Mausam, (1988), **39**, 1, 65-70

551.509.313

A stochastic approach for monthly streamflow forecast

S. R. PURI, S. N. KATHURIA*, D. S. UPADHYAY and SURENDRA KUMAR

Meteorological Office, New Delhi

(*Received 14 September 1984*)

सार — पिछले 40 वर्षों (1925–1964) के जल प्रवाह के आंकड़ों का प्रयोग करते हुए भाखड़ा बांध पर सतलुज के जल का मासिक प्रवाह (x) का पूर्वानुमान देने के लिए (0, 1, 1) × (0, 1, 1) क्रम के गुणात्मक मौसम एरिमा (Arima) मॉडल का प्रयोग किया गया है।

पूर्वानुमान की यथार्थता सात वर्षों (1965–71) के आंकड़ों से जांची गई है। पूर्वानुमान की अवधि के लिए वर्ग माध्य मूल त्रुटि का परिकलन किया गया है। पाया गया है कि यह त्रुटि 3 प्रतिशत (दिसम्बर) से 43 प्रतिशत (सितम्बर) तक बदलती है।

ABSTRACT. A multiplicative seasonal ARIMA model of order (0, 1, 1) × (0, 1, 1) has been applied for the prediction of monthly flow (x) in *Sutlej* at Bhakra dam site, using 40 years (1925-64) discharge data.

The accuracy of prediction has been tested by using seven years (1965-71) observations. The root mean square error for the forecast period was calculated. It is found that the root mean square error varies from 3 per cent in December to about 43 per cent in September.

1. Introduction

In mountaneous watershed, the streamflow has two components (*i*) rainfall and (*ii*) snow and glacier melt. In such cases, the runoff prediction by physical processes is far too complex to be realistic. Besides, routing of streamflow is also difficult owing to the scarcity of discharge data at various points and the complex character of physiography. Thus, if we have sufficiently long-term stationary time series of discharge, an ARIMA (Auto Regressive Integrated Moving Average) model may be fairly justified at least for the prediction of monthly discharge.

A general ARIMA model consists of a deterministic component and a stochastic component. The analysis of long-term monthly discharge time series shows a dominating seasonal factor accompanied by random fluctuations. The nature of variability involved itself suggests the applicability of seasonal ARIMA model of certain order. It may, however, be mentioned that in such cases the extreme values occurring in the time series cannot be covered fully. Therefore, the prediction values achieved by this process may further be revised in cases where acute conditions of floods or droughts prevail. The main advantage of the use of this model is its simplicity and a small number of parameters required to be estimated.

After the development of ARIMA model (Box & Jenkins 1970) it has been widely used in the analysis of various meteorological time series in respect of rainfall (Thapliyal 1981) the discharge and temperature (Mcmichael & Hunter 1972) and 500 mb flow pattern (Puri *et al.* 1981). Rao *et al.* (1982) studied the performance of ARIMA models with different orders using Bayesian decision theory. Using about 500 monthly flow data of *Krishna* and *Godavari* rivers, a seasonal ARIMA $[(1, 0, 0) \times (0, 1, 1)_{12}]$ model was found to be best among 11 models considered (without any transformation of given data). However, a seasonal ARIMA $[(5, 0, 0) \times (0, 1, 1)_{12}]$ model fitted to long transformed data was found to be best among 33 models considered.

In the present study, a prediction technique for monthly flow in *Sutlej* at Bhakra dam site has been developed using the Box-Jenkins seasonal ARIMA model (Clarke 1973).

*Present address : Central Water Commission, New Delhi.

