Streamflow forecasting of Astore River with Seasonal Autoregressive Integrated Moving Average model

Rana Muhammad Adnan
Xiaohui Yuan
School of Hydropower and Information Engineering, Huazhong University of Science & Technology, 430074 Wuhan, China
Ozgur Kisi
Center for Interdisciplinary Research, International Black Sea University, Tbilisi, Georgia
Yanbin Yuan
School of Resource and Environmental Engineering, Wuhan University of Technology, China

doi: 10.19044/esj.2017.v13n12p145 URL:http://dx.doi.org/10.19044/esj.2017.v13n12p145

Abstract

Simulation of streamflow is one of important factors in water utilization. In this paper, a linear statistical model i.e. Seasonal Autoregressive Integrated Moving Average model (SARIMA) is applied for modeling streamflow data of Astore River (1974 – 2010). On the basis of minimum Akaike Information Criteria Corrected (AIC_c) and Bayesian Information Criteria (BIC) values, the best model from different model structures has been identified. For testing period (2004-2010), the prediction accuracy of selected SARIMA model in comparison of auto regressive (AR) is evaluated on basis of root mean square error (RMSE), the mean absolute error (MAE) and coefficient of determination (R^2). The results show that SARIMA performed better than AR model and can be used in streamflow forecasting at the study site.

Keywords: Streamflow forecasting; Astore River; SARIMA; AR

Introduction

Streamflow forecasting is a key step in planning of water projects, irrigation systems, hydropower system and optimize utilization of water resources (Zhang et al., 2011). Due to continuous increase of population growth, industrial uses and irrigation needs, the streamflow forecasting has received great attentions of researchers for operational River management (Xu et al., 2014). The importance of water measurement compelled

researchers to apply various types of forecasting models to estimate and forecast streamflow. These models consist of rainfall-runoff model, lumped conceptual models, black box models and stochastic models (Bahremand and De Smedt, 2010; Tayyab et al., 2016; Tayyab et al., 2015; Tingsanchali and Gautam, 2000; Wang, 2006).

Seasonal autoregressive integrated moving average (SARIMA) models have been used by various researchers for modeling different variables in hydrology. Rabenja et al. (2009) applied both stochastic models to forecast monthly precipitation and discharge of Namorona River in Madagascar. They concluded that the SARIMA model is more preferable for runoff forecast. Otok (2009) forecasted runoff data in Indonesia by applying SARIMA model in comparison of autoregressive integrated moving average (ARIMA) and transfer function model (TFM) statistical models. They suggested that SARIMA stochastic models perform better than ARIMA models and TFM in forecasting streamflow. Mirzavand and Ghazavi (2015) applied SARIMA and AR models to forecast groundwater levels in semi-arid environment. He compared the results of both models with ARIMA model. Psilovikos and Elhag (2013) used the Seasonal ARIMA model in comparison of non-seasonal ARIMA model for forecasting daily evapotranspiration over Nile delta region. Valipour (2015) in another research used SARIMA and ARIMA stochastic models to analyze the streamflow in different districts of United States. He used the annual flow data of Rivers and concluded that SARIMA models show better performance than ARIMA models. Dastorani et al. (2016) applied AR model in comparison of ARIMA, moving average (MA) and auto regressive moving average (ARMA) models for predicting monthly rainfall. Ghanbarpour et al. (2010) applied the two stochastic models by analysing karstic river flow in the Sansoorakh karst drainage basin. His study results showed that SARIMA models perform better than deseasonalized ARIMA models

Methodology

A stochastic model explains the probability structure of sequences of observation. Box and Jenkins developed ARIMA stochastic models that describe a wide class of models forecasting a univariate time series that can be made stationary by applying transformations – mainly differences for Trend and Seasonality, and power function to regulate the variance (Box and Jenkins, 1970; Box and Jenkins, 1976; Box et al., 1967) The model, word "ARIMA" consists of three terms i.e. i) AR ii) I and iii) MA terms. Lags of differenced time series in the forecasting equations are called "autoregressive(AR)" term, whereas lags of the forecasted errors are called "moving average (MA)" term and the time series which requires

differenceation to become stationary should be "Integrated (I)" (Ghafoor and Hanif, 2005).

