Multivariate flexible least squares analysis of hydrological time series

A. RAMACHANDRA RAO

School of Civil Engineering, Purdue University, West Lafayette, Indiana 47907, USA

Abstract Models with coefficients which change with time can be developed by using Kalman filter techniques and these have been applied to model river flow, rainfall and other hydrological time series. However, application of Kalman filter techniques requires assumptions about the statistics of processes and state variables which may depend on unknown factors, and these assumptions may not be valid. A new approach to this problem is the flexible least squares (FLS) method. In this approach, the parameters are assumed to evolve slowly over time. The parameters of the model are estimated by minimizing the sums of squared measurement and dynamic errors conditional on the given observations. The method is based on a cost efficient frontier and is called the generalized flexible least squares method. The objective of this study is to apply the FLS method to hydrological time series. Data from the Green River basin in Kentucky in the USA are used in the study. The method is found to perform well.

INTRODUCTION

Models whose coefficients change with time can be developed by using Kalman filter techniques (Kalman, 1960). Amirthanathan (1989) and Wood *et al.* (1979) applied the filtering techniques to analyse river flow and rainfall. Sallas & Harville (1981) developed a best linear recursive estimation method for mixed linear models by using the Kalman filter to obtain recursive estimators for the model. In this method, a state-space model is defined and it consists of observation and state equations. Based on these equations, a recursive algorithm is developed. However, application of Kalman filter techniques requires assumptions about the statistics of processes and state variables which may be unknown or which may not be valid.

A new approach to least squares method is developed by Kalaba & Tesfatsion (1989). In this approach, instead of assuming that the models have constant parameters, the parameters are assumed to evolve slowly over time. This method is called the FLS method. The parameters of the model are estimated by minimizing the sums of squared measurement and dynamic errors conditional on the given observations. The FLS method has been tested extensively by using linear, quadratic, sinusoidal and regime shift data with noisy observations and has been found to work satisfactorily. The objective of the research reported here is to investigate the feasibility of application of the FLS method to hydrological data.

DATA USED IN THIS STUDY

The data used in this study consist of daily river flows, precipitation and temperature from the Green River basin in the USA (Fig. 1). The Green River basin is located in west central Kentucky and has a gently rolling terrain. The Green River is a tributary of the Ohio River and has a drainage area of 9230 miles^2. The Green River basin is affected by frequent temperature changes. High flows in these rivers are mainly due to thunderstorms. The average annual precipitation in the Green River basin is 47 in. The daily streamflow data are collected at Greensburg, Gresham, Munfordville, Falls of Rough and Dundee in Kentucky. The precipitation data are collected at a station between Greensburg and Munfordville and another station between Falls of Rough and Dundee. Daily data measured during water years 1970-1972 are used in the present study.

MODEL SELECTION

Theory

Let x_t, $t = 1, 2, \dots$ be the $(n \times 1)$ state vector available at time t. Then the process can be modelled by the state transition equation as $x_{t+1} \approx F(t) + a(t)$ $t = 1, 2, \dots$ where $F(t)$ is a known $(n \times n)$ square matrix and $a(t)$ is a known n-dimensional column vector. Let y_t, $t = 1, 2, \dots$ be the $(m \times 1)$ vector of observations at time t. Then the observation

