

Long-Term Streamflow Prediction Using Large-Scale Climate Indices and Historical Weather Data (Case Study: Bukan Dam)

Razieh Taraghi Delgarm¹, Masoud Tajrishy²

1- Water Resources specialist, Environmental and Water Research Center (EWRC), Sharif University of Technology, Tehran, Iran.

2- Professor of Civil Engineering, Sharif University of Technology, Tehran, Iran.
Email: Razieh.Taraghi@gmail.com

Abstract

One of the most important principles of water resources planning and management in any country from a hydrological point of view is rainfall and streamflow prediction and understanding its temporal variations to build provide storage capacity[1]. Bukan dam is located upstream of Lake Urmia basin and is one of the most important water resources place in the country, so predicting it's streamflow, especially during drought periods, will greatly assist in calculating the volume of water released to restore Lake Urmia. In this study, using large-scale climate signals and Ocean- Atmospheric climate variables including SST and geopotential 500 hPa and precipitation, the streamflow of Bukan Dam during three periods, Feb-June, March-June and April-June was predicted by two main methods, Principal Component Regression and Genetic Programming. The results showed "R square" increases by adding precipitation variable between 26 and 39 percent and the use of the nonlinear (genetic programming) model increases it by 31 to 49 percent, so, which indicates the need for accurate and up-to-date data. It should be noted that although empirical linear regression models may not represent all the physical conditions and processes of modeling streamflow, they are still used due to their features such as ease of use and calibration as well as high accuracy of prediction equations.

Keywords: Long-Term Streamflow Prediction, Ocean-Atmospheric Climate Variables, Bukan Dam Basin, Principal Component Regression, Genetic Programming.

1. INTRODUCTION

In our country, due to the arid and semi-arid climate, the increasing need for water, and its proper management, research is needed to accurately predict the undeniable flow volume. Since decisions about storage and water release require long-term time horizons, these forecasts are of great importance for a resource management system and this leads to greater utility of electricity generation, supply needs, and reduction of floods and droughts, so that a slight increase in the accuracy of these forecasts will benefit the operating system.

For example, the US Army Corps of Engineers has earned an average annual profit of \$ 1 billion a year from operating properly using long-term hydrological forecasts (as opposed to in comparison with disregarding them)[2]

Hamlet's research also showed that for a 1% improvement in forecasts for of the Columbia River, a profit of \$ 6.2 million a per year would be generated for energy production[3].

Lake Urmia has suffered significant reductions in its water level in recent years. These declines continued until at the end of 1393-94 water year, the lake's water volume reached about 10 percent of its maximum reserve volume. Bukan Dam, built on the ZarrinehRoud River, is the largest water supplier to Lake Urmia, with an estimated volume of 40% of the total inflow to Lake Urmia[1]. Bukan Dam reservoir is the largest reservoir of Lake Urmia basin with a volume of 810 million cubic meters and accurate prediction of the inflow to it, can reduce over-cautious releases ,and it drives a significant amount of streamflow of ZarrinehRoud River to Lake Urmia and helps to revive it. On the other hand, it helps to reduce the amount of out of favor releases and minimizes downstream damage.

In recent decades, the identification of large-scale climatic signals as hydrological predictors in the absence of up-to-date and accurate ground measurements, especially in Iran, has produced a huge change in forecasts. In addition to considering hydrological and regional meteorological statistics in streamflow forecasting models, many researchers have investigated the effects of climate change and fluctuations using large-scale climate signals.

For example, some pieces of research such as Moradkhani & Meier[4], Xun Sun et al (2014)[5], Yadav(2013)[6], Davey et al (2014)[7] and Ali İhsan Marti(2014)[8] can be cited. Mekanik et al. (2013) predicted long-term spring precipitation in Victoria Australia using signals from the Bipolar Indian Ocean (IOD) and the Southern Oscillation. Using neural network techniques and multiple regression analysis, they concluded that multiple regression in central and western Victoria and neural network in eastern Victoria provided better results[9].

Yadav et al. (2013) also examined the relationship between ENSO index and winter precipitation in north and central India. The results of their research showed that the variations of seasonal precipitation and winter precipitation in India depend on the El Niño South Oscillation Index (ENSO)[6]

In Iran, also, valuable research has been done to investigate the relationship between large-scale climate signals such as SOI and NAO with precipitation and discharge, including the research by Zahraie and Karamouz (2004)[10], Nikzad et al. (2011) [11] and Ashouri et al. (2008)[12].

