International Seminar on Hydrology of Extremes

1-3 December, 1988
Roorkee (India)

FLOOD FORECASTING FOR THE LOWER INDUS RIVER

H. N. Phien
Associate Professor

N. C. Austriaco
Director, CEC

A. Pornprasertsakul
Senior Lab. Supervisor

Asian Institute of Technology, Bangkok, Thailand

SYNOPSIS

A statistical approach was used to develop operational models for forecasting 6 hourly discharges at six stations on the Lower Indus River in Pakistan. The general form of the forecasting equation for a station was obtained by expressing the discharge at a time unit t as a linear function of the discharges at preceding time units at that station as well as at the immediately upstream station. In spite of their simple form, the models so obtained can produce very accurate forecast values for forecasting lead times from one to eight units (of 6 hours).

1.0 INTRODUCTION

There exist many conceptual models for simulating the rainfall-runoff process. Most of these models include a river routing component, see for example, [2], [5], [6]. When areal rainfall data are not accessible (due to the large coverage of the river basin), only discharge records are available for modelling or forecasting purposes. In this case, the only tool to be used is the river routing scheme built in the models. Since data on lateral flow and local rainfall are not available, the values obtained by river routing are far from the observed ones. What has just been mentioned is precisely the case of the lower part of the Indus River in Pakistan (Fig. 1). For all the reaches from Terbela to Kotri, no records are available for the lateral flows and rainfall. In such a case, conceptual models cannot be used. In response to a request by the Federal Flood Commission, a statistical approach was proposed. Supporting reasons of this approach are presented in Section 2, and the resulting models for the different stations are provided in Section 3.

2.0 MODEL DEVELOPMENT

2.1 The Proposed Approach

For the reaches of the Lower Indus River, only discharge data during the flood season, June to September, were provided by the Federal Flood Commission. These data, expressed in 1000 cfs (28.32 m /s), are given on a 6 hourly basis, and the longest record obtained is for ten years (1976-1985). For all the stations considered, namely Terbela, Kalabagh, Chasma, Taunsa, Gudu, Sukkur and Kotri, there exist many missing values. However, no attempt was made to fill in these.

Within the context of a statistical approach, time-series models and regression models are most popular.

