Current World Environment Vol. **4**(1), 79-87 (2009)

Time series modeling and forecast of river flow

RASHMI NIGAM[1], SOHAIL BUX[2], SUDHIR NIGAM,
K.R. PARDASANI, S.K. MITTAL and RUHI HAQUE

[1]Department of Mathematics, MANIT, Bhopal - 462 003 (India).
[2]Department of Mechanical Engineeering, MANIT, Bhopal - 462 003 (India).

(Received: March 10, 2009; Accepted: May 03, 2009)

ABSTRACT

Changing climate, human interventions to natural water flow pattern, haphazard urbanization etc., are the reasons for intense flood even after development of so many structural measures of overflow control. Kulfo River basin is situated in relatively dry southern area of the Ethiopia and is still under geographical modification with hilly topography and impervious soil texture. The concern of the present research is to simulate flood episode in order to develop flood management strategies to reduce disaster. The complexicity of natural hydrological phenomenon and dependent random variables can be better expressed considering it as stochastic process. Flood (maximum river flow) forecasting on the Kulfo River with monthly runoff data using stochastic ARIMA, Time Series model was developed for warning purposes. The analysis of seasonally varying time series of discharge data has revealed that a higher order ARIMA model may produce excellent results for three to six months forecast.

Key words: Stochastic, Flow pattern, ARIMA model, Flood,
Perennial River, Inundation, Rainfall, Runoff, Time Series

INTRODUCTION

Hydroinformatics focus on applications of advanced information technologies and statistical tools for better understanding and management of the hydrological phenomenon. Hydrological phenomenons are cyclic and stochastic in nature. In hydroinformatics the river is considered as a water-based asset, with flows patterns largely as stochastic. River can be considered both as a generic object with properties pertaining to the flow behavior and as a particular object with its own unique characteristics. The significant information needed for river flood management is about past and present runoff in the river and the governing rainfall data covering the river catchment area coupled with the derived information about the human dimension, the historical, sociological, legal, economic and even political aspects (Patel and Shete, 2007).

The physical models are most efficient to estimate a flood event because such models have built within it knowledge of the physics, like the dimensions of the flood plains, variations in meteorological parameter, runoff coefficients, roughness values, head losses, etc.. However, when there is a need for very rapid predictions of flow patterns, say, in a forecasting situation, or when long time series are being run, or a Monte Carlo analysis is required, then the physicallybased model is cumbersome (Beven, 2001). There are now a number of such situations where time series analyses are being used to forecast such extreme events in rivers. Time series analysis allows identification of hidden deterministic behavior and thus understanding of cause and effect relationship in problems (Schwartz and marcus1990). The univariate model provides an increased understanding of the behavior of the system. The changes in the measured values are the function of shocks (the deviation from the expected results provided system dynamicsheld constant) to the time process occurring during the past months. So that the residuals are not correlated more than 5% insignificant level and normalize the residuals as much as possible. Thus the fitted model is the better choice (Murray and Farber, 1982).

The paper attempts to describe the occurrences of the river floods using stochastic time series forecasting method. The particular emphasis has been given for the accurate flood forecasting and warning for an effective management of flood disaster if required. The form of modeling used to make sense of the acquired data is Time Series ARIMA modeling. The results of an analysis of acquired and generated information supports decision to regulate the real world, the river, its environments and the people associated with it.

Study area and data

The applicability of data based stochastic analysis is studied for a perennial medium size hilly river named Kulfo. The river spans in great Abaya - Chamo basin of southern Ethiopia. The Abaya - Chamo basin has a catchments of about 16,400 square kilometers with river drain area consisting of about 3500 square kilometers. Rainfall pattern is mostly indefinite imparting frequent inundation to Kulfo river basin despite an average yearly rainfall figure. The river has response time of about six to eight hours depending upon the rainfall intensity.