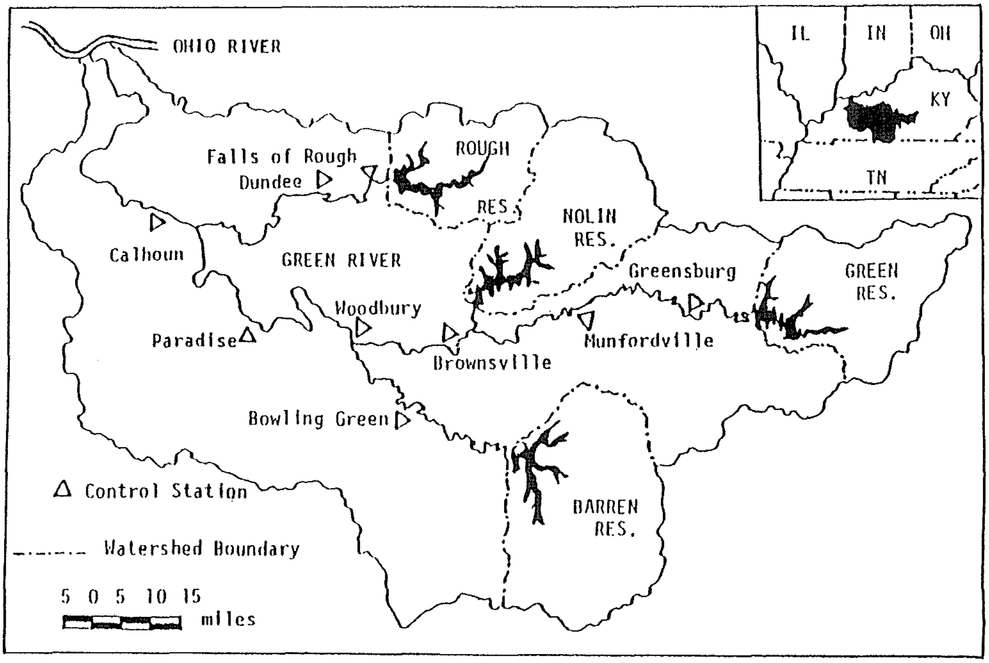

Fig. 1 Green River basin, Kentucky, USA.

equation for the approximately linear model is given by $y_t \approx H(t)x_t + b(t)$ where $H(t)$ is a known $(m \times n)$ rectangular matrix and $b(t)$ is a known m-dimensional column vector.

Each possible estimate $\hat{x}_1$ $\hat{x}_2$ is associated with two distinct types of model specification errors. The first type of error is related to the discrepancies between the actual (observed) and estimated value of y_t at each time period t and is the **measurement error.** The second type of error is the **dynamic error** which is related to discrepancies due to the mis-specification of the state transition equations.

There is a cost related to each of these model specification errors. For any given state sequence estimate $\hat{X}_T = (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_T)$ the dynamic cost is given by $C_D(\hat{X}_T, T)$:

$$C_D(\hat{X}_T, T) = \sum_{t=1}^{T-1} [\hat{x}_{t+1} - (F(t)\hat{x}_t + a(t))]^T D(t)[\hat{x}_{t+1} - (F(t)\hat{x}_t + a(t))] \tag{1}$$

and the measurement cost by $C_M(\hat{X}_T, T)$:

$$C_M(\hat{X}_T, T) = \sum_{t=1}^{T} [y_t - (H(t)\hat{x}_t + b(t))]^T M(t)[y_t - (H(t)\hat{x}_t + b(t))] \tag{2}$$

where $D(t)$ and $M(t)$ are scaling matrices of orders n and m, respectively.

If the conditions in equations (1) and (2) are exactly satisfied, then both C_D and C_M would be equal to zero. But in practical cases, it is not known whether equations (1) and (2) are exactly satisfied so that each state sequence estimate $\hat{x}_t$ is associated with both measurement and dynamic costs.

The sets of state sequence estimates that minimize both C_D and C_M are called the FLS estimates. Each of these estimates shows the time evolution of the state vector in a manner minimally incompatible with the specifications in equations (1) and (2).

The assumptions underlying this method are:

(a) the system consists of approximately linear dynamic and measurement relationships; and

(b) the state vector evolves with time in such a way that it is minimally incompatible with prior dynamic and measurement specifications.

The cost efficiency frontier is obtained by a minimization procedure in which C_M is minimized subject to a constant C_D. Each state sequence estimate, $\hat{X}_T$ can be assigned an incompatibility cost:

$$C(X_T; \mu, T) = \mu C_D(X_T, T) + C_M(X_T, T) \tag{3}$$

where μ is the Lagrange multiplier which takes a value between 0 and $+\infty$. The parameter μ determines the trade-off between the dynamic and measurement costs along the cost efficiency frontier.

The recursive equation for the cost estimation is written as:

$$\phi(x_{T+1}; \mu, T) = x_{T+1}^T Q_T(\mu) x_{T+1} - 2p_T(\mu)^T x_{T+1} + r_T(\mu) \tag{4}$$

where $Q_T(\mu)$, $p_T(\mu)$ and $r_T(\mu)$ are given in equations (5), (6) and (7). These equations are derived in Tirtotjondro & Rao (1992).

$$\begin{aligned} Q_T(\mu) &= \mu D(T)F(T)G_T(\mu) - 2\mu D(T)F(T)G_T(\mu) + \mu D(T) \\ &= \mu D(T)[I - F(T)G_T(\mu)] \end{aligned} \tag{5}$$

$$p_T(\mu) = G_T(\mu)^T\{H(T)^T M(T)[y_T - b(T)] + p_{T-1}(\mu)\} + Q_T(\mu)^T a(T) \tag{6}$$

$$\begin{aligned} r_T(\mu) &= r_{T-1}(\mu) + [y_T - b(T)]^T M(T)[y_T - b(T)] + \mu\, a(T)^T D(T) a(T) \\ &\quad - s_T(\mu)^T [V_T(\mu)^T]^{-1} s_T(\mu) \end{aligned} \tag{7}$$

