Available online www.jsaer.com

***Journal of Scientific and Engineering Research*, 2017, 4(4):322-326**

Research Article

ISSN: 2394-2630
CODEN(USA): JSERBR

Application of Linear Stochastic Models to Monthly Streamflow Data

Tariq Mahgoub Mohamed[1*], Ette Harrison Etuk[2]

[1]Department of Civil Engineering, Jazan University, Jazan, KSA
[2]Department of Mathematics, Rivers State University Nigeria

Abstract Time series analysis and forecasting has become a major tool in different applications in hydrology and environmental management fields. Linear stochastic models known as multiplicative Seasonal Autoregressive Integrated Moving Average (SARIMA) model were used to simulate and forecast monthly streamflow of Rahad River, Sudan. For the analysis, monthly streamflow data for the years 1972–2009 were used. A visual inspection of the time plot gives the expected impression of a generally horizontal trend and 12-month seasonal periodicity. The seasonality observed in Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) plots of monthly streamflow data was removed using first order seasonal differencing prior to the development of the SARIMA model. Interestingly, the SARIMA (2,0,0)x(0,1,1)12 model developed was found to be most suitable for simulating monthly streamflow for Rahad River. The model was found appropriate to forecast three years of monthly streamflow and assist decision makers to establish priorities for water demand.

Keywords Streamflow, Rahad River, Sudan, SARIMA models.

1. Introduction

The Rahad River, which catchment is in the Ethiopian uplands, is entirely seasonal. It rise to the west of Lake Tana, Ethiopia, and flows westwards across the Sudanese border joining the Blue Nile below Wad Madani, Sudan. The basin is characterized by highly rugged topography and considerable variation of altitude ranging from about 410 meters above sea level (masl) at Wad Madani to over 4,250 (masl) in the Ethiopian highlands [1]. The flow in the river starts in July; the flood reaches its peak in the last week of September and dries out by the end of November. Rahad River has been measured at Abu Haraz, Sudan, near its mouth from 1908 to 1951, with a record at El Hawata from 1972. The gap in the record between 1951 and 1972 was filled by means of a statistical model. The average annual flow for the Rahad river is 1.076 km^3 (1972-2009). The range of annual flows is great; the maximum recorded in the early years was 1.96 km^3 in 1909 for the river, compared with low flows in 1941 of 0.53 km^3 .This low flow has been canceled in 1984 by flows of 0.29 km^3 [2].

The Rahad agricultural project, which is semi-arid region, lies along the east bank of the Rahad River about 160 km southeast of Khartoum in the central part of the Sudan. ELFau town is the headquarters of the project which is about 280 km from Khartoum along Khartoum – Port Sudan highway .The project area of the scheme is about 25 km wide and 160 km long. It is situated in a vast clay plain at an elevation of 400-430 meters above sea level [3]. The annual rainfall ranges from 350 mm in the northern part of the project to about 600 mm in the south.

The length of rainy season fluctuates around five months i.e. from June to October and the peak of rainfall is in August. Temperatures are highest in April and May, and lowest in January. The water supply resources for the Rahad project are the Blue Nile River and the Rahad seasonal river. During a normal year the Rahad could supply the full requirements of the project during August and September, but not during the peak month of October [4]. Therefore, the monthly flow forecasting for Rahad River plays an important role in the planning and management of Rahad agricultural scheme.

During the last decades, several studies have developed methods of analyzing stochastic characteristics of streamflow time series (Yurekli [5]; Modarres [6] and Can [7]). The most widely used model is the ARIMA model. For instance, Can [7] fitted an ARIMA (0,1,1) model to mean monthly streamflows at Asagıkagdaric gauging station on Karasu River, Turkey. Yurekli [5] examined monthly streamflow data in Cekerek stream watershed, Turkey, and fitted a SARIMA (1,0,0)x(0,1,1)12 to it. In this study, linear stochastic models known as multiplicative seasonal autoregressive integrated moving average (SARIMA) models were used to model monthly flow for Rahad River, Sudan.