Seasonal ARIMA Model

Seasonal ARIMA model, which are commonly known as seasonal autoregressive integrated moving average (SARIMA) models are used to deal with seasonality (Reinsel et al., 1994). the SARIMA model can be explained as ARIMA(p, d, q)(P, D, Q)$_L$, where "p" represents the non seasonal autoregressive term, "q" represents the non seasonal moving average term, "d" represents the non seasonal differencing terms and (P, D, Q)$_L$ represents the seasonal auto regressive, seasonal moving average and seasonal difference terms respectively. The general form of Seasonal ARIMA $(p,d,q)(P,D,Q)_L$ model can be written as follows(Shunway and Stoffer 2001) by using the backshift operator ($B^n(S_t)= S_{t-n}$):

$$\Phi_{NAR}(B)\,\Phi_{SAR}(B^L)(1-B)^d\,(1-B^L)^D\,S_t = \theta_{NMA}(B)\,\theta_{SMA}(B^L)\,e_t \quad (1)$$

Whereas the $\Phi_{NAR}(B)$, $\Phi_{SAR}(B^L)$, $\theta_{NMA}(B)$ and $\theta_{SMA}(B^L)$ parameters can be expressed in detailed form as:

$$\Phi_{NAR}(B) = 1- \Phi_{1NAR}(B) - \ldots - \Phi_{pNAR}(B^p) \quad (2)$$

$$\Phi_{SAR}(B^L) = 1- \Phi_{1SAR}(B^L)- \ldots - \Phi_{pSAR}(B^{pL}) \quad (3)$$

$$\theta_{NMA}(B) = 1- \theta_{1NMA}(B) - \ldots - \theta_{qNMA}(B^q) \quad (4)$$

$$\theta_{SMA}(B^L) = 1- \theta_{1SMA}(B^L) - \ldots - \Phi_{qSAR}(B^{qL}) \quad (5)$$

Where L= sesonality lag, Φ_{NAR} = non seasonal autoregressive parameter, Φ_{SAR} = seasonal autoregressive parameter, θ_{NMA} = non seasonal moving average parameter, θ_{SMA} = seasonal moving average parameter, D = seasonal difference, d = non seasonal difference, S_t = Streamflow at time t.

AR Model

ARs models started to predict a time series in the start of 19th century when Yule introduced first AR model to predict wolfer's sunspot data in 1927. The auto regressive (AR) of order p can be given as;

$$S_t= \Phi_1 S_{t-1}+ \Phi_2 S_{t-2} + \Phi_3 S_{t-3} + \ldots\ldots\ldots + \Phi_p S_{t-p} + e_t \quad (6)$$

Where Φ_1,Φ_2,Φ_3 and Φ_p are the coefficient of AR model , e_t indicates the error term at time period t and S_t refers the value of forecasted streamflow at time period t.

The Box –Jenkins Stochastic models building methodology

Box & Jenkins linear stochastic models building is based on three steps i.e. Identification, Estimation and Diagnostic check (Box and Jenkins 1976; (Box and Jenkins, 1976; Mishra and Desai, 2005; Modarres, 2007). Identification stage involves two steps. In the first step, time series is analyzed for stationarity in "mean" and "variance". If the variance is not stable, it can be made stable by power transformation i.e. log transformation for lambda=0. Appropriate seasonal or non seasonal differencing of the series and sometimes both seasonal and non seasonal differencing is performed to obtain stationarity and normality. Second step of the identification require the autocorrelation function (ACF) and partial autocorrelation function (PACF) after applying seasonal and non seasonal difference. The ACF is a useful tool to measure the relation of earlier value on later values where PACF measures the amount of correlation between a variable and a lag of itself. The information obtained from both correlation functions is used to determine an initial guess for the non seasonal p, q parameters and seasonal P, Q parameters (Durdu, 2010).

Study area

The Astore River basin (Figure 1) is located in Northern Pakistan. Astore River basin is situated in the high mountains of Hindukush-Karakoram-Himalaya (HKH) region. The whole data set was divided into two periods; implementation period and testing period. The implantation period covered the data values from 1974 to 2003 and has been used for building of SARIMA models. The testing period covers the streamflow data values from 2004 to 2010 and has been used to evaluate the performance of both selected models.