For development of the flood forecast model at least one set of continuous flow measurement in each sub-catchments associated with a significant risk area, capable of capturing extreme flood discharges is desirable. Further, a standardized recording of key information, facilitating quick flood response should be maintained. In order to simplify the study only two hydrological variables are selected for the analysis. Data pertaining to the rainfall has been collected from Meteorological Station situated within 1 km range of river and river flow has been measured at rain gauge installed at river discharge point near Abaya Lake. The time plot of individual rainfall Mean Monthly Rainfall (MMR) and Mean Monthly Discharge (MMD) time series have been given in fig 1 in a comparable form to reveal quantitative consistency of MMD with MMR data set. The data ranges have been taken as monthly cumulative and eight year data beginning from January 1990 and continuing till the end of 1998.

Fig. 1 shows an examination of the quality of data base and the degree of independence for each variable. From these plots we infer that the data behave in a normalized manner and spread in data is uniform over the time with cyclic nature. The relationship between rainfall and river discharge is linear or very nearly so. There is some inconsistency in observed data towards the end of year 1996, where the runoff and rainfall are not concurrent in terms of quantity. The exact reason for which cannot be inferred however, the gap may be attributed to observational mistake or sudden release of dammed water.

Methodology of model development

Real time flood forecasting can be done using statistical, stochastic, deterministic and soft computing techniques. When the occurrence and outcome of a phenomenon, as in natural processes, are random or uncertain the process is characterized as stochastic (Priyan and Dalwadi, 2007). In hydrological phenomenon the rainfall is the main occurrence with runoff as its foremost associated outcome. Both rainfall and runoff are function of space and time and are covary geographically and temporarily (and seasonally). Therefore, the flood which too is a consequence of rainfall and runoff can smartly be represented and forecasted by use of stochastical modeling of historical runoff data. Stochastic model are good enough to capture sudden changes in natural flood however the gaps in flood data impair forecasting results.

The observation of rainfall and runoff taken at temporal order constitute the time series. The inherent cause effect relationship in the hydrological phenomenon of a stochastic time series can be analysed by applying of the Box – Jenkins approach (Box and Jenkin, 1994). The Box- Jenkins approach usages the concept of AutoRegression Integration and Moving Average (abbreviated as ARIMA) modeling, where the dependent variable is lag regressed onto itself and smoothened thus giving rise to the ARMA and related ARIMA and SARIMA models (S stands for the seasonally regressed time series). These models are applicable to stationary series, where there is no systematic change in mean (i.e. the series has been detrained) and variance is constant over time (Kendall and Ord, 1992).

The degree of dependence of variables is analyzed by estimating the autocorrelation function. for MMD series. These results are significant at the 95% confidence level or twice the standard deviation which is 0.068 (using $1/n^{½}$, where n = 84) the total numbers of observed variables. In general a non seasonal ARIMA model can be written as

$$(1-\phi_1 B-\phi_2 B^2-..-\phi_p B^p)\nabla^{\mathrm{d}}\mathrm{z}_{\mathrm{t}}=(1\text{-}\theta_1\mathrm{B}\text{-}\theta_2\mathrm{B}^2\text{-}..\text{-}\theta_{\mathrm{q}}\mathrm{B}^{\mathrm{q}})a_t$$

where at, denotes the residual series, B backward shift operator defined as $BZ_t = Z_{t-1}$, $B^2Zt = Z_{t-2}$ and so on and the terms of ϕ and θ denotes coefficient value of an autoregressive and moving average process of order p and q respectively. When an observation zt of a particular month has some relation with the observation made in the same month of the previous year, the seasonal dependency modify the equation as:

$$(1-\Phi_1 B-\Phi_2 B^{2s}-..-\Phi_p B^{ps})\nabla_s^{\mathrm{D}}\mathrm{z}_{\mathrm{t}}=(1\text{-}\Theta_1\mathrm{B}^{\mathrm{s}}\text{-}\Theta_2\mathrm{B}^{2\mathrm{s}}\text{-}..\text{-}\Theta_{\mathrm{q}}\mathrm{B}^{\mathrm{Qs}})e_t$$

where et is a normal random deviate, and seasonality s = 12 and the terms given by Θ and Φ, represents corresponding seasonal moving average and autoregressive operators of order Q and P. As et of seasonal ARIMA equation is not necessarily independent, therefore combining non seasonal and seasonal equations we get the general multiplicative seasonal ARIMA model of order (p,d,q) x (P,D,Q) of the form

$$\Phi_p(B^s)\phi_p(B)\nabla_{\mathrm{s}}^{\mathrm{D}}\mathrm{z}_{\mathrm{t}}=\Theta_{\mathrm{Q}}(\mathrm{B}^{\mathrm{s}})\text{-}\theta_{\mathrm{q}}(\mathrm{B})\mathrm{a}_t$$