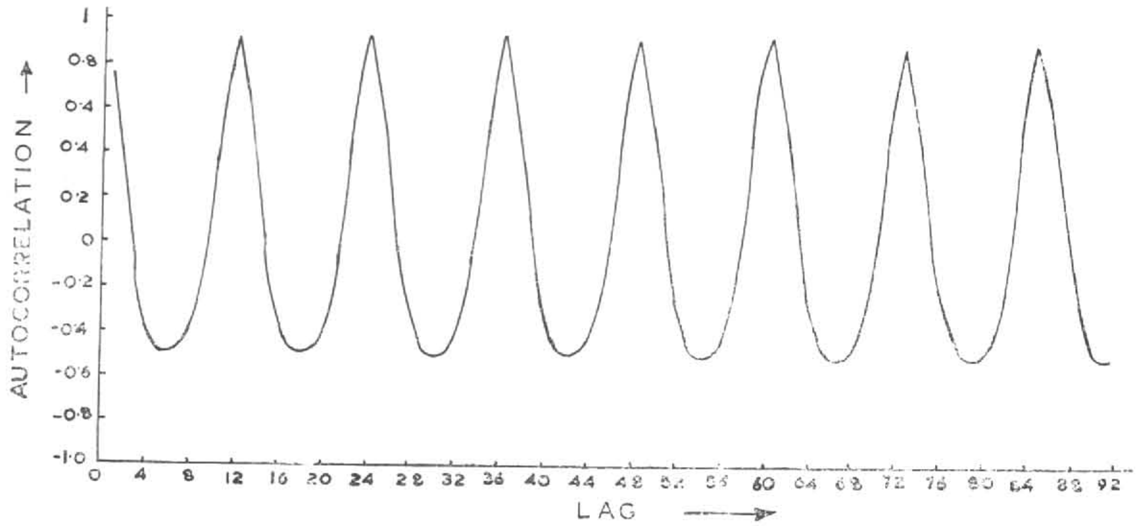

Fig. 1. Correlogram analysis of monthly run-off in *Sutlej* at Bhakra

Sutlej catchment has an area of 57,224 km^2 of which only 22,310 km^2 lies in India. Only about 8,000 km^2 of this is below 3,000 m altitude and responsible for rainfall component of the flow. 11% of the entire catchment is permanently glaciated area and snow line descends down to 1600-1800 m asl during high winter period.

2. Data used

The actual observations of monthly inflow recorded at Bhakra dam site for a period of 47 years (1925-1971) have been utilised in this study.

The statistics of annual observed discharge (1925-71) is as follows:

Mean	17511 cusec
S.D.	2658 cusec
C.V.	15.2%
Skewness (γ_1)	0.84
Kurtosis (β_1)	2.43
S.E. of γ_1	0.128
S.E. of β_1	0.512

3. Methodology

3.1. *Symbols used*

Q_t — Monthly discharge series

ϕ — Random variate

B — Backward difference operator

B_S — Seasonal backward difference operator

a_t — Uncorrelated random variable series

S — Seasonal subscript

p — Order of autoregressive component

d — Differencing operator

q — Order of moving average component

P — Order of seasonal autoregressive component

D — Differencing operator (seasonal)

Q — Order of seasonal moving average component

α, β — Parameters of the model

∇ — Grade operator

Z — Sequences

e_t — Errors series

σ — Standard deviation

s_e — Standard error

c_v — Coefficient of variation

Let $x_{11}, x_{12}, \ldots\ldots, x_{112}$; $x_{21}, x_{22}, \ldots\ldots, x_{212}$; $x_{n1}, x_{n2}, \ldots\ldots, x_{n12}$ represent monthly