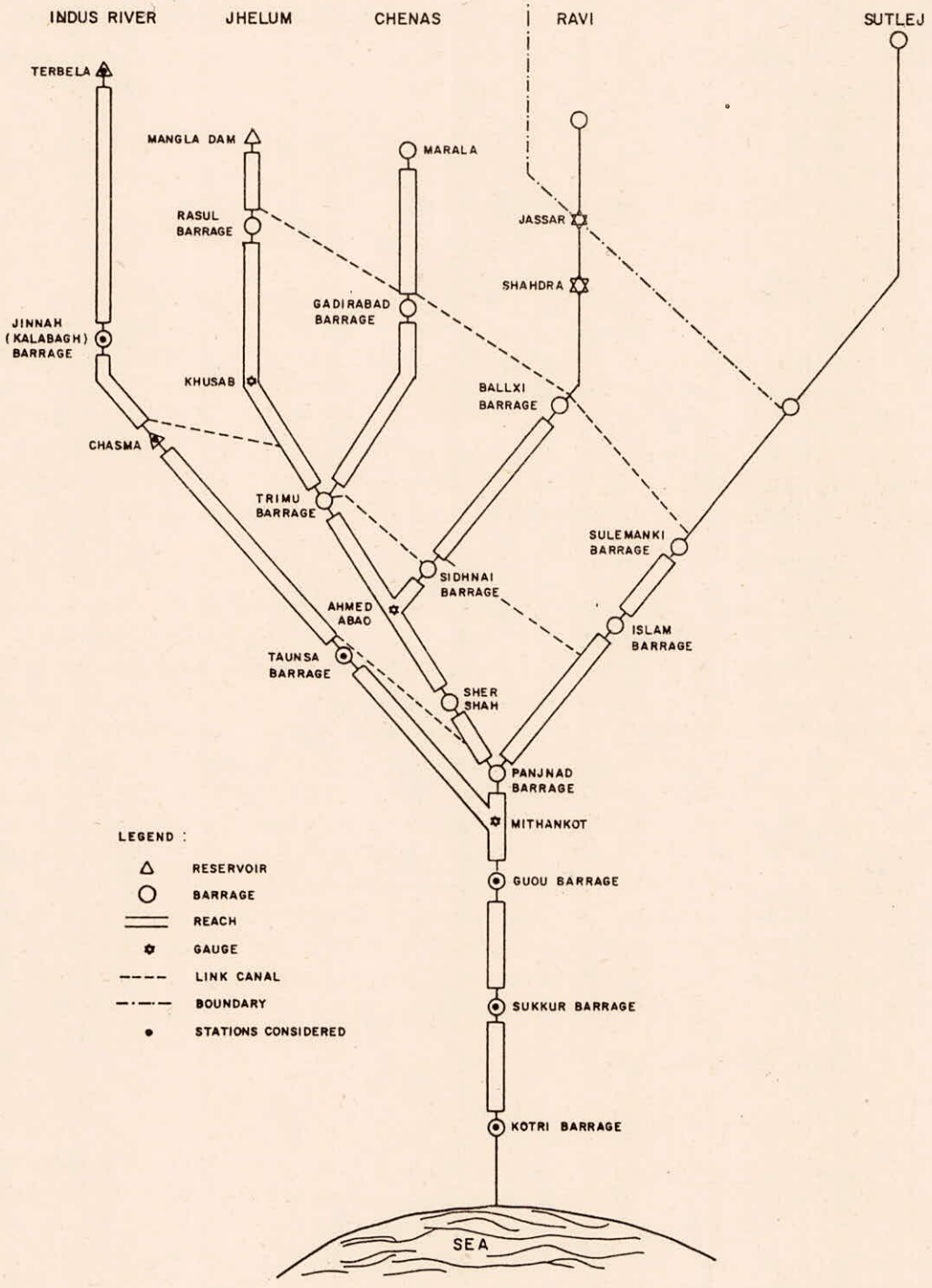

Fig. 1 Stations Considered in the Lower Indus Basin

2.1.1 Time-series models

These models are commonly based on those introduced by Box and Jenkins [1]. For a time unit of six hours, it seems that some form of the seasonal Autoregressive Integrated Moving Average (ARIMA) or Transfer Function models would be appropriate. The ARIMA models can be used when only the discharge data at a station are employed, while Transfer Function models can incorporate the discharge data at an upstream station. Unfortunately, these models cannot be applied to the Lower Indus River, because of the following two reasons, namely, the interruption and missing values in the records available.

- Data Interruption : As previously mentioned, the data provided are available only for the flood season. As such, the data for the different years of records at a station cannot be combined together to form a non-interrupted time series. If only the data in each flood season were employed, there would correspond several different models for different years. The problem of selecting which model for operational forecasting would then arise.
- Missing Values : Since there are many missing values in the records obtained, it is extremely difficult to use Box-Jenkins models. At some stations, the number of missing values is even larger than that for existing ones. In such cases, techniques to deal with missing values in the Box-Jenkins approach would provided poor forecasting capabilities.

2.1.2 Regression analysis

This approach is not new. Practitioners frequently use regression analysis in flood forecasting. Typically, the discharge at a station is expressed as a linear relationship of the discharges of some upstream stations with the time-lags between upstream stations and the stations concerned explicitly incorporated.

Regression analysis was adopted in this study with the following considerations.

(a) The time lags, which were used in the selection of lagged variables, were based upon the estimated travel time available for the different reaches in the Lower Indus River.

(b) A simple autocorrelation analysis would reveal that the discharge at a station at a time unit t depends heavily on its values on preceding times t-1, t-2, etc. (In fact this is the basic assumption of the Box-Jenkins time series approach.) These values must be taken into account as well in the development of forecasting models.

With all the foregoing discussions in mind, the forecasting model proposed for the Lower Indus River should be of the following form.

$$Y(t+L) = A + \sum_{i=0}^{m} a(i)Y(t-i) + \sum_{i=0}^{n} b(i)U(t-i) + \sum_{i=0}^{p} c(i)V(t-i) + \ldots \qquad (1)$$

where A, a(i), b(i), c(i) are for regression coefficients; Y(t) is the (transformed) value of discharge at time t for the station concerned; U(t), V(t), ... are the (transformed) values of discharge at time t for upstream station; m, n, p, ... are structural parameters; and L is the forecasting lead time.