RESULT AND DISCUSSION

Model development process

The objectives of this study are firstly to identify a suitable ARIMA model based on Box-Jenkins approach. Since rainfall and hence river runoff is a seasonal phenomenon, we need to identify the order (p,d,q) x (P,D,Q) for a seasonal univariate model and also to find out the degree of best fit seasonality, which provides a parsimonious representation for both the stochastic component and the total series under consideration. Finally the least square estimates of the parameters of time series models are used for forecasting the river flow. For identification of ARIMA model parameters (p, d, q) and (P, D, Q), the Auto correlations coefficient (ACC) and Partial autocorrelations coefficient (PAC) of MMD time series have been plotted for various combinations of differencing (d=0 and d=1) and lags. The graphical representation of ARIMA model building process is given in the fig. 2. The original and differenced data has been plotted to check the stationarity in the MMD time series. Identification of model parameter is mainly based on ACC and PAC plots of time series. The runoff of the Kulfo River shows a strong seasonal pattern, the same can be seen in the ACC and PACC plots and hence the flow pattern requires a seasonal model.

Table 1: Values of seasonal arima model parameters

Parameter	Coefficient	Std. Error Coefficient	t value	P
AR1 (ϕ_1)	-0.9997	0.0033	-307.51	0
SAR1 (Φ_1)	0.1887	0.1984	0.95	0.245
SAR2 (Φ_2)	-0.411	0.1744	-2.36	0.021
SAR3 (Φ_3)	-0.6456	0.1684	-3.83	0
MA1 (θ_1)	-1.018	0.0001	-6923.29	0
SMA1 (Θ_1)	1.708	0.2624	6.51	0
SMA1 (Θ_2)	-1.1789	0.3877	-3.04	0.003
SMA1 (Θ_3)	0.3167	0.3796	0.83	0.107
Constant	2.0979	0.1025	20.47	0

An analysis of significant ACC and PAC plots implies to the first order non seasonal ARMA and third order seasonal ARIMA parameterization of MMD series. Compared with 95% confidence limits, few of the partial autocorrelations (three) are found significant. The finally selected seasonal ARIMA forecast model for the mean monthly discharge of Kulfo river is (p,d,q, P,D,Q)S = (1,0,1,3,1,3)12. The other ARIMA model tried with different parametric value could not result in an

