RIVER FLOW FORECASTING USING MULTIVARIATE TIME SERIES MODELING

Sudheer, K. P., *Scientist 'B'*[1], Gosain, A. K., *Professor*[2] and Ramasastri, K. S., *Scientist 'F'*[3]
[1]National Institute of Hydrology, Deltaic Regional Centre, Siddartha Nagar, Kakinada
[2]Department of Civil Engineering, Indian Institute of Technology, Delhi
[3]National Institute of Hydrology, Roorkee

ABSTRACT

Making efficient forecasts of floods is one of the major tasks in flood hydrology. In situations, where the major concern is to accurately predict the river flows or floods, a hydrologist may make use of time series modeling approach instead of developing a conceptual model for the basin. In this paper, river flow forecasting is approached assuming that daily flows follow an auto regressive moving average (ARMA) process. Based on auto correlation and partial auto correlation functions, various ARMA models were considered and evaluated for their performance. The ARMA(3,1) model was found suitable, while others were discarded based on statistical analysis. The model has been used for forecasting daily flows using a one step ahead procedure. The forecasted series was analyzed for model performance and found satisfactory. The model was developed using the data of the Baitarani river basin, Orissa.

1.0 INTRODUCTION

One of the major tasks in flood hydrology is to make efficient forecasts of the occurrence of flood events. Flood forecasts are useful for issuing flood watches, alerts and warnings, so that flood emergency procedures can be implemented effectively. Flood forecasts can be made by various approaches, which among other factors depend on the size of the basin of interest, the available networks of hydro-meteorological stations etc. One approach for forecasting the flood is to transform the rainfall data into runoff using rainfall-runoff models. Numerous conceptual rainfall-runoff models are available (Burnash, et. al., 1973; Brazil and Hudlow, 1980). However, the implementation and calibration of such a model can typically present various difficulties (Duan et. al., 1992). They require sophisticated mathematical tools (Duan et. al., 1992, 1993, 1994; Sorooshian, et. al., 1993), significant amount of calibration data (Yapo et. al., 1995) and some degree of expertise and experience with the model.

While such conceptual models are of importance in the understanding of the hydrological processes, there are many situations such as stream flow forecasting where the major concern is with making accurate predictions at specified watershed locations. In such a situation, a hydrologist may prefer not to spend the time and effort to develop and implement a conceptual

model, and instead implement a simpler linear time series model. Short term forecasts and synthetic sequences of hydrologic data are obtained generally through time series modeling, because they are relatively easy to develop and implement. They have been found to provide satisfactory predictions in many applications (Bras and Rodrignez-Iturbe, 1985; Salas et. al., 1988; Wood, 1980). The present paper reports a research work conducted to develop an ARMA model for forecasting the daily flows during monsoon season in the Baitarani River of Orissa, India. Various steps involved in developing and implementing such a model is described in brief.

2.0 MODEL DEFINITION

For the purpose of present investigation, the general form of ARMA(p,q) model with p auto regressive terms and q moving average terms is used. The model is represented as,

$$Q_t = \sum_{i=1}^{p} \phi_i Q_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \qquad (1)$$

where, Q_t and Q_{t-1} are the observations of stochastic component of the time series at time t and t-i respectively; ϕ_i and θ_i are auto regressive and moving average coefficients respectively, ε_t is assumed to be white noise normally distributed with zero mean.

2.1 Data base and analysis

The data used in this work refers to the flow in the river Baitarani in Orissa State, shown in Fig 1. The total area of the basin up to Anandpur gauge discharge site is 8570 sq. km. The monitoring system consists of a network of rain gauges and a gauge discharge station. Daily values of stream flow during monsoon season (June to October) for 22 years (1972-1994) has been employed in the present time series analysis.

In representing various components of a hydrologic time series, the first problem, which may be encountered, is due to the possible presence of inconsistency and non-homogeneity of the historical data. These transient components can be identified by statistical analysis of the historical data, then properly described and removed. The resulting series will maintain the periodic and stochastic components of the original process. The periodic components are induced by the annual astronomic cycles. This phenomenon explains within the year periodicity in mean and variance. Salas et. al. (1988) suggested the use of a Fourier series analysis, and has been employed in the present study to remove the periodicity from the time series.