Since the discharge at a station is actually the integrating factor (a term borrowed from Professor Masami Sugawara - personal communication) of all components producing discharge, it is believed that the proposed equation would be able to incorporate the lacking information on lateral flows.

Because of the collinearity among Y, U, V, etc. the coefficients A, a(i), b(i), c(i), ... can be calculated from several techniques. In this work, the method of stepwise regression was used; as a result of which, some of these coefficients will have zero values.

The transformations which are commonly used in practice are square root and logarithmic transformations. These were adopted in the present work. However, they were found to provide worse results as compared to those corresponding to untransformed data. Consequently, no transformation was made.

2.2 Accuracy Indicators

In this work, since regression analysis was used, the relevant accuracy indicators are the multiple correlation coefficient R, the coefficient of multiple determination R^2, the Adjusted R^2 and Standard Error, s. These are readily available in Textbooks, e.g. Draper and Smith [3].

It should be noted that R^2 is a relative measure of how much all the terms in the regression equation that involves predictor ("independent") variables explain the total variation of the observations.

The multiple correlation coefficient R is in fact the correlation between the observed and estimated values. As such, it reveals the match in the pattern (shape) of the estimated flow with that of the observed data.

Generally, the value of R^2 can be increased by adding new terms into the model. The Adjusted R^2 incorporates some penalty for such attempt. Finally, the Standard Error is obtained from the sum of squared errors. As such, s should be small for the fitted model to be good.

2.3 Calibration and Verification

Traditionally, when a model is developed, the data available should be split into two sets. One set, called training set, normally having a longer record, is for calibration thereby the values of the parameters, i.e. the regression coefficients in our case, are obtained. The other set, called checking set, is used for verification.

In this study, the first eight years of records (1976-1983) were used as training sets whereas the last two years (1984-1985) were used as checking sets.

Within this context, the aforementioned accuracy indicators are mainly applied to the calibration stage. For the verification stage, R, R^2 can still be applied, while the standard deviation should be replaced by the root mean squared error defined as

$$RMSE = \{(1/K) \sum_{t=1}^{K} [\hat{Y}(t) - \hat{Y}(t)]^2\}^{\frac{1}{2}} \qquad (2)$$

where K is the number of data points used in the verification stage and $\hat{Y}$ denotes the estimate of Y.

3.0 APPLICATION RESULTS

3.1 General Considerations

As mentioned in Section 2, both calibration and verification stages were adopted in this work. Considering the usefulness of the developed models for future use, all the 10 years of available records were employed in arriving at the equations to be reported in the following. In the development of models, the stepwise regression procedure with missing values option was used. Replacement of missing values by the corresponding long-terms means did not give good results as those with missing values option.

The following reaches were considered : Terbela-Kalabagh, Kalabagh-Chasma, Chasma-Taunsa, Taunsa-Gudu, Gudu-Sukkur, and Sukkur-Kotri. It was found that for all these reaches, only one upstream station was used in deriving the forecasting equation for the corresponding downstream station. Thus eq. 1 can be written as:

$$D(t+L) = A + \sum_{i=0}^{m} a(i)D(t-i) + \sum_{i=0}^{n} b(i)U(t-i) \qquad (3)$$

where D and U stand for downstream discharge and upstream discharge, respectively.

From the available information on travel time on reach reach, it was decided to take n=12. Appropriate value sof m can be obtained using the autocorrelation function at each station [4] and sensitivity analysis. It was found that m=12 is generally applicable for all stations concerned.

3.2 Results

The maximum forecasting lead time adopted in this study was 8 units (or 48 hours or two days). Therefore, there are altogether 48 (=6x8) equations obtained. They are collected in Tables 1 through 6 in a summarized form.

As expected, among the predictors considered, i.e. the discharge values (at the station concerned and at the immediately upstream station) for 13 time units (i=0 to 12 in eq. 1), many of them did not appear in the resulting forecasting equations. This is due to the use of stepwise regression analysis, which selects the most contributive predictors.

