JCARI   IFAC Conference, Lisbon

MULTIVARIATE WEEKLY STREAMFLOW FORECASTING MODEL

G.C. de Oliveira
J.P. da Costa
J.M. Damazio
J. Kelman
CEPEL - Electric Energy Research Center
Cidade Universitaria - Ilha do Fundao
P.O. Box 2754 - Rio de Janeiro, Brasil

Abstract. In the operation of a reservoir system, it is of interest to know an estimate of the confidence interval of the future inflow volumes for each reservoir. In this paper, the one-step ahead forecasts of weekly inflows to each reservoir are obtained through the use of a multivariate autoregressive stochastic model (MAR) which, besides serial dependency, represents the spatial inflow dependency structure. The continuous incorporation of measurements favours a recursive estimation procedure for model's parameters. The stochastic model was formulated within the framework of the state-space representation where the state (MAR parameters) is modelled by a simple random walk process to allow for time dependency of parameters. The recursive algorithm employed is the extended Kalman Filter, in which the noise covariance is estimated at each step. The forecast error covariance matrix is used to build the confidence regions for inflows. A case study with a South Brazilian reservoir system for hydropower production is presented. The order of the MAR model was chosen based on performance indexes related to the forecast error obtained within the historical data. The comparison between the MAR model and the best univariate fit to each site series indicates that the multivariate scheme produces forecasts similar to the univariate ones, besides providing multivariate confidence regions for the future inflows.

Keywords. Prediction; Kalman Filter, Multivariate Systems, Streamflow Modelling, State-Space Methods, Hydrology.

INTRODUCTION

Weekly streamflow forecasting is a useful technique for the short-term operation of a hydropower reservoir system (e.g. Pereira, 1985). When the reservoirs that constitute the system are owned by different utilities, it is necessary to check the compatibility of the several at-site forecasts, which are usually done using models developed specifically for each site either through a rainfall-runoff relationship or through a time series approach. As the forecasts may be obtained without considerations to information in nearby sites, the coordinating organism for the operation of the whole system needs a tool to detect whenever the set of at-site forecasts do not fit together. The multivariate forecast confidence region obtained by some simple model is a reasonable "detection device".

It is necessary to deal with the confidence region for the one-week ahead streamflows rather than with the set of the confidence intervals, because a set of at-site forecasts may not be not located on the tails of the univariate distributions and therefore be considered a reasonable prediction, while they are in fact located in the "tail" of the multivariate distribution, and therefore should be considered suspicious. Figure 1 shows an example for two sites, where it can be seen that point A is a suspicious forecast, although it could not be detected by the univariate approach. On the other hand, points B, C or D would be considered unacceptable forecasts under one or both of the univariate confidence intervals and acceptable forecasts by the multivariate confidence region.

The multivariate modelling of the streamflow process may have the further advantage of producing more accurate forecasts than those produced by the use of a set of univariate models of the same type. On the other hand the univariate models are generally selected specifically for each site, even if restricted to the time series approach. It is not obvious which of the two alternatives is more accurate and this question must be examined in a case by case basis.

MAR(p) MODEL - STATE SPACE FORMULATION

Let z_t be a n-vector of standardized normal variables, and v_t be a n-vector of normal variables at instant t such that:

$$E\,(v_t) = 0,$$

$$Cov\,(v_t\, v_t') = R,$$

$$E\,(v_t\, v_s') = 0, \quad t \neq s,$$

where E (.) stands for expectation, Cov (.) stands for covariance and ' stands for transpose. The MAR (p) model is:

$$z_t = A_1 z_{t-1} + \ldots + A_p z_{t-p} + v_t \tag{1}$$

where $A_1, \ldots, A_p$ are the nxn parameter matrices.

It is known (e.g. Ledolter, 1978) that individual series from a MAR model follow an ARMA model. Since univariate forecasting streamflow studies usually deal with low-order ARMA models, the MAR(p) family is a reasonable framework. Note that the total number of parameters of MAR(p) model is pxnxn (matrices $A_1, \ldots, A_p$) plus $n(n+1)/2$ (symmetric R matrix).

An alternative to the moments (Salas, 1980) or maximum likelihood (e.g. Salas and Pegram, 1979) estimates of the MAR parameters, which can present difficulties when one has short records, is the Kalman Filter algorithm.

A state-space formulation of equation (1) considering the matrices $A_1,\ldots,A_p$ as the state x_t is given by:

$$x_t = x_{t-1} + w_t \tag{2}$$

$$z_t = H_t' x_t + v_t \tag{3}$$

where x_t is a n^2p-vector defined as:

$$x_t = [a^1_{11}\ldots a^1_{nn} \quad a^2_{11}\ldots a^2_{nn} \quad \ldots \quad a^p_{11}\ldots a^p_{nn}]' \tag{4}$$

and

$$A_k = \{a^k_{ij}\} \; j=1,\ldots,n;\; i=1,\ldots,n\ ,$$

w_t is a n^2p-vector of gaussian random system noise such that:

$$E\,(w_t) = 0,$$

$$\mathrm{Cov}\,(w_t) = Q,$$

$$E\,(w_t\, w_s') = 0 \qquad s \neq t\ ,$$

$$E\,(w_t\, v_s') = 0 \qquad \forall\ s,t\ ,$$

and H_t is a $n \times n^2p$ *matrix defined as:*

$$H_t = [H_1\ H_2\ \ldots\ H_p] \tag{5}$$

where H_i is a $n \times n^2$ matrix given by

$$H_i = \begin{bmatrix} z'_{t-i} & 0 & \ldots & 0 \\ 0 & z'_{t-i} & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & z'_{t-i} \end{bmatrix}$$

where 0 is a $1 \times n$ vector of zeroes.