The recursive equations (8) and (9) are used to find the FLS estimates of the state vector at time $T \geq 1$. By defining:

$$U_T(\mu) = H(T)^T M(T) H(T) + Q_{T-1}(\mu) \tag{8}$$

and

$$z_T(\mu) = H(T)^T M(T)[y_T - b(T)] + p_{T-1}(\mu) \tag{9}$$

the FLS estimate is given by equation (10).

$$\hat{x}_T^{FLS} = [U_T(\mu)]^{-1} z_T(\mu)$$

Further details of the procedure are found in Kalaba & Tesfatsion (1989) and Tirtotjondro & Rao (1992).

DISCUSSION OF RESULTS

The data are normalized by subtracting the mean and dividing by the standard deviation. The data are analysed by using the observations from a year as a unit. The model used in the study has flows y_t from two stations and rainfall r_t as the input variables. The measurement equation for this model is given by equation mentioned under "Theory" with the measurement equation matrix given by equation (11).

$$H(T) = \begin{bmatrix} y_{1,t-1} & y_{1,t-2} & r_{t-1} & 0 & 0 & 0 \\ 0 & 0 & 0 & y_{2,t-1} & y_{2,t-2} & r_{t-1} \end{bmatrix} \tag{11}$$

The vector of state variables is given by equation (14).

$$x_t^T = [x_{1,t}, x_{2,t}, \ldots x_{6,t}] \tag{12}$$

The cost efficiency frontier $C^F(N)$ is generated by varying the weight parameter μ in the cost function (3). In this study, the penalty weights are varied from small values close to 0 to 10 000. Once the cost efficiency frontier is established, a weight μ is chosen such that the point corresponding to it has the minimal distance to the origin. The μ values are given in Table 1.

The μ values in Table 1 and Fig. 2 are not small. Consequently, it appears that the variations in the system equations cannot be ignored. The μ values generally varied between 1 and 10. Figure 2 shows the cost efficiency frontier for the model. Figure 3 is an example of the variation of the first element of the state vector with time. It can be

Table 1 Optimal values of μ.

Flow 1	Flow 2	Rainfall	μ
*d*70	*f*70	*p*870	3
*d*71	*f*71	*p*871	3
*d*72	*f*72	*p*872	3
*m*70	*g*70	*p*270	5
*m*71	*g*71	*p*271	3

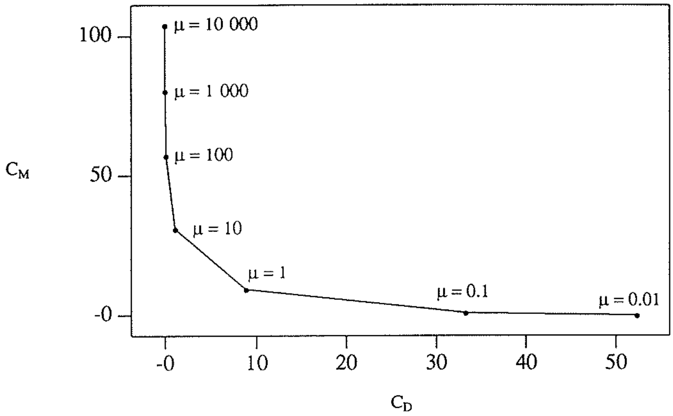

Fig. 2 Cost efficiency frontier for model II for the year 1970.

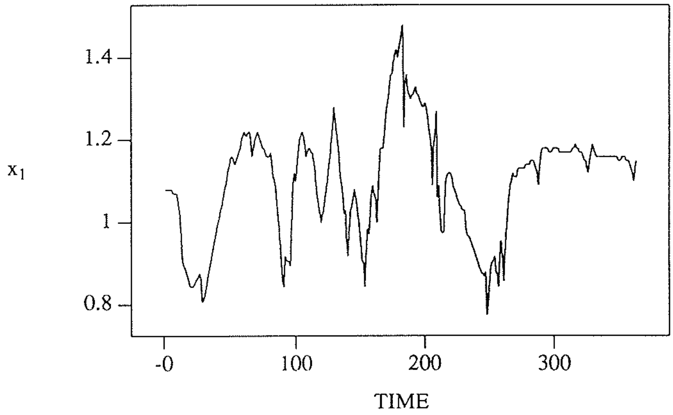

Fig. 3 Variation of state variable x_1 with time for the year 1970.

seen that the state vector varies significantly with time. In Figs 4 and 5 the observed and forecast normalized flows for Dundee and Falls of Rough, respectively, are shown. The forecast performance of the model is good. In general, the forecasts adapt very well to the fluctuations in the observed flows.