Table 1 Forecasting equations for Kalabagh with Terbela as upstream station (6 years of record)

Coefficients	Lead time (units of 6 hours) 1	2	3	4	5	6	7	8
A	7.868	13.643	17.971	23.010	27.036	34.228	35.680	39.738
a(0)	0.888	0.750	0.666	0.583	0.551	0.551	0.482	0.449
a(2)	0	0	0	0.167	0.155	0	0.122	0.144
a(3)	0	0.115	0.148	0	0	0.128	0	0.141
a(4)	0	0	0	0	0	0	0.158	0.141
a(5)	0	0	0	0	0	0.160	0	0
a(6)	0.075	0	0	0	0.163	0	0	0
a(7)	0	0	0	0.139	0	0	0	0
a(8)	0	0	0.100	0	0	0	0	0
a(9)	0	0.071	0	0	0	0	0	0
a(11)	0	0	0	0	0	0	0.066	0.078
R	0.941	0.894	0.859	0.832	0.813	0.792	0.774	0.755
R^2	0.885	0.799	0.738	0.693	0.660	0.627	0.599	0.570
Adjusted R^2	0.885	0.798	0.737	0.692	0.659	0.626	0.597	0.568

Table 2 Forecasting equations for Chasma with Kalabagh as upstream station (7 years of record)

Coefficients	Lead time (units of 6 hours) 1	2	3	4	5	6	7	8
A	13.475	16.216	23.464	30.247	33.951	40.282	44.027	47.651
a(0)	0.379	0.402	0.274	0.410	0.393	0.317	0.325	0.305
a(1)	0.152	0	0.151	0	0	0	0	0
a(2)	0	0.147	0	0	0	0	0	0
b(0)	0.270	0.488	0.494	0.478	0.473	0.524	0.491	0.492
b(1)	0.162	0	0	0	0	0	0	0
b(3)	0	-0.089	0	0	0	0	0	0
R	0.955	0.942	0.917	0.890	0.860	0.834	0.802	0.778
R^2	0.913	0.888	0.814	0.793	0.740	0.696	0.643	0.605
Adjusted R^2	0.912	0.887	0.804	0.792	0.739	0.694	0.641	0.604

Table 3 Forecasting equations for Taunsa with Chasma as upstream station (8 years of record)

Coefficients	Lead time (units of 6 hours) 1	2	3	4	5	6	7	8
A	0.697	4.008	2.535	0.691	5.799	10.325	9.611	10.109
a(0)	0	0	0.163	0	0	0	0	0.220
a(1)	0.362	0	0	0.095	0	0	0	0
a(5)	0	0	0	0	0	0	-0.069	-0.318
a(7)	0	0	0	0	0	0.058	0	0
a(8)	0	0	0	0	0.057	0	0	0
a(9)	0	0	0	0.048	0	0	0	0
a(10)	0	0.058	0.049	0	0	0	0	0
a(11)	0	0	0	0	0	0	0.103	0.266
b(0)	0	0	0	0	0.051	0.173	0.320	0.239
b(1)	0	0	0	0	0.138	0.270	0.143	0
b(2)	0	0	0	0.165	0.277	0.209	0.169	0.210
b(3)	0	0.069	0.171	0.250	0.204	0.112	0.098	0
b(4)	0	0.182	0.265	0.195	0.114	0.072	0.072	0.301
b(5)	0.228	0.259	0.183	0.133	0.076	0	0.107	0
b(6)	0.223	0.191	0.107	0.056	0	0	0	0
b(7)	0.155	0.105	0	0	0	0	0	0
b(8)	0	0.065	0	0	0	0	0	0
R	0.985	0.994	0.995	0.994	0.995	0.995	0.992	0.978
R	0.969	0.987	0.989	0.989	0.989	0.989	0.983	0.957
Adjusted R	0.966	0.985	0.987	0.986	0.987	0.987	0.979	0.947

As seen from these tables, the forecasting accuracy is very high, particularly when the lead time is equal to 6 hours and 12 hours. Except for some minor irregulations for larger lead times, there corresponds a decreasing accuracy in the forecasting models. This is commonly expected, since large lead times mean more randomness or uncertainty involved.

It can also be observed that as one moves downstream, the accuracy of the forecasting models improves. This may be due to two reasons. First, the discharges at the stations close to the Terbela Reservoir are regulated by the operation of that reservoir. Regulated flows are commonly unsuitable for a statistical approach. Since no information on the flows released from the reservoir is available, it seems that no methods would be able to provide better forecast values. Second, upstream stations have smaller catchment areas. Correspondingly, the fluctuations of the discharges are higher.