For example, take p=2 and n=2 so that

$$A_1 = \begin{bmatrix} a^1_{11} & a^1_{12} \\ a^1_{21} & a^1_{22} \end{bmatrix}$$

$$A_2 = \begin{bmatrix} a^2_{11} & a^2_{12} \\ a^2_{21} & a^2_{22} \end{bmatrix}$$

Then the state-vector becomes:

$$x_t = [a^1_{11}\ a^1_{12}\ a^1_{21}\ a^1_{22} \quad a^2_{11}\ a^2_{12}\ a^2_{21}\ a^2_{22}]'$$

and the H_t matrix:

$$H_t = \begin{bmatrix} z^{(1)}_{t-1} & z^{(2)}_{t-1} & 0 & 0 & z^{(1)}_{t-2} & z^{(2)}_{t-2} & 0 & 0 \\ 0 & 0 & z^{(1)}_{t-1} & z^{(2)}_{t-1} & 0 & 0 & z^{(1)}_{t-2} & z^{(2)}_{t-2} \end{bmatrix}$$

Equation (2) represents a random walk for the MAR model parameters, to allow for time variation. Equation (3) is just another way of writing equation (1).

The Kalman Filter algorithm is a set of equations which allows an estimate to be updated once a new observation becomes available.

The forecasting equations (6)-(9) below give the optimal forecast $\hat{x}(t|t-1)$ of x_t and the optimal forecast $\hat{z}\,(t|t-1)$ of z_t given all the information currently available, besides the uncertainty of these forecasts:

$$\hat{x}(t|t-1)=E(x_t|z_1,\ldots,z_{t-1})=\hat{x}(t-1|t-1) \tag{6}$$

$$P(t|t-1)=\mathrm{cov}(x_t-\hat{x}(t|t-1)|z_1,\ldots,z_{t-1})=P(t-1|t-1)+Q \tag{7}$$

$$\hat{z}(t|t-1)= H_t\ \hat{x}(t|t-1) \tag{8}$$

$$Z(t|t-1)=\mathrm{cov}(z_t-\hat{z}(t/t-1)|z_1,\ldots,z_{t-1})=H_t\, P(t|t-1)H_t'+R \tag{9}$$

At each new observation z_t, define the $n \times 1$ innovation vector u_t as:

$$u_t = z_t - H_t\ \hat{x}(t|t-1) \tag{10}$$

The updating equations (11)-(13) below incorporate the observation z_t into the estimate $\hat{x}(t|t)$ of x_t:

$$\hat{x}(t|t) = \hat{x}(t|t-1) + K_t\ u_t \tag{11}$$

where $K_t = P(t|t-1)\ H_t'\ Z(t|t-1)^{-1}$ is the Kalman gain, and the uncertainty of $\hat{x}(t|t)$ is given by

$$P(t|t) = \mathrm{cov}(x_t - \hat{x}(t|t)|z_1,\ldots,z_t)=(I-K_t\ H_t)\ P(t|t-1) \tag{12}$$

These estimates are conditioned on initial values x_o, $P(0|0)$, Q_o and R_o. In the case of unknown noise covariance matrices Q and R, O'Connel (1980) derives recursive equations for Q_t and R_t that are updated at each new measurement:

$$R_t=((t-1)R_{t-1} + (u_t\ u_t' - H_t\ P(t|t-1)\ H_t'))/t \tag{13}$$

$$Q_t=((t-1)\ Q_{t-1} + (K_t\ u_t\ u_t'\ K_t' + P(t|t)-P(t-1|t-1)))/t \tag{14}$$

where now:

$$K_t = P(t|t-1)\ H_t'\ (H_t\ P(t|t-1)\ H_t' + R_t)^{-1} \tag{15}$$

The measurement forecast error covariance matrix $Z(t|t-1)$ can be used to build at each time t-1 a multivariate confidence region for the forecast $\hat{z}(t|t-1)$.

MODEL FITTING

~~It was~~ were selected Weekly data from three cascaded gauging sites at Iguaçu River, South Brazil, two of them associated with hydropower plants (see Table 1). In order to obtain normality in the data, a logarithm transformation was first applied to the incremental inflow volumes, resulting in a 3-vector y_t of weekly data, t=1, ... ,1040 (20 years of concurrent data). Using the first 18 years of data, the weekly means and *standard* deviations of each site were estimated, and their periodic behaviour ~~were~~ was represented by adjusted Fourier functions.

