Fariborz A.K., Shahram, S,K., Reza, N.A. (2018), Gharehsou River flow prediction using Time Series Theory. pp. 153-161. In Gastescu, P., Bretcan, P. (edit, 2018), *Water resources and wetlands*, 4th International Conference Water resources and wetlands, 5-9 September 2018, Tulcea (Romania), p.312

Available online at http://www.limnology.ro/wrw2018/proceedings.html

4th International Conference Water resources and wetlands, 5-9 September 2018, Tulcea (Romania)

GHAREHSOU RIVER FLOW PREDICTION USING TIME SERIES THEORY

Fariborz Ahmadzadeh Kaleybar [1*], Shahram Shahmohammadi Kalalagh[2], Reza Alipour Nikmanesh[3]

[1*] Department of Water Sciences and Engineering, Tabriz Branch, Islamic Azad University, Tabriz, Iran. Azad University, Tabriz, Iran, Tel: (+98)9144268417, E-mail*: f.ahmadzadeh@iaut.ac.ir*

[2] Department of Water Sciences and Engineering, Tabriz Branch, Islamic Azad University, Tabriz, Iran. Azad University, Tabriz, Iran, Tel: (+98)9141046214, E-mail: *Shahmohammadi_sh@iaut.ac.ir*

[3] Department of Civil Engineering, Ahar Branch, Islamic Azad University, Ahar, Iran, Azad University, Ahar, Iran,Tel: (+98)9144106533, E-mail:*reza.alipour.nikmanesh@gmail.com*

Abstract

In recent decades, due to shortage of water resources, the problem of optimal management of these resources is very important. On the other hand, different patterns, such as regression analysis and time series for predicting water resources, have been commonly used by various researchers .In the present study, with the help of different methods of time series (forecasting and modeling), including trend analysis, models of Holt and Winters and various auto moving average method suggested by Box and Jenkins were fitted to the average monthly discharge (years1973 to 2012) of Samian hydrometric station at the Gharehsou river (northwest of Iran). The accuracy of each method in predicting the average monthly discharge was studied by Akaike information criterion (AIC), root mean square error (RMSE), mean absolute error (MAE) and correlation coefficient (R^2). Trend analysis method with AIC= 1028, RMSE=3.1, MAE=2 and R^2=0.9 is the best method for short-term forecasting and model ARIMA (1,1,1) (2,0,2), with AIC=916, RMSE=2.7, MAE=1.9 and R^2=0.93 is the best method for long-term forecast of Gharehsou average monthly discharge.

Keywords: Gharehsou river, Time series Theory, Holt and Winters method, Box and Jenkins method, Akaike Information Criterion

1 INTRODUCTION

River behavior is one of the most important issues in the design, operation and studies of water engineering. Therefore, the use of modern methods such as available techniques at time series, artificial neural networks, fuzzy logic, genetic programming, chaos theory and other methods for prospecting for prediction of water resources due to their innovation and capabilities have recently received attention. One of the issues that is being considered in this area is the long-term forecasting model of the average monthly discharge of the river. The existence of such models makes decision-making on the exploitation of water resources, especially in droughts, more precise and leads to optimal management.

In hydrology, the average daily, monthly and annual discharge of the river, the level of groundwater in one area, the river water quality (the amount of oxygen dissolved in the river water), etc. can be considered as examples of the time series. Since the time series values are measured at different times, there is a time correlation between the observed time series values, which we can analyze using statistics as well as probabilistic rules governing time dependence. Therefore, the use of time series theory as one of the most common methods for predicting river flows has been proposed. The use of time series in hydrology has begun five decades ago and has reached its peak by providing Box and Jenkins models. Considering the importance of using the theory of time series in water resources engineering and the positive results in helping to study, design and exploit water resources, researchers have been asked to do more research in this regard.

Thomas and Fiering (1962) were among the first to use autoregressive model (AR) for analyzing river flows in different seasons, and examined the trend component in the river discharge time series.

Chow and Kareliotis (1970) analyzed the one-dimensional time series of rainfall and temperature. In this analysis, the existence of severe one-year periodic rotational components and a weak frequency with a six-month period was recognized in the data series.

Perhaps the first important step in the practical application of time series in hydrology was taken by McKerchar and Delleur (1974). According to the seasonal characteristics of the river flow parameters, ARIMA seasonal and multiplicative models were selected to simulate the river. The ARIMA model provides a more precise component of the trend than data extraction and more accurate prediction. Burlando (1996) has used ARIMA models to predict hourly rainfall, and the values obtained are compared with rain measurement data, and it is concluded that with increasing rainfall continuity, predictions are more accurate and shorten the durability Rainfall, the difference in the amount of precipitation predicted more than its actual value.

