Linear	-	-	0.56
SPI5	Latian	29	0.8	Linear	-	-	0.57
	Taleghan	66	0.06	RBF	γ = 1.16	-	0.23
SPI6	Karaj	963	0.4	Linear	-	-	0.14
	Latian	497	0.1	Linear	-	-	0.33

*See Table 2 for the equations of the Kernel functions

Table 6: Comparison between the performances of the SVR and ANN model in the test dataset.

		SVR			ANN		
Scenario	Basin	MSE	NMSE	R^2	MSE	NMSE	R^2
SPI1	Latian	0.59	0.53	59	0.6	0.53	52
	Mamloo	0.15	0.11	92	1.15	0.87	32
	Taleghan	0.30	0.33	73	0.15	0.17	81
SPI2	Latian	0.76	1.19	15	1.02	1.58	9
	Mamloo	0.73	1.29	76	0.87	1.54	31
SPI3	Karaj	0.29	0.49	75	0.55	0.92	41
SPI4	Taleghan	2.48	1.30	54	0.7	0.36	65
SPI5	Latian	0.34	0.17	89	0.57	0.30	67
	Taleghan	0.22	0.11	78	0.7	0.36	65
SPI6	Karaj	0.05	0.04	96	0.29	0.26	85
	Latian	0.1	0.08	97	0.01	0.03	99

Table 7:Lead-times of the SPI forecasts

	Forecast Lead-time (month)					
Watershed	SPI1	SPI2	SPI3	SPI4	SPI5	SPI6
Karaj	-	-	2	-	-	2
Latian	5	1	-	-	2	3
Mamloo	5	2	-	-	-	-
Taleghan	5	-	-	2	4	-

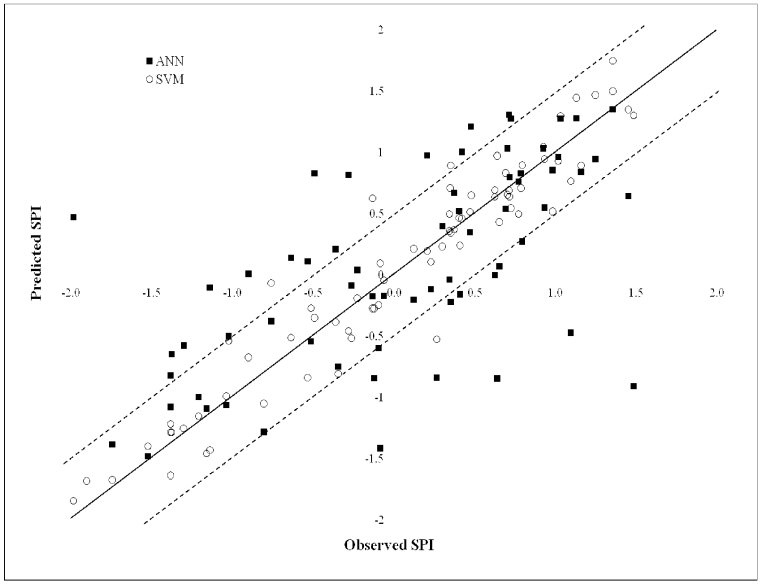

Fig. 5: Predicted versus observed SPI values for all scenario-basins in the test dataset

The network configuration with different training steps and neurons have been used and according to MSE criteria and sigmoid transfer function, a three layer ANN with back-propagation training paradigm and two neurons in hidden layer was selected and trained by the same set of datasets used for SVR training and testing. The best achieved learning rate is 0.1 and the number of iterations has been fixed at 1000. The results of the ANN model for the test dataset are presented in Table 6 which shows that based on derived NMSE, SVR outperforms ANN in almost all scenarios, but ANN works better than SVR in SPI1 and SPI4 for Taleghan basin and SPI6 for Latian basin. The results of the ANN model for the test dataset are presented in Table 7 which shows that based on derived NMSE, SVR outperforms ANN in almost all scenarios, but ANN works better than SVR in SPI1 and SPI4 for Taleghan basin and SPI6 for Latian basin. Figure 5 shows the scatter plot of predicted versus observed SPI values for all scenario-basins in the test dataset. It also indicates the higher accuracies of SVR predictions when available data for training are limited. Table 8 shows the model lead-time for different SPI scenario-basins based on the selected predictors. As it can be seen in this Table, the longest lead-time is 5 months for the SPI1 scenario. This scenario which is among those selected for the Taleghan Basin provides useful information to the reservoir operators.