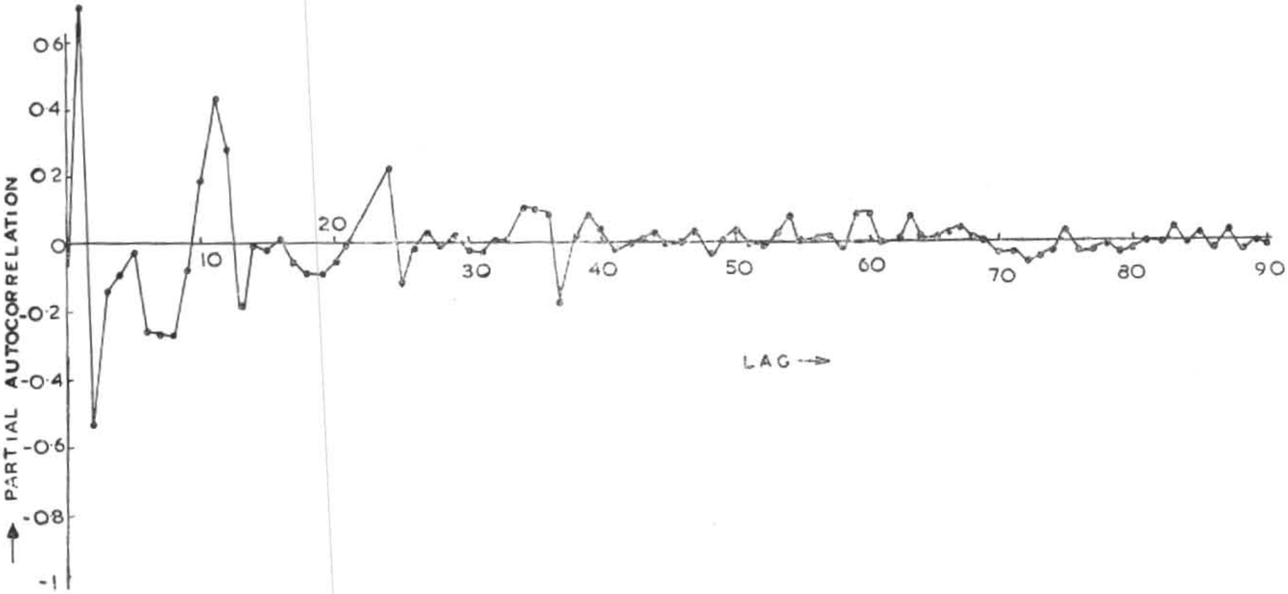

Fig. 2. Partial autocorrelation of original data

time series of discharge for n years recorded at a particular point of the river, where x_{ij} ($i=1,\ldots\ldots, n$: $j=1,\ldots\ldots\ldots, 12$) is the discharge of ith year and jth month. If the series is stationary with reference to variance, that is, variance in a particular month remains practically constant with time, we may express the series as a regression model having random variable ϕ. Now in the present series, we have two types of relationships: (1) between x_{ij} and $x_{i,j+1}$ and (2) between x_{ij} and $x_{i+1,j}$. To account for both the relationships Box-Jenkins seasonal model:

$$\phi(B^s) \nabla_s^D x_t = \theta(B^s) a_t \tag{1}$$

may be used. Here the monthly flow data are being studied.

Hence, $$\phi(B^s)=1-\phi B^{12} \tag{2}$$

$$\theta(B^s)=1-\theta B^{12} \tag{3}$$

$$\nabla_s=1-B^{12} \tag{4}$$

3.2. *Autocorrelation and partial autocorrelation analysis*

About 40 years (1925-1964) monthly inflow data have used been to compute autocorrelations and partial autocorrelations of lags 1 to 150. The correlograms of these are given in Figs. 1 and 2 respectively. The presence of strong seasonal component is apparent from Fig. 1. The behaviour of partial autocorrelation shows a rapid convergence after the second peak at lag 24. However, only one significant peak at lag 12 is observed suggesting the use of order one in respect of moving average component. Thus, a multiplicative seasonal model $(p, d, q) \times (P, D, Q)$ as suggested by Box and Jenkins (1970) appears to be appropriate with $p=P=0$, $d=D=1$ and $q=Q=1$.

As we are considering the differences of first order in time series and also in seasonal terms, the autoregressive element of order one is automatically included in the prediction model even though the order of autoregressive component is taken as zero.

The $(0, 1, 1) \times (0, 1, 1)_{12}$ model may be expressed as :

$$\nabla(\nabla_{12}\, x_t) = (1-\alpha B)(1-\beta\; B^{12})a_t$$

where a_t is a sequence of uncorrelated random variance with mean zero and variance σ^2.