2. Materials and Methods

2.1. Data

In this study, streamflow data for the Rahad River at El Hawata gauging station were obtained from the Ministry of Water Resources and Electricity, covering the period 1972–2009. It includes a length of 38-years 456 monthly observations.

2.2. Modeling by Sarima Methods

A stationary time series can be modeled in different ways: an autoregressive (AR) process, a moving average (MA) process, or an autoregressive and moving average (ARMA) process. However, an ARMA model can be used when the data are stationary, ARMA models can be extended to non-stationary series by allowing differencing of data series. These models are called autoregressive integrated moving average (ARIMA) models. A time series is said to be stationary if it has constant mean and variance.

The general non-seasonal ARIMA model is AR to order p and MA to order q and operates on d^{th} difference of the time series X_t; thus a model of the ARIMA family is classified by three parameters (p, d, q) that can have zero or positive integral values. The general non-seasonal ARIMA model may be written as

$$\phi(B)\nabla^d X_t = \theta(B)\varepsilon_t \qquad (1)$$

Where:

$\phi(B)$ and $\theta(B)$ = Polynomials of order p and q, respectively.

$$\phi(B) = (1 - \phi_1 B - \phi_2 B^2 - \cdots \phi_P B^p) \quad (2)$$

And

$$\theta(B) = \left(1 - \theta_1 B - \theta_2 B^2 - \cdots \theta_q B^q\right) \quad (3)$$

Often time series possess a seasonal component that repeats every s observations. For monthly observations s = 12 (12 in 1 year), for quarterly observations s = 4 (4 in 1 year). Box et al [8] has generalized the ARIMA model to deal with seasonality, and define a general multiplicative seasonal ARIMA model, which are commonly known as SARIMA models. In short notation the SARIMA model described as ARIMA (p, d, q) x (P, D, Q) s, which is mentioned below:

$$\phi_p(B)\Phi_p(B^s)\nabla^d\nabla_s^D(X_t) = \theta_q(B)\Theta_Q(\mathrm{B}^s)\varepsilon_\mathrm{t} \qquad (4)$$

Where p is the order of non-seasonal autoregression, d the number of regular differencing, q the order of nonseasonal MA, P the order of seasonal autoregression, D the number of seasonal differencing, Q the order of seasonal MA, s is the length of season, Φ_p and Θ_Q are the seasonal polynomials of order P and Q, respectively.

2.3 Statistical Software

The econometric and statistical software Eviews-6 was used for all the analytical work. It is based on the least squares optimization criterion.

2.4. Performance Evaluation

The following measures were used to evaluate the performance of the models:

1. Mean Absolute Error:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - F_i| \qquad (5)$$

2. Root Mean Squared Error:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - F_i)^2} \qquad (6)$$

3. Theil Inequality Coefficient:

$$TIC = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - F_i)^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i)^2} + \sqrt{\frac{1}{n}\sum_{i=1}^{n}(F_i)^2}} \qquad (7)$$

4. Coefficient of Determination:

$$R^2 = \left[\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(F_i - \bar{F})}{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \sum_{i=1}^{n}(F_i - \bar{F})^2}}\right]^2 \qquad (8)$$

5. Coefficient of Efficiency:

$$E = 1 - \frac{\sum_{i=1}^{n}(Y_i - F_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \qquad (9)$$

3. Results and Discussion

The time series model development consists of three stages: identification, estimation and diagnostic check [8]. In the identification stage, data transformation is often needed to make the time series stationary. During the estimation stage the model parameters are calculated. Finally, diagnostic test of the model is performed to reveal possible model inadequacies to assist in the best model selection.

3.1. Model Identification

Model computation was made with streamflow monthly data from between January 1972 and December 2006. The data set from January 2007 to December 2009 was considered in forecasting estimations of the model.

The time series plot was conducted using the monthly streamflow data for Rahad River at El Hawata gauge station to assess the stability of the data, and Figure 1 was obtained. The plot shows that there is a seasonal cycle of the series and the series is non-stationary. The seasonal fluctuations occur every 12 months, resulting in period of time series S =12. The time-plot shows no noticeable trend.