CONCLUSION

In this paper, SVM has been utilized for prediction of meteorological drought at seasonal time-scale using specific variables of NCEP/NCAR reanalysis dataset as the model predictors. MI has been used as the feature selection filter to decrease the input space dimensions. One of the major obstacles against development of the prediction models for hydrologic variables is the lack of long records of the predictors for calibration and validation of these models. Using NCEP/NCAR reanalysis dataset which is accessible easily through web has made the calibration and future updating of this model a relatively easy task. The results of this study have shown that the choice of the predictor variables can significantly affect the accuracy of the model results. MI which has been suggested by other researchers for feature selection has also proven to be useful in this study.

The implemented statistical learning method, SVM, is more sensitive to selection of kernel function and its parameters; it is found that linear kernel function outperformed other kernel functions in SPI prediction. In assessing the performance of the proposed models in prediction of seasonal SPIs, the prediction lead-time should also be taken into account.

In assessing the performance of the proposed models in prediction of seasonal SPIs, the prediction lead-time should also be taken into account. The longest lead-time is for SPI1 scenario and the rest of the selected scenario-basins have two or three months prediction lead-time which provides enough time for the decision makers to adjust the water allocation policies. The predictions for the SPI6 scenario in Karaj and Latian Basins which covers all the reservoir refill season issued at the end of July also provide valuable information for the decision makers. Two important issues leads up to good predictions, regression based on SVM concept and selected meteorological large scale signals with MI.

The proposed approach to predict SPI can be extended to the use of a variety of other meteorological variables such as wind speed and relative humidity. Other feature selection techniques in combination with various optimization techniques can also be used for automatic feature selection which can be another direction for further research.

REFERENCES

1. Oguntoyinbo. J.S., 1986. Drought prediction. Climatic Change, 9: 79-90.
2. Nazemossadat, M.J. and I. Cordery, 2000. On the relationship between ENSO and autumn rainfall in Iran, International J. Climatol., 1: 42-67.
3. Nazemossadat, M.J., 1998. Persian Gulf sea surface tempreture as a drought diagnostic for southern parts of Iran, J. Drought News Network, 10: 12-14.
4. Zahraie, B. and Karamouz, 2004. Seasonal Precipitation Prediction Using Large Scale Climate Signals, Proceedings of EWRI-2004 Conference, Salt lake City, USA.
5. Steinemann A., 2003. Drought Indicators and Triggers: a Stochastic Approach to Evaluation. Journal of American Water Resources Association, 39(5): 1217-1233.
6. Loukas A. and L. Vasiliades, 2004. Probabilistic Analysis of Drought Spatiotemporal Characteristics in Thessaly Region, Greece. Natural Hazards and Earth System Sci., 4: 719-731.
7. Paulo, A.A. and L.S. Pereira, 2006. Drought concepts and characterization: comparing drought indices applied at local and regional Scales. Water Int., 31(1): 37-49.
8. Moreira, E.E., A.A. Paulo, L.S. Pereira and J.T. Mexia, 2006. Analysis of SPI drought class transitions using loglinear models. J. Hydrol., 331: 349-359.
9. Mishra, A.K. and V.R. Desai, 2006. Drought forecasting using feed-forward recursive neural network. Ecol. Modell., 198: 127-138.
10. Karamouz, M., M. Fallahi, S. Nazif and M. Rahimi Farahani, 2009. Long Lead Rainfall Prediction Using Statistical Downscaling and Artificial Neural Network Modeling. Scientia Iranica, 16(2): 165-172.
11. Zahraie, B. and A. Roozbahani, 2007. Climate Signal Clustering Using Genetic Algorithm for Precipitation Forecasting: A Case Study of Southeast of Iran Proceedings of the Word Environmental and Water Resources Congress (ASCE), Tampa, Florida, USA.
12. Raziei, T., B. Saghafian, A. Paulo, L.S. Pereira and I. Bordi, 2009. Spatial Patterns and Temporal Variability of Drought in Western Iran. Water Resources Manage., 23(3): 439-455.
13. Cancelliere, A., G. Di Mauro, B. Bonaccorso and G. Rossi, 2007. Drought forecasting using the Standardized Precipitation Index. Water Resour. Manage., 21(5): 801-819.
14. Kamban P. and A. Elshorbagy, 2007. Cluster-Based Hydrologic Prediction Using Genetic Algorithm-Trained Neural Networks. Journal Hydrologic Engineering, 12: 52-62.
15. Barros A.P. and G.J. Bowden, 2008. Toward long-lead operational forecasts of drought:An experimental study in the Murray-Darling River Basin. J. Hydrol., 357: 349- 367.
16. Tripathi, S.H., V.V. Srinivas and R.S. Nanjundiah, 2006. Downscaling of precipitation for climate change scenarios: A support vector machine approach. J. Hydrol., 330: 621- 640.
17. Han, D., L. Chan and N. Zhu, 2007. Flood forecasting using support vector machines. J. hydroinformatics. 9(4): 267-276.
18. Wang, W., C. Men and W. Lu, 2008. Online prediction model based on support vector machine. Neurocomputing, 71: 550-558.
19. Behzad, M., K. Asghari, M. Eazi and M. Palhang, 2009. Generalization performance of support vector machines and neural networks runoff modeling. Expert System with Applications, 36: 7624-7629.