Fig. 1 Astore River Basin in Pakistan map

Results and discussion

The first step is to check whether the monthly streamflow time series is stationary and has seasonality. Monthly streamflow data shows that there is a strong seasonal pattern and it is not stationary. The summary of statistical indexes of the monthly streamflow time series is shown in table 1. The historical data of streamflow of Astore River showed positive skewness (i.e. 1.407). In addition, the data table indicates that testing period extremes (minimum value and maximum value) are within the range of the implementation period, which helps for better model prediction performance.

Table 1. statistical summary of streamflow time series

Duration	Min. value	Max. value	Mean value	variance	Standard deviation	Coefficient of kurtosis	Coefficient skewness
1974-2010 (whole data set)	19.34	654.9	136	19768.8	140.6	1.151	1.407
1974-2003 (implementation period)	19.34	654.9	135	20029.2	141.5	1.286	1.453

2004-2010 (testing period)	29.63	612.3	140.2	18858.2	137.3	0.641	1.219

Seasonal ARIMA models

The plot of the monthly streamflow time series shows that it requires a seasonal difference to obtain stationarity in mean and log transformation to make stable variance. ACF, PACF graphs can be obtained after applying seasonal difference to determine the p, q, P and Q parameters for SARIMA model. The streamflow data have a strong seasonal pattern and also non-stationarity in mean, so seasonal differencing was performed. The seasonally differenced monthly streamflow data with ACF and PACF plots is shown in figure 2. According to the plots of the seasonally differenced streamflow data, spikes can been seen in the ACF plot at lags 12, whereas in the PACF plot, the spike can been seen at lag12 and lag 24. These plots suggest the seasonal AR(2) and MA (1) term. There is only one significant spike in the PACF plot at the non seasonal lags, whereas the pattern of the ACF plot indicates three significant spikes. So the non seasonal lags of both plots suggested a possible AR (2) and MA (3) term.