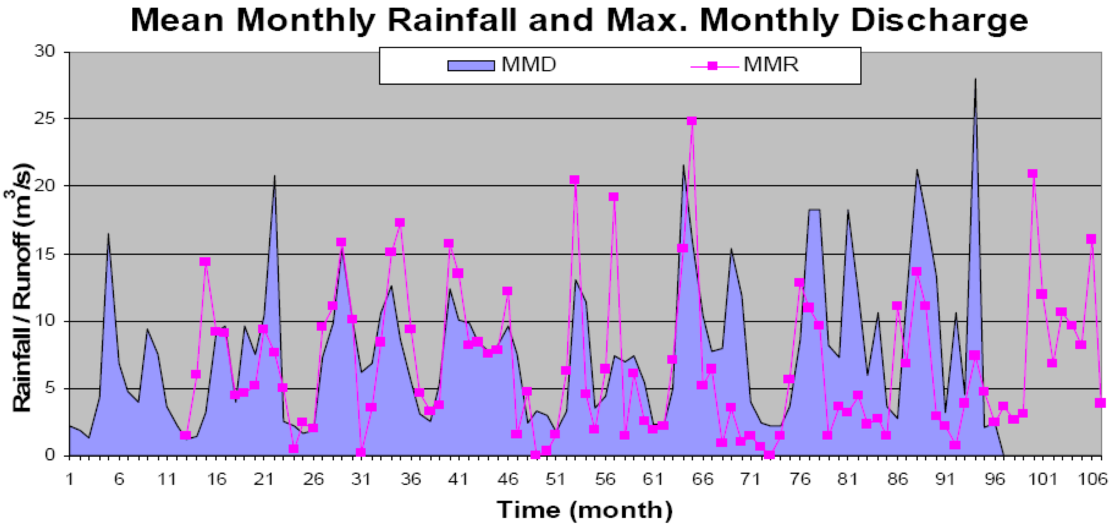
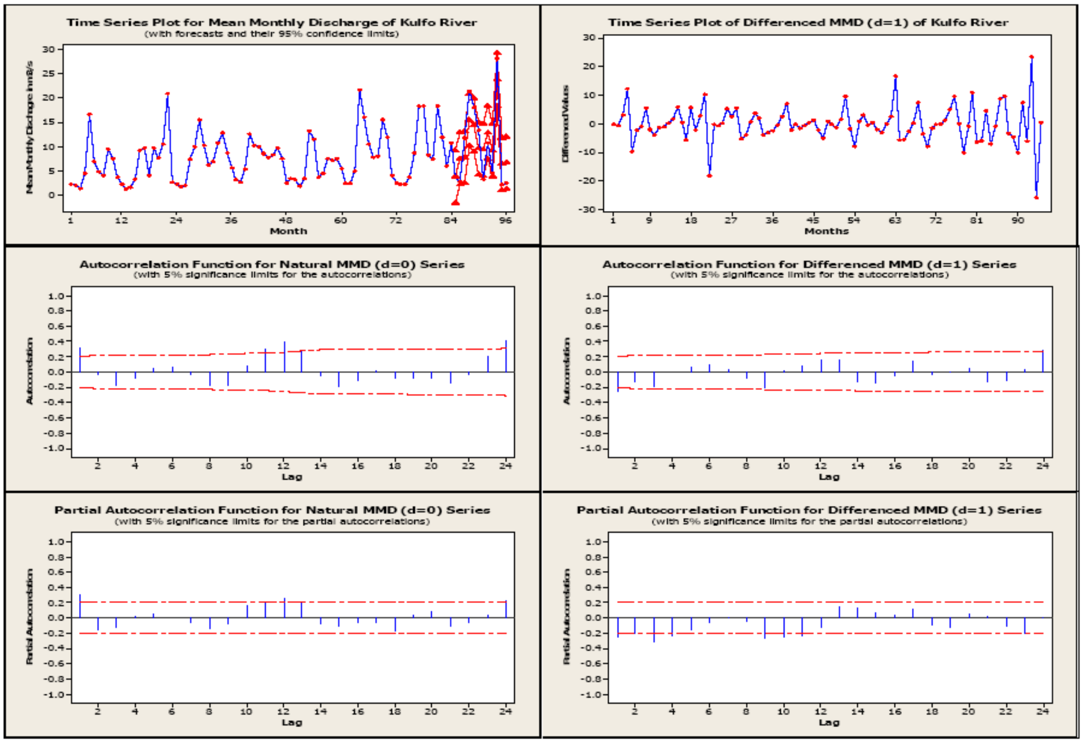

Mean Monthly Rainfall and Max. Monthly Discharge

Fig. 1: Rainfall and runoff of Kulfo River (Jan. 1990 - Dec. 1998)


Time Series Plot for Mean Monthly Discharge of Kulfo River
(with forecasts and their 95% confidence limits)

Time Series Plot of Differenced MMD (d=1) of Kulfo River

Autocorrelation Function for Natural MMD (d=0) Series
(with 5% significance limits for the autocorrelations)

Autocorrelation Function for Differenced MMD (d=1) Series
(with 5% significance limits for the autocorrelations)

Partial Autocorrelation Function for Natural MMD (d=0) Series
(with 5% significance limits for the partial autocorrelations)

Partial Autocorrelation Function for Differenced MMD (d=1) Series
(with 5% significance limits for the partial autocorrelations)

Fig. 2: Model building process of MMD time Series of River

impressive forecast. In table 1, the values of various AR and MA parameters for seasonal and non seasonal cases are given with corresponding standard error (SE) of coefficient, and t and p values. Small SE coefficient and p value with corresponding t statistic indicate the significance and the preciseness of the coefficient estimation.