20. Dibike, Y.B., S. Velickov, D. Solomatine and M.B. Abbott, 2001. Model induction with support vector machines: Introduction and applications. J. Computing in Civil. Engineering. 15(3): 208-216.
21. McKee, T.B., N.J. Doesken and J. Kleist, 1993. The relationship of drought frequency and duration to time scales. In: Proceedings of the Eighth Conference on Applied Climatology. Am. Meteor. Soc. Boston, pp: 179-184.
22. Guttman, N.B., 1998. Comparing the Palmer drought index and the Standardized Precipitation Index, J. Amer. Water Resour. Association. 34: 113-121.
23. Keyantash, J. and J.A. Dracup, 2002. A Regional Multivariate Drought Index Applied to California." Eos Trans. AGU, 82(20), Spring Meet. Supplement. Abstract H22D-04.
24. Poggio T., V. Torre and C. Koch, 1985. Computational vision and regularization theory. Nature, 317: 314-319.
25. Hadamard, J., 1923. Lectures on the Cauchy Problem in Linear Partial Differential Equations. Yale University Press.
26. Vapnik, V.N. and C. Cortes, 1995. Support vector networks. Machine Learning, 20: 273-297.
27. Cover, T.M., 1965. "Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition," IEEE Trans. Elec. Comp., EC-14: 326-334.
28. Aizerman M., E. Braverman and L. Rozonoer, 1964. Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control., 25: 821-837.
29. Vapnik, V.N., 1998. Statistical Learning Theory. Wiley, New York.
30. Bowden, G.J., G.C. Dandy and H.R. Maier, 2005. Input determination for neural network models in water resources applications. Part 1 - background and methodology. J. Hydrol., 301(1-4): 75-92.
31. Abedini, M.J. and M. Nasseri, 2008. Inverse Distance Weighted Revisited. Proceedings of the 4th APHW Conference, Beijing, China.
32. Chang, C.C. and C.J. Lin, 2009. LIBSVM: A Library for Support Vector Machines (Version 2.9.1, 2009). Software available at: http://www. csie. ntu. edu. tw/~cjlin/libsvm.
33. Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K.C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, Roy Jenne and Dennis Joseph, 1996. The NCEP/NCAR Reanalysis 40-year Project. Bull. Amer. Meteor. Soc. 77: 437-471.