Komornık et al. (2006) compared the efficiency of hydrological models of time series in the Czech Republic, whose results show the high efficiency of these models in predicting hydrological processes. Damle and Yalcin (2007) predicted flood using time series in the Mississippi River of America. Their results also indicate the ability of time series in generating daily flow data and the accuracy of predicted results. Gorbanpour et al. (2010) used the ARIMA and DARMA time series models to simulate the flow of river flow of the hole rock, a karstic river. The results showed that ARIMA model has a better performance in monthly and weekly modeling of this river. Considering the positive history of time series theory in modeling of stochastic phenomena and the importance of forecasting the flow of Gharehsou river at the entrance to Sabalan Dam, this study has long-term predicted by this method.

2 METHODS

2.1 Case Study

The catchment area of Gharehsou River, which is one of the main sub basins of the Aras River, is up to 5326 km^2 to the site of Sabalan dam construction. The average total annual runoff of the Gharehsou river is 276MCM (million cubic meters), of which 120 MCM are produced by the other dam at the upper reaches of the dam. Of the remaining 155 MCM, 115 MCM are save by Sabalan dam and the rest of runoff is released to the bottom of the dam. The present study was carried out to predict the short-term and long-term average monthly discharge of Gharehsou River at Samian Hydrometric Station at the entrance to Sabalan Dam (Latitude 48-14-48, Latitude 38-22-53). The 39-year-old monthly flow of the river has been used (1973-2012).

2.2 Data Retrieval

Data retrieval is often done in the following ways:

a) Mean of nearby points
b) Median of nearby points
c) Linear interpolation
d) Linear trend at point

During the 468 months of the statistical period (39 years), the samian hydrometric station was only 14 months (dispersed), which is approximately 3%, without the average monthly flow rate, which was then retrieval using the linear trend at point method (the appropriate method for data dispersed with seasonal trend) and ready for use in time series.

2.3 Investigating of Normality and Correlation of Data

The second step in analyzing the time series is to examine the normality of the data. Normality of data is important because time series theory is developed based on the normalization of data. In SPSS software, the standardization and normalization of data can be done in the computation section. One of the methods for data normalization is using the logarithm of data.In this method, the D_{max} value is obtained from the equation 1.

$$D_{max} = sup_{x \in R}(F_n - F) \qquad (Eq. 1)$$

D_{max}: Maximum difference in data from the corresponding value in normal distribution.
F: The cumulative data function in the cumulative distribution chart.
F_n: The value of the normal cumulative function of the data in the normal cumulative distribution chart.

2.4 Investigation of Trend and Its Separation - Analysis of Decomposition

The next step is to determine the trend in the time series and also to remove it in order to static the data. After the data is static, the appropriate models are fitted to the data. The software (SPSS) adjusts a line to the data to determine the trend component. The slope of this line is equal to the trend component. Obviously, if the slope of this line is zero and the line is horizontal, the data is missing the trend and static.

In case of a trend, it is necessary to use the residual analysis method to convert the time series to the static series. For this purpose, the amount of trend function is calculated at any time, and the amount of time series data is deducted from the corresponding data in the trend function and this difference is considered as a new value for the time series (Capilla 2008). The property of this method is that the mean of data is zero and the data trend is also deleted.

2.5 Predictive Methods in Time Series
2.5.1 Box and Jenkins Model

The Box-Jenkins model is a three-step repeat method. The first step is an experimental identification step, which is performed using autocorrelation function (ACF) and partial autocorrelation function (PACF). In the second step or the estimation stage, the model parameters are estimated, and the third stage is the phase of recognition of fit fitness, which at this stage is sufficient and the appropriateness of the experimental identification and estimation of the model is evaluated by statistical tests such as the independence of the model error values. When the final model is obtained after adjustment and correction, we use it to predict the future values of the time series. To determine the type and rank of the model, it should first be judged on the ACF and PACF charts. By drawing a confidence interval, the last consecutive point outside of the probability limit is X perpendicular to it. From this point, the line is entered on the axis and the corresponding delay is read. This delay specifies the model rank. Equations 2 and 3 show the functions of ACF and PACF with delay of k.

$$\rho_k = \frac{\sum_{i=1}^{n-k}(Z_i - \bar{Z})(Z_{i+k} - \bar{Z})}{\sum_{i=1}^{n}(Z_i - \bar{Z})^2} \qquad -1 \leq \rho_k \leq 1 \qquad (Eq. 2)$$

$$\phi_k(K) = \frac{P_k - \sum_{i=1}^{k-1} \phi_i(K-1)\rho_{k-1}}{1 - \sum_{i=1}^{k-1} \phi_i(K-1)\rho_i} \qquad (Eq. 3)$$

Z_i , Z_{i+k}: The values of variables or time series data at time i and in the stage with time delay k

A model for fitting data is selected to be applicable to the following tests.