The Fourier series fit procedure requires the selection of the number of the significant harmonics. Salas et. al. (1988) provides a procedure for selecting the number of significant harmonics by plotting the periodogram. However, this procedure added too many harmonics to the function (Aboitiz et. al., 1986). Thus the selection of the number of significant harmonics was done by visual inspection of the resulting function. The number of selected harmonics, h*, was chosen

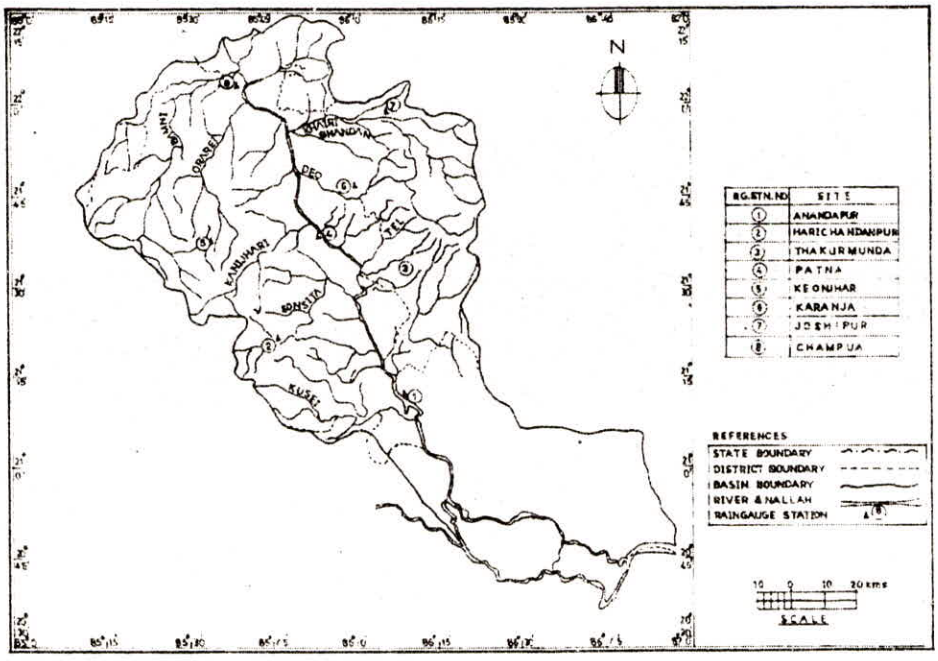

Fig. 1 : The Basin map and location of rain gauges in the Baitarani River Basin

by plotting the periodic parameters and the Fourier series function for several values of h*, and inspecting these plots. As climatic conditions should not change drastically in the basin from day to day over the season, it can be expected that the population daily mean and standard deviation will be reasonably smooth function over time. Therefore, the value of h* selected was that which produced a smooth function without much fluctuations.

2.2 Identification of the candidate models

In the model identification stage, the relationship between two time series may be studied by examining the cross correlation structure of the series (Salas et. al., 1988). The cross correlation function of an auto regressive [AR(p)] model is infinite in extent and the partial auto correlation is zero after lag *p*. The autocorrelation and partial autocorrelation analysis of individual series complements the cross-correlation analysis. In the present investigation, identification of candidate models is done by means of sample auto correlation and partial auto correlation functions. These functions reveal the correlation structure of the time series and, thus, are helpful in determining the underlying stochastic process.

2.3 Forecasting equations

Because the interest here is to forecast daily flow values one day in advance, explicit expressions are given here for such forecasts based on ARMA model (Box and Jenkins, 1976).

$$Q_t(1) = \phi_1 Q_{t-1} + ... + \phi_p Q_{t-p} - \theta_1 \varepsilon_{t-1} - ..... - \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (2)$$

in which

$$\varepsilon_{t+k} = \begin{cases} Q_{t+k} - Q_{t+k-1}(1) & k = 0,-1,-2,.. \\ 0 & k = 1,2,3,..... \end{cases} \quad (3)$$

The one step ahead forecast procedure was adopted for forecasting the standardized flow series in question. In order to get the forecasted values of flow, a reverse standardization was performed on the residual series.