Nevertheless, with the accuracy obtained, the models developed in this study should provide satisfactory forecasting values.

Table 4 Forecasting equations for Gudu with Taunsa as upstream station (8 years of record)

Coefficients	Lead time (units of 6 hours) 1	2	3	4	5	6	7	8
A	-8.140	-9.359	-16.033	-25.576	-27.281	-30.576	-30.576	-37.643
a(0)	0.813	0.800	0.678	0.773	0.771	0.767	0.765	0.797
a(1)	0.203	0	0.322	0	0	0	0	0
a(2)	0	0.254	0	0.334	0.303	0.345	0.325	0.308
a(6)	0	0	0	0	0	0	-0.243	-0.391
a(7)	0	0	0	-0.268	0	-0.374	0	0
a(8)	0	-0.102	0	0	-0.320	0	0	0
a(9)	-0.051	0	0	0	0	0	0	0
a(10)	0	0	-0.073	0	0	0	-0.177	0
a(11)	0	0	0	0.069	0	0.130	0	0.111
b(0)	0	0	0	0	0.187	0.173	0.293	0.379
b(1)	0	0	0	0.192	0	0.154	0.187	0.264
b(2)	0	0	0.105	0	0.205	0.220	0.246	0.283
b(3)	0.087	0.123	0.157	0.246	0.217	0.232	0.225	0.245
b(4)	0	0.139	0.172	0.250	0.232	0.221	0.163	0
b(5)	0	0.105	0.107	0	0	0	0	0
b(8)	0	0	0	0	0	-0.125	-0.156	-0.180
b(9)	0	0	0	-0.107	-0.147	-0.151	-0.156	-0.185
b(10)	0	-0.096	-0.150	-0.152	-0.173	-0.166	-0.177	-0.207
b(11)	0	-0.156	-0.206	-0.184	-0.224	-0.217	-0.248	-0.163
R	0.994	0.991	0.988	0.984	0.981	0.978	0.974	0.970
R	0.989	0.981	0.975	0.968	0.962	0.956	0.945	0.942
Adjusted R	0.989	0.981	0.975	0.968	0.962	0.956	0.949	0.941

3.3 Discussions

- From the forecasting equations developed, one can adopt the following scheme also. At time t,
  - forecast the discharge at Kalabagh with lead time equal to 6 hours (one unit)
  - consider the forecast value at Kalabagh as the observed value at time t+1 and forecast the discharge at Chasma with lead time equal to 6 hours, hence one obtains the forecast for Chasma at time t+2, i.e. the forecasting lead time is two units.

  By moving dowstream and repeating the same procedure, one can obtain the forecast values for Taunsa at t+3, for Gudu at t+4, for Sukkur at t+5, and for Kotri at t+6, i.e. 6 time units ahead. Since the accuracy for the forecasting equation with lead time equal to 6 hours at any station is very high, this scheme leads to some slight improvement as compared to the accuracy obtained directly from regression analysis.

Table 5 Forecasting equations for Sukkur with Gudu as upstream station (10 years of record)

Coefficients	Lead time (units of 6 hours) 1	2	3	4	5	6	7	8
A	-2.531	-1.870	-1.941	-2.118	0.957	-1.332	-3.768	-2.993
a(0)	0.237	0.234	0.214	0.199	0.147	0.144	0.166	0.125
a(1)	0.179	0.146	0.162	0.104	0.105	0.612	0	0
a(2)	0.104	0.112	0.085	0.080	0.062	0	0.076	0.057
a(3)	0.079	0.061	0	0	0	0	0	0
a(11)	0	-0.035	0	-0.033	0	0	0	0
b(0)	0	0.252	0	0.380	0	0	0.859	0.954
b(1)	0.225	0	0.359	0	0	0.815	0	0
b(2)	0	0	0	0	0.795	0	0	0
b(3)	0	0	0	0.265	0	0	0	0
b(4)	0	0	0.236	0	0	0	0	0
b(5)	0	0.227	0	0	0	0	0	0
b(6)	0.176	0	0	0	0	0	0	0
b(8)	0	0	0	0	0	0	0	-0.142
b(9)	0	0	0	0	-0.120	0	0	0
b(10)	0	0	0	0	0	0	-0.104	0
b(11)	0	0	-0.059	-0.033	0	-0.072	0	0
R	0.986	0.985	0.984	0.968	0.982	0.981	0.986	0.984
R^2	0.972	0.970	0.964	0.937	0.963	0.963	0.971	0.968
Adjusted R^2	0.972	0.970	0.969	0.967	0.963	0.963	0.971	0.968