VALIDATION TESTS

The residuals of the models are tested for independence using the Portmanteau Lack of Fit test (Box & Jenkins, 1976) and the Box & Ljung (1978) test. In the Portmanteau Lack of Fit test, the statistic Q is computed using equation (13) and compared with the chi-squared value $\chi_\alpha^2\ (L - p - q)$ corresponding to a 95% significance level.

$$Q1 = N\sum_{k=1}^{L} r_k^2(\epsilon) \tag{13}$$

$$Q2 = N^2\sum_{k=1}^{L} (N-k)^{-1} r_k^2(\epsilon) \tag{14}$$

$r_k^2(\epsilon)$ is the square of the correlation coefficients of the residuals and L is the maximum lag considered. p and q are the number of autoregressive and moving average parameters used in the model. The residuals are independent when $Q < \chi_\alpha^2\ (L - p - q)$.

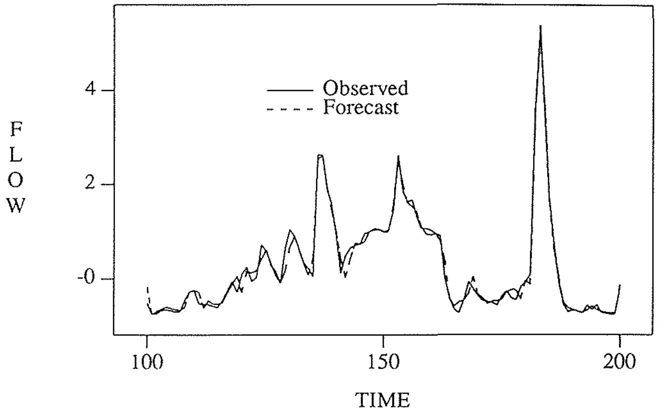

Fig. 4 Observed and forecast normalized flows at Dundee (1970).

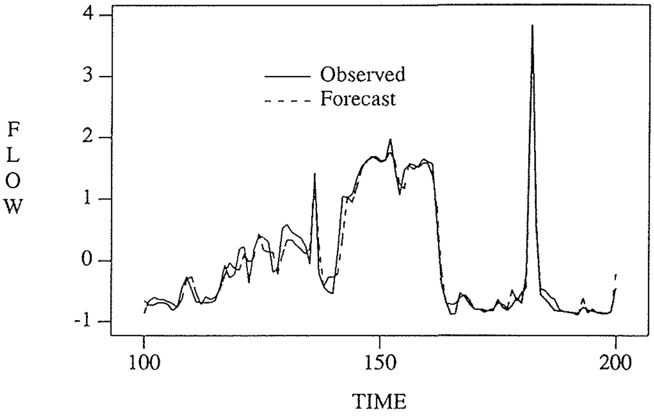

Fig. 5 Observed and forecast normalized flows at Falls of Rough (1970).

The results are given in Table 2. The Box & Ljung (1978) test statistic $Q2$ is given by equation (14). The results show that in all cases, the residuals can be considered to be uncorrelated. Therefore it can be concluded that the models are valid.

FORECASTING

As an example, the multiple days-ahead forecast is computed for the Model with Green River flows at Dundee and Falls of Rough in 1970 as flow inputs and reach 8 precipitation in 1970 as rainfall data. The results of forecasting are shown in Table 3 where MSE is the mean square error between the observed and forecasted values, $\bar{e}$ is the mean of the errors and σ_e is the standard deviation of the errors. These results show that the 1 and 2 days ahead forecast is quite good with very small mean square errors. Figures 6 and 7 show the comparison of the observed and the multiple days-ahead forecast for the example. These figures show that the 1 and 2 days-ahead forecasts are quite accurate. The forecast accuracy decreases in performance with the increase in the number of steps.

CONCLUSIONS

The FLS method is a very good method for screening multivariate models. With different choices for μ, the best μ value can be easily determined from the Cost Efficiency Frontier $C(x_T; \mu, T)$ by choosing the μ value which corresponds to the point with the smallest distance to the origin. The μ parameter obtained in this study are not small. Hence, the variation in the state variables cannot be ignored. The models are found to perform very well in forecasting.