Then a standardized 3-vector z_t *is obtained as*

$$z_t(i) = (y_t(i) - \mu_t(i))/\sigma_t(i) \tag{16}$$

where $\mu_t(i)$, $\sigma_t(i)$ are the Fourier functions for the weekly mean and standard deviations for sites i=1,2,3. For the correct identification of the dependence *structure* of z_t it is important to remove all the periodicity in the means. Since overremoval of harmonics in the standard deviations does not modify significantly the identification of the dependence model (Yevjevich and Obeysekera, 1985), the same number of relevant harmonics was used to remove periodicities in the means and standard deviations in all sites.

TABLE 1. Hydrologic Sites Characteristics

Site	Name	Drainage Area(km^2)	Hydropower Plant	Installed Capacity (MW)
1	P. Amazonas	3662	-	-
2	U. Vitoria	24211	F. Areia	2508
3	S. Osorio	45824	S. Osorio	1998

Table 2 shows that the z_t series have a high spatial dependency, which justifies a multivariate modelling attempt.

TABLE 2. Sample z_t cross correlation estimates between sites

	1	2	3
1	1.	.808	.700
2		1.	.831
3			1.

MAR Model

The first autocorrelation coefficients for the 3 sites do not present a periodic pattern. Figures 2, 3 and 4 show the autocorrelation function of $z_t(i)$, for sites i=1,2,3. Thus it can be inferred that the MAR parameters are time invariant, that is, Q=0. The Kalman Filter algorithm was first applied to MAR(1) and MAR(2) in order to identify the value of p. In both cases, a forecast experiment was performed with the remaining two years of z_t data. In order to choose the best forecasting model, the performance index selected was the root mean squared forecast error of the incremental inflow volumes. The forecast $\hat{q}_t(i)$ at time t, site i is given by

$$\hat{q}_t(i) = \exp\left(\tilde{y}_t(i) + 0.5\,\sigma_y^2(i)\right) \qquad (17)$$

where

$$\tilde{y}_t(i) = \hat{z}_t(i)\,\sigma_t(i) + \mu_t(i), \qquad (18)$$

$$\sigma_y^2(i) = \sigma_t^2(i)\,\sigma_z^2(i) \qquad (19)$$

and

$\sigma_z^2(i)$ is the i-th diagonal element of $Z(t|t-1)$,

$\hat{z}_t(i)$ is the i-th component of $\hat{\hat{z}}(t|t-1)$

Table 3 presents the performance indices at each site for both cases, and also the performance index for the sum over the sites of individual forecast errors.

TABLE 3. Root Mean Squared Forecast Error-MAR Model

Site	MAR(1)	MAR(2)
1	30.	29.
2	133.	156.
3	318.	321.
1+2+3	423.	437.

By this criteria, MAR(1) is the best choice. Table 4 shows its parameters (A_1 matrix) estimated by the Kalman Filter and their associated *uncertainties*.

TABLE 4. MAR(1) Parameter Estimates

	Parameter	Standard Deviation
a_{11}	.779	.033
a_{12}	-.179	.045
a_{13}	.194	.037
a_{21}	.262	.032
a_{22}	.472	.043
a_{23}	.171	.035
a_{31}	.060	.033
a_{32}	.074	.045
a_{33}	.694	.036

Figures 5,6 and 7 show the residual autocorrelation functions for sites 1,2 and 3 and their 95% confidence interval. Except for site 2, they can be considered as white noise. Figures 8,9 and 10 show the forecasted and measured inflows for sites 1,2 and 3 as well as their 68% confidence intervals.

Univariate ARMA Model

In order to compare the performance of the multivariate scheme with at-site model forecasts, individual ARMA models were fitted to each site with parameters estimated by maximum likelihood method (Hipel, McLeod and Lennox, 1977) using the first 18 years of data. The model order at each site was chosen by the same criterion used earlier.

The ARMA (p,q) model can be written as

$$z_t - \phi_1 z_{t-1} - \dots - \phi_p z_{t-p} = a_t - \theta_1 a_{t-1} - \dots - \theta_q a_{t-q} \qquad (20)$$

where ϕ_j, j=1,...,p are the AR parameters, θ_j, j=1,...,q the MA parameters, and a_t is a normally independently distributed white noise residual with zero mean and variance σ_a^2, where the site subscript i has been dropped for notational convenience.

Table 5 presents the parameters of the best ARMA (p,q) *model* fitted for each site.