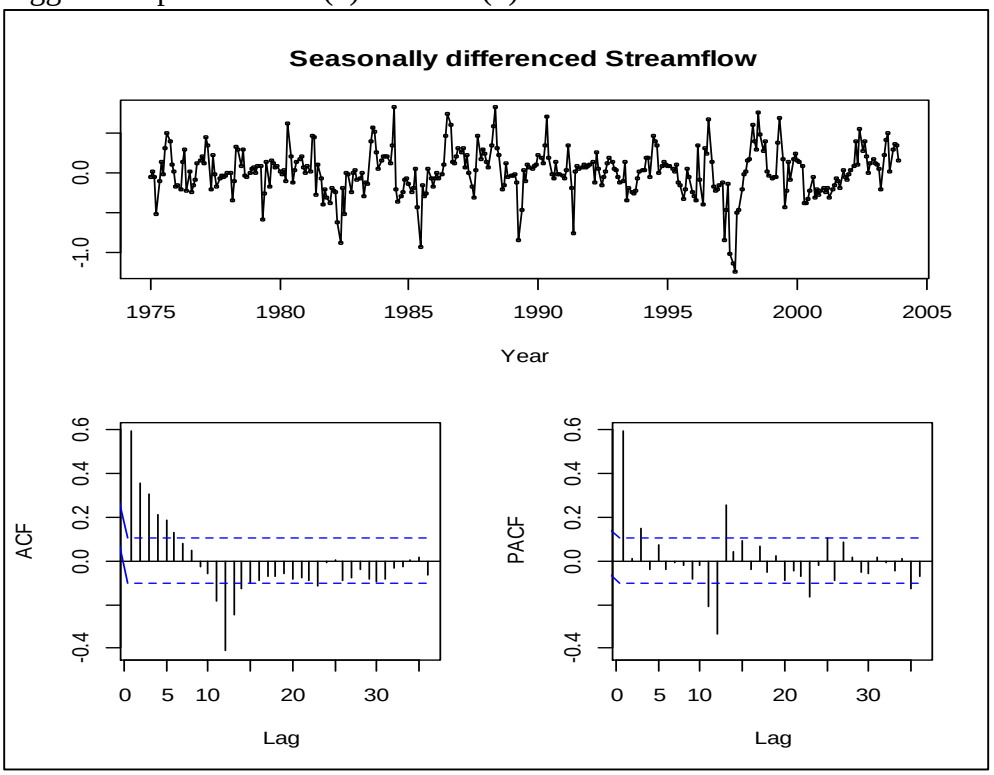

Fig.2 ACF, PACF graphs after seasonal difference

Table 2. AIC_c and BIC values for different SARIMA model structures

Model Structure	AIC_c	BIC
$ARIMA(2,0,1)(2,1,1)_{12}$	-1226.86	-1200.22
$ARIMA(2,0,0)(2,1,0)_{12}$	-1187.61	-1168.52
$ARIMA(2,0,2)(2,1,0)_{12}$	-1204.87	-1178.23
$ARIMA(1,0,1)(2,1,0)_{12}$	-1191.28	-1172.19
$ARIMA(2,0,1)(1,1,0)_{12}$	-1181.31	-1162.23
$ARIMA(2,0,2)(0,1,1)_{12}$	-1238.66	-1215.8
$ARIMA(2,0,3)(0,1,2)_{12}$	-1237.39	-1203.25
$ARIMA(1,0,2)(1,1,1)_{12}$	-1238.43	-1215.56
$ARIMA(2,0,2)(2,1,1)_{12}$	-1226.86	-1200.22
$ARIMA(1,0,2)(0,1,2)_{12}$	-1237.81	-1214.95
$ARIMA(2,0,1)(1,1,2)_{12}$	-1225.31	-1198.67
$ARIMA(1,0,2)(1,1,2)_{12}$	-1238.59	-1211.95