Roygar and Golian (2014) investigated the role of large-scale climate signals on precipitation and discharge in upstream of Golestan Dam watershed. Correlation analysis and correlation coefficients confirmed the relationship between climate and precipitation signals and discharge and showed that NINO 1 + 2 has the most effect on the study area[13].

KianiFalavarjani et al. (2011) predicted long-term streamflow of ZayandehRoud River using large-scale climate signals. Climate signals with SLP¹ and SST² indices were collected from 18 stations and then ZayandehRoud River Runoff was predicted using Support Vector Machin (SVM) method. The results show that the predictions are improved both in terms of increasing the time interval up to one year and the accuracy of predictions[14].

Recently, various statistical methods, such as PCA³, have been used to investigate the correlation and to explain the relationship between sea surface temperature of water zones and precipitation. These methods are particularly applicable in large geographical areas that have decades of data (or more than a century). In these areas, the usual statistical methods are not sufficient to reduce the volume of data and analyze them. One of the advantages of the PCA method is that by deleting similar data, it significantly reduces the available time series[15].

For example, PCA is capable of converting 100 interconnected data sets into a small number of new standalone data sets (eg 4 series). This new data set, while being statistically independent of each other, represents a large percentage of the total variance of the original data. In this study PCA method (Principal Component Analysis) was used[16, 17]

Nazem Sadat et al. (2005) used PCA method to extract the main components and reduce the amount of Persian Gulf water surface temperature data[18]. Other researchers such as Behrangi et al. (2009)[19] and Nazem Sadat and Shirvani (2006) [20] have also used the mentioned statistical methods in climate studies.

Conventional modeling techniques, such as regression and time series modeling, fail more in modeling nonlinear processes such as streamflow prediction. Therefore, nonlinear techniques can be a more efficient tool for predicting these processes.

The use of neural networks, gene expression models (Genetic Programming) and Bayesian networks have been prominent models in recent decades that have found many applications in various fields such as weather forecasting. In gene expression models, the input variables, the target, and the set of functions must be defined first, and the optimal model structure and the coefficients will be identified during the training process.

In recent years, varieties of studies have been done in field of streamflow forecasting using this model worldwide. ZareAmini et al. (2014) investigated the ability of genetic programming to estimate soil temperature and concluded that the genetic programming method outperforms the neural network method by providing an explicit solution. Other researches include estimation of daily discharge of Karun River [21].

Following the above and the importance of accurate streamflow estimation in water resources management, the volume of inflow to Bukan Dam was studied using two main methods, Principal Component Regression and Genetic Programming. Finally, the accuracy of the two methods is compared and evaluated.

¹Sea Level Pressure

²Sea Surface Temperature

³Principal Component Analysis

2. MATERIALS AND METHODS
2.1 INTRODUCTION TO THE STUDY AREA

Urmia Lake basin is about 51460 square kilometers located in northwest of Iran and Bukan dam is one of the important areas of this basin for drinking water supply and agriculture of local community. The useful usable capacity of the reservoir is 480 million cubic meters and its total capacity is 650 million cubic meters. The most important rivers in this basin are ZarinehRoud (the most rugged), SiminehRoud, Mahabadchay, Godarchay, Baranduzchay, Shahrchay, Rozehchay, Nazluchay, Ajichay and Sufichay. Much Most of water in these rivers is supplied by precipitation and snow, and thus is greatly reduced in summer. Lake Urmia has an area of about 5000 square kilometers, which changes during the Watery and low water years as well as dry and wet seasons, The location of Bukan Dam are shown below (Fig. 1)[1].

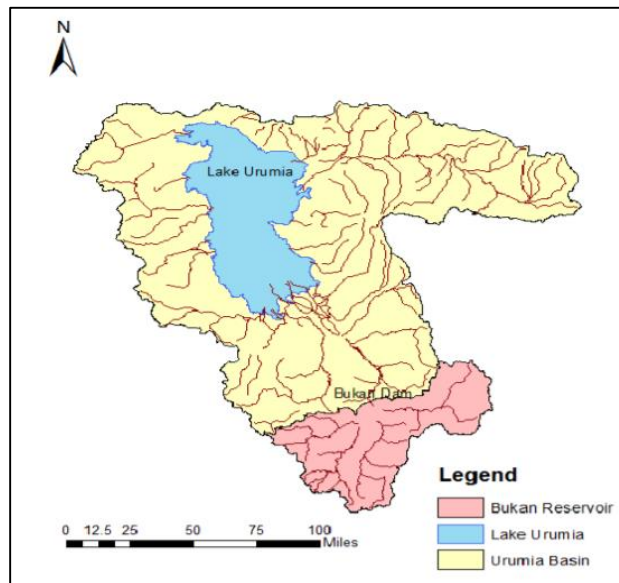


Figure 1. The location of Bukan Dam[1]

