Forecasting the monthly inflow rate of the Palar-Porundalar dam in Tamil Nadu using SARIMA model

K. Mano Chitra
Department of Physical science and Information Technology, Tamil Nadu Agricultural University, Coimbatore-641003 (Tamil Nadu), India

R. Pangayar Selvi*
Department of Social Science, Agricultural College and Research Institute, Killikulam, Tamil Nadu Agricultural University, Coimbatore-641003 (Tamil Nadu), India

K. Mahendran
Department of Agriculture and Rural Management, Tamil Nadu Agricultural University, Coimbatore-641003 (Tamil Nadu), India

*Corresponding author. E-mail: pangai@rediffmail.com

Abstract
Dam inflow forecasting information is essential for planning and management of the dam system. Time series analysis is the most commonly employed technique to forecast the future values based on historical information. In this study, Palar-Porandalar dam in Tamil Nadu inflow series were forecasted in R software package using ARIMA model with seasonal factors. The monthly inflow series of the dam from 2003 January to 2017 December were used as an input source for modeling and forecasting process. Mann-Kendall's trend test and various Stationarity test were performed to verify the Stationary nature of the data set. From the Correlogram plot, different models were identified; their parameters were optimized and residuals were diagnostically tested using Autocorrelation plot and Ljung Box test. Finally, the best model was selected based on minimum Akaike Information Criteria (AIC), BIC, RMSE and Theil's U statistic values. From various models, SARIMA $(0, 0, 1) (1, 0, 2)_{12}$ model was selected as the best one for forecasting the inflow series.

Keywords: AIC, Correlogram, forecast, SARIMA, Stationarity test

INTRODUCTION

To attain the socio-economic development, the basic requirement of a country is to have an adequate amount of water with appropriate quality for domestic, industrial and agricultural purposes. Since water is one of the renewable sources of energy, is now becoming a more valuable commodity. Seasonal fluctuations and climatic irregularities make complex behavioral changes in the natural hydrological cycle leading to the construction of dams and reservoirs to store water and provide a consistent discharge. [International Commission on Large Dams (ICLD), 2018]. Around half of the world's global river systems are regulated by dams which provide irrigation and other economic activities to the command area people (Richter and Thomas, 2007). One of the important applications of time series model is to predict the future action based on historical data. Time series technique namely Autoregressive Integrated Moving Average (ARIMA) is a flexible; relatively systematic and a powerful class of model that can be applied for real time situations. To predict both the seasonal and non-seasonal factors in the data set, multiplicative SARIMA model is used (Box et al., 2008). This SARIMA model is capable of forecasting monthly reservoir inflow, even for low values and short-term forecasting than the hybrid Artificial Neural Network-Genetic Algorithm model (Moeeni et al., 2017). Also Tadesse et al. (2017) employed the SARIMA model to forecast the inflows of Waterval River. Based on this past research information, the present study aims to forecast the inflow rate of the Palar-Porundalar dam in the Dindigul district of Tamil Nadu using seasonal ARIMA model.

MATERIALS AND METHODS

The Palar-Porandalar dam is found in between the Palar and Porandalar stream flow, situated in Balasamudhram village of Palani taluk of Dindigul district in Tamil Nadu. It is an earthen type of dam with the geographical location of $10^0 22'38.622"$ North latitude and $77^0 29.5'7516"$ East longitude. The total capacity of the dam is 1524 M.C.ft and the catchment area is 259 km^2. The main function

of the dam is to provide water source for domestic, agricultural, fishery and ecological conservation.

The data for the present study was accessed from the Water Resource Department (WRD) and Public Work Department (PWD), Palani. The monthly dam inflow series for the period of 2003January to 2017December were used for analysis. In these, 2003 to 2015 data were used as training set to identify the most suitable model and 2016-2017 data were used as a test series to validate the fitted model. Thus, from the selected model, the inflow series of the dam for2018 to 2019 were forecasted using the software package R.

Seasonal ARIMA model: The ARIMA models can be applied only for non seasonal and stationary data. For seasonal time series, Box and Jenkins have proposed a model called the seasonal ARIMA or SARIMA model. These seasonal ARIMA model is an adaptation of autoregressive integrated moving average (ARIMA) models to specific fit seasonal time series data. Their formulation includes the seasonal part in model fitting. A combination of non-seasonal and seasonal process yields the multiplicative Seasonal ARIMA, (p, d, q)

$\times$ (P, D, Q)$_s$ where s is the period per season. The general form as:

$$\Phi_P(B^S)\phi_p(B)(1-B)^d(1-B^S)^D Y_t = \Theta_Q(B^S)\theta_q(B)\varepsilon_t$$

Where: $\phi_p(B), \theta_q(B)$ and (1-B)^d are non-seasonal autoregressive operator of order p, non-seasonal moving average operator of order q and non-seasonal differencing operator of order d, respectively.