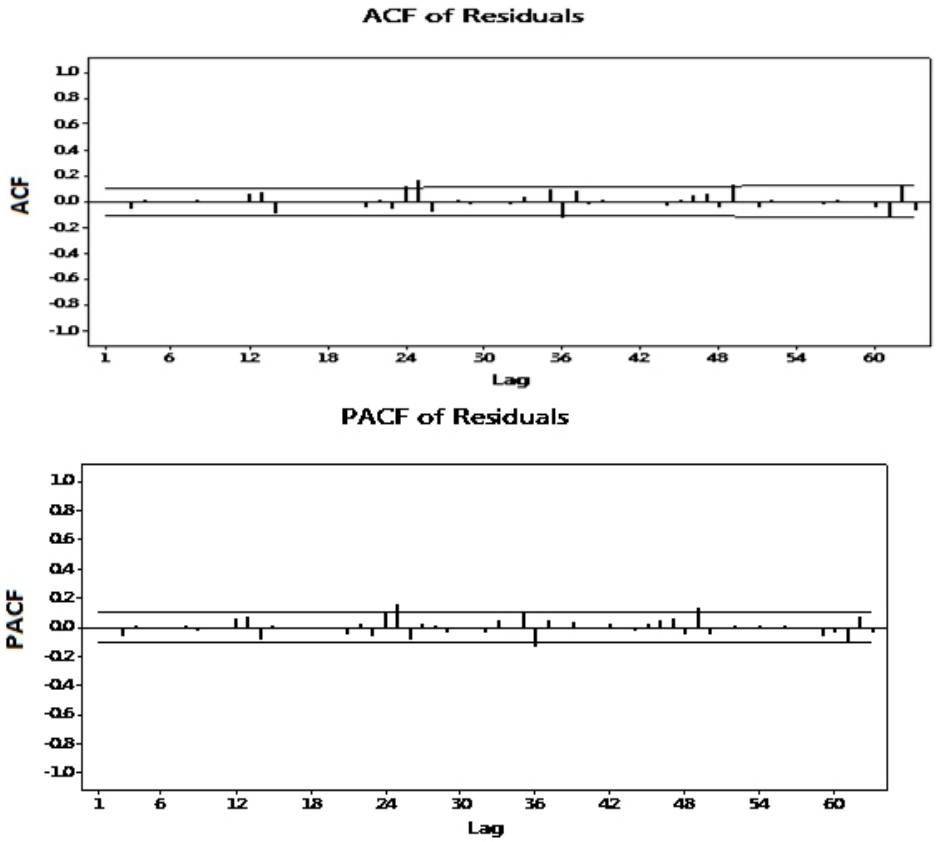

Fig.3 Residual graphs of ACF,PACF of SARIMA model

Consequently, in the identification stage, the possible model for this monthly streamflow time series is SARIMA (2, 0, 3)$(2,1,1)_{12}$. This model is fitted, its parameters are estimated. As the estimation of p, q, P and Q by ACF and PACF is based on empirical data, other model thus with values of

p, q, P and Q around the empirical estimated – need also to be considered, to determine which model is most suitable to represent the time series. The best model structure is selected on basis of the minimum value of AIC_c and BIC i.e SARIMA $(2,0,2)(0,1,1)_{12}$. The values of AIC_c and BIC of different model structures are shown in table 2. After selecting the best model structure, the diagnostic checks are performed on this model. The residuals of ACF and PACF from this model are shown in figure 4. It can be seen that mostly residuals are uncorrelated; few spikes are significant at higher lags.

The selected model also passed the second test of diagnostic check. The value of the probabilities (p) in the Ljung Box test shows that the residuals have no remaining autocorrelations. The summary of this test is listed in table 3. The model passed all required checks and then used to forecast the next 7 years, i.e. testing period. From eq (1), in back shift notation, SARIMA $(2,0,2)(0, 1, 1)_{12}$ model can be written as

$$[1- \Phi_{1NAR}(B)- \Phi_{2NAR}(B)-B^{12}+ \Phi_{1NAR}(B^{13})+ \Phi_{2NAR}(B^{14})]S_t = [1- \Phi_{1NMA}(B)- \Phi_{2NMA}(B^2)- \Phi_{1SMA}(B^{12})+ \Phi_{1NMA}\Phi_{1SMA}(B^{13})+ \Phi_{2NMA}\Phi_{1SMA}(B^{14})]e_t. \quad (7)$$

By substituting the coefficients, one obtains the model;

$$St-0.5262St-1-0.3328St-2-St-12+0.5262St-13+0.3328St-14 = et+0.0934et-1+0.3737et-2+0.7564t-12-0.0706et-13+0.2827et-14. \quad (8)$$

Table 3 Summary of Ljung Box test for SARIMA Model

lag	Lag 12	Lag 24	36	48
Chi square	2.90	15.00	38.80	46.40
Df	6	18	30	42
p-value	0.823	0.664	0.130	0.297

Model Performance Evaluation

In order to evaluate the performance of the selected SARIMA model in comparison of AR model, one month ahead forecasts were generated for the testing period from January 2004 to December 2010. By examining the ACF and PACF graphs, the order P of AR model is 2.Thus AR (2) model is used for predicting monthly streamflow data. The statistical indexes used for this purpose are root mean square error (RMSE), the mean absolute error (MAE), the mean absolute percentage error (MAPE), the Nash efficiency (NE) and coefficient of determination (R^2). They can be defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum (S_O - S_f)^2} \quad (9)$$

$$MAE = \frac{1}{N}\sum \left|S_o - S_f\right| \quad (10)$$

$$R^2 = \left[\frac{\sum (S_o - \overline{S_o})(S_f - \overline{S_f})}{\sqrt{\sum (S_o - \overline{S_o}) \sum ((S_f - \overline{S_f})^2)}} \right] \quad (11)$$

Where N is the total number of observations, S_o is observed flow, S_f is forecasted streamflow,

$\overline{S_o}$ is average of streamflow and $\overline{S_f}$ is average forecasted flow.

The summary of all the statistical indices applied to evaluate prediction performance of both models are shown in table 4. Figure 4 and 5 shows the hydrographs of original streamflow data and forecasted streamflow data by SARIMA and AR models, respectively. It can be seen from hydrographs that SARIMA model forecasted streamflow are in better in line with the original streamflow than the forecasted streamflow of AR model and also having large value of coefficient of determination. SARIMA models also provide less errors values of forecasted streamflow with respect to original streamflow.