Therefore,

$$\nabla(x_t-x_{12}) = (1-\alpha B-\beta B^{12}+\alpha\beta B^{13})\, a_t \tag{5}$$

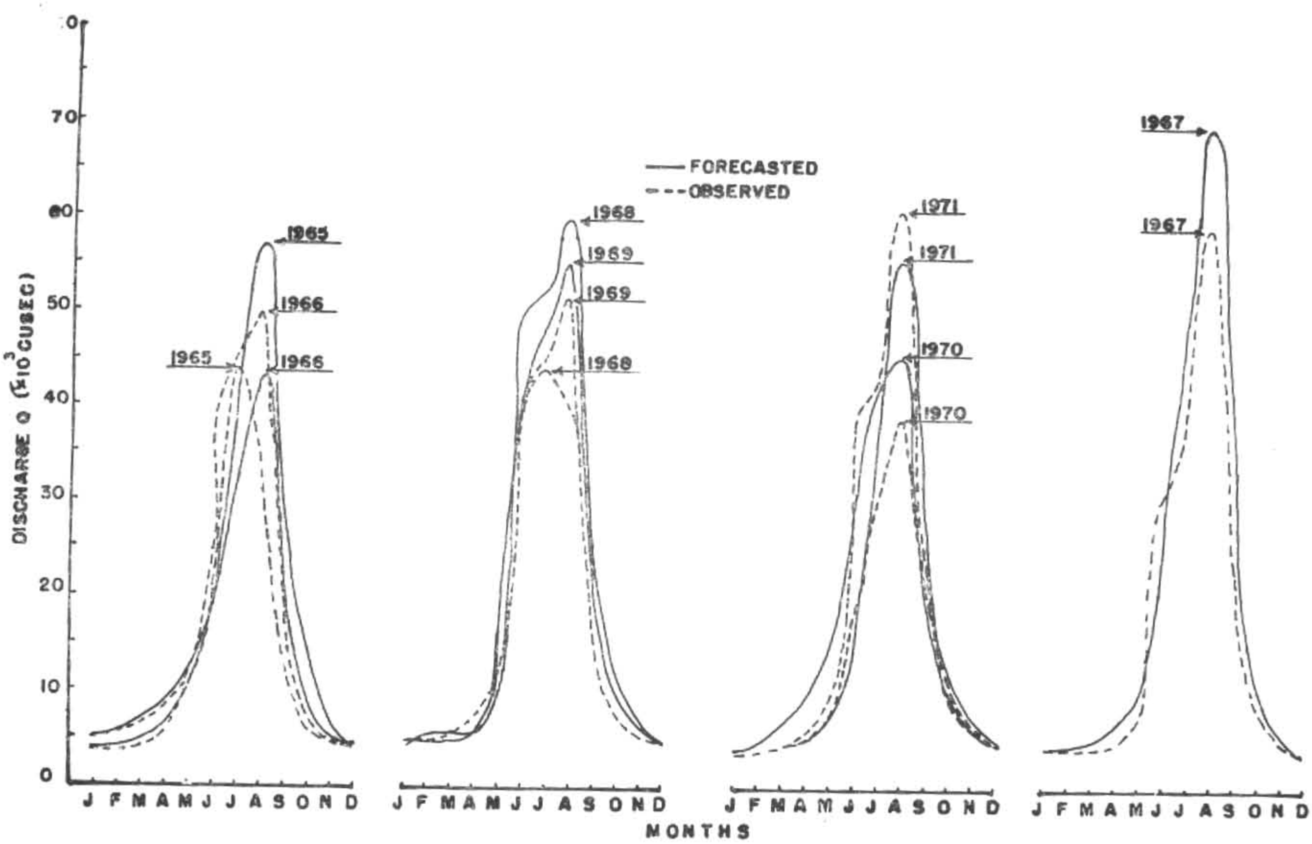

Fig. 3. Comparison of forecasted and observed discharges

TABLE 1

Estimation of parameters

Iteration	α	β
1	0.46	0.86
2	0.53	0.95
3	0.58	0.95
4	0.61	0.96
5	0.63	0.96
6	0.64	0.96
7	0.65	0.96

TABLE 2

Residual sum of squares

α	β 0.7	0.8	0.9	1.0
0.4	26.05	27.00	25.97	31.54
0.5	27.67	26.61	25.56	31.07
0.6	27.57	26.47	25.39	30.87
0.7	27.81	26.64	25.51	30.96
0.8	28.62	27.34	26.09	31.53
0.9	30.52	29.01	27.55	33.65