1. In the PACF chart to select the model rank, a significant difference between the delay value and the zero value and its range from the ±2/√ N ratio is considered.
2. Akaike information criterion (AIC) is also to be considered for modeling, which is explained in the following paragraphs.

Typically Autoregressive Integrated Moving Average (ARIMA), Autoregressive Moving Average (ARMA), Moving Average (MA) and Autoregressive (AR) models can be selected. The reason for the widespread use of this model is its ability to create a correlation between the values of the present time and the earlier times, as well as the simplicity of the structure of these models.

Equations 4 and5 show the functions of AR and MA.

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots \phi_p Z_{t-p} + a_t \qquad (Eq. 4)$$
$$Z_t = a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots \theta_q a_{t-q} \qquad (Eq. 5)$$

$\phi_1, \phi_{2,\ldots} \phi p$: Coefficients and parameters of the AR model

a_t : A random and independent value of time that follows the normal distribution with the mean zero

$\theta_{1,} \theta_2, \ldots \theta_q$: Coefficients and parameters of the MA model

The ARMA model is a combination of two models and the ARIMA model is an integral part of the ARMA model. In most cases, this model is shown as ARIMA (p, d, q), in which p, d, q are real non-negative integers that determine the degree of regression, integrity, and moving average. In its models, the ARIMA correlated with ARIMA (p, d, q), (P, D, Q) the (p, d, q) is used to define non-seasonal orders and (P, D, Q) are used to define seasonal rankings.

2.5.2 Holt and Winters Method

One of the other methods of forecasting is the Holt and Winters method. Using this method, it is easy to extend the representation to series that include trends and seasonal variations. The Holt and Winters model consists of two linear smoothing (single and doble) models that are known for the Holt method and the Winters method. The Winters method can be used for short-term forecasting as well as for medium-term forecasts. This procedure provides dynamic estimates of the components of the trend, level, and seasonal. The smoothing equations are as follows (Eq.6, 7 , 8):

$$\bar{X}_t = a(\bar{X}_{t-1} + T_{t-1}) + (1-a)\frac{X_t}{F_{t-s}} \qquad (Eq.6)$$

$$T_t = \beta T_{t-1} + (1-\beta)(\bar{X}_t - \bar{X}_{t-1}) \qquad (Eq.7)$$

$$F_t = \gamma F_{t-s} + (1-\gamma)\frac{X_t}{\bar{X}_t} \qquad (Eq.8)$$

That;
Tt: Trend
Xt: The series levels at time t (are real values that are seasonal and trendy)
Ft: Seasonal component
α: A fixed value of 1> α>0

2.5.3 Trend Model and Decomposition Model

The trend is the long-term changes in the mean of time series. In other words, the natural course of the long-term series is called a trend. In this case, the time slots are ignored and the overall view is taken into consideration. Studying data over a long period of time can provide a general idea of the behavior of the phenomenon that helps predict the future. The trend modeling is done in four linear, 2degree, exponential, and S curve methods. That linear and quadratic approach is more practical. The equation 10, 11, 12, 13 shows the linear, second degree, exponential and S curve method.

$Y_t = b_0 + b_1 t + e_t$ (Eq.10)
$Y_t = b_0 + b_1 * t + b_2 t^2 + e_t$ (Eq.11)
$Y_t = b_0 * b_1^t * e_t$ (Eq.12)
$Y_t = 10^a / (b_0 + b_1 b_2^t)$ (Eq.13)

That;
b: represents the average change from one period to the next and e_t: is the model error section.

The process decomposition model varies with the trend model. This model has a linear periodic pattern. In the sense that it can be estimated without error from the past. Analysis of the residues is performed after the trend is removed from the process.

2.6 Verification and Validation of Models

This step involves proving that the selected model works correctly for known conditions. The results of the model during the simulation period should be compared with the data set. If the acceptable level of compatibility between the data and the simulation results of the model is reached, the model's accuracy can be verified at least for the conditions defined by the verification step data set. In the event that at least acceptable adaptation is not achieved, then analysis should be performed to identify possible reasons for the revision and refinement of the model for the differences between the data and the simulation results of the model.
In order to investigate and determine the best prediction model of time series and validation, the Akaike information criterion (AIC)(Equation 14) and root mean squared errors(RMSE) (Equation 15), mean absolute errors(MAE) (Equation 16) and the regression coefficient (R2) (Equation 17) can be used. The corresponding equations to each one are given below (Barnston 1992).

