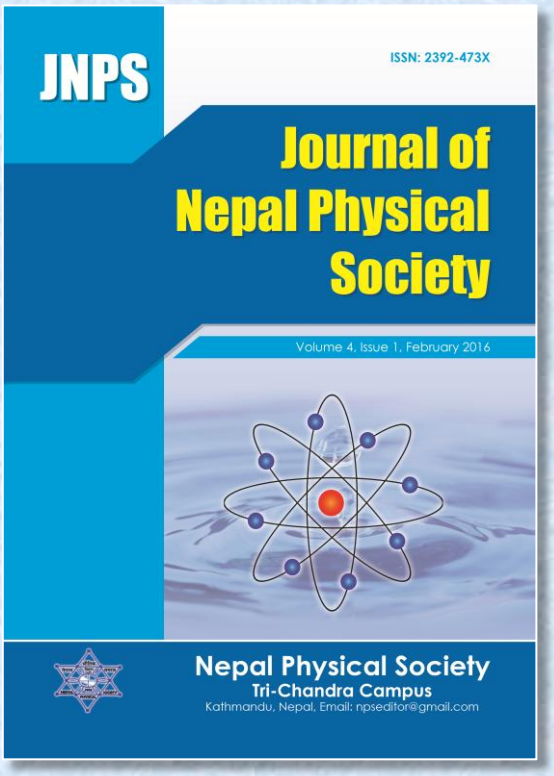
Application of ARIMA Model for River Discharges Analysis

Bhola NS Ghimire

Journal of Nepal Physical Society

Volume 4, Issue 1, February 2017

ISSN: 2392-473X



Editors:

Dr. Gopi Chandra Kaphle

Dr. Devendra Adhikari

Mr. Deependra Parajuli

JNPS, **4** (1), 27-32 (2017)

Published by:

Nepal Physical Society

P.O. Box : 2934

Tri-Chandra Campus

Kathmandu, Nepal

Email: npseditor@gmail.com

JNPS ***4*** *(1), 27-32 (2017)*
ISSN: 2392-473X

Research Article

Application of ARIMA Model for River Discharges Analysis

Bhola NS Ghimire
Department of Civil Engineering, Pulchowk Campus, Institute of Engineering, T. U., Lalitpur, Nepal
Corresponding Email: bholag@ioe.edu.np

ABSTRACT

Time series data often arise when monitoring hydrological processes. Most of the hydrological data are time related and directly or indirectly their analysis related with time component. Time series analysis accounts for the fact that data points taken over time may have an internal structure (such as autocorrelation, trend or seasonal variation) that should be accounted for. Many methods and approaches for formulating time series forecasting models are available in literature. This study will give a brief overview of auto-regressive integrated moving average (ARIMA) process and its application to forecast the river discharges for a river. The developed ARIMA model is tested successfully for two hydrological stations for a river in US.

Keywords: Time Series Analysis, ARIMA Model, Hydrological Process, Autocorrelation, Seasonal Variation.

INTRODUCTION

Auto-Regressive Integrated Moving Average (ARIMA) method is widely used in field of time series modeling and analysis. These models were described by Box and Jenkins (1976) and further discussed by Walter (Chatfield, 1996). The Box-Jenkins approach in hydrological modeling is used by several researchers. Chew et al. (1993) conducted a comparison of six rainfall-runoff modeling approaches to simulate daily, monthly and annual flows in eight unregulated catchments. Langu (1993) used time series analysis to detect the changes in rainfall and runoff patterns. Kuo and Sun (1993) used the time series model for ten days stream flow forecast and generate synthesis hydrograph caused by typhoons in Tanshui River in Taiwan. Naill and Momani (2009) used the time series analysis for rainfall data in Jordan.

This paper is aimed to show the usefulness of this popular technique ARIMA for a typical case study.

ARIMA MODEL

Trend and prediction of time series can be computed by using ARIMA model. ARIMA (*p*,*d*,*q*) model is a complex linear model. In statistics, normally in time series analysis, ARIMA model is generalization of autoregressive moving average (ARMA) models, sometimes called Box-Jenkins models after the iterative Box-Jenkins methodology. Given a time series of data X_t, the ARMA model is a tool for understanding and, perhaps, predicting future values in this series. The model consists of two parts, an autoregressive (AR) part and a moving average (MA) part. And that of in ARIMA model the third part integrated (I) included. The model is usually then referred to as the ARIMA (*p*,*d*,*q*) model where *p* is the order of the autoregressive part, *d* is the order of non seasonal differences and *q* is the order of the moving average part.