Non-stationary is also confirmed by the Augmented Dickey- Fuller Unit Root Test (ADF) on the monthly streamflow data in Table 1. The ADF Test was done on the entire streamflow data. The table displays results of the test: statistic value -1.04065 greater than critical vales -2.57019, -1.94154, -1.61621 all at 1%, 5%, and 10% respectively. This indicates that the series is non-stationary and also confirm that the data needs differencing in order to be stationary.

From the plot of the ACF and PACF of the monthly data, Figure 2, it has been found that the data must be differenced by one seasonal degree of differencing to achieve stationary (D = 1, S = 12). Differencing for non-seasonal ARIMA was not done due to absence of trends in the data sets. Figure 3 confirms that the ACF and PACF plots for the differenced and de-seasonalized data were stable and the SARIMA model $(p,0,q)(P,1,Q)_{12}$ could be identified for further analysis.

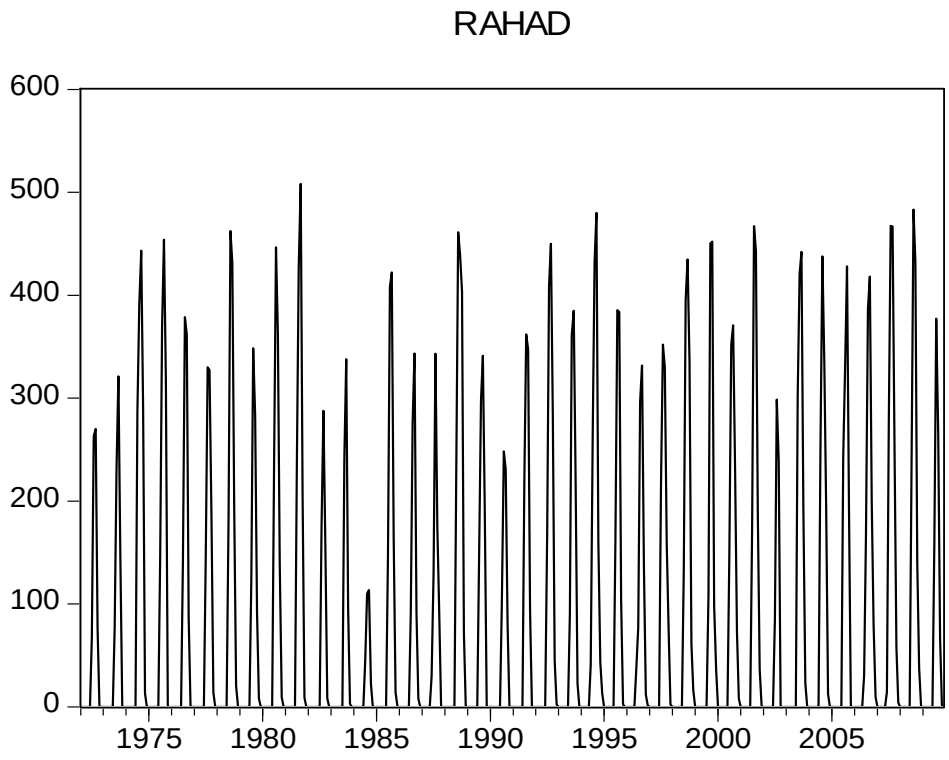

Figure 1: Time series of monthly streamflow of Rahad River (1970–2009) in (Mm^3)

Table 1: ADF-Unit Root Test for Rahad River Monthly Flow

Station	Variable	ADF test	Level of Confidence	Critical Value	Probability	Result
EL Hawata	Monthly Flow	-1.04065	1%	-2.57019	0.2687	Non-stationary
			5%	-1.94154		
			10%	-1.61621		

Once the time series was adjusted for stationarity, the order of autoregressive and moving average was estimated using the autocorrelation and partial autocorrelation function plots, Figure3. The autocorrelation structure suggests many multiplicative SARIMA models.

The optional models, the Akaike Information Criterion (AIC) and the Schwarz Criterion (SC) values are shown in Table 2. The model that gives the minimum AIC and SC is selected as best fit model. Obviously, model SARIMA (2,0,0) $(0,1,1)_{12}$ has the smallest values of AIC and SC, then one would temporarily have a model SARIMA (2,0,0)x$(0,1,1)_{12}$.