$$AIC = 2K + n[Ln(RSS/n)] \qquad (Eq.14)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Z(x_i) - Z^*(x_i))^2}{n}} \qquad (Eq.15)$$

$$MAE = \frac{\sum_{i=1}^{n}|Z(x_i) - Z^*(x_i)|}{n} \qquad (Eq.16)$$

$$R^2 = \left[\frac{\sum_{i=1}^{n}(Z(x_i) - \overline{Z(x_i)})(Z^*(x_i) - \overline{Z^*(x_i)})}{\sqrt{\sum_{i=1}^{n}(Z(x_i) - \overline{Z(x_i)})^2 \sum_{i=1}^{n}\left(Z^*(x_i) - \overline{Z^*(x_i)}\right)^2}}\right]^2 \qquad (Eq.17)$$

In which
RSS: Sum of squares remaining

K: Number of model parameters
N: Number of data
Z: Estimated values
Z*: Observation Values

2.7 Prediction Model

After selecting the model for the AIC criterion and the error estimation, the best models are selected and the end section of the modeling process is predicted. At this stage, fitted patterns are used to predict the future of the series. Short-term forecasts are seasonal and long-term forecasts for the next 10 years.

3 RESULTS AND DISCUSSION

The first step in analyzing time series after data reconstruction is the graphical drawing of that series. Figure 2 shows the time series graph of the mean monthly discharge of Gharehsou river at the Samian hydrometric station during the studied period.

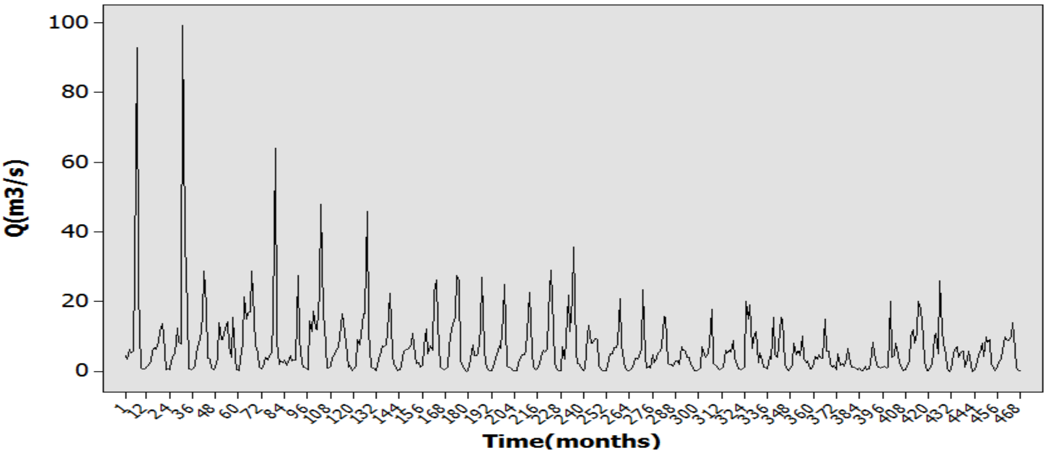

Figure 1. Monthly mean discharge time series of Gharehsou river at Samian hydrometry station

Regarding the ability of SPSS software in the computing part, this software was used to perform normalization and normalization tests. Figure 3 shows the graph of the normal test. The logarithm standardization method has been used to test the normality.

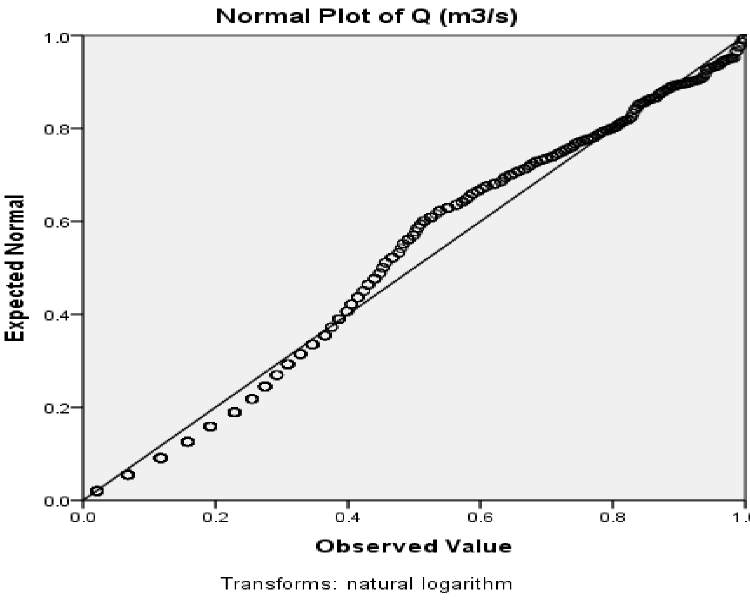

Figure 2. Diagram of the normal test of monthly discharge data of Samian hydrometry Station

Linear trend analysis is used in this study. The software for fitting the trend component, fit a line to the data.