The notation AR(*p*) refers to the autoregressive model of order *p*, which can be written as:

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t \quad (1)$$

where $\varphi_1,..,\varphi_p$ are the parameters of the model, *c* is a constant and ε_t is white noise. The constant term is omitted by many authors for simplicity.

Similarly, the notation MA (*q*) refers to the moving average model of order *q*. This can be written as:

$$X_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \quad (2)$$

Where the $\theta_1,.. ,\theta_q$ are the parameters of the model, μ is the expectation of X_t (often assumed equal to 0), and the ε_t, ε_{t-1}, …, ε_{t-q} are the white noise error terms. The moving average model is a finite impulse response filter with some additional interpretation placed on it. While combing these two models, the ARMA (*p*,*q*) is obtained.

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \quad (3)$$

The error term ε_t are generally assumed to be independent identically-distributed random variables sampled from a normal distribution with zero mean: $\varepsilon_t \sim N(0,\sigma^2)$ where, σ^2 is the variance. Some researchers have used the equation in a lag operator form. In lag operator form, AR (*p*) model is given by-

$$\varepsilon_t = \left(1 - \sum_{i=1}^{p} \varphi_i L^i\right) X_t = \varphi X_t \quad (4)$$

And MA(*q*) model is given by-

$$X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t = \theta \varepsilon_t \quad (5)$$

Where, φ and θ are defined by the parameters containing inside the parenthesis of each model. Combining these models we can manipulate and write in the following form.

$$\left(1 - \sum_{i=1}^{p} \varphi_i L^i\right) X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t \quad (6)$$

Assume that the polynomial of first term of above equation has a unitary root of multiplicity *d*. Then this equation can be updated including the difference term, which can be expressed as,

$$\left(1 - \sum_{i=1}^{p} \varphi_i L^i\right) = \left(1 + \sum_{i=1}^{p-d} \psi_i L^i\right)(1-L)^d \quad (7)$$

An ARIMA (*p, d, q*) process expresses this polynomial factorization property, and is finally written as:

$$\left(1 - \sum_{i=1}^{p} \varphi_i L^i\right)(1-L)^d = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t \quad (8)$$

More precisely, the ARIMA (*p, d, q*) model can be written as:

$$\varphi_p(B)(1-B)^d y_t = \theta_q(B) \varepsilon_t \quad (9)$$

The model while used with seasonal fluctuation, with seasonal length *s*, the process is called SARIMA (p, d, q)(P, D, Q)$_s$, where p, d, q represents the order of process AR, order of difference (I) and order of process (MA) for non seasonal part and P, D, Q, represents the order of seasonal process AR, order of seasonal difference and order of seasonal MA and *s* is the length of seasonal period.

The general equation of SARIMA model is:

$$\varphi_p(B)(1-B)^d \Phi_P(B^s)(1-B^s)^D y_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t \quad (10)$$

Where, $\varphi_p(B)$ is auto regressive operator, $\theta q(B)$ is the operator of moving average; $\Phi P(Bs)$ is seasonal autoregressive operator, $\Theta Q(Bs)$ is seasonal operator of moving averages, ε_t is white noise.

STATISTICAL TESTS FOR MODEL PERFORMANCE

There are several statistical tests for model performance. In this study some of these tests are used which are easy to understand and use.

Coefficient of Correlation

A very important part of statistics is describing the relationship between two (or more) variables. One of the most fundamental concepts in research is the concept of correlation. If two variables are correlated, this means that it can use information about one variable to predict the values of the other variable. The coefficient of correlation is given by following equation.

$$r = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}} \quad (11)$$

Where, r is correlation coefficient; x_i and y_i are independent (observed) and dependent (predicted) variables and $\bar{x}$ and $\bar{y}$ are their corresponding means.

Root Mean Square Error

The root mean square error (RMSE)) (also root mean square deviation (RMSD)) is a frequently-used measure of the differences between values predicted by a model or an estimator and the values actually observed from the thing being modeled or estimated. RMSE is a good measure of accuracy. These individual differences are also called residuals, and the RMSE serves to aggregate them into a single measure of predictive power. The mathematical form of RMSE is given by:

$$RMSE = \sqrt{\frac{(x_i - y_i)^2}{N}} \quad (12)$$

Where N is total number of data set and other variables are same as earlier equation. These two

equations are used for the comparison of observed and predicted values.