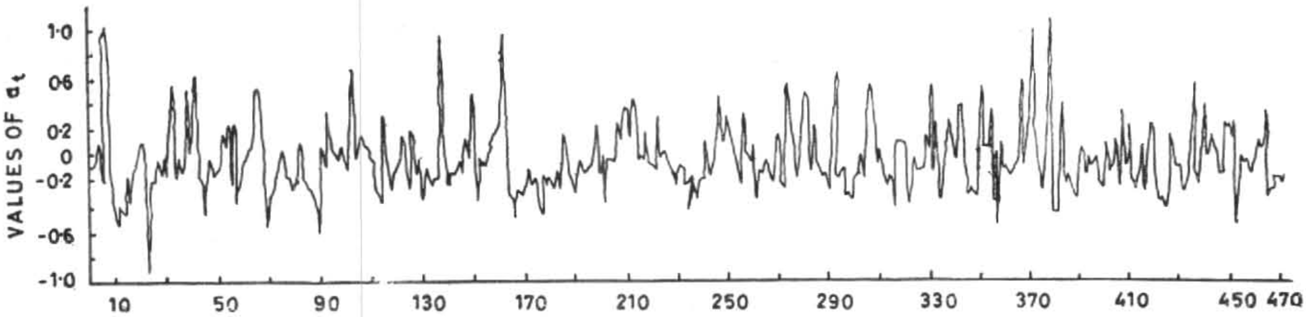

Fig. 4. Values of a_t *versus* time (months)

Therefore, $x_t = x_{t-1} + x_{t-12} - x_{t-13} + a_t - \alpha a_{t-1} - \beta a_{t-12} + \alpha\beta a_{t-13}$ (6)

The 40 years monthly data of discharge(x_t) have been used to form the series :

$$x_{-12}, x_{-11}, \ldots, x_{-1}, x_0, x_1, \ldots\ldots, x_{467}$$

The process of estimation of parameters α and β as suggested by Box and Jenkins (1970) is summarised below:

The initial estimate α_0 and β_0 are worked out as:

$$\alpha_0 = [-1+\sqrt{(1-4\, r_1^2)}]/2\, r_1, \quad (7)$$

$$\beta_0 = [-1+\sqrt{(1-4\, r^2_{12})^2}]/2r_{12} \quad (8)$$

where r_1, r_{12} are autocorrelations of lags 1 and 12 respectively. Now the error series e_t is generated by

$$e_t = x_t - x_{t-1} - x_{t-12} + x_{t-13} + \alpha_0\, e_{t+1} + \beta_0\, e_{t+12} - \alpha_0\, \beta_0\, e_{t+13} \quad (9)$$

Here t will vary from 467 to 1. The higher order unknown errors are taken as zero.

The series of $a_t(a_{-12}, a_{-11}, \ldots\ldots\ldots\ldots, a_{467})$ is also generated by using :

$$a_t = x_t - x_{t-1} - x_{t-12} + x_{t-13} + \alpha_0\; a_{t-1} + \beta_0 a_{t-12} - \alpha_0\beta_0\, a_{t-13} \quad (10)$$

replacing unknown a_t's $= 0$

The five sequence of a_t need to be generated for the following sets of parameters :

$$(\alpha_0, \beta_0), (\alpha_0, \beta_0 \pm 0.1\, \beta_0), (\alpha_0 \pm 0.1\, \alpha_0, \beta_0).$$

Let the sequences $Z(1, t)$ and $Z(2, t)$ be defined as :

$$Z(1, t) = \frac{[a_t(\alpha_0 + 0.1\alpha_0, \beta_0) - a_t(\alpha_0 - 0.1\alpha_0, \beta_0)]}{0.2\alpha_0} \quad (11)$$

$$Z(2, t) = \frac{[a_t(\alpha_0, \beta_0 + 0.1\,\beta_0) - a_t(\alpha_0, \beta_0 - 0.1\beta_0)]}{0.2\beta_0} \quad (12)$$