3.2. Parameter Estimation

After the identification of model using the AIC and SC criteria, estimation of parameters is done. The value of the parameters, associated standard errors, t-ratios and p-values (< 5 %) are listed in Table 3. The result indicated that the parameters are significant since its p-values are smaller than alpha level (0.05) and should be retained in the model.

Table 2: Comparison of AIC and SC for the Selected Models

Variable	Station	Model	AIC	SC
Monthly Flow	EL Hawata	SARIMA(2,0,0)x$(0,1,1)_{12}$	10.4142	10.4438
		SARIMA(2,0,0)x$(1,1,1)_{12}$	10.4306	10.4710
		SARIMA(1,0,0)x$(0,1,1)_{12}$	10.4290	10.4487
		SARIMA(2,0,0)x$(0,1,0)_{12}$	10.9879	11.0076

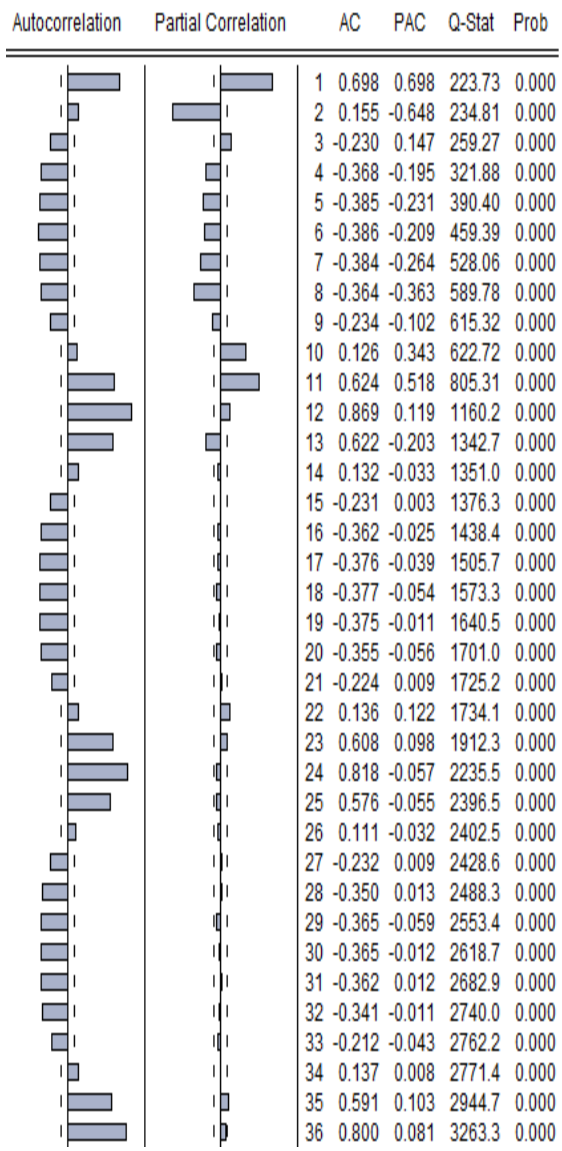

Figure 2: ACF and PACF Plots for Rahad River Monthly Flow

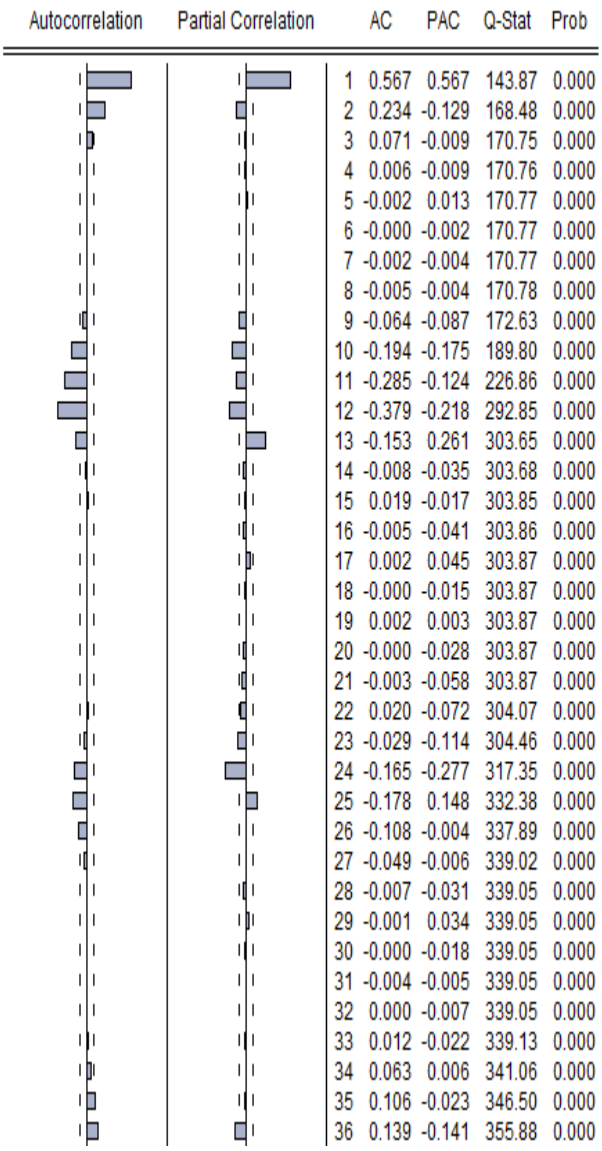

Figure 3: ACF and PACF Plots after one Seasonal Difference

Table 3: Estimation of the SARIMA (2, 0, 0)x(0, 1, 1)$_{12}$ Model

Dependent Variable: D(RAHAD,0,12)
Method: Least Squares
Sample (adjusted): 1973M03 2006M12
Included observations: 406 after adjustments
Convergence achieved after 14 iterations
MA Backcast: 1972M03 1973M02

Variable	Coefficient	Std. Error	t-Statistic	Prob.
AR(1)	0.631616	0.049298	12.81226	0.0000
AR(2)	-0.148952	0.049431	-3.013314	0.0027
MA(12)	-0.966407	0.008976	-107.6645	0.0000

R-squared	0.624067	Mean dependent var	1.522149
Adjusted R-squared	0.622202	S.D. dependent var	71.60754
S.E. of regression	44.01379	Akaike info criterion	10.41424