3.0 RESULTS AND DISCUSSION

The foregoing approach for time series prediction has been applied to the monsoon flow data of the Baitarani River. The daily flow data during monsoon season for 22 years has been employed for identifying a suitable ARMA model and for estimating its parameters. The Fourier fit of the daily means and standard deviations for different harmonics are presented in Fig 2. The Fourier series model with three harmonics fitted well to the periodic mean, except at the beginning and at the end of the season (Fig 2). The values of Fourier coefficients for both the parameters are depicted in Table 1. The model explained about 67% of the variance in the sample mean series (Table 1). In case of the periodic standard deviation, the Fourier series fit showed similar fluctuations as that of the periodic standard deviation series (Fig 2), but was only able to explain about 28% of the variance (Table 1). However, for both the mean and standard deviation, the fitted models resulted in smooth functions, which can be expected with a large sample size.

Table 1 : Parameters of Fourier series models for daily mean and standard deviation

Parameters	Mean	Standard deviation
Seasonal mean, $\bar{u}$	330.03	382.86
Seasonal variance	203.35	355.37
Fourier coefficients		
A_1	-231.43	-227.54
B_1	6.98	53.15
A_2	-38.96	-87.42
B_2	1.61	-11.62
A_3	-44.11	-114.1
B_3	16.41	-1.46
Overall explained variance (% of total)	67.99	28.59

Estimates of the periodic mean and standard deviation obtained from the fitted Fourier series models were utilized to obtain the standardized flow series. The mean and standard deviation of the resulting standardized series were found to be 0.0 and 1.63 cumec, respectively, which are sufficiently close to the theoretical values (0.0 and 1.0 respectively).