TABLE 3

Root mean square error

Month	RMSE value (cusec)	% of average estimate
January	937	19
February	1477	28
March	1034	18
April	822	14
May	3274	33
June	12375	42
July	14779	40
August	17227	30
September	12280	43
October	3630	35
November	956	15
December	114	3

If the first sequence (based on α_0, β_0) is denoted by $a_{0,t}$ (a multiple regression) :

$$a_{0,t} = C + b_\alpha\, Z(1, t) + b_\beta Z(2, t)$$

where b_α and b_β are the multiple regression coefficient and C is a constant. We may now modify the initial estimate α_0 and β_0 by $\alpha_1 = \alpha_0 + b_\alpha$ and $\beta_1 = \beta_0 + b_\beta$. The process of modification as described above is repeated till b_α, b_β become negligible as compared to the initial estimates α_0 and β_0.

4. Results and discussions

The monthly discharge series Q_t in *Sutlej* recorded over Bhakra dam site for the years 1925-1964 has been used for this analysis :

$$x_t = \ln Q_t, \quad t = 1, 2, \ldots\ldots\ldots\ldots\ldots, 480.$$

The autocorrelation and partial autocorrelation up to lag 150 have been computed and the correlograms are provided in Figs. 1 and 2. Examining the autocorrelations, we see that r_L is maximum at $L=12$ and minimum at $L=6$. This establishes an annual cycle in the series. The initial estimtes $\alpha_0=0.32$ and $\beta_0=0.73$. For 7 iterations the values of the estimates α, β are given in Table 1. The final estimates of the α and β after 7 iterations are :

$$\alpha = 0.65 \text{ and } \beta = 0.96.$$

Therefore, the prediction model may be writen as :

$$x_t = x_{t-1} + x_{t-12} - x_{t-13} + a_t - 0.65\ a_{t-1} - 0.96\ a_{t-12} + 0.62\ a_{t-13}.$$

The predictive values of discharge for the years 1965 to 1971 has ben verified with the actual observation and comparisons are given in Fig. 3. The predicted peak discharge though tallies in respect of the period of its occurrence in most cases, its magnitudes are generally higher than the observed ones. The deviation is particularly striking in 1967. The general pattern and the magnitude of discharge for other months are reasonably accurate.

The variation of residuals sum of squares for different values of α and β is given in Table 2 showing a minimum at final values of the estimates considered in the paper.

The behaviour of sequence at unautocorrelated random variable for $t=1, 467$ is shown in Fig. 4. It may be regarded as a probability distribution with mean zero and variance σ^2.

The root mean square error of the forecast period was worked out and is given for all the months in Table 3. It may be seen from Table 3, root mean square error is minimum (3 %) with respect to average forecast flow in the month of December and maximum (43 %) for the month of September.

5. Conclusions

The technique presented above establishes its superiority over the regression models in the sense that it contains a systematic component of estimating errors. But essentially it is a statistical model which largely depends upon the pattern observed during past. It is not capable to govern the fluctuations occurring due to any other type of variability.

Acknowledgement

The authors are grateful to Dr. R.P. Sarker, Director General of Meteorology for his keen interest in this paper.

References

Box, G. E. P. and Jenkins, G. M., 1970, *Time Series Analysis, Forecasting and Control*, Holden-Day Inc., California.

Clarke, R. I., 1973, *Mathematical Models in Hydrology*, Food and Agriculture Orgn. of U. N. Rome, Ch. 4.

Mcmichael, F. C. and Hunter, J. S., 1972, *Water Resources Res.*, **8**, p. 87 (Publishers : American Geophysical Union).

Puri, S. R., Bansal, R. K., Datta, R. K. and Hingorani, J. U., 1981, *Proc. Computer Soc. India*, **3**, p. 369.

Rao, A. R., Kashyap, R. L. and Mao, L., 1982, *Water Resources Res.*, **18**, p. 1097.

Thapliyal, V., 1981, Met. Monograph—Climatology, No. 12 India Met. Dep., Pune.