2.2 HOW TO CHOOSE FORECAST PERIODS

In this study, using two methods: gene expression programming and multivariate regression based on principal components and taking into account the meteorological variable (precipitation) and large scale climate signals into account, streamflow of ZarrinehRoud in upstream of Bukan Dam during Feb- June, March – June and April- June are forecasted.

Due to large amount of runoff in the western part of the country during late winter (late winter to late spring), the choice of the three predicted periods seems reasonable. As can be seen in Fig. 2, more than 80% of total annual volume of Bukan Dam streamflow flows occurs during from February to June. In other words, by choosing these three periods, we can predict about 80% of annual volume of streamflow.

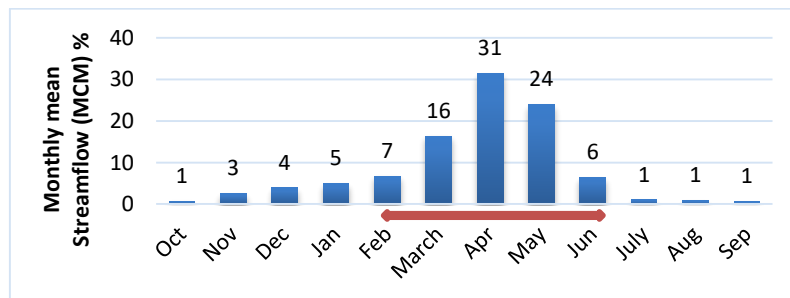


Figure 2. Percentage of Monthly mean streamflow of Bukan Dam Reservoir

2.3 LARGE-SCALE CLIMATE SIGNALS

Climate signals are indicators to quantify the intensity and to indicate fluctuations in climate phenomena. Phenomena such as ENSO, PDO⁴ and NAO⁵ that have fluctuated annually and have been shown to affect temperature and rainfall variations around the world. Information about climate indicators is available on Active Climate Change Center's website⁶. Some of these signals, known as temperature indices, reflect changes in ocean surface temperature over specified geographic ranges. Information about these signals is available on US National Oceanic and Atmospheric Administration (NOAA)' website. Others are atmospheric indicators that take into account changes in air pressure at different levels, wind components (intensity, velocity, etc.) and other atmospheric components. In general, the large-scale climate indicators are divided into 5 categories as follows:

- Teleconnections: It is used to explain how changes in climate patterns in one part of the world can cause changes in other parts of the world. Among these climate indices of this section are NAO, PDO, PNA⁷, WP⁸ and NOI⁹ indices.
- Atmospheric patterns such as SOI¹⁰ index
- Precipitation
- ENSO and Pacific surface temperature such as ONI, NINO and WHWP¹¹ indices
- Atlantic surface temperature such as TNA¹² and TSA

2.4 WATER SURFACE TEMPERATURE IN ADJACENT WATER ZONES AND 500 HPAGEOPOTENTIAL HEIGHT

The National Center for Environmental Prediction (NCEP), in conjunction with the National Center for Atmospheric Research (NCAR)¹³, is leading and supporting a project called "Reanalysis" to collect global data (over 50 years). These efforts include the collection, quality control and simulation of ground data, Radio and anemometer data and data are obtained from ships, planes, and satellites. The water surface temperature and geopotential height data used in this study are also pixels with 2 ° longitudinal accuracy at 2 ° transverse (about 190 km * 220 km) from the same data centers[22].