Sum squared resid	780697.1	Schwarz criterion	10.44385
Log likelihood	-2111.092	Hannan-Quinn criter.	10.42596
Durbin-Watson stat	1.993533		

Inverted AR Roots	.32-.22i	.32+.22i		
Inverted MA Roots	0.99	.86+.50i	.86-.50i	.50+.86i
	.50-.86i	.00+1.00i	-.00-1.00i	-.50+.86i
	-.50-.86i	-.86-.50i	-.86+.50i	-0.99

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.001	0.001	0.0003	
		2	-0.024	-0.024	0.2499	
		3	0.033	0.033	0.7307	
		4	-0.022	-0.022	0.9387	0.333
		5	0.004	0.006	0.9474	0.623
		6	-0.000	-0.003	0.9474	0.814
		7	-0.013	-0.011	1.0198	0.907
		8	0.007	0.006	1.0411	0.959
		9	0.013	0.012	1.1145	0.981
		10	-0.125	-0.125	8.2343	0.312
		11	0.021	0.022	8.4422	0.392
		12	0.087	0.082	11.871	0.221
		13	-0.007	0.002	11.891	0.292
		14	0.026	0.023	12.211	0.348
		15	0.015	0.012	12.318	0.421
		16	-0.036	-0.033	12.924	0.454
		17	0.013	0.009	12.998	0.527
		18	-0.007	-0.006	13.022	0.601
		19	-0.003	0.004	13.027	0.671
		20	0.003	-0.016	13.030	0.734
		21	-0.028	-0.024	13.401	0.767
		22	0.002	0.022	13.403	0.817
		23	0.088	0.084	16.992	0.654
		24	-0.102	-0.105	21.862	0.407
		25	-0.072	-0.070	24.314	0.331
		26	0.022	0.003	24.548	0.374
		27	-0.029	-0.021	24.941	0.409
		28	0.013	0.017	25.023	0.461
		29	-0.007	-0.011	25.048	0.516
		30	-0.001	0.003	25.049	0.572
		31	-0.005	-0.016	25.059	0.625
		32	-0.004	-0.000	25.067	0.675
		33	-0.013	0.016	25.147	0.718
		34	0.047	0.021	26.231	0.710
		35	0.056	0.028	27.725	0.683
		36	-0.022	-0.000	27.968	0.716