Table 6 Forecasting equations for Kotri with Sukkur as upstream station (10 years of record)

Coefficients	Lead time (units of 6 hours) 1	2	3	4	5	6	7	8
A	-0.693	-1.496	-1.987	-2.980	-3.791	3.144	3.918	4.937
a(0)	1.063	1.144	1.248	1.426	1.512	1.828	1.908	1.992
a(4)	0	0	0	0	0	-0.843	-0.925	-0.013
a(5)	0	0	0	-0.452	-0.547	0	0	0
a(7)	0	0	-0.267	0	0	0	0	0
a(8)	-0.070	-0.157	0	0	0	0	0	0
b(1)	0	0.012	0.017	0.024	0.031	0	0	0
b(2)	0.006	0	0	0	0	0	0	0
R	0.999	0.998	0.997	0.996	0.994	0.992	0.990	0.987
R^2	0.998	0.996	0.995	0.992	0.989	0.984	0.979	0.974
Adjusted R^2	0.998	0.996	0.995	0.992	0.989	0.984	0.979	0.974

Although stepwise regression analysis is capable of selecting the best predictors, the resulting equation lacks apparent physical meanings. For example, in the forecasting equation for Chasma:

$$C(t+L) = A+a(0)C(t)+a(1)C(t-1)+a(2)C(t-2)+b(0)K(t)+b(1)K(t-1) +b(3)K(t-3)$$

where C and K stand for Chasma and Kalabagh, respectively, one may question why K(t-2) does not appear while K(t-3) does. In this case, it should be understood, from a statistical point of view, that due to a high correlation between K(t-1) and K(t-2), the contribution of K(t-2) has been well taken care by that of K(t-1).

However, to overcome this lack of physical interpretation, several other alternatives like ridge regression on regression on principal components are being attempted. Results of the study will soon be reported.

4.0 CONCLUSIONS

Regardless of its simple form and its less demand for input data, the proposed statistical approach leads to forecasting models with high accuracy for all the stations on the lower part of the Indus River. The resulting models, being expressed as simple linear equations, can be conveniently used for operational forecasting purposes because forecast values can be computed easily. The high accuracy obtained by the developed models is believed to result from the fact that past values of discharges at the station considered are taken into account and these lagged variables would be able to take care of all factors affecting the variation of the discharge in the following time units. Due to the use of stepwise regression, the forecasting models may seem to lack some physical interpretation but this, it is hoped, would be overcome by adopting an alternate technique.

ACKNOWLEDGEMENTS

The authors wish to express their sincere thanks to:

- the US Agency for International Development, Pakistan for its partial financial support of the study,
- The Federal Flood Commission, Pakistan for the supply of data employed, and to
- Mrs. Nantawan Nakasen for her care in typing this paper.

REFERENCES

1. Box, G.E.P. and Jenkins, G.M., (1976), Time Series Analysis : Forecasting and Control, Revised Ed., Holden-Day, San Francisco.
2. Crawford, N.H. and Linsley, R.K., (1966), Digital Simulation in Hydrology, Stanford watershed Model IV, Technical Report, No. 39, Department of Civial Engineering, Stanford University.

3. Draper, N. and Smith, H., (1981), Applied Regression analysis, John Wiley & Sons, New York.

4. Phien, H.N. and Lee, S.T., (1986), Forecasting of Daily Discharges of burmese Riovers, International Journal for Development Technology, Vol. 4, pp. 173-188.

5. Rockwood, D.M., (1968), Application of SSARR Program to the Lower Mekong Riover, Proceedings of the Symposium on the Use of Analogy and Digital Computer in Hydrology, UNESCO, Vol. 1, Paris.

6. Sugaward, M., Watanabe, I., Ozaki, E., and Katsuyama, Y., (1984), Thank Model with Snow Component, National Research Center for Disaster Prevention, Tokyo.