2.5 STREAMFLOW FORECASTING MODELING FRAMEWORK

In order to identify the variables affecting the Streamflow changes of Bukan Dam, first, the correlation between large-scale climate signals with streamflow in different forecast periods was investigated. Then the correlation between the meteorological variable of precipitation and streamflow in different forecast periods is investigated. In addition, since autumn streamflow indicates previous rainfall status and previous soil moisture, it can be significantly correlated with spring volume, so the relationship between autumn streamflow and streamflow in three forecast periods is also studied. In order to identify large-scale climate signals affecting the streamflow, the correlation between the mean monthly discharge in the first period (February to June), the second period (March to June) and the third forecast period (April to June) was correlated with large-scale signals such as: AO, BEST, MEI, NAO, NINO1 + 2, NINO3,4, NINO4, NOI, ONI, PDO, PNA, SOI, WHWP, WSA, TSA, TNA in prior months¹⁴.

The following relationship (Eq. 1) is used to test the significance of the cross-correlation of the data. In this formula, R is the correlation coefficient and n is the number of data. The parameter t has an approximate distribution of T with n-2 degrees of freedom[1].

⁴Pacific Decadal Oscillation

⁵North Atlantic Oscillation

⁶ <http://www.esrl.noaa.gov/psd/> - <http://idn.ceos.org/> - <http://jisao.washington.edu/> - <http://portal.iri.columbia.edu/>

⁷Pacific North American Index

⁸Western Pacific Index

⁹Northern Oscillation Index

¹⁰Southern Oscillation Index

¹¹Western Hemisphere Warm Pool

¹²Tropical Northern Atlantic Index

¹³<https://www.esrl.noaa.gov/psd/cgi-bin/data/timeseries/timeseries1.pl>

¹⁴<https://www.esrl.noaa.gov/psd/data/climateindices/list/>

$$t = \frac{R}{\sqrt{\frac{1-R^2}{n-2}}} \tag{1}$$

If the value of t calculated from the above formula is less than the critical value of t with n-2 degrees of freedom at the significant level desired, the hypothesis of two-time correlation (independence) is accepted and otherwise rejected. For example, the correlation coefficients for the first forecast period (February-June) at the inlet of Bukan Dam with PDO signal in the previous months are presented in Fig 3. As shown in fig 3, the correlation coefficient between streamflow and PDO index value in November, October, September and August exceeds the acceptable correlation coefficient of 95%. This means that there is a significant correlation between this period and the PDO index in November, October, September and August. Therefore, the PDO index value in these months are selected as one of the input parameters to for forecasting model of Bukan Dam streamflow in the first forecast period.

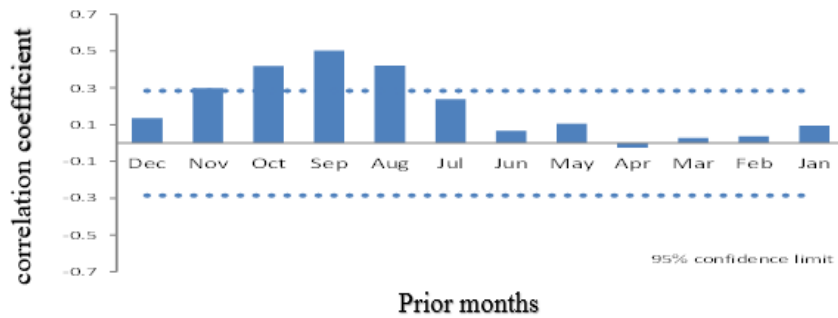


Figure 3. Correlation coefficient at the inlet of Bukan Dam with PDO signal in prior months (Feb-June)

Regarding the geographical location of Iran, it can be said that the amount of rainfall in Bukan Dam basin is due to the activity of Low Latitude Rainfall Systems (Sudanese systems), which is influenced by changes in surface temperature of Red Sea and Persian Gulf, High Latitude Rainwater Systems (Mediterranean systems). Therefore, surface water temperatures in these areas were also added as predictors.

2.6 GENE EXPRESSION PROGRAMING (GEP)

The gene expression programming method was first proposed by Ferreira (2001). This method follows the evolution of intelligent models and is based on Darwin's theory of evolution.

The advantage of using this method over other methods e.g neural network is explicit mathematical relationship between predictors and predictors. Also in this method, first the structure of blocks including input variables, target function and set of functions are defined and then the optimal structure of the model is determined in the training process. On the other hand, this model is capable of performing sensitivity analysis and selecting the most effective variables from the input variables for the model structure [23]. For modeling in order to build the model based on genetic programming, the first step is to choose the fitting function. In this study, the root mean square error was used. In step two, the set of input variables and the set of functions are selected to produce the chromosomes. In this study, operators (+, -, ×) were used. Head size and the number of genes are determined by trial and error. In the following, the link function should be specified, which was used +.