AIC and BIC criteria

The two most commonly used penalized model selection criteria, the Akaike's information criterion (AIC) and the Bayesian information criterion (BIC), are examined and compared for ARIMA model selection.

- **AIC** - In general case,

$$AIC = 2k + n\ln(SSE/n) \quad (13)$$

Where k is the number of parameters in the statistical model, n is the number of observations and SSE is square sum of error given by -

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 \quad (14)$$

- **BIC** - In general BIC is given by-

$$BIC = k\ln(n) + n\ln(SSE/n) \quad (15)$$

The minimum values of these *AIC* and *BIC* criteria give the better model performance.

CASE STUDY DESCRIPTION

Discharge Data

For the application demonstration of ARIMA model, the time series daily discharge data two stations in Schuylkill River at Berne (Station no: 01470500, Lat. 40°31'21" and Long. 75°59'55") and Philadelphia (Station no: 01474500, Lat. 39°58'04" and Long. 75°11'20"), USA are taken. The catchments area of Berne station is about 919.45 km^2 and that of Philadelphia station is 4902.85 km^2. This information was obtained from USGS website.

The data from the period October 01, 2000 to September 30, 2006 were taken for both of the stations. Initial all six years data were taken for ARIMA model development and finally using model last one year data (October 01, 2006 to September 30, 2007) were predicted for both the stations. Some of the statistical parameters for these sites are shown in Table 1 of the discharge data. The parameters μ, σ, σ/μ, C_{sk}, C_{kr}, X_{max}, X_{min} are mean, standard deviation, variance, skew-ness, kurtosis, maximum and minimum values respectively. The discharge limits of Berne station are 2.13 to 972.01 m^3/s and that of Philadelphia station are 2.24 to 1484.94 m^3/s.

Table 1. The daily statistical parameters for Schuylkill River.

Station	Basin Area (Km^2)	μ	σ	σ/μ	C_{sk}	C_{kr}	X_{max}	X_{min}
Berne 01470500	919.45	22.09	33.90	1.53	12.18	270.59	972.01	2.13
Philadelphia 01474500	4902.85	99.20	118.55	1.195	4.54	33.27	1484.94	2.24

Development of ARIMA Models

From the time series plot for the given data (figure 1), it can be observed that there is no seasonality for daily data. In fact, it is very difficult to fix the seasonality for daily data and due to the large span of time (365 days), it is unreliable too. So, the model formulation has done without seasonality. The ARIMA models for the both stations are developed by using SPSS. Initially, the several models were tested based on the AIC and BIC criterion. The AIC and BIC values for few models for these are given in the following Table 2.

Table 2. The AIC and BIC values for some testing ARIMA models for Berne data and Philadelphia data.

	ARIMA - Berne Data								
	(1,0,0)	(1,1,0)	(2,0,0)	(1,0,1)	(1,1,1)	(1,1,2)	(1,2,1)	(1,2,2)	(2,2,2)
AIC	23425	23772	23425	23424	23413	23411	23798	23810	23939
BIC	23437	23784	23442	23442	23431	23438	23805	23813	23968
	ARIMA – Philadelphia Data								
	(0,1,0)	(0,2,0)	(1,0,0)	(1,1,0)	(1,1,1)			(2,11)	(2,2,1)
AIC	29897	31744	29568	29896	29544			29502	29772
BIC	29903	31750	29580	29908	29561			29525	29796

From Table 2, it can be judged that on the principle of AIC and BIC test, ARIMA (1, 1, 2) is suitable for Berne station and ARIMA (1, 1, 2) is suitable for Philadelphia station.

However while we observed the correlation matrix of the ARIMA parameters for Berne Station from Table 3, the parameters MA(1) and MA(2) has very high correlation approaching to unity. So that their effects in ARIMA model are negligible and we can reduce this MA parameter. Then the ARIMA (1, 1, 1) is proposed for the further analysis even though ARIMA (1, 1, 2) has fairly less AIC and BIC values. The correlation matrix and parameters for final model for the Berne station are given in the following Table 4.

Table 3. The correlation matrix for parameters of Berne data.