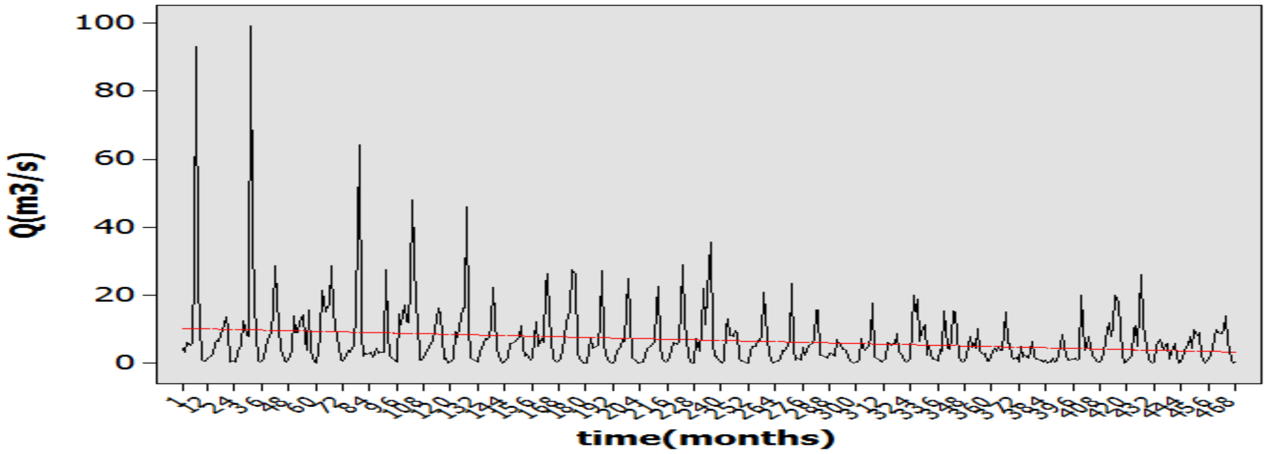

Figure 3. Graph of linear trend analysis at Samian hydrometry station

The slope of this line is equal to the trend. According to Figure 4, the slope is negative, so the trend for this series is a downward trend. Obviously, if the slope of this line is zero and the line is horizontal, the data is missing the trend component and static.

Using the residual analysis method, the time series of the average Gharehsou river was estimated to be static conversion and modeling (Fig. 4). The property of this method is that the mean of the data is zero (the horizontal line of Trend represents this), in other words, the slope is equal to zero (that is, the trend line, which has a negative slope in Fig.4, has been turned into a horizontal line)and the data trend is also eliminated.

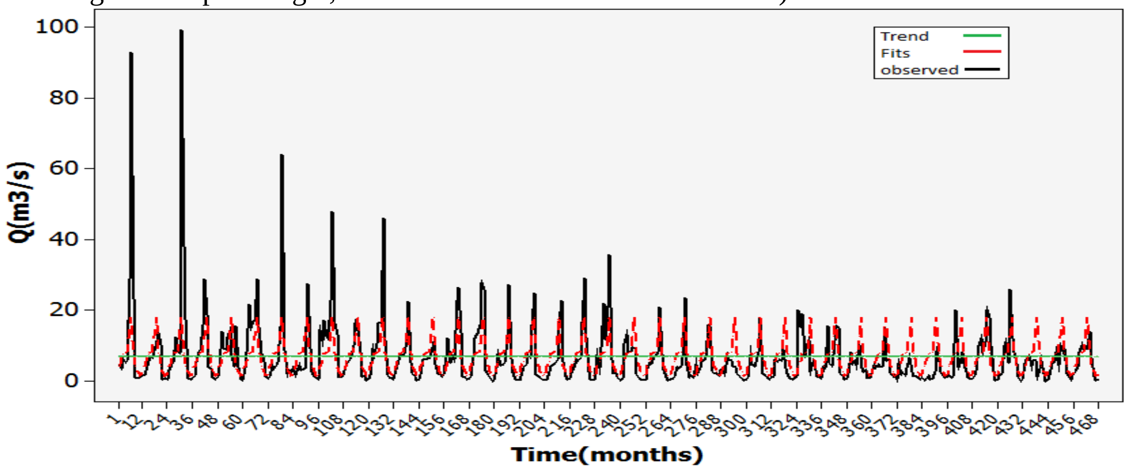

Figure 4. Analysis of residuals after deletion of trend component

Figure 5 depicts the diagram of the Winters method. The Winters method can be used for short-term forecasting as well as for medium-term forecasts.