2.7 SCENARIOS USED

After identifying the appropriate variables, modeling was performed was built using linear regression and genetic programming. And if the linear relationship between the predictor variables is high, a more appropriate model is developed. The statistical period is from 1344-45 to 1394-95. The final years have been separated for verification purposes. Modeling was performed twice, Once in the presence of precipitation variable and again in the presence of meteorological variable (precipitation). Although the use of precipitation meteorological

variables can greatly increase the prediction accuracy, since in Iran we have a lack of up-to-date data, modeling without the presence of precipitation variable can yield an acceptable estimate. In order to predict streamflow of Bukan Dam with different probability levels, according to the regression equation, the values of equality or exceeds with probabilities of 95, 90, 30,70, 10 and 5% are prepared as a table (Table. 1).

3. MODEL VERIFICATION CONTROL

After estimating the prediction equation, it is necessary to calculate the confidence level of the model. In general, the closer the actual value of the series to its predicted value, the greater the accuracy of the prediction model. Various statistical indices can be used to evaluate the accuracy of the model (e.gRoot Mean Square Errors (RMSE) and Mean Absolute Percentage Errors (MAPE)).

Since the purpose of this study is to use forecasting models to predict future status, the results of the model for the recent water years including 1395-96 and 1396-97 are presented below.

4. RESULTS AND DISCUSSION

The results of multivariate regression based on principal components are presented in Table 1 and the results of the validation are presented in Table 2 .It is found that by adding rainfall variables, the results are improved.

Table 1-The results of multivariate regression based on principal components

streamflow prediction with Different Levels of Probability or Exceedance Probability (million cubic meters)							Obsrved Streamflow (million cubic meters)	Predicted Streamflow (million cubic meters)	R squared	MAPE	prediction Period	Year	Type of Model
0.95	0.9	0.7	0.5	0.3	0.1	0.05							
1914	1739	1374	1121	869	504	329	985	1121	0.55	0.3	Feb-june	1395-96	
2459	2299	1965	1733	1502	1168	1007	945	1733	0.58	0.4	March-June		
1380	1245	965	771	578	298	163	740	771	0.56	0.3	Apr-June		
2065	1937	1672	1487	1303	1037	910	985	1487	0.74	0.3	Feb-june	1395-96	Rainfall added
1764	1645	1396	1225	1053	804	685	945	1225	0.73	0.3	March-June		
1095	1005	818	689	559	372	282	740	689	0.78	0.3	Apr-June		
1315	1140	775	522	270	0	0		522	0.55	0.3	Feb-june	1396-97	
1835	1674	1340	1109	878	543	383		1109	0.58	0.4	March-June		
1052	972	803	687	571	403	322		687	0.56	0.3	Apr-June		
1899	1771	1505	1321	1137	871	744		1321	0.7	0.3	Feb-june	1396-97	Rainfall added
1519	1400	1152	980	808	559	440		980	0.73	0.3	March-June		

Table 2-The results of model validation (multivariate regression based on principal components)

		Feb-June		Feb-June (Rainfall Added)		March-June		March-June (Rainfall Added)		Apr-June		Apr-June (Rainfall Added)	
		Calibration	Verification	Calibration	Verification	Calibration	Verification	Calibration	Verification	Calibration	Verification	Calibration	Verification
$RMSE = \sqrt{\frac{\sum_{t=1}^n (A_t - F_t)^2}{n}}$	RMSE	272	723	230	395	297	1074	223	377	230	571	189	286
$MAPE = \frac{\sum_{t=1}^n A_t - F_t }{\sum_{t=1}^n A_t} \times 100$	MAPE	17	47	18	61	24	57	21	27	29	39	26	22
	R squared	0.82		0.88		0.76		0.87		0.81		0.88	

Also the results of modeling with genetic programming method for years 1395-96 and 1396-97 are presented in Figure 4 and a brief comparison between Genetic Programing and Linear Regression methods is presented in Figures 5 and 6.