Figure 4: ACF and PACF Plots for SARIMA (2, 0, 0)x(0, 1, 1)$_{12}$ Residuals

3.3. Diagnostic Check

Once an appropriate model is selected and its parameters are estimated, the Box Jenkins methodology requires examining the residuals of the model to verify that the model is an adequate one for the series. An adequate model should have uncorrelated residuals. This is the minimal condition. For a good forecasting model, the residuals must satisfy the requirements of a white noise process. Several tests were carried out on the residual series. The tests are summarized briefly in the following paragraphs.

3.3.1. ACF and PACF of Residuals

The ACF and PACF of residuals of the model SARIMA (2,0,0)x(0,1,1)$_{12}$ are shown in Figure 4. Most of the values of the RACF and RPACF lies within confidence limits except very few individual correlations appear large compared with the confidence limits. The figure indicates no significant correlation between the residuals.

3.3.2 Portmantateau Lack-Of-Fit Test (The Ljung–Box Test)

The goodness-of-fit of the selected model was tested using the Ljung-Box statistic test. The test is employed for checking independence of residual. From Figure 4, the goodness of fit values for the autocorrelations of residuals from the model up to lag 24was ≥ 0.05. The result proves the acceptance of the null hypothesis of model adequacy at the 5% significance level and the set of autocorrelations of residuals was considered white noise.

3.3.3 The Breusch-Godfrey Serial Correlation LM Test

The Breusch-Godfrey Serial Correlation LM test accepts the hypothesis of no serial correlation in the residuals, as shown in Table 4.

The graph showing the observed and fitted values is presented in Figure 5. The Figure shows a very close agreement between the fitted model and the actual data. Since the model diagnostic tests show that all the parameter estimates are significant and the residual series is white noise, the estimation and diagnostic checking stages of the modeling process are complete.

4. Forecasting Of Monthly Streamflow

SARIMA model can also be used for forecasting future values based on the historical data. The SARIMA $(2,0,0)x(0,1,1)_{12}$ model was tested for its validity to forecast 36 observations obtained for the years 2007−2009 for Rahad river. The observed streamflow was found to be closely aligned to the forecasted values, Figure 6.

Table 4: The Breusch-Godfrey Serial Correlation LM Test

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	0.210056	Prob. F(2,401)	0.8106
Obs*R-squared	0.424327	Prob. Chi-Square(2)	0.8088

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	0.840840	Prob. F(12,391)	0.6082
Obs*R-squared	10.21304	Prob. Chi-Square(12)	0.5973