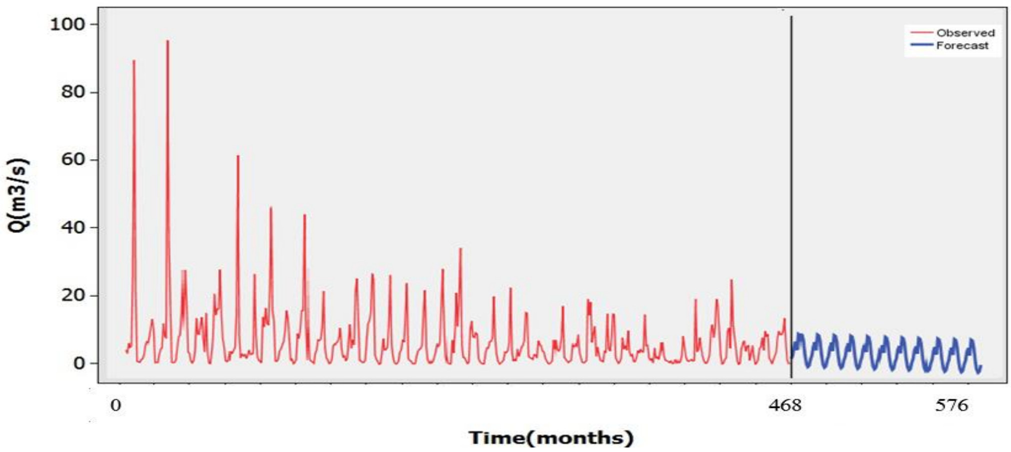

Figure 5. Chart of Winters method for short-term forecasting of the Samian station discharge

In Box and Jenkins method, first, the ACF and PACF charts were used to identify the type and order of the model (Fig. 6 and 7).

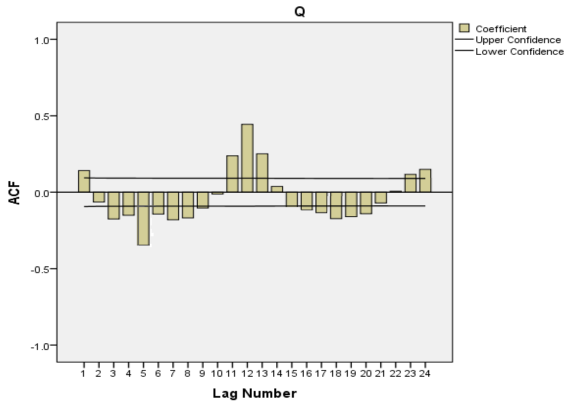

Figure 6. ACF Chart for Samian station discharge

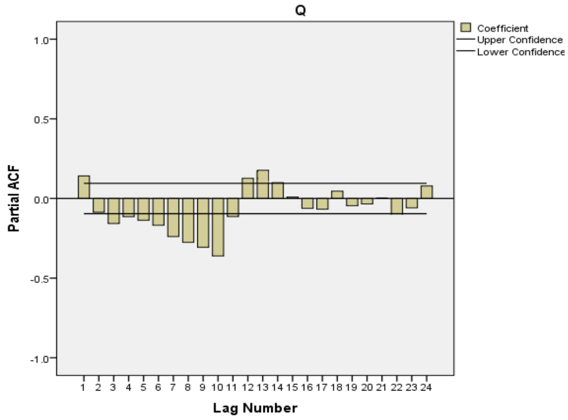

Figure 7. PACF Chart for Samian station discharge

The SPSS model provides the best predictive models of ARIMA based on the ACF and PACF results through the output filter screen and the Best-fitting models option. Sometimes it only shows a model and sometimes dozens of models for prediction. In these studies, the SPSS model predicts 20 models for ARIMA.

Validation is performed for **25%** of the existing data (the last 10 years) and the forecast for the next 10 years is considered.Table 1 Shows the AIC, RMSE, MAE and R2 between actual and predicted values of generated models output of trend curve, Winters and the Box and Jenkins (ARIMA) methods.