The fitting of the properly transformed data to the time series model is accomplished by obtaining least square estimate of the parameters. The residuals from the univariate process are used

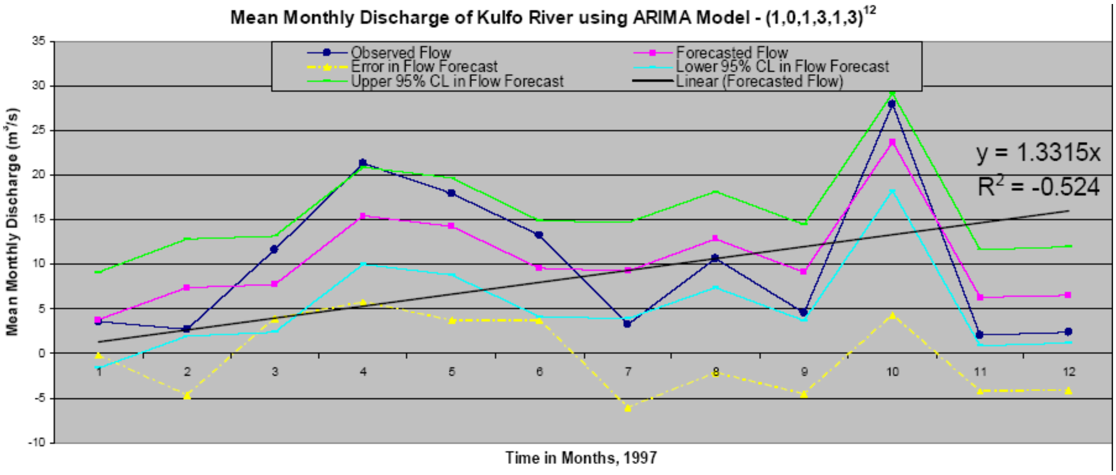


Fig. 3: Actual and simulated mean monthly discharge

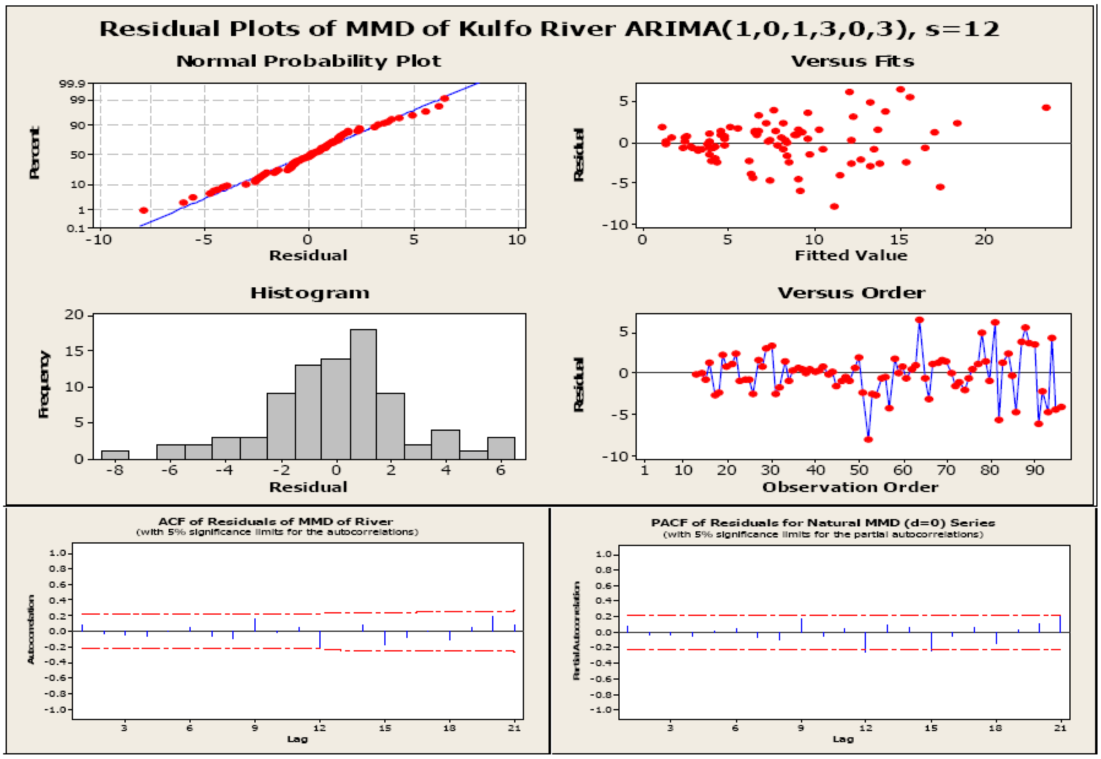


Fig. 4: Residual analysis of fitted arima model

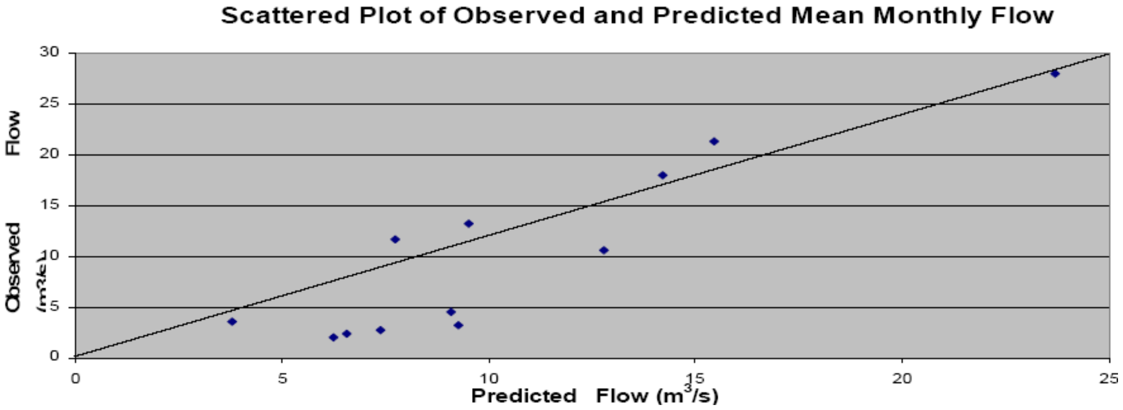


Fig. 5: Scattered analysis of fitted arima model

in the final selection of the complete dynamic model. For each iteration / phase of model fitting, the criteria for adequacy of model is that the residuals should be independent (i.e. no or negligible autocorrelation exist) and that the model exhibits parsimony (least number of parameters). The residuals should also exhibit a symmetric distribution (e.g. a normal distribution) (Murray et. al., 1982). The negative values of fitted auto regression parameter explains that the time series variables are related in a quash manner, means their effect is to reduce the discharge value.

The forecast

The process of creating reliable and robust analysis mostly results in the models of underpredicted extreme river discharge. This is in view of the fact that the computational modeling is limited to situations where the theoretical analysis is valid and sufficient data are available for proper calibration and verification. Monthly runoff discharge is forecasted for 12 months lead time taking January 1997 as the origin. It is obvious from fig. 3 that the results of forecast for all twelve months are reasonably close to the actual river flow and following the mean values. The forecasts are sensibly fair in capturing runoff pattern even in the case of astonishing peaks as in the case of October, 1997. The forecast error equally fall both sides of zero mean value and the forecasted values have a linear increasing trend.

Forecast evaluation

There are two methods to evaluate the performance of model results first one is using established statistical formulae and the other is stochastic, through the application of data itself i.e. residual analysis. The importance of the later one is justified because it emphasis more to judge the inherent data characteristics. The main advantage of stochastic method is that one can evaluate model performance by the same methods used for model building and any discrepancies in model formulation can be easily recognized so that modeler can take immediate decision for improvement of model. Former one is applicable for all models related to the natural process and are adequately explained in the literature.

Preliminary statistical evaluation

The results of Modified Box-Pierce (Ljung-Box) Chi-Square statistic, t test and on the forecasted values of model are as under:

Lag	12	24	36	48
Chi-Square	10.4	24.9	42.2	56.3
Degree of Freedom	3	15	27	39
P-Value	0.015	0.052	0.032	0.036

The accuracy of fit hypothesis is judged using chi square value corresponding to the lag and degree of freedom (DF), smaller p value corresponding to the adequacy of model fit.