TABLE 5. Parameters of Univariate ARMA Models

Site	Model Order	AR Parameters		MA Parameters
1	(2, 1)	.285	.312	.624
2	(1, 1)	.764	-	.226
3	(1, 0)	.797	-	-

The forecast $\hat{z}_t$ is given by

$$\hat{z}_t = \phi_1 z_{t-1} + \dots + \phi_p z_{t-p} - \theta_1 \hat{a}_{t-1} - \dots - \theta_q \hat{a}_{t-q} \qquad (21)$$

where

$$\hat{a}_{t-j} = z_{t-j} - \hat{z}_{t-j}, \qquad j=1,\dots,q \qquad (22)$$

For this case equation (17) becomes

$$\hat{q}_t = \exp(\tilde{y}_t + 0.5\,\sigma_y^2), \qquad (23)$$

where

$\tilde{y}_t = \hat{z}_t\,\sigma_t + \mu_t$,

$\sigma^2_y = \sigma^2_t\,\sigma^2_a$.

Note that the site subscript i has also been dropped.

Table 6 presents the performance indices at each site obtained in this case, as well as the performance index for the sum over the sites of individual forecast errors.

TABLE 6. Root Mean Squared Forecast Error-Unisite ARMA Models

Site	Performance Index
1	29.
2	121.
3	313.
1+2+3	399.

Figures 11, 12 and 13 show the residual autocorrelation functions for sites 1, 2 and 3 and their 95% confidence intervals. Figures 14, 15 and 16 show the forecasted and measured incremental inflows for each site as well as their 68% confidence interval.

COMPARISON BETWEEN MULTIVARIATE AND UNIVARIATE APPROACHES

The simplest forecasting model is the "naive" formulation $\hat{z}_t = z_{t-1}$ which gives a lower bound to the prediction experiment. With this "naive" model, the residuals did not conform to a white noise and the root mean squared forecast error is given in Table 7.

TABLE 7. Root Mean Squared Forecast Error Naive Model

Site	Performance Index
1	40.
2	144.
3	380.
1+2+3	480.

Tables 3 and 6 show that multivariate and univariate approaches have similar performance indices for the forecast period, representing 10% to 20% of gain in relation to the naive formulation.

The residuals' autocorrelation functions from both approaches have very similar patterns. All values, except for site 2 lag 1, lie inside the 95% confidence intervals for the white noise process.

The forecasted values shown in Figures 8,9,10,14,15 and 16 demonstrate that both approaches give satisfactory results. The multivariate approach is advantageous because it produces a confidence region which can be constructed by means of the forecast error covariance matrix $Z(t|t-1)$, in the following way:

$$(z_t-\hat{z}(t|t))'\, Z(t|t-1)^{-1}\, (z_t-\hat{z}(t|t)) \le \chi^2(n,\alpha) \qquad (24)$$

Figure 1 compares the 95% confidence region for the incremental inflows of two reservoirs with the two univariate confidence intervals that were obtained only using the diagonal elements of $Z(t|t-1)$.

CONCLUSIONS

The short-term scheduling of a hydropower system is based on the inflow forecast volumes ~~to~~ for each reservoir. These forecasts are usually obtained through univariate models. However in a river basin the inflows are not only serially but also spatially correlated, and by using a multivariate formulation one can produce a vector of forecasts, as well as the corresponding multivariate confidence region that take these effects into account. This confidence region can also be used to validate forecasts produced by specific at-site models.

In this paper ~~is~~ was presented a multivariate autoregressive stochastic model, MAR(p), within a state-space formulation, such that a Kalman Filter algorithm can be employed to estimate the model parameters and to produce forecasts. It was shown that in a 3-site case study with two-year weekly inflows, the Kalman Filter sucessfully estimated the MAR model parameters and produced one-step ahead forecasts with a performance equivalent to the at-site best ARMA forecasts.

ACKNOWLEDGEMENTS

This work was supported by the Supervision and Control Group of ELETROBRAS (SINSC).

REFERENCES

Hipel, K.W., McLeod, I.A. and Lennox, W.C. (1977). Advances in Box-Jenkins Modelling. 1. Model Construction. Water Resources Research, 13(3) 567-568.

Ledolter, J.(1978). The Analysis of Multivariate Time Series Applied to Problems in Hydrology. Journal of Hydrology, 36, 327-352.

O'Connell, P.E. (Ed.) (1980). Real-Time Hydrological Forecasting and Control. Proceedings of 1st International Workshop, Institute of Hydrology.

Salas, J.O., Delleur J.W., Yevjevich, V. and Lane, W. L. (1980). Applied Modelling of Hydrologic Time Series. Water Resources Publications. Fort Collins, Colorado, USA.

Salas, J.D. and Pegram, G.G.S. (1979). A Seasonal Multivariate Multilag Autoregressive Model in H.J. Morel-Seytoux (Ed.), Modelling Hydrologic Processes. Water Resources Publications. Fort Collins, Colorado, USA.

Pereira, M.V.F. (1985). Optimal Scheduling of Hydrothermal Systems - An Overview. IFAC Symposium on Planning and Operation of Electric Energy Systems, Rio de Janeiro, Brasil.

Yevjvich, V. and Obeysekera, J.T.B.(1985). Effects of Incorrectly Removed Periodicity in Parameters on Stochastic Dependence. Water Resources Research, 21(5), 685-690.