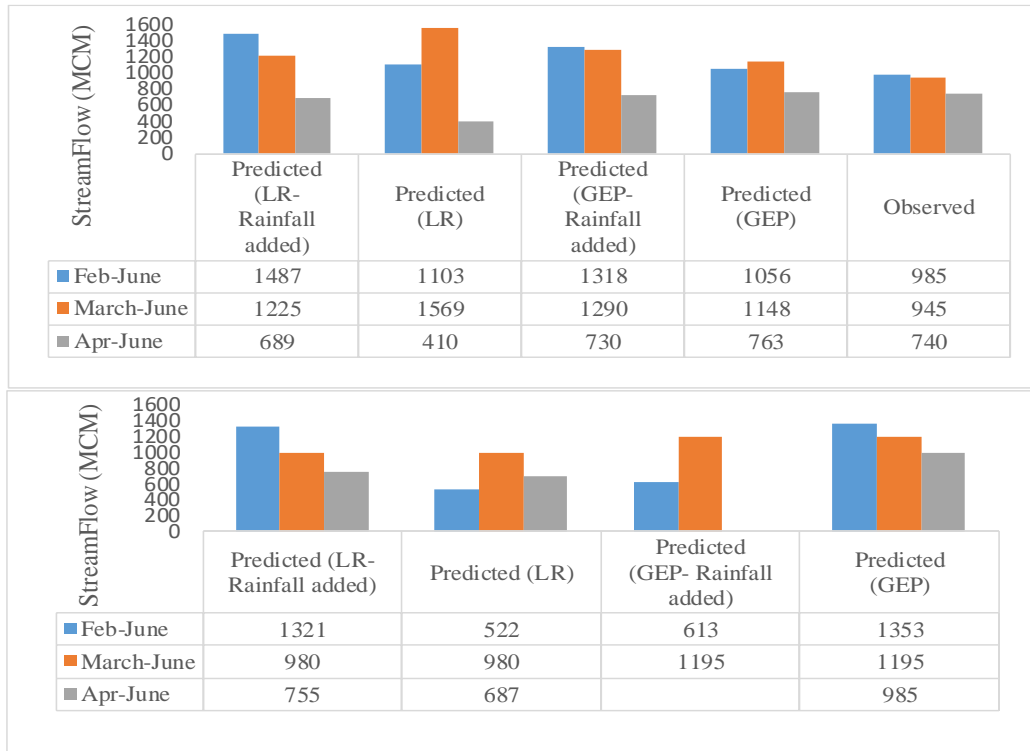


Figure 4. The results of modeling with genetic programming method for years 1395-96 (up) and 1396-97 (down)



Figure 5. Brief comparison between Genetic Programming and Linear Regression methods (Feb-June), (March- June) and (Apr-June)



Figure 6. Brief comparison between Genetic Programming and Linear Regression methods (Feb-June), (March- June) and (Apr-June) – Rainfall added

Modeling results show that both multivariate regression models based on principal components as well as genetic programming method have acceptable accuracy. Although the results of predictions by genetic programming method are somewhat closer to reality due to nonlinear modeling, regression models are still applicable due to low computational cost and acceptable accuracy. The results also show that just adding rainfall variables in the forecast period from February to June improved the R squared or Coefficient of Determination of the model from 0.55 to 0.74 (35%) and the use of nonlinear models (here GEP) improved 49% from 0.55 to 0.82. In the forecast period of March to June, addition of precipitation variable increased the Coefficient of Determination from 0.58 to 0.73 (26% increase) and the use of GEP model increased the forecast results by 31%. These values have increased by about 40% and 45% during the forecast period from April to June, which reveals the need for up-to-date and accurate data and also at the right time for prediction.

5. REFERENCES

1. Taraghi Delgarm, R. (2016), "Long term Seasonal Rainfall and Streamflow Prediction using Ocean-Atmospheric Climate Variables (case study: Bukan Dam)", Sharif University of Technology, M.Sc thesis, Civil Engineering, (in Persian).
2. Azimi, M. (2011), "Long term Seasonal Rainfall and Streamflow Prediction using regionalization of Ocean-Atmospheric Climate Variables" Sharif University of Technology, M.Sc thesis, Civil Engineering, (in Persian).
3. Alan F., Hamlet., HUPPERT, D., Lettenmaier, D., (2002). "Economic value of longleadstreamflow forecasts for Columbia River hydropower", Journal of water resources planning and management, 128(2): 91-101.
4. Moradkhani, H., Meier, M., "Long-Lead Water Supply Forecast using Large-scale Climate Predictors and Independent Component Analysis", J. of Hydrologic Engineering, 15(10).