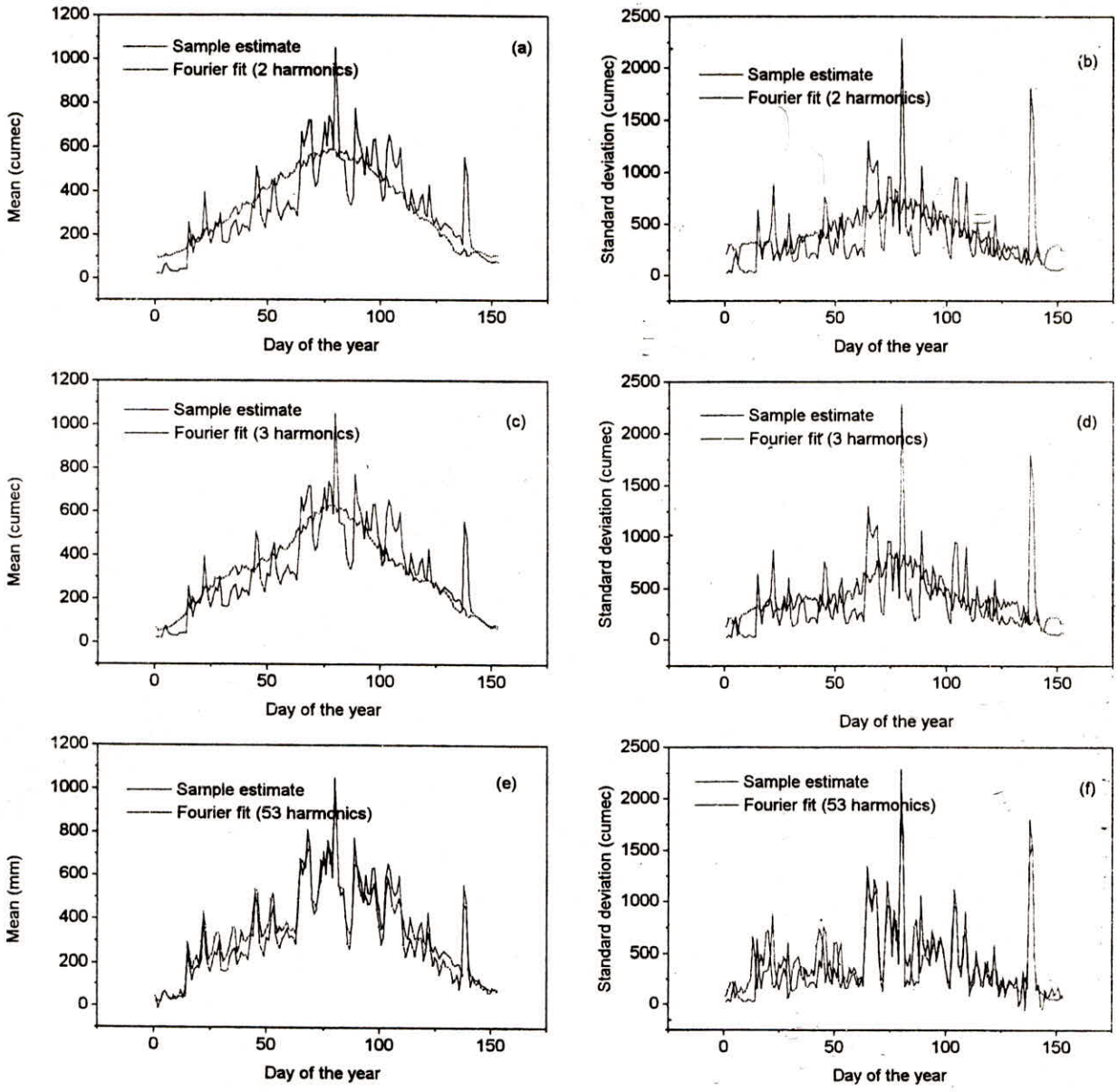

Fig 2 Plot of Fourier fit periodic means and standard deviation
(a,b): 2 harmonics (c,d): 3 harmonics (e,f): 53 harmonics

The auto correlation function (ACF) and the corresponding 95% confidence bands from lag 0 to lag 16, (0 to 16 days) were estimated for the standardized flow series. The results are shown in Fig 3. Lag 0 auto correlation is always unity as it is correlation of the variable with itself. However, as the lag increases the correlation between the variable and the same variable at specified lag decreases, i.e., covariance decays. The auto correlation function showed significant correlation, at 95% confidence level (Anderson and Jenkins, 1970), up to lag 7 (7 day), and thereafter, fell below the confidence band. The gradual decaying pattern of auto correlation exhibits the presence of a dominant auto regressive process. Similarly the partial autocorrelation function (PACF) and corresponding 95% confidence (Barlett, 1946) were estimated for lag 0 to lag 16, and are shown in Fig 4. The PACF showed significant correlation at lag 4 (4 day) and thereafter fell below the confidence band. The rapid decaying pattern of the PACF also confirms the dominance of auto regressive process, relative to the moving average process.

Based on the behavior of ACF and PACF, the underlying stochastic process can be characterized by AR(1), AR(2), AR(3), AR(4), AR(5), ARMA(1,1), ARMA(2,1), ARMA(3,1), ARMA(4,1), ARMA(5,1). Even though the PACF only showed significant correlation at lag 4, the higher order AR and ARMA models were also considered as candidate models, as the values of the PACF at lag 5 was very close to the significant level, as can be seen in Fig 4. Furthermore, the ACF and PACF are only helpful in identifying the most significant processes. It is difficult, however, to determine the final model structure from these functions. Thus, each of the above selected candidate models is to be subjected to further analysis. The best model can only be determined by comparing the performance of each of the candidate models, evaluated using various numerical indices reported for performance evaluation.