Model	(p,d,q)	(P,D,Q)	R^2	RMSE	MAE	AIC
ARIMA	(1,1,1)	(1,0,1)	0.93	2.7	1.9	875.5
ARIMA	(1,1,0)	(1,0,1)	0.92	2.8	1.9	955.5
ARIMA	(1,1,1)	(2,0,2)	0.93	2.7	1.9	916.0
Trend curve	–	–	0.90	3.1	2.0	1028.0
Holt & Winters	–	–	0.89	4.1	2.7	1064.7

Table 1. Criteria for comparison between different models of prediction of Samian hydrometric station

Three models of ARIMA that They have the lowest AIC, RMSE , MAE and high R2 (shown in the table1) have been selected as superior models of Box and Jenkins methods, and given that the model ARIMA(1,1,1) (1.0,1) considers the negative average for the average monthly discharge, this model can not be used for this parameter (Q). Considering that, according to the researchers, the AIC is a very suitable criterion for assessing the box and Jenkins models [12], model ARIMA(1,1,1) (2,0,2) has the lowest AIC, so this model has been selected as the best predictor of long-term average monthly discharge of Gharehsou river. Figures 8 and 9 show the ARIMA (1.1.1) (2.0.2) validation and forecasting model.

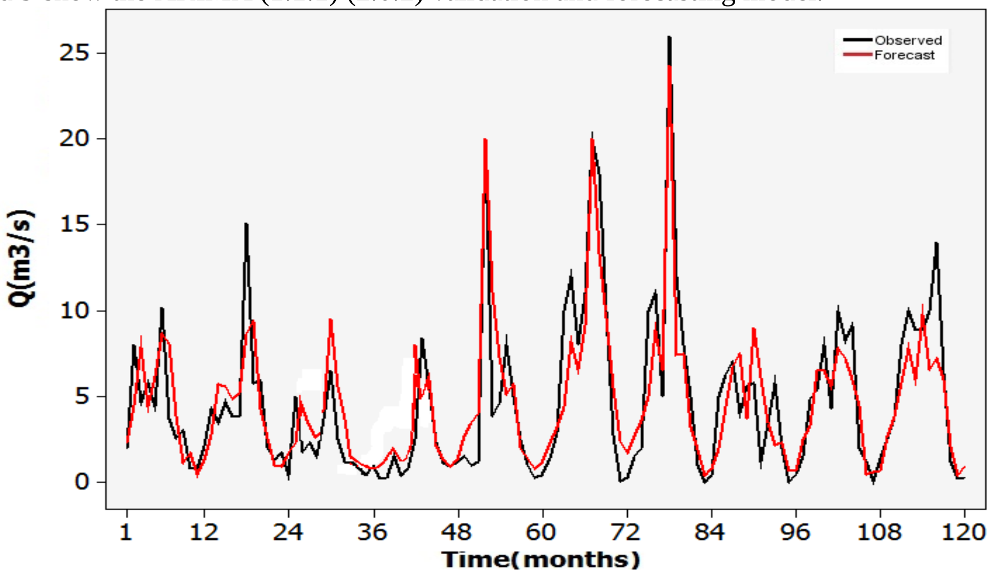

Figure 8. Validation chart of ARIMA (1,1,1) (2,0,2) model (10 years old)

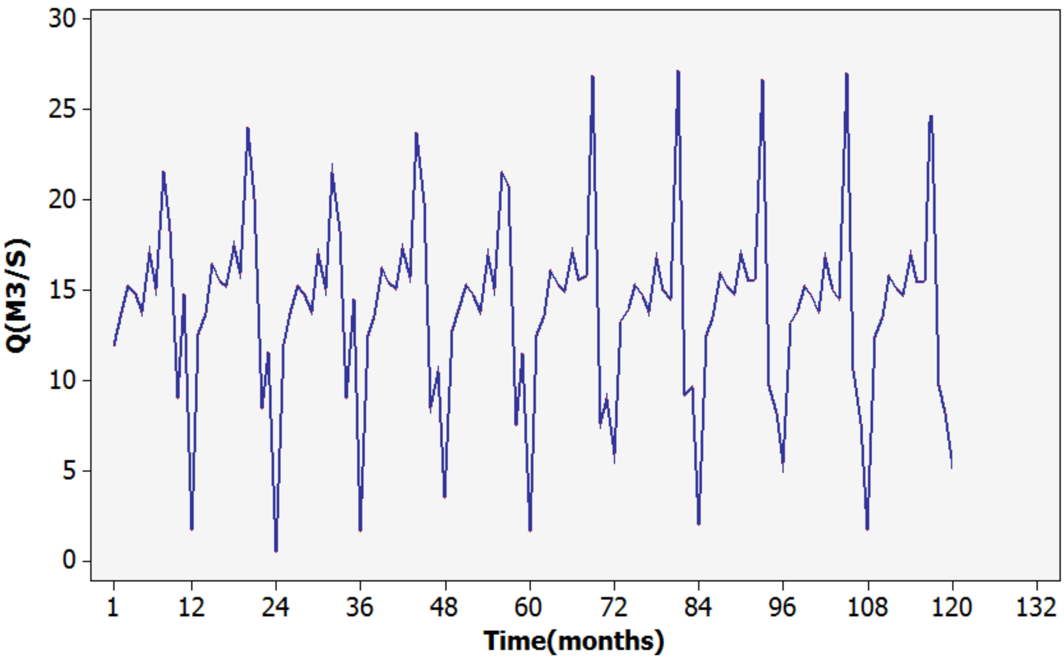

Figure 9. Forecast chart of ARIMA(1,1,1) (2,0,2) model(10 years later)