While, $\Phi_P(B^S), \Theta_Q(B^S)$ and (1-B^S)^D are seasonal autoregressive operator, seasonal moving average operator and seasonal differencing operators with their orders P, Q and D respectively, are representedas follows;

$$\Phi_P(B^S) = 1 - \Phi_1 B^S - \Phi_2 B^{2S} - ... - \Phi_P B^{PS}$$

$$\Theta_Q(B^S) = 1 - \Theta_1 B^S - \Theta_2 B^{2S} - ... - \Theta_Q B^{QS}$$

Here S is the length of seasonality. The four stages in the SARIMA model were given as follows:

Identification of the model: The first step was to plot the time series data and examine its Stationarity. The stationarity condition is verified by various tests like Augmented-Dickey Fuller test (Dickey and Fuller., 1979), Phillips–Perron (pp) (Phillipsand Perron., 1988) and Kwiatkowski–Philips–Schmidt–shin (KPSS) (Kwiatkowski et

al.,1992) were performed at 0.5 significance level (α = 0.05). The trend of the data set is further tested by Mann–Kendall's trend test (Kendall. 1975). Sample Autocorrelation Function (ACF) and Partial Autocorrelation Functions (PACF) (Singh *et al.*, 2011) were plotted to identify the order of (seasonal or non-seasonal) AR (P, p), MA (Q, q) model (Cryer and Chank 2008; Salas et al. 1980). Based on the stated principles, many SARIMA models were identified and the best one is selected by comparing its AIC and BIC values.

Parameter estimation: The parameter of the model is estimated and optimized by using Maximum likelihood function method (Hamilton. 1994).

Diagnostic checking: The residual correlograms (ACF and PACF), Ljung–box Q tests (Ljung and Box 1978) were done to check the correlation nature of the residuals. Residual ACF and PACF are plotted with various lags to the tolerance limit of 95 percent to find its behavioral pattern over the time period.

Evaluation of the model: As a final step, precision of the model were evaluated against performance measures: AIC, BIC values (Akaike, 1974). Forecasting performance of the model was tested by Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and MAPE. The forecasting accuracy was tested by Theil's U statistic.

RESULTS AND DISCUSSION

The monthly time series plot of the inflow series is shown in Fig.1. It shows that the data remains stationary but it should be further tested by Augmented Dickey-Fuller test, KPSS, PP test and Mann Kendall's trend test (Tadesse et al, 2017). Dam inflow series is usually a random factor, where the maximum inflows are recorded in the rainy season and minimum in the summer season. The results of Mann-Kendall's trend test and Stationarity test results are given in the Table 1.

The tau value of trend test was -0.0928 and it shows there is no trend factor in data set. This is because the p value of the test is 0.06 which is
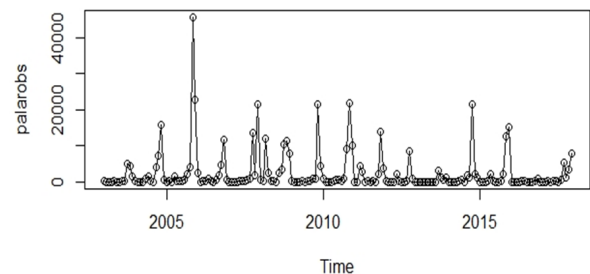

Fig.1. *Time series plot of monthly inflow series of dam.*

Table 1. Stationarity test for monthly inflow series.

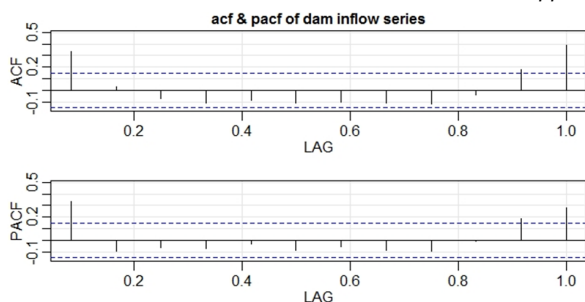| Unit root and Stationarity test	Tau	P value	Alpha
Mann-Kendall's trend test	-0.0928	0.06	0.05
KPSS	0.1806	0.1	0.05
ADF (Augmented Dickey-Fuller test)	-6.4269	0.01	0.05
PP (Philips-Perron test)	-110.81	0.01	0.05

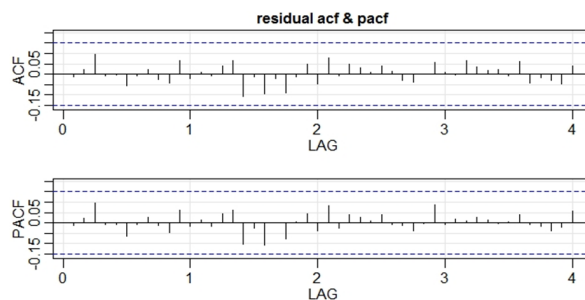**Fig. 2.** *ACF and PACF of the mean monthly inflow data.*

Fig. 3. *ACF and PACF of SARIMA (0, 0, 1) (1, 0, 2) 12 residuals.*

Table 2. Identified models with AIC and BIC values.