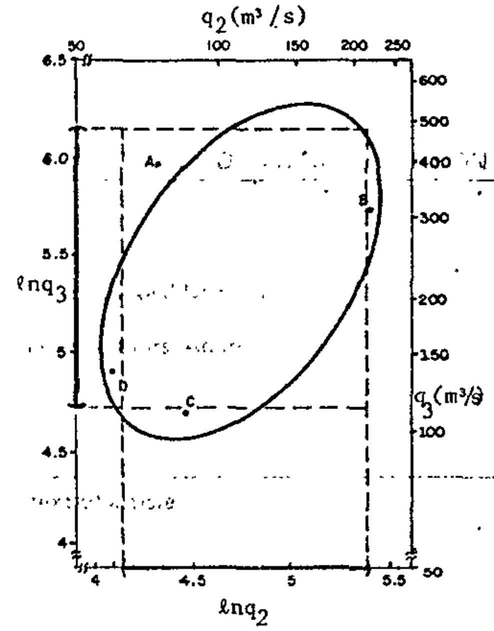

Fig. 1. 95% confidence region and confidence intervals for week 1040 at U.Vitoria (q_2) and S.Osorio (q_3)

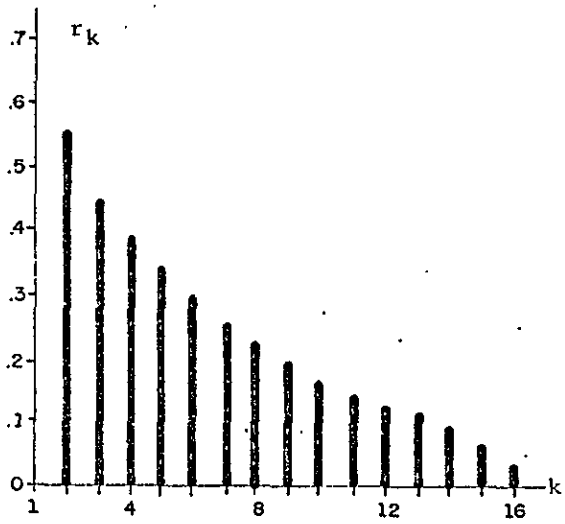

Fig. 2. Autocorrelation Function P. Amazonas

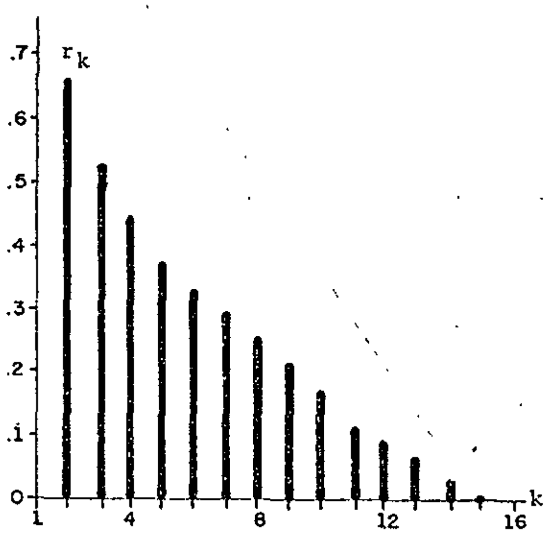

Fig. 3. Autocorrelation Function U. Vitoria

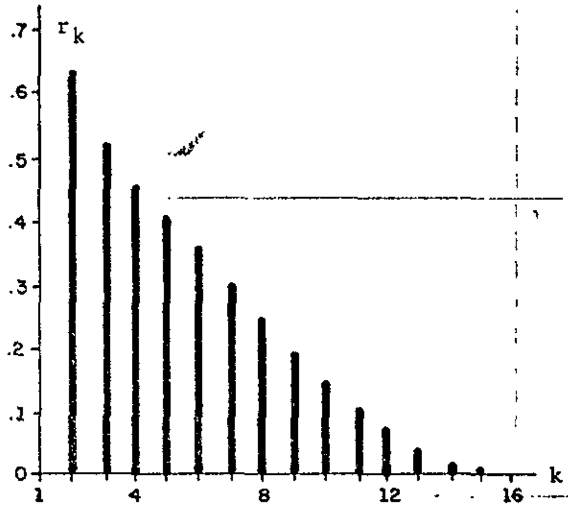

Fig. 4. Autocorrelation Function S.Osorio

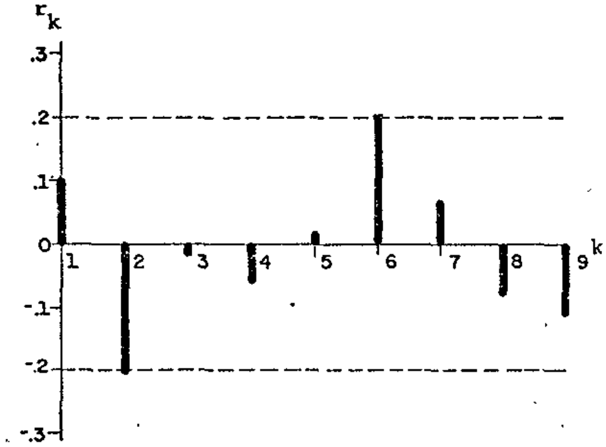