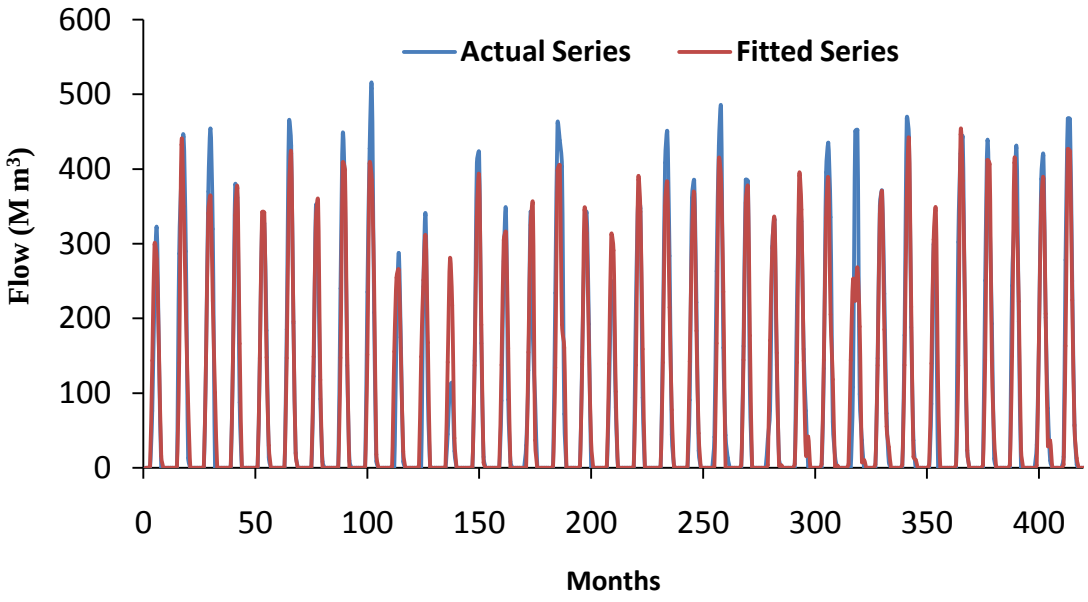

Figure 5: Comparison of Observed Data and SARIMA Model Flow (1972-2006)

4.1. Forecasting Accuracy

If the fitted SARIMA (2, 0, 0)x(0, 1, 1)$_{12}$ model has to perform well in forecasting, the forecast error will be relatively small .To check goodness of the prediction, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Theil inequality coefficient, Coefficient of Determination (R^2) and Nash Sutcliffe Efficiency Criteria (E) were used. Table 5 illustrates all of the statistic measures. From the statistics measurement, Table 5, it is observed that the model has lower values of RMSE and MAE. Theil inequality coefficient turns out to be 0.149, which is relatively close to zero. The Theil inequality coefficient always lies between zero and one, where zero indicates a perfect fit. The Coefficient of Determination (R^2) value of 0.91, Figure 9, and Nash Sutcliffe Efficiency Criteria (E) value of 89.3% showed the very good performance of the model.

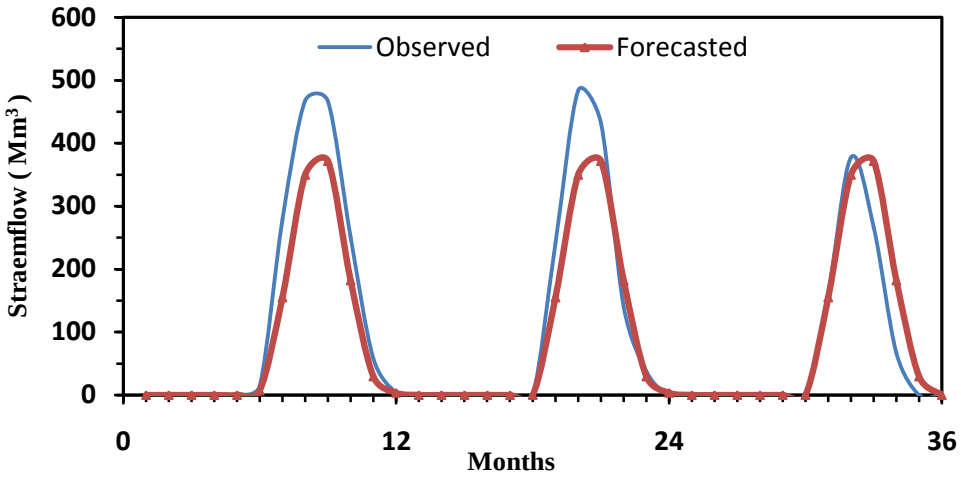

Figure 6: Forecasting of monthly streamflow using developed SARIMA model (2,0,0)x(0,1,1)$_{12}$, (2007−2009)

Table 5: Forecasting Accuracy Statistic

Statistic Measures	Value
MAE	30.25
RMSE	52.87
Theil inequality coefficient	0.149
R^2	0.91
E	89.3%