ARIMA (1,1,2)		Non-Seasonal Lags			Constant
		φ_1	θ_1	θ_2	
Non-Seasonal Lags	φ_1	1.0	.737	-.716	0
	θ_1	.737	1.00	-.993	0
	θ_2	-.716	-.993	1.0	0
Constant		0	0	0	1.0

Table 4. The correlation matrix and final parameters for Berne Station.

a. The correlation matrix

ARIMA (1,1,1)		Non-Seasonal Lags		Constant
		φ_1	θ_1	
Non-Seasonal Lags	φ_1	1.0	.320	0
	θ_1	.320	1.0	0
Constant		0	0	1.0

b. The final parameters

ARIMA (1, 1, 1)		Estimates	Std Error
Non-Seasonal Lags	φ_1	.698	.015
	θ_1	.991	.003
Constant		-.001	.014

Similarly, for Philadelphia station, AIC and BIC values for some models, correlation matrix for selected model and final ARIMA model parameters are given in Table 5.

Table 5. The correlation matrix and final parameters for Philadelphia data.

a. The correlation matrix

ARIMA (2,1,1)		Non-Seasonal Lags			Constant
		φ_1	φ_2	θ_1	
Non-Seasonal Lags	φ_1	1.0	-.598	.253	0
	φ_2	-.598	1.0	.214	0
	θ_1	.253	.214	1.0	0
Constant		0	0	0	1.0

b. The final parameters

ARIMA (2,1,1)		Estimates	Std Error
Non-Seasonal Lags	φ_1	.787	.020
	φ_2	-.141	.020
	θ_1	.968	.006
Constant		-.009	.140

As discussed in earlier, for ARIMA (1, 1, 1), the general equation can be reduced as:

$$\varphi_1(B)(1-B)^1 y_t = \theta_1(B)\varepsilon_t \quad (16)$$

While substituting the model parameters in the above equation and simplify it, we get the final model for Berne station as:

$$y_t = 1.698\, y_{t-1} - 0.698\, y_{t-2} - 0.991 e_{t-1} - 0.001 \quad (17)$$

Similarly, for ARIMA (2, 1, 1) model, the general equation is given by:

$$\varphi_2(B)(1-B)^1 y_t = \theta_1(B)\varepsilon_t \quad (18)$$

Based on the model parameters, the final model for Philadelphia station is given by:

$$y_t = 1.787\, y_{t-1} - 0.928\, y_{t-2} + 0.141\, y_{t-3} - 0.968\, e_{t-1} - 0.009 \quad (19)$$

RESULT AND DISCUSSIONS

The time series plot of original data and predicted data from the final models for both the stations are given in figure 1. This shows that the general trend is followed by predicted data to that of observed data. In the plot, the six years observed data followed by last one year predicted data for the both stations.

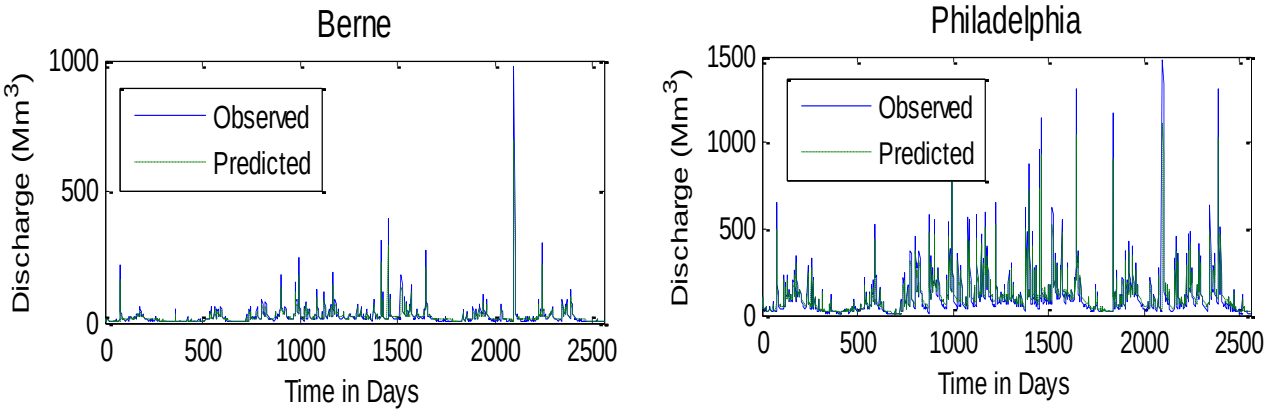


Fig. 1. Time series plot of observed and predicted daily discharge at Berne and Philadelphia Stations.

The scatter plots of these two stations are also given in figure 2. This shows that they are quite good models. The R^2 for Berne station and Philadelphia stations are: 0.9901 and 0.9688 respectively.