Fig. 5. P. Amazonas - Residual Autocorrelation Function MAR (1) Model

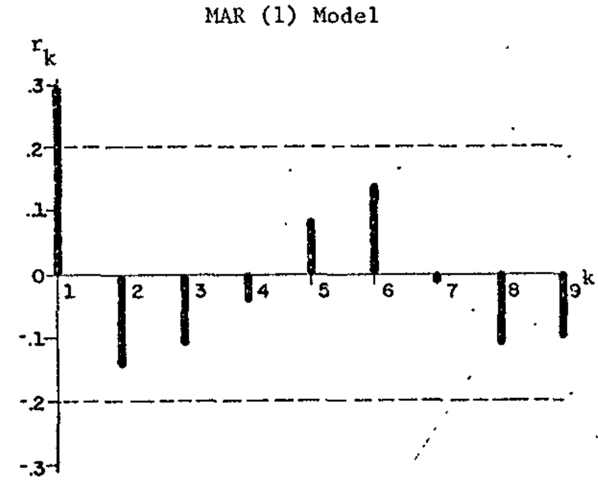

Fig. 6. U. Vitória - Residual Autocorrelation Function MAR (1) Model

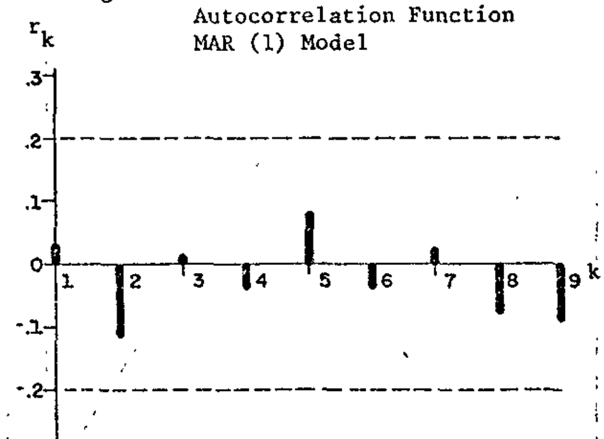

Fig. 7. S. Osório - Residual Autocorrelation Function MAR(1) Model

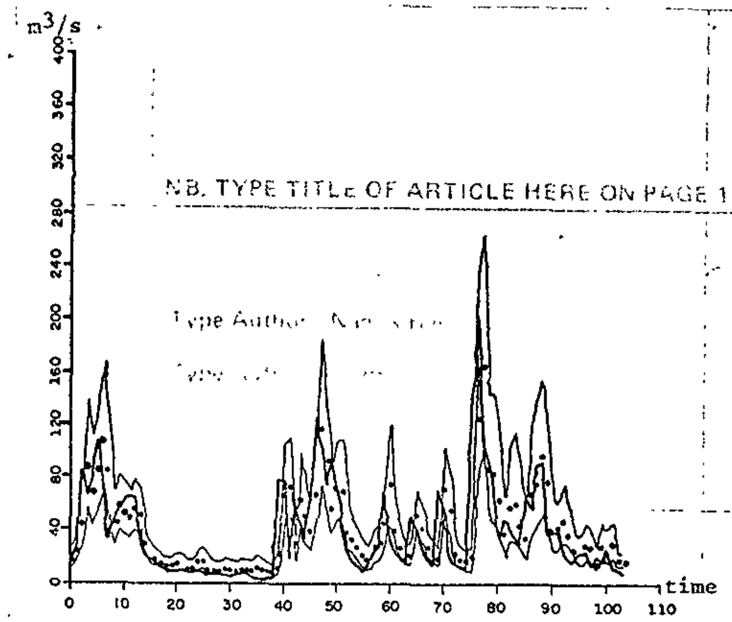

Fig. 8. P. Amazonas - two years of weekly inflow measured (-), forecasted (.) and 68% confidence interval MAR(1) model