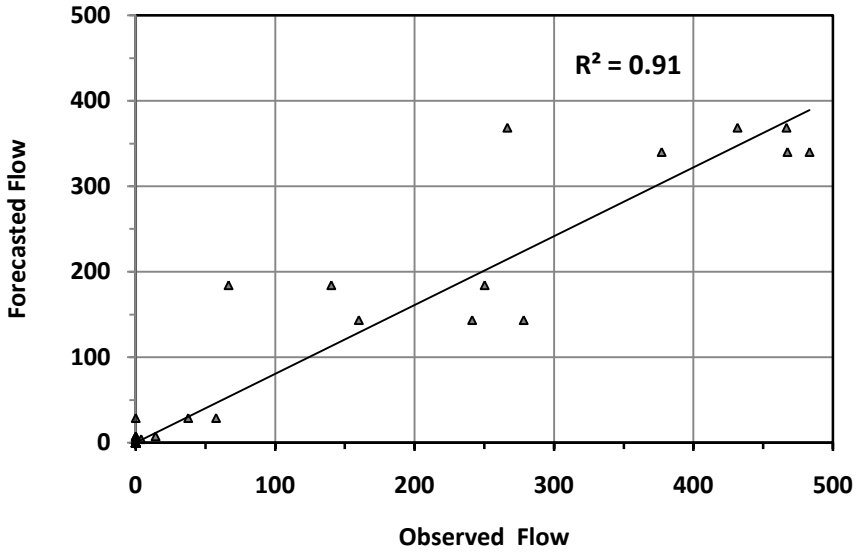

Figure 7: Calibration results of SARIMA model (2,0,0)x(0,1,1)$_{12}$

Conclusion

In this paper, linear stochastic model known as Multiplicative Seasonal Autoregressive Integrated Moving Average model, SARIMA, was used to simulate and forecast monthly streamflow for Rahad River, Sudan. The tentative model that best fits the criteria and meets the requirement is model SARIMA $(2,0,0)x(0,1,1)_{12}$. By analyzing the forecasted values, it was found that use of SARIMA model for forecasting monthly streamflow is admirably good. The fitting of stochastic ARIMA models to streamflow time series could result in a better tool which can be used for water resource planning. SARIMA model has the ability to predict accurately the future monthly streamflow for all streamflow gauge stations in Sudan.

References

[1]. Melesse A., M., (2011). Nile River Basin: Hydrology, Climate and Water Use, Springer Dordrecht Heidelberg, London.

[2]. Sutcliffe et Al (1999). The Hydrology of the Nile, IAHS Special Publication no. 5, IAHS Press, Institute of Hydrology, Wallingford, Oxfordshire OX10 8BB, UK.

[3]. Benedict et Al (1982) Sudan: The Rahad Irrigation Project, U.S.Agency for International Development (AID).

[4]. Document of international Bank for Reconstruction and Development - International Development Association (1973). Appraisal of the Rahad Irrigation Project, Sudan, Agriculture Projects Department Eastern Africa Regional Office.

[5]. Yurekli, K., Kurunc, K., Ozturk, F., (2005) Application of linear stochastic models to monthly flow data of Kelkit Stream, Ecological Modeling 183 , 67–75.

[6]. Modarres, R., (2007), Streamflow Drought Time Series Forecasting. Stoch Environ Res Risk Assess, 21:223–233.

[7]. Can., I., Selim, S., (2009) Stochastic modeling of mean monthly flows of Karasu River, in Turkey, Water and Environment Journal. doi:10.1111/j.1747-6593.2009.00186.x

[8]. Box, G. E. P., Jenkins, G. M., Reinsel, G. C. (1994). Time Series Analysis Forecasting and Control, 3rd ed., Englewood Cliffs, N.J. Prentice Hall.

[9]. Nash, J. E., Sutcliffe, J. V. (1970). "River flow forecasting through conceptual models: 1. A discussion of principles." Journal of Hydrology, 10, 282–290.

[10]. Shamseldin, A. Y., O'Connor, K. M., Liang, G. C. (1997). Methods for combining the output of different rainfall-runoff models. Journal of Hydrology, 197, 203–229.