Stochastical Evaluation (The Residual Analysis)

Residual (difference between the observed and predicted / fitted values) represents that part of the observation which is not explained by the fitted model. After fitting the model residual analyses have been done using a set of five residual plots:

1. Anormal probability plot.
2. The residuals versus the fitted values,
3. The residuals versus the order of the data
4. The histogram of the residuals
5. The residuals versus predictors (ACF and PACF for MMD time series).

A well fitted time series model is represented by the normally distributed but constant residual, which do not exhibit any pattern (trend, seasonality, cyclic etc.) as a function of the response variable. The residual analysis of runoff data is shown in figure 4 reveals that that

´ Histogram of residual clearly shows that most of the residuals are concentrated within a narrow range of zero. A large fraction of residuals are negative and within a narrow limit, indicating the model suitability to drag the optimum filtering of data.

´ The normal probability plot of residual also confirms the central tendency of residuals being between the ±1.5, i.e. that more than 95% residuals are normally distributed and are constant, that is, they do not exhibit a trend as a function of the response variable.

´ The plot of residual versus the order of the data reveals that most of the residual values fall within ±5% till 50 observations.

´ The plot of residual versus the fitted values shows that about 99% residual values falls within ±5% and the range of residuals lying centered within 10 of fitted values.

For a best fit time series model the residuals should be insignificant and their autocorrelation should be weak within 95% confidence limits. The plot of the residual's autocorrelation and partial autocorrelation clearly reveal that there is no dominant ACF and PACF of residuals till lag 12, i.e. ACF and PACF of residuals are insignificant confirming that the appropriateness of the fitted model. Thus the residual analyses establish that the residual sequence behaves as a white noise series and the fitted model perform well in the domain of observations.

Comprehensive statistical evaluation

Model evaluation analysis begins with univariate analysis. A quantitatively close output in

Table 2: Statistical evaluation of river runoff forecasts

Univariate Analysis				Error Analysis								Performance Measures					
		S. D. Skewness		Errors					RMSE								
O	**P**	**O**	**P**	**MPE**	**MBE**	**MAE**	**SDE**	**NMSE**	**T**	**S**	**U**	**CD**	**IA**	**FB**	**MG**	**VG**	**FAC**
8.6	5.4	0.93	1.41	-0.003	0.4	3.9	4.4	0.1232	4.2	0.5	4.5	0.952	0.898	-0.04	0.74	1.56	4
8.6	5.9	0.93	1.26	0.003	-0.3	3.2	3.7	0.0866	3.5	0.4	3.3	0.970	0.935	0.034	0.829	1.273	0

standard deviation and skewness in both observed and simulated data predict preliminary worthiness of model prediction. A scatter plot between the paired observations and predictions, figure 5, reveals the magnitude and spread of the model's over or under-predictions.

Statistical comparisons of model estimates or predictions with pair wise matched observations remain among the most basic means of assessing model performance in the hydrological studies. Hanna and Chang (2004) and ASTM (2000) propose some comprehensive statistical model performance measures which includes the fractional bias (FB), the geometric mean bias (MG), the normalized mean square error (NMSE), the geometric variance (VG), the correlation coefficient (R), Index of Agreement and the fraction of predictions within a factor of two of observations (FAC2). The detail discussion over the subject can be referred at Nigam *et al*., (2008). The Mean error or bias is the fundamental to judge the over or under predictive nature of model. The general relations among errors are as MBE ≤ MAE ≤ RMSE.

According to Oreskes et al (1994), evaluation (verification and validation) of mathematical models of natural systems are impossible, because natural systems are never closed and because model solutions are always non-unique. The random nature of the process leads to a certain irreducible inherent uncertainty. Thus models can only be confirmed or evaluated by the demonstration of good agreement between several sets of observations and predictions.

The numerical values of these parameters are given in the table 2 for the two cases first with the real forecast and another for the even out forecast within factor of 2. These criteria provide more information on the systematic and dynamic errors inherent in the model simulation. A perfect model would have MG, VG, R, and FAC2=1.0; and FB and NMSE=0.0. The modified performance values can be attributed to the implied mean forecasts (Boyle *et al.*, 2000).