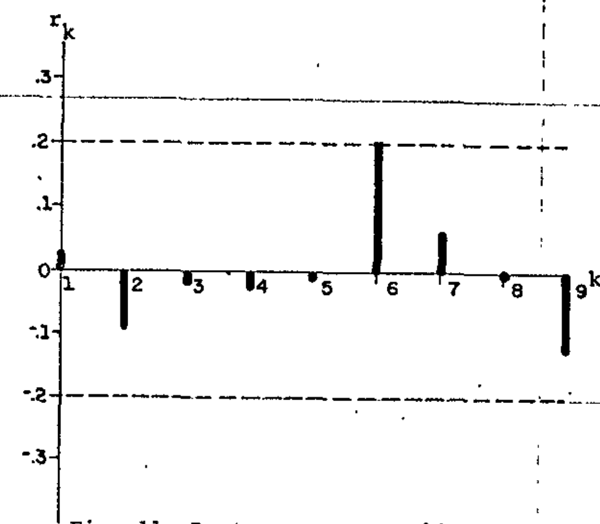

Fig. 11. P. Amazonas - Residual Autocorrelation Funcion ARMA (2,1) Model

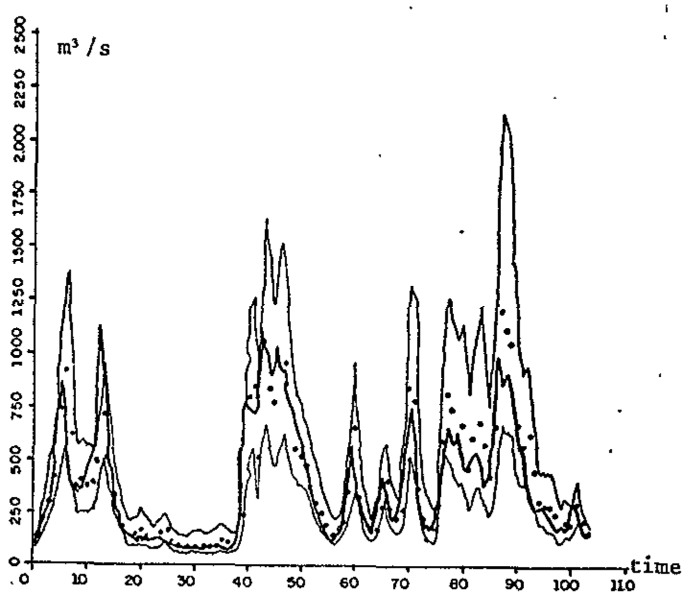

Fig. 9. U. Vitoria - two years of weekly inflow measured (-), forecasted (.) and 68% confidence interval. MAR(1) model

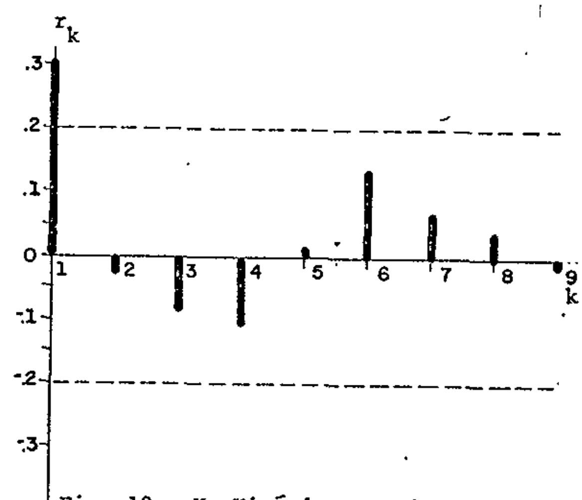

Fig. 12. U. Vitória - Residual Autocorrelation Function ARMA (1,1) Model

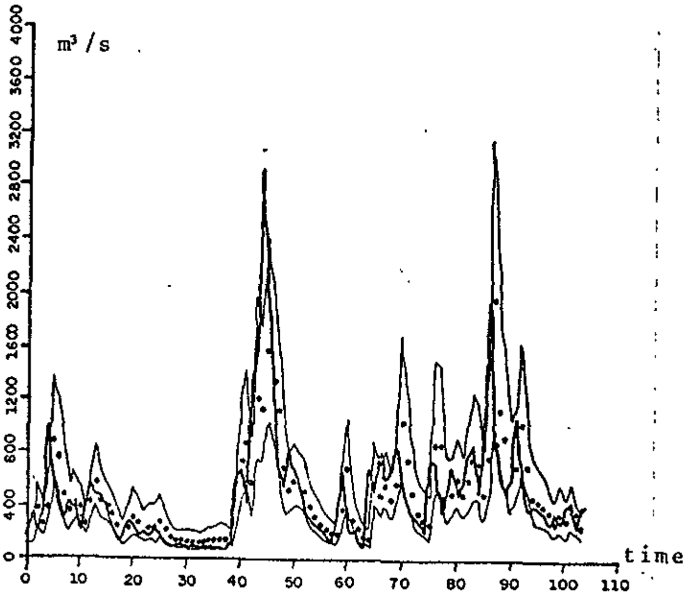

Fig. 10. S. Osorio-two years of weekly inflow measured(-), forecasted(.) and 68% confidence interval MAR(1) model