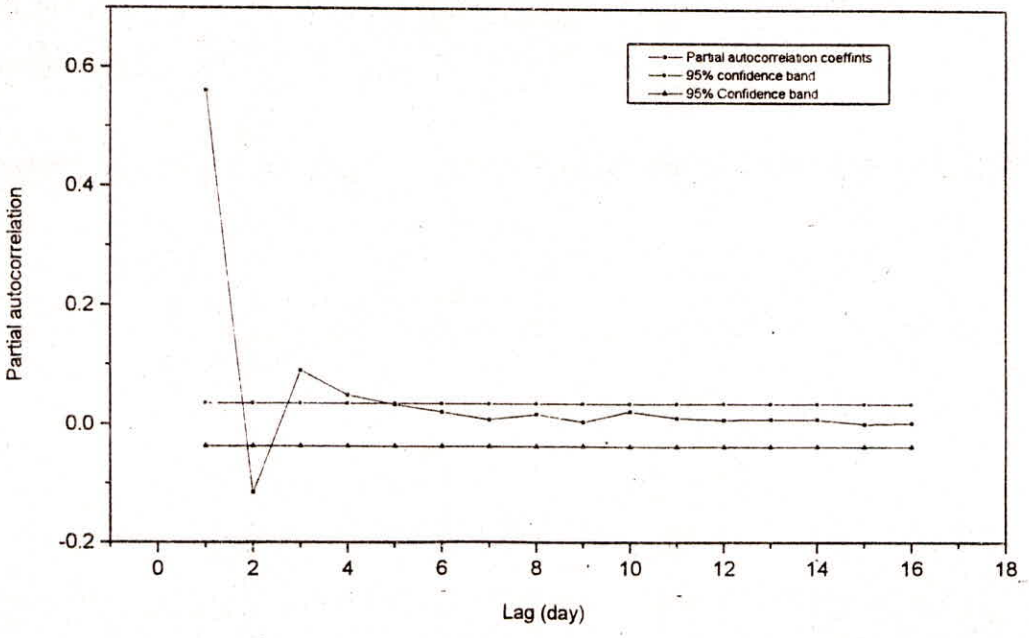

Fig. 3 : ACF plot of the standardized flow series

3.1 Parameter estimation

The method of least squares was employed to obtain the parameter estimates for all chosen candidate models. This method, which minimizes the residual sum of square, was based on a constraint optimization method proposed by Marquardt (1963). Estimates of various parameters

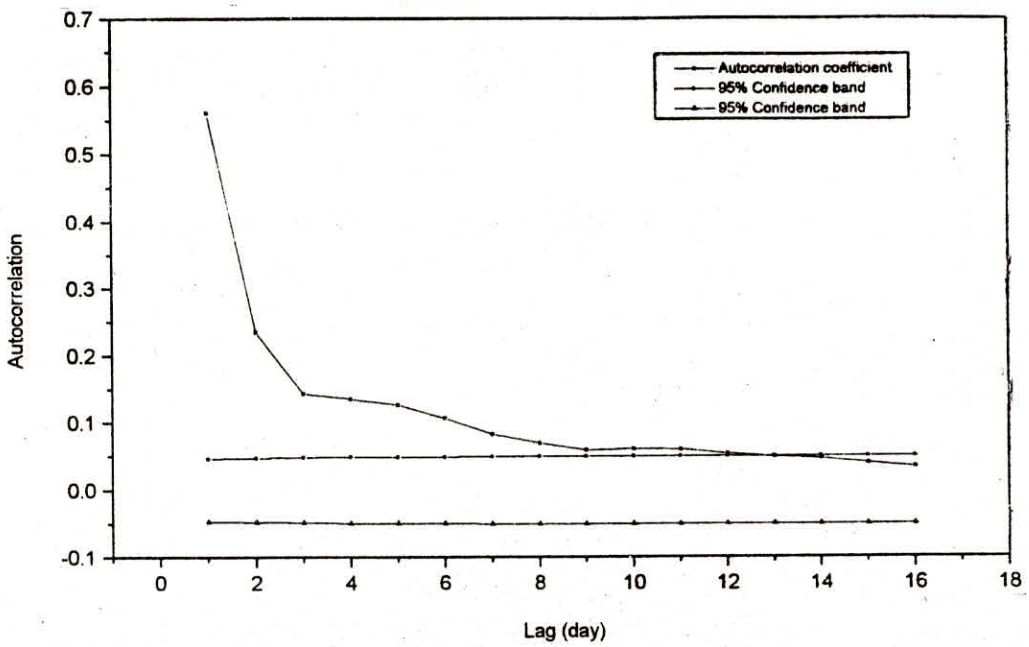

Fig. 4 : PACF plot of the standardized flow series

are given in Table 2. The residual variance of all the models were very similar suggesting that higher order model may not improve the results. In order to select the correct model form among the similar competing ARMA models, the Akaike Information Criterion (AIC) was used. Using this criterion, the model, which gives the minimum AIC, is the one to be selected. The AIC is calculated using the following relationship suggested by Shumway, (1988):

$$AIC = \ln(RMSE) + \frac{2n}{N} \qquad (4)$$

where RMSE is the root mean square error, N is the sample size and n is the number of parameters estimated. The AIC was minimum for ARMA(3,1) model (Table 2), suggesting the best ARMA model is ARMA(3,1). However, AIC has a tendency of overfitting parameters