Table 2 Results of residual tests.

$Q1$	$Q2$	$\chi^2_{95\%}$	Remarks
7.979	15.864	22.4	independent
7.472	5.086	22.4	independent
19.024	12.068	22.4	independent
16.331	12.723	22.4	independent
7.452	12.327	22.4	independent

Table 3 Multiple steps-ahead: mean and mean square error statistics.

Flow	Step	MSE	$\bar{e}$	σ_e
d70	1	0.057	0.011	0.239
	2	0.187	0.026	0.432
	3	0.510	0.039	0.714
f70	1	0.057	−0.062	0.264
	2	0.187	−0.067	0.440
	3	0.510	−0.062	0.619

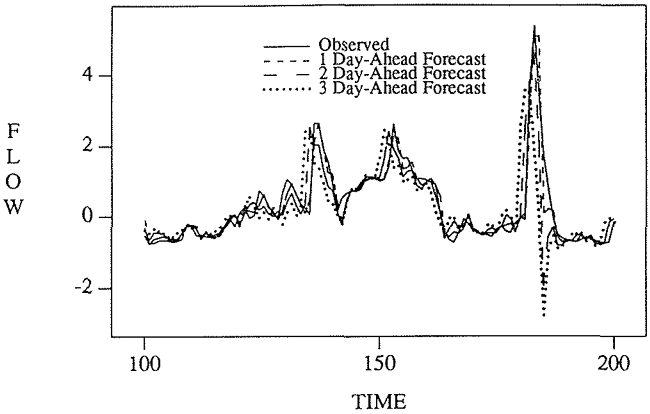

Fig. 6 Observed and months-ahead forecasts of normalized flows at Dundee (1970).

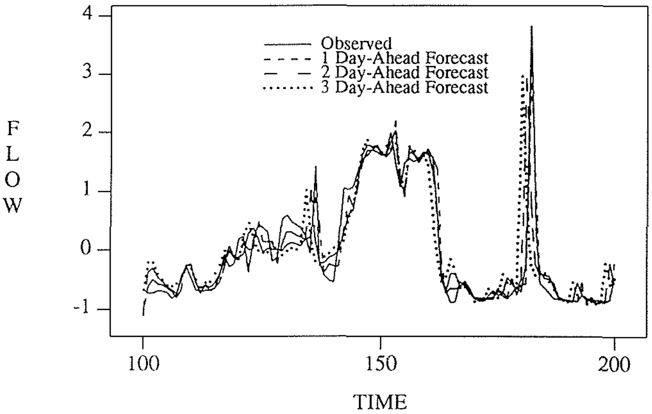

Fig. 7 Observed and months-ahead forecasts of normalized flows at Falls of Rough (1970).

REFERENCES

Amirthanathan, G. E. (1989) Optimal filtering techniques in flood forecasting. In: *FRIENDS in Hydrology* (ed. by L. Roald, K. Nordseth & K. A. Hassel) (Proc. FREND Symp., Bolkesjø, Norway, April 1989), 13-25. IAHS Publ. no. 187.

Box, G. E. P. & Jenkins, G. M. (1976) *Time Series Analysis: Forecasting and Control.* Holden-Day, Oakland, California.

Box, G. E. P. & Ljung, G. M. (1978) On a measure of lack of fit in time series models. *Biometrika* **65**(2), 297-303.

Kalaba, R. & Tesfatsion, L. (1989) Time-varying linear regression via flexible least squares. *Int. J. Comp. Maths. Appl.* **17**(8/9), 1215-1245.

Kalman, R. E. (1960) A new approach to linear filtering and prediction problems. *J. Basic Engng. Trans. Am. Soc. Mech. Engng.* **82** Ser. D(1), 35-45.

Sallas, W. M. & Harville, D. A. (1981) Best linear recursive estimation for mixed linear models. *J. Am. Stat. Assoc.* **76**(376), 860-869.

Tirtotjondro, W. W. & Rao, A. R. (1992) Flexible least squares analysis of hydrologic data. *Tech. Rep. CE-EHE-92-4, School of Civil Engineering, Purdue Univ., W. Lafayette, Indiana 47907, USA.*

Wood, E. F., Szollosi-Nagy, A. & Todini, E. (1979) An adaptive algorithm for analyzing short-term structural and parameter changes in hydrologic prediction models. In: *Modelling Hydrologic Processes* (ed. by H. J. Morel-Seytoux, J. D. Salas, T. G. Sanders & R. E. Smith), 801-815. Water Resources Publications, Fort Collins, Colorado, USA.