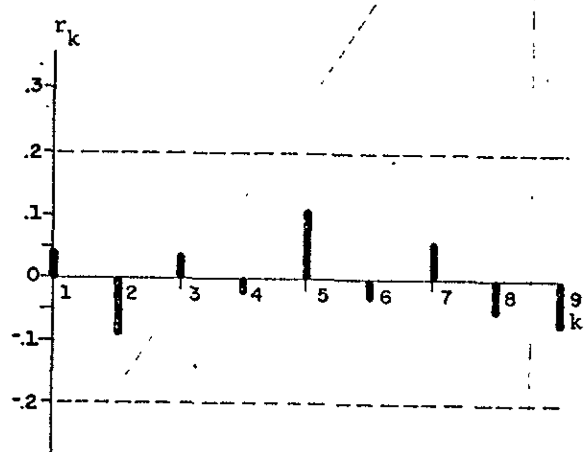

Fig. 13. S. Osório - Residual Autocorrelation Function ARMA (1,0) Model

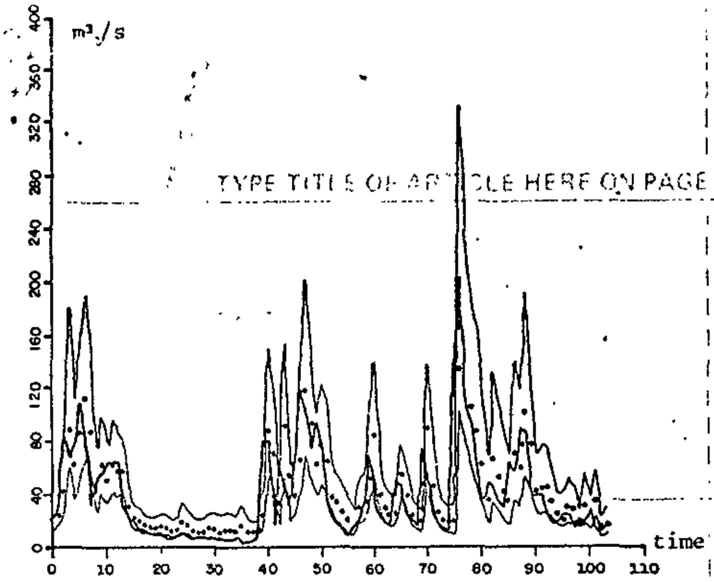

Fig. 14. P. Amazonas two years of weekly inflows measured(-), forecasted(.) and 68% confidence interval. ARMA(2,1) model

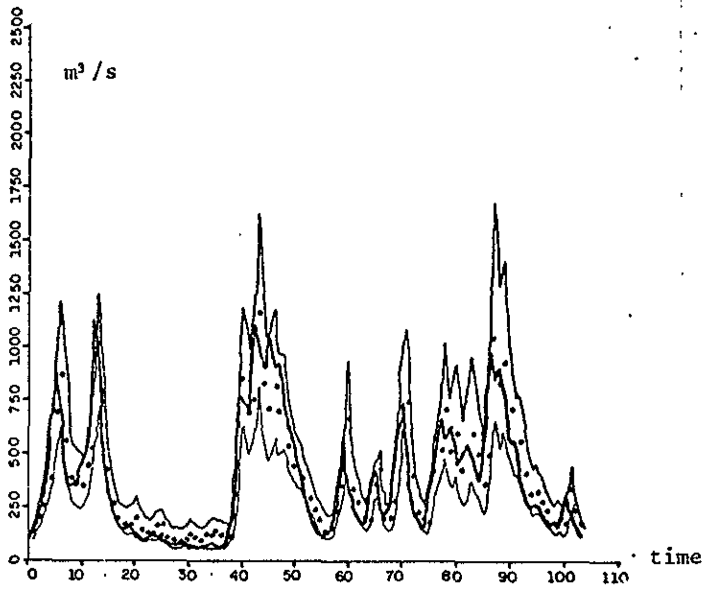

Fig. 15. U. Vitoria two years of weekly inflow measured(-) forecasted(.) and 68% confidence interval. ARMA(1,1) model

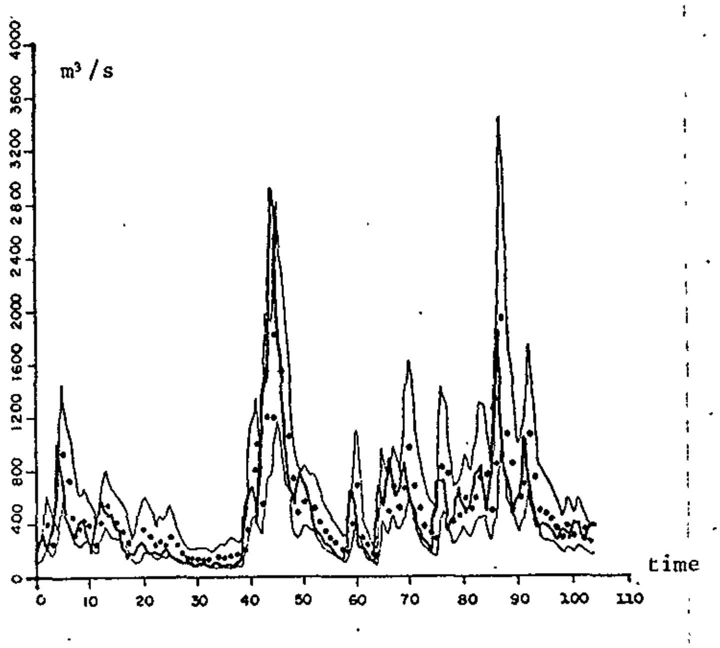

Fig. 16. S.Osório two years of weekly inflow measured (-) forecasted (.) and 68% confidence interval. ARMA(1,0) model