5. Xun.Suna,M.Thyrb, B. Renarda, M.Lang (2014). "A general regional frequency analysis framework for quantifying local-scale climate effects: A case study of ENSO effects on Southeast Queensland rainfall", Journal of Hydrology Volume 512, 6 May 2014, Pages 53–67.
6. R.K. Yadava, D.A. Ramua, A.P. Dimrib (2013). "On the relationship between ENSO patterns and winter precipitation over North and Central India", Global and Plane tary Chang e 107,50 – 58.
7. M.K. Davey a,b, A. Brookshaw a, S. Inesona (2013). "On the relationship between ENSO patterns and winter precipitation over North and Central India", Climate Risk Management ,5–24.
8. Ali Ihsan Marti.(2013). "ENSO Effect on Black Sea Precipitation". Global and Planetary Change 107, 50 – 58.
9. Mekanik , M. A. Imteaz (2013). "Analysing lagged ENSO and IOD as potential predictors for long-term rainfall forecasting using multiple regression modelling", International Congress on Modelling and Simulation, Australia, December 2013.
10. Zahraie, B., and Karamouz, (2004), "Seasonal Precipitation Prediction Using Large Scale Climate Signals", Proceedings of EWRI-2004 Conference, Salt lake City, USA.
11. Nikzad, M. Behbahani, M.R andRahimi. A (1390). "Evaluation of Large Scale Signals and Sea Surface Temperature of Persian Gulf and Red Sea in Prediction of Drought by Artificial Neural Network in Khuzestan Province"Second National Conference on Applied Water Resources Research of Iran ,(in Persian).
12. Ashouri, H., A. Abrishamchi, H., Moradkhani, and M. Tajrishy (2008). "Assessment of Interannual and Interdecadal Climate Variability Effects on Water Supply in Zayandehrood River Basin, Iran", The First International Conference on Water Resources and Climate Change, Sultanate of Oman.
13. Golian. S, and Roygar. H, (1393), " Investigating the Relationship between Large-Scale Climate Indices and Monthly Precipitation and Monthly mean streamflow of Golestan Dam Basin", 8th National Congress of Civil Engineering, Ahvaz, Iran ,(in Persian).
14. Kiani F, M. and Ahmadi. A, (1390), "Long-term streamflow forecasting using climate signals and intelligent computing methods" 6th National Congress of Civil Engineering, Semnan. Iran ,(in Persian).
15. Song-Weon Lee (2004). "Investigation of Techniques for Improvement of Seasonal Streamflow Forecasting the Upper Rio Grande Basin", Texas A&M University, Doctor of Philosophy, Civil Engineering.
16. Drosdowsky, W. (1993). "An analysis of Australian seasonal rainfall anomalies 1950- 1987, Spatial pattern". Intelligent Journal of Climatology. 13: 1-30.
17. Yatagai, A. and T. Yasunari. (1995), "Inter-annual variations of summer precipitation in the arid/semi-arid regions in China and Mongolia, Their regionality and relation to the Asian summer Monsoon". J. Meteor. Soc. Japan 73: 909-923.
18. Nazemossadat, M.J. & A. Shirvani, (2005), "Forecast of Winter Precipitation of South Iran by Persian Golf Sea Surface Temperature", Scientific-Agriculture magazine, 29:2 pp .77-65
19. Behrangi, A., Kuo-lin H., Bisher I.,Sorooshian S., Huffman G. J., Kuligowski R.J., (2009), "PERSIANN-MSA: A Precipitation Estimation Method from Satellite-Based Multispectral Analysis", Journal of Hydrometeorology, Vol. 10, PP..1414-1429.
20. Nazemossadat, M.J. & A. Shirvani, (2005), "Forecast of Winter Precipitation of South Iran by Persian Golf Sea Surface Temperature", Scientific-Agriculture magazine, 29:2 pp .77-65.
21. Zare. F, Ghorbani.F (2014), "Evaluation of Genetic Programming in Estimation of Soil Temperature"Geographical Space magazine, 19-38: (47).
22. Kistler, R. (2001), "The NCEP-NCAR 50 year Reanalysis: Monthly Means CDROM and Documentation", Bulletin of the American Meteorological Society, Vol, 82, No. 2.
23. Ferreira, C. (2001). Gene expression programming: A new adaptive algorithm for solving problems. Complex Systems. 13: 2. 87-129.