According to table 1, the trend curve method has the lowest AIC, RMSE, MAE, and high R^2 criteria more than the Winters method. Accordingly, according to the researchers' recommendation, the trend curve method for the extraction of the seasonal short-term prediction was selected.

4 CONCLUSIONS

The prediction of the flow into the storage dams is one of the most important issues in their planning and optimal management for the production of hydroelectric power and the allocation of water to resource consumption. The prediction of hydrological variables is a very efficient tool for managing water resources. On the other hand, The use of concepts governing time series is highly evaluated in hydrological forecasting. In this research, for estimating the discharge rate to the Sabalan Dam, the average monthly discharge data of Samian hydrometric station located on Gharehsou River was used from 1973 to the year 2012 (39 years). Different methods of prediction and modeling in time series including trend analysis, Winters' method, and different ARIMA models recommended by Box and Jenkins were fitted to the data and verified. The summary of the results is as follows:

1. In the box and jenkins method, both the function is in the form of a sinusoidal wave, with the ACF and the PACF (in order to identify the type and order of the model). Therefore, the component of the seasonal trend is available. Therefore, in the time series of the average monthly discharge of the Ghareh Souz river, a differential separation method is used for the removal of the non-resident, so the component d must exist at one of the seasonal and non-seasonal levels of the models.
2. The amount of PACF is reduced after 2 delays and is within the confinement range. In the case of ACF, high mutations are observed at initial delays, which decrease after several delays. Therefore, the rank of the model does not exceed 2.
3. Three models ARIMA (1,1,1) (2,0,2), ARIMA (1,1,1) (1,0,1) and ARIMA (1,1,0) (1,0,1), which has the lowest AIC, RMSE, MAE and high R2, have been selected as superior models of Box and Jenkins methods, and given that the model (1.0,1) (1.1 , 1) ARIMA considers negative values for forecasting average monthly discharge, this model can not be used to predict discharge. Considering that, according to researchers, the AIC is a very suitable criterion for evaluating box and Jenkins models, ARIMA (2.0,2) (1,1,1) has the lowest AIC value, Therefore, this model was selected as the best method for predicting long-term average monthly discharge of Gharehsou river.
4. According to the table (1-4), the method for the extraction of the trend curve has a lower AIC, RMSE and MAE, and higher R2 than the Winters method. Accordingly, according to the researchers' recommendation, the method for the extraction of the trend curve for short prediction Seasonal selected.

5 REFERENCES

Barnston, A.G. (1992). *Correspondence among the Correlation, RMSE, and Hiedke Forecast Verification Measures, Refinement of the Heidke Score*. Climate Analysis Center. Washington D.C, 11pp.

Burlando P., Montana A. and Raze R. (1996), *Forecasting of storm rainfall by combined use of radar, rain gages and liner models,* Atmospheric Research, **42**: 199-216.

Capilla, C. (2008). *Time series analysis and identification of trends in a Mediterranean urban area. Global and Planetary Change,* **63**: 275–281.

Chow, W.T. and S.J. Kareliotis. (1970). *Analysis of Stochastic Hydrologic Systems*. Water Resources Research, **16**: 1569-1582.

Damle, C. and A. Yalcin. (2007). *Flood prediction using time series data mining*. Journal of Hydrology. **333, (2-4)**: 305-316.

Ghorbanpoor M. A., Abbaspoor K., Jalalvand G. and Ashtiani Moghadam G. (2010), *Stochastic modeling of surface stream flow at different time scales: Sangsoorakh karst basin, Iran.* Journal of Cave and Karst studies. **72**: 1-10.

Komornık, J. Komornıkova, M. Mesiar, R. Szokeova, D. and J. Szolgay. (2006). *Comparison of forecasting performance of nonlinear models of hydrological time series*. Physics and Chemistry of the Earth, **31**: 1127–1145.

McKerchar, A.I. and L.W. Delleur. (1974). *Application of seasonal parametric linear stochastic models to monthly flow data.* J. Water resource reservoir, **10**: 246-255.

Thomas, H.A. and M.B. Fiering. (1962). *Mathematical synthesis of stream flow sequences for the analysis of river basin by simulation*. Harward University Press. Cambridge, 751pp.