MODEL	AIC	BIC
SARIMA(0,0,2)(1,0,2)12	3032.67	3055.03
SARIMA(0,0,1)(2,0,1)12	3031.55	3050.71
SARIMA(0,0,1)(2,0,0)12	3046.99	3062.97
SARIMA(0,0,1)(1,0,2)12	3030.89	3050.04

Table 3. Parameter values of SARIMA (0,0,1) (1,0,2)12.

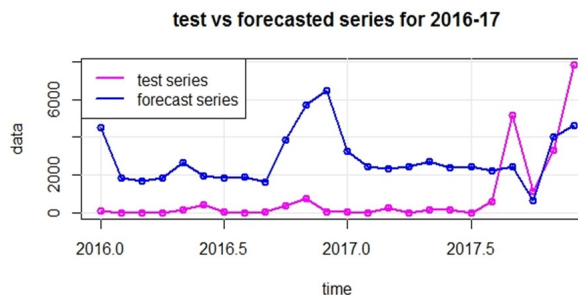| Parameter	MA1	SAR1	SMA1	SMA2	Intercept
Values	0.5214	0.9522	-0.9938	0.2173	994.929

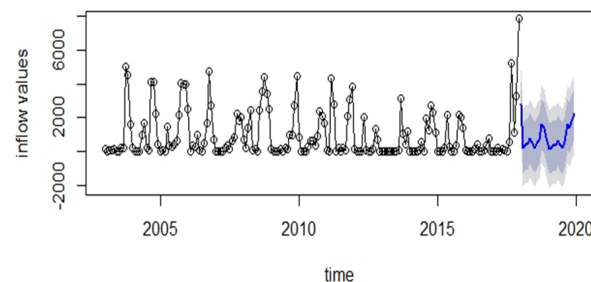**Fig. 4.** *Time series plot of test series vs. forecasted series.*

Fig. 5. *Forecasted Value of the model ARIMA (0,0, 2) (1,0, 2)$_{12}$.*

by practice and suitable model is selected based on lowest AIC and BIC values. Training data set of 2003 to 2015 were used in model identification.

From ACF graph, there is no negative spike and ACF cuts off after I lag andit shows significant spike at lag 12. So the Q value would be 1. Similarly, PACF plot has zero negative spikes and positive spike similar to ACF. So, the value of p was

greater than the significant value 0.05. Augmented Dickey-Fuller (ADF) and Phillips-Perron (PP) test statistic valueswere -6.4269 and -110.81. Their p value were0.01, lesser than the alpha value 0.05. It confirms that the data is stationary. In addition, KPSS test was also performed to prove stationarity in the dataset. But the inference of KPSS is contrast to that of ADF and PP test. The p value of KPSS test is 0.1 which is greater than 0.05, confirms that the data is Stationary.

Model Identification: Since the data set is stationary, there is no need of differencing (d=0). SARIMA (p, 0, q) (P, 0, Q) $_{12}$ were the suggested model. The order of AR and MA are determined by examining the ACF and PACF plots shown in Fig.2. Different P, p, Q, q values were determined

selected as 1. Thus, by practice various models were identified and shown in Table 2 with their AIC and BIC values.

From Table 2, lowest AIC and BIC values are recorded for SARIMA (0, 0, 1) (1, 0, 2)$_{12}$. So, SARIMA (0, 0, 1) (1, 0, 2)$_{12}$ model was selected as the suitable model for further process.

Parameter estimation: The parameters of the selected model are estimated by Maximum-likelihood function and the values are given in Table 3.

Diagnostic checking of the model: To verify the significance of the identified model, ACF and PACF of the residuals are plotted and cross checked against white noise. From Fig.3, ACF and PACF spikes were settled down within the 95 per cent confidence interval which concluded that the residuals were very small in relation to its standard error and there is no significant correlation between them. So, the residuals of the model were independent or it is simply a white noise term. In addition to this, Ljung-Box Q' statistic is performed to confirm the independency nature of residuals.

Table 4. Ljung box test statistic at lag = 20.