CONCLUSION

The first order autoregressive parameter indicates a substantial degree of variability and dependence in the stochastic component; the domination of the first order parameter being the highest for the smallest catchment. The values of higher order persistence on the other hand speak of a fairly uniform degree of dependence in the stochastic components on past events. It is found that the simulated values are in good agreement with the observed runoff for the first six months. In the first three months the simulated values are much closed to the actual one and over-predicted which a desirable outcome from the hydrological models.

From third months to the sixth months forecasts three simulated values are in good concurrence with the observed pattern of the actual runoff hence can be considered an affirmative model contribution. After six months the pattern of simulated values are sensible to capture the peak discharges as well as the pattern of the past flow. However the quantitative results of model predictions are inconsistent and owing to under-predictive nature cannot be used for forecasting purposes. Though residual study confirms model suitability up to twelve month ahead forecast to a great extent but considering the practical requirements of forecasted values, model can be considered fair to six month forecasts only.

It is demonstrated that ARIMA modeling is a appropriate approach to model hydrological data which often exhibits autocorrelation with time and need proper explanation of underlying dynamics which cannot be done by simple statistical forecasting methods like regression analysis etc. The Box-Jenkins approach considers autocorrelations among the variables as well as laglead relationship between variables. If we add effect of rainfall data too in MMD series, i.e. multivariate approach, this will help in to attain a solid presumption of a cause-effect relationship in time series analysis.

REFERENCES

1. ASTM, "Standard Guide for Statistical Evaluation of Atmospheric Dispersion Model Performance," D6589, American Society for Testing & Materials, Conshohocken, PA 19428-2959 (2000)
2. Beven, J.K., "*Rainfall-Runoff Modelling*-The Primer," John Wiley & Sons Ltd., Chichester, 319 (2001).
3. Box, G.E.P., Jenkins, G.M., and Reinsel, G.C., "Time Series Analysis, Forecasting and Control," third edition, Prentice Hall, Englewood Cliffs (1994).
4. Boyle, D.P., Gupta, H.V., and Sorooshian, S., "Toward Improved Calibration of Hydrologic Models: Combining the Strengths of Manual and Automatic Methods," *Water Resources,* AGU, **36**(12): 3663-3674 2000.
5. Hanna, S.R. and Chang, J. C., "Air Quality Model Performance Evaluation. *Journal of Meteorology and Atmospheric Physics*," **87**: 167-196 (2004).
6. Kendall, M., and Ord, J.K., "Time Series, International Journal of forecasting," **4**: 532-533 (1992)
7. Legates, D. R. and McCabe Jr., G. J., "Evaluating the use of Goodness-of-Fit Measures in Hydrologic and Hydro Climatic Model Validation," *Water Resource*, **35**(1): 233-241 (1999).
8. Murray, L.C., and Farber, R.J., "Time Series Analysis of an historical visibility data base," *Atmospheric Environment*, **16**(10): 2299-2308 (1982).
9. Nigam, S., Kulshrestha, M., Mittal, S., and Singh, K., "Computational Modeling of Air Pollutants in Congested Traffic Condition," International Conference on Best Practices to Relieve Congestion on Mixed-Traffic Urban Streets in Developing Countries, IIT Chennai, Allied Publisher Pvt. Ltd., India, 389-400 (2008).
10. Patel, N.R., and Shete, D.T., "Probability Distribution Analysis of Consecutive Days Rainfall Data for Sabarkantha District of North Gujrat Region, India", National Conference on Hydraulics and Water resources, SVNIT, Surat Published by the Elite Publishing House Pvt. Ltd., India, 86-93 (2007).
11. Priyan, K., and Dalwadi, H.J., "Statistical Analysis of Hydrological Data: A study of meghal river basin," National Conference on Hydraulics and Water resources, SVNIT, Surat Published by the Elite Publishing House Pvt. Ltd., India, 16-23 (2007).
12. Schwarz, J., and Marcus, A., "Mortality and Air Pollution in London: A Time Series Analysis," *American Journal of Epidemiology,* **105**: 1273-1281 (1990).
13. Willmot, C.J., "Statistics for Evaluation and Comparison of Models," *Journal of Geophysics Research,* **90**(5): 8995-9005 (1985)