Table 2 : Parameter estimates of ARMA(p,q) models

	AR(1,0)	AR(2,0)	AR(3,0)	AR(4,0)	AR(5,0)	AR(1,1)	AR(2,1)	AR(3,1)	AR(4,1)	AR(5,1)
ϕ_1	0.5571	0.6223	0.6334	0.6289	0.6273	0.3840	0.2495	1.2692	1.5376	1.5578
ϕ_2		-0.1171	-0.1761	-0.1677	-0.1670		0.0843	-0.5682	-0.7453	-0.7541
ϕ_3			0.0947	0.0646	0.0705			0.1628	0.2353	0.2244
ϕ_4				0.0474	0.0254				0.0558	-0.0262
ϕ_5					0.0351					-0.0227
θ_1						-0.2533	-0.3850	0.6450	0.9118	0.9330
RV*	1.8820	1.8568	1.8408	1.8373	1.8357	1.8491	1.8482	1.8335	1.8324	1.8332
AIC	0.6331	0.6203	0.6124	0.6112	0.6111	0.6162	0.6164	0.6091	0.6093	0.6104
BIC	0.6360	0.6263	0.6213	0.6231	0.6260	0.6221	0.6253	0.6211	0.6242	0.6283

RV: Residual variance

(Shumway, 1988). Modifications to AIC have been proposed to improve the large sample performance, as Shibata (1976) showed that a consistent estimator for the order is not found using AIC. Therefore another criterion for selection of the model order, the Bayesian Information Criterion (BIC), was also used.

According to Shumway (1988), BIC leads most often to correct estimates of the model order and has the smallest prediction error. It is defined as,

$$BIC = \ln(RMSE) + \frac{n \ln(N)}{N} \qquad (5)$$

where all the terms are as previously defined. Using BIC as well as AIC as model order selection criteria, ARMA(3,1) was selected as best model, since it gave minimum value of BIC and AIC. The t-statistics were computed for all the model parameters and compared with the t-critical at 95% significant level. The t-test indicated that all parameters except for ϕ_1, ϕ_2, ϕ_3 for all AR and ARMA models and θ_1 for all ARMA models were not statistically different from zero at the 95% significance level, suggesting that the use of higher order ARMA models may not improve the results. This also confirms the suggestion of selecting ARMA(3,1) model for forecasting the standardized flow series.

3.2 Evaluation of the ARMA(3,1) performance

The ARMA(3,1) model was employed to forecast the river flows during monsoon season in the Baitarani river. One step ahead forecasting procedure was adopted for the present study (Box and Jenkins, 1976). In this procedure, the lead-time is advanced by one time step only and the first forecast is obtained. Before making the next forecast, the actual value is substituted in place of forecast and this is used to forecast the next one. Here the parameters of the model are assumed to be the same while forecasting. This is helpful in assessing the capability of the model to reproduce original series.

The foregoing forecast approach has been applied to the flow series data of Baitarani river basin so as to forecast the flood. The ARMA(3,1) model was evaluated for statistical moment preservation, namely: the first, second, third and fourth moments about the mean and the results are presented in Table 3. The table presents the values of these moments for three years. The model was able to preserve first and second moment in all the years. The third moment was also preserved well, except in the year 1980. The fourth moment was not preserved very well. The total volume of flow (area under the hydrograph) was preserved very well in all the years.

The main criteria for judging the performance of the model is a visual comparison between the recorded and model fitted stream-flow hydrograph. However, visual comparison may involve judgement of the modeler. To avoid this, several numerical criteria are described in the literature to evaluate the performance of the model. A plot of the observed and forecasted flow during different years is depicted in the Fig 5. A visual inspection of the hydrograph leads to satisfactory performance of the model. The figure reveals that the model was able to forecast the