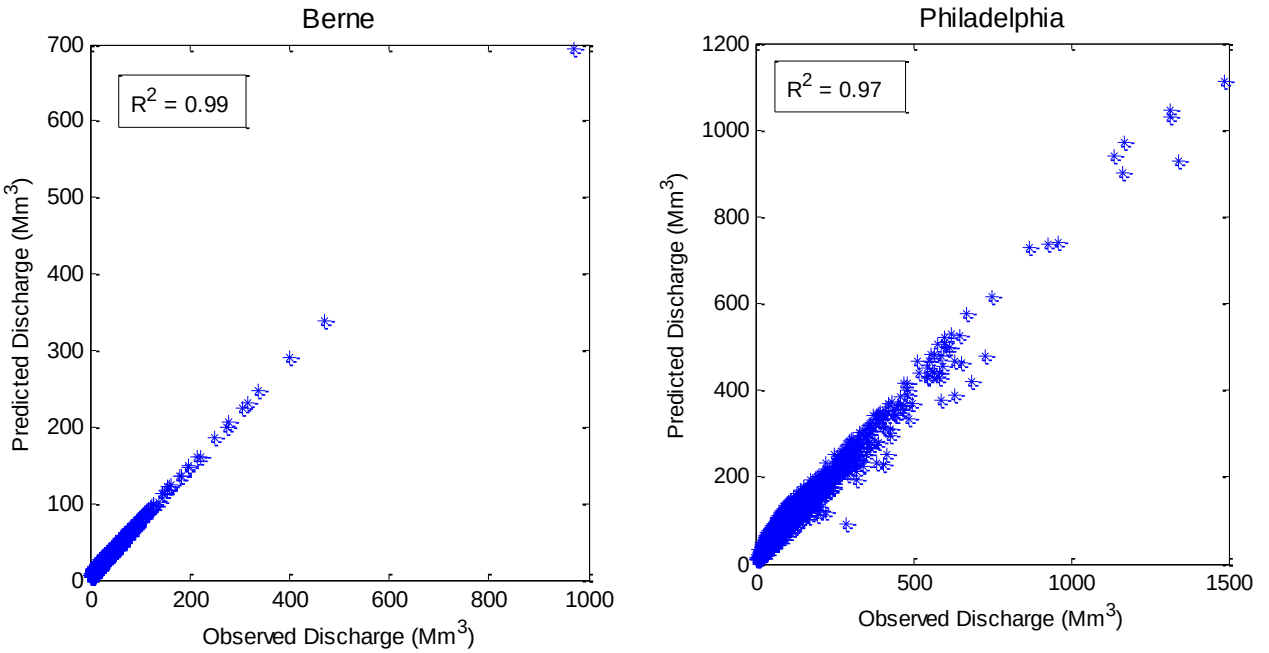


Fig. 2. Scatter plot of observed and predicted daily discharge at Berne and Philadelphia Stations.

CONCLUSIONS

Time series analysis for the river discharge shows that it is an important tool for modeling and forecasting. ARIMA (1, 1, 1) model is fitted for Berne station and ARIMA (2, 1, 1) is fitted for Philadelphia station. Both the stations are lies in the same rivers but they have different catchments coverage. So it should be noted that even the river is same, depending upon the catchments characteristics, applicable models are different

individual sites. The coefficient of determinations (0.99 for Berne and 0.969 for Philadelphia) shows that the model is useful for runoff forecasting.

REFERENCES:

Box, G. E. P., and Jenkins, G. M. (1976). Time series analysis: forecasting and control. *Revised Edn. Holden-Day*, San Francisco. ....

Chatfield, C. (1996). Analysis of time series: an introduction. *5th ed. Chapman and hall*, Boca Raton.

Chew, F. H. S.; Stewardson, M. J., and McMahon, T. A. (1993). Comparision of six rainfall-runoff modelling approaches. *J. of Hydrology*, **147**(1-4): 1-36.

Langu, E. M. (1993). Detection of changes in rainfall and runoff patterns. *J. of Hydrology*, **147** (1-4): 153-167.

Kuo, J. T., and Sun, Y. H. (1993). An intervention model for average 10 day stream flow forecast and synthesis. *J. of Hydrology*, **151**(1): 35-56.

Nail, P. E., and Momani, M. (2009). Time series analysis model for rainfall data in Jrdan: case study for using Time series analysis. *American J. of Environmental Sciences*, **5**(5): 599-604.