Table 4. Evaluation of model performance for testing period on basis of statistical indexes

Statistical Index	SARIMA $(2,0,2)(0,1,1)_{12}$	AR (2)
MAE	22.414	24.651
RMSE	42.765	47.572
R^2	0.903	0.884

Conclusion

Streamflow forecasting is a vital component of planning and management of water resources. In this study, SARIMA applied for streamflow forecasting of Astore River in northern Pakistan. The performance of the both stastistical models is compared by generating one-month-ahead forecast for testing period from 2004 to 2010. The results shows that SARIMA model perform better than AR model due to having less value of error and greater value of similarity index. It can be concluded that SARIMA model can be for modeling streamflow data at this site.

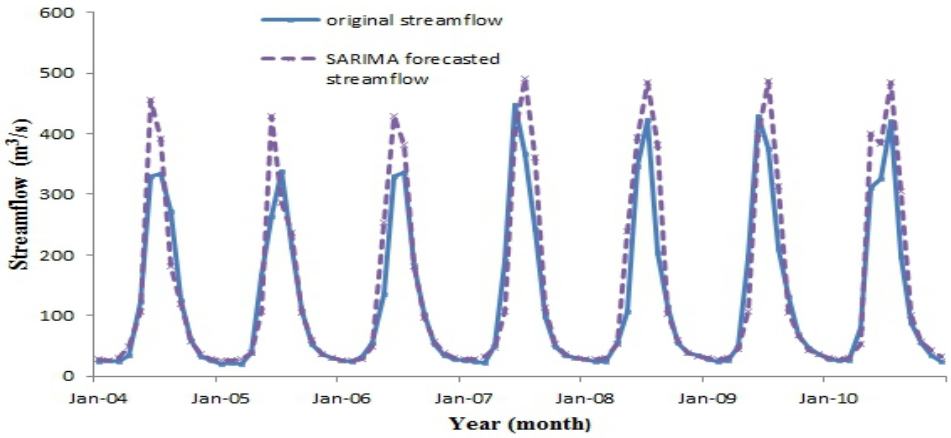

Fig.4 Observed and forecasted streamflow hydrograph using SARIMA model for testing period

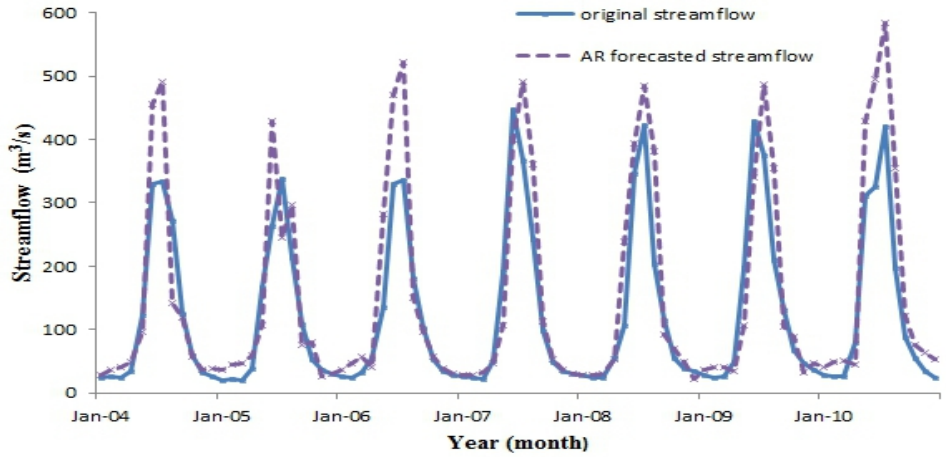

Fig.5 Observed and forecasted streamflow hydrograph using AR model for testing period

References:

1. Bahremand, A., and De Smedt, F. (2010). Predictive analysis and simulation uncertainty of a distributed hydrological model. *Water resources management* 24, 2869-2880.
2. Box, G., and Jenkins, G. (1970). (1970). Time series analysis; forecasting and control. Holden-Day, San Francisco(CA).
3. Box, G. E., and Jenkins, G. M. (1976). Series analysis forecasting and control. *Holden-Day, San Francisco*, 575.
4. Box, G. E., Jenkins, G. M., and Bacon, D. (1967). "MODELS FOR FORECASTING SEASONAL AND NON-SEASONAL TIME SERIES." DTIC Document.
5. Dastorani, M., Mirzavand, M., Dastorani, M. T., and Sadatinejad, S. J. (2016). Comparative study among different time series models

applied to monthly rainfall forecasting in semi-arid climate condition. *Natural Hazards* 81, 1811-1827.