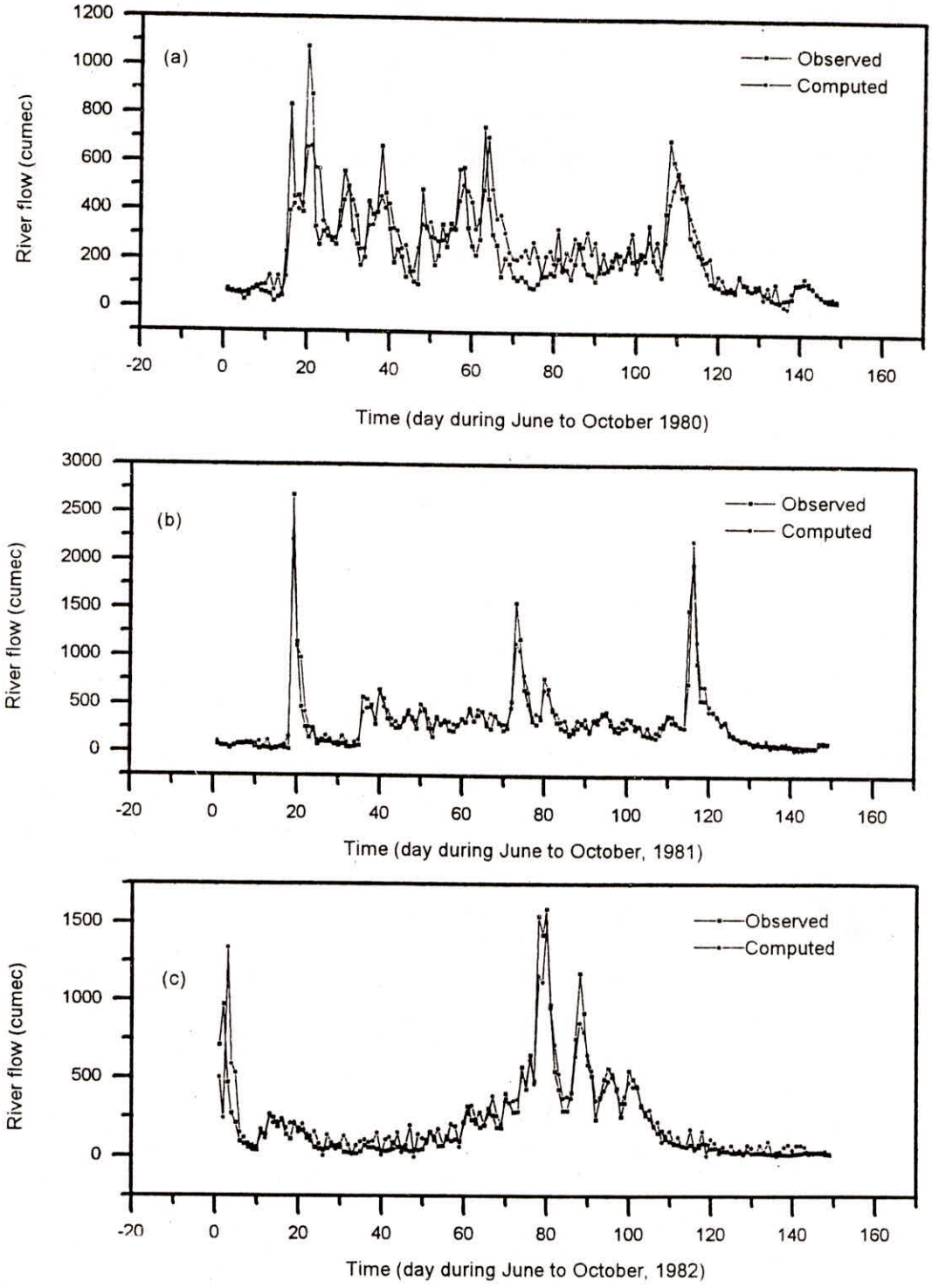

Fig. 5 : Comparative plot of observed and forecasted monsoon flows : (a) 1980 (b) 1981 (c) 1982

peak flow and time to peak almost accurately. However, to avoid subjectivity during visual inspection, various numerical indices were considered to statistically evaluate the performance of the model. The statistical indices considered are the root mean square error (RMSE), the percentage error in maximum flood (%MF) and the correlation (CORR) between the observed and forecasted series. The values of these indices during different years are presented in Table 3 and are discussed below.

Table 3 : The descriptive statistics of the recorded and ARMA(3,1) forecasted river flows

	1980		1981		1982	
	Recorded	Computed	Recorded	Computed	Recorded	Computed
Mean	220.97	242.17	302.49	314.16	214.92	231.24
Standard deviation	185.76	149.73	342.14	311.50	285.32	260.31
Skewness	1.77	0.49	3.83	3.51	2.66	2.28
Kurtosis	10.11	4.22	54.63	33.84	40.12	34.25
Peak flow (cumec)	1079.40	664.64	2678.40	2212.78	1592.90	1514.47
Total flow (million cubic meters)	14609.55	16092.09	19858.56	20628.93	14491.53	15671.85
RMSE (Std. Series)		0.40		1.20		0.77
%MF		38.43		17.38		4.92
Correlation		0.84		0.95		0.90