Ljung Box test	Values
Q statistic	9.2049
P value	0.9548
Alpha	0.05

Table 5. Forecasting performance of SARIMA (0, 0, 1) (1, 0, 2)$_{12}$ model.

Parameters	RMSE	MAE	MAPE	MPE	Theil's U statistic
Values	2818.036	2509.354	90.4739	68.5449	0.8497

From Table 4, Ljung Box Q' test the statistic is given as 9.2049 whose p value is 0.95 greater than significant value 0.05, concluded that the residuals of the identified model are not correlated with each other. Now, the performance of the selected model is cross-checked by forecasting the training series and comparing it with the available test series set. The model is validated and its forecasting efficiency is calculated by comparing the forecasted and available data set.

Thus, the forecasting performance of the selected model is confirmed by RMSE, MAE, MAPE and Theil's U statistic values. The RMSE, MAE, MAPE and Theil's U statistic value are shown in Table 5.

Theil's U statistic is the test procedure commonly employed to validate the forecasting accuracy of the selected model. If the value of Theil's U is greater than 1, the obtained model is not the best one to forecast the series. If the value is equal to 1, the selected model is equal to the naive method. For the best fit model, the U's statistic value should be less than 1. Since, the U's statistic value of SARIMA (0, 0, 1) (1, 0, 2)$_{12}$ is 0.8497* which is lesser than 1, indicating that the model is having greater forecasting accuracy to predict the future activity. Thus, the SARIMA (0, 0, 1) (1, 0, 2)$_{12}$ is the best suitable model to be used for forecasting process. Fig.4 shows the performance of the selected model by plotting the forecasted series and test series in a single plot.

Forecasting of inflows: Since the selected SARIMA model is considered to be the best suitable model for forecasting the dam inflow series. So, the inflow series of 2018 to 2019 were forecasted using SARIMA (0, 0, 1) (1, 0, 2)$_{12}$model and are shown in Fig.5.

From Fig.5, it is verified that the forecasted inflow series values were lying within the confidence interval and so the values can be further used for water allocation planning and dam management practices.

Conclusion

The inflow series of the Palar-Porundalar dam were forecasted in R statistical package. From various model, SARIMA (0, 0, 1) (1, 0, 2)$_{12}$ was selected as the suitable model for forecasting the series. These forecasted series may used for future irrigation planning and dam functioning process.

REFERENCES

1. Akaike, H. (1974). A new look at the stationary model identification. *IEEE Transactions on Automatic Control*, 19(6), 716-723.
2. Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (2008). Time series analysis: forecasting and control, John Wiley and Sons, New Jersey.
3. Commission Internationale Des Grands Barrages, International Commission on Large Dams (2018). 'Role of dams' from www.icold-cigb.org/GB/dams/role_of_dams.asp
4. Cryer, J.D and Chank, S. (2008). Time series analysis: with applications in R, Springer- New York.
5. Dickey, D.A and Fuller, W.A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366), 427–431. https://doi.org/10.2307/2286348.
6. Hamilton, J.D. (1994). Time Series Analysis. Princeton University Press, New Jersey.
7. Kendall, M. (1975). Multivariate analysis. Charles Griffin and Company, London.
8. Kwiatkowski, D., Phillips, P.C.B., Schmidt, P and Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 54, 159-178. https://doi.org/10.1016/0304-4076 (92)90104-Y.
9. Ljung, G.M. and Box, G.E.P. (1978). On a measure of a lack of fit in time series models. *Biometrika*, 65 (2), 297–303. https://doi.org/10.1093/biomet/65.2.297.
10. Moeeni, H., Bonakdari, H. and Ebtehaj, I. (2017). Monthly reservoir inflow forecasting using a new hybrid SARIMA genetic programming approach. *Indian Academy of Sciences*, 126(18). DOI: 10.1007/s12040-017-0798-y.
11. Phillips, P.C.B and Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, 75(2), 335–346. https://doi.org/10.1093/biomet/75.2.335.
12. Richter, B.D and Thomas, G.A. (2007). Restoring Environmental Flows by Modifying Dam operations. Ecology and Society, 12(1).
13. Salas, J.D., Delleur, J.W., Yevjevich, V. and Lane, W.L. (1980). Applied modeling of hydrologic time series. *Water Resource Publications*.
14. Singh, M., Singh, R. and Shinde, V. (2011). Application of software packages for monthly stream flow forecasting of Kangsabati River in India, *International Journal of Computer Applications*, 20(3):7–14.
15. Tadesse, K.B. and Dinka, M.O. (2017). Application of SARIMA model to forecasting monthly flows in Waterval River, South Africa *Journal of Water and Land development*, 35(X-XII), 229-236. https://doi.org/10.1515/jwld-2017-0088