6. Durdu, Ö. F. (2010). Stochastic approaches for time series forecasting of boron: a case study of Western Turkey. *Environmental monitoring and assessment* 169, 687-701.
7. Ghafoor, A., and Hanif, S. (2005). Analysis of the trade pattern of Pakistan: Past trends and future prospects. *Journal of Agriculture & Social Sciences* 1, 346-349.
8. Ghanbarpour, M. R., Abbaspour, K. C., Jalalvand, G., and Moghaddam, G. A. (2010). Stochastic modeling of surface stream flow at different time scales: Sangsoorakh karst basin, Iran. *Journal of Cave and Karst Studies* 72, 1-10.
9. Mirzavand, M., and Ghazavi, R. (2015). A stochastic modelling technique for groundwater level forecasting in an arid environment using time series methods. *Water Resources Management* 29, 1315-1328.
10. Mishra, A., and Desai, V. (2005). Drought forecasting using stochastic models. *Stochastic Environmental Research and Risk Assessment* 19, 326-339.
11. Modarres, R. (2007). Streamflow drought time series forecasting. *Stochastic Environmental Research and Risk Assessment* 21, 223-233.
12. Otok, B. W. (2009). Development of rainfall forecasting model in Indonesia by using ASTAR, transfer function, and ARIMA methods. *European Journal of Scientific Research* 38, 386-395.
13. Psilovikos, A., and Elhag, M. (2013). Forecasting of remotely sensed daily evapotranspiration data over Nile Delta region, Egypt. *Water resources management* 27, 4115-4130.
14. Rabenja, A. T., Ratiarison, A., and Rabeharisoa, J. (2009). Forecasting of the Rainfall and the Discharge of the Namorona River in Vohiparara and FFT Analyses of These Data. *In* "Proceedings, 4th International Conference in High-Energy Physics, Antananarivo, Madagascar", pp. 1-12.
15. Reinsel, G., Box, G., and Jenkins, G. (1994). Time Series Analysis: Forecasting and Control. Englewood Cliffs, NJ: Prentice Hall.
16. Tayyab, M., Zhou, J., Zeng, X., and Adnan, R. (2016). Discharge Forecasting By Applying Artificial Neural Networks At The Jinsha River Basin, China. *European Scientific Journal* 12.
17. Tayyab, M., Zhou, J., Zeng, X., Zhao, N., and Adnan, R. (2015). Integrated Combination of the Multi Hydrological Models by Applying the Least Square Method. *Research Journal of Applied Sciences, Engineering and Technology* 10, 107-111.

18. Tingsanchali, T., and Gautam, M. R. (2000). Application of tank, NAM, ARMA and neural network models to flood forecasting. *Hydrological Processes* 14, 2473-2487.
19. Valipour, M. (2015). Long- term runoff study using SARIMA and ARIMA models in the United States. *Meteorological Applications*.
20. Wang, W. (2006). "Stochasticity, nonlinearity and forecasting of streamflow processes," Ios Press.
21. Xu, J., Chen, Y., Li, W., Nie, Q., Song, C., and Wei, C. (2014). Integrating wavelet analysis and BPANN to simulate the annual runoff with regional climate change: a case study of Yarkand River, Northwest China. *Water resources management* 28, 2523-2537.
22. Yule, G. U. (1927). On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 226, 267-298.
23. Zhang, Q., Wang, B.-D., He, B., Peng, Y., and Ren, M.-L. (2011). Singular spectrum analysis and ARIMA hybrid model for annual runoff forecasting. *Water resources management* 25, 2683-2703.