The RMSE statistic measures the residual variance; the optimal value is 0.0. The value of RMSE was minimal compared to the flood values observed in the series. The RMSE value of the standardized series was minimal which corresponds to about 150 cumecs. This value is small compared to the high flows during the season. Hence the model could be said to be performing satisfactorily. The percentage error in maximum flow (%MF) measured the percent error in matching the maximum flow of the data record; 0.0 is the best, positive values indicates over estimation, and negative values indicate under estimation. The model slightly over estimated the peak flow during all the years analyzed, however the estimation was only 5% high in the year 1982. The correlation statistic measured the linear correlation between the observed and simulated flows; the optimal value is 1.0. The correlation coefficient (CORR) is satisfactory for all the years under analysis. These analysis reveal that the ARMA(3,1) model was satisfactorily forecasting the river flows in the Baitarani river during monsoon season.

4.0 SUMMARY AND CONCLUSIONS

A research study has been conducted to develop a time series model to forecast the daily flows during monsoon season in the Baitarani River, Orissa. The historical flow series exhibited significant correlation (95% level) at lag 4 days. On the basis of auto correlation and partial auto correlation functions, this series could be characterized by autoregressive moving average (3,1) model. Parameters of the model were estimated using the least square method. A t-statistic analysis revealed that higher order ARMA models might not improve the results significantly. The estimated parameters of the ARMA(3,1) model were used to forecast the flow series in a one step ahead procedure. The results showed significant correlation with the recorded values (0.83

to 0.94). The performance of the model has been evaluated using various statistical indices too and the results of the analysis revealed that ARMA(3,1) model was able to forecast the stream flow to a satisfactory level. This model could be employed in efficient flood management of the Baitarani River.

REFERENCES

Aboitiz, M., J. W. Labadie, and D. F. Heerman, (1986). *Stochastic soil moisture estimation and forecasting for irrigation fields.* Water resources research, 22(2):180-190.

Akaike, H. (1974). *A new look at the statistical model identification.* IEEE transactions of automation and control, AC-19:716-723.

Bartlett, M. S. (1946). *On the theoretical specification and sampling properties of autocorrelated time series.* Supplement to the Royal Statistical Society. 8:27-41.

Box, G. E. P. and G. M. Jenkins, (1970). *Time series analysis, forecasting and control.*, Holden day, San Fransisco.

Bras, R. L., and I. Rodriguez_Iturbe. (1985). *Random functions in hydrology.* Addison-Wesley, Reading, Mass.

Brazil, L. E. and M. D. Hudlow (1980). *Calibration procedures used with the National Weather Service Forecast system,* Paper presented at the IFAC Symposium on water and land resources systems. Int. Federation of automation and control, Cleveland, Ohio.

Burnash, R. J. E., R. L. Ferral, and R. A. McGuire. (1973). *A generalized stream flow simulation system,* Rep. 220, Jt. Fed. State River Forecast centre, Sacramento, California.

Duan, Q., S. Sorooshian, and V. K. Gupta. (1992). *Effective and efficient global optimization for conceptual rainfall runoff models.* Water resources research, 28(4):1015-1031.

Duan, Q., S. Sorooshian, and V. K. Gupta. (1994). *Optimal use of SCE-UA global optimization method for calibrating watershed models.* Journal of hydrology, 158:265-284.

Duan, Q., V. K. Gupta, and S. Sorooshian, (1993). *A shuffled complex evolution approach for effective and efficient global minimization.* Journal of optimization: Theory Application, 73(3):501-521.

Salas, J. D., J. W. Delleur, V. Yevjevich, and W. L. Lane, (1988). *Applied modeling of hydrologic time series*, Littleton, Colarado, Water Resources Publications.

Shibata, R. (1976). *Selection of the order of an autoregressive model by Akaike's information criteria.* Biomtrika. 63:117-126.

Shumway, R. H. (1988). *Applies statistical time series modeling*. Eaglewood Cliffs, New Joursey, Prentice Hall.

Sorooshian, S., Q. Duan, and V. K. Gupta. (1993). *Calibration of rainfall runoff models: Application of global optimization to Sacramento soil moisture accounting model,* Water resources research, 29(4): 1185-1194.

Wood, E. F. (Ed.) (1980). *Workshop on real time forecasting control of water resource system,* Pergamon, New York.

Yapo, P. O., V. H. Gupta., and S. Sorooshian. (1996). *Automatic calibration of conceptual rainfall runoff models: sensitivity to calibration data*. Journal of hydrology, 181:23-48.