Stochastic models of streamflow: some case studies

P. P. MUJUMDAR* & D. NAGESH KUMAR

Department of Civil Engineering, Indian Institute of Science,
Bangalore, 560 012, India.

Abstract Ten candidate models of the Auto-Regressive Moving Average (ARMA) family are investigated for representing and forecasting monthly and ten-day streamflow in three Indian rivers. The best models for forecasting and representation of data are selected by using the criteria of Minimum Mean Square Error (MMSE) and Maximum Likelihood (ML) respectively. The selected models are validated for significance of the residual mean, significance of the periodicities in the residuals and significance of the correlation in the residuals. The models selected, based on the ML criterion for the synthetic generation of the three monthly series of the Rivers Cauvery, Hemavathy and Malaprabha, are respectively AR(4), ARMA(2,1) and ARMA(3,1). For the ten-day series of the Malaprabha River, the AR(4) model is selected. The AR(1) model resulted in the minimum mean square error in all the cases studied and is recommended for use in forecasting flows one time step ahead.

Modèles stochastiques de l'écoulement - quelques études de cas

Résumé Dix modèles test de la famille ARMA ont été éxaminés pour représenter et prévoir les débits mensuels et décadaires de trois rivières Indiennes. Les meilleurs modèles pour prévoir et représenter les données ont été choisis en utilisant respectivement le critére du Moindre Carré (MMCE) et le Maximum de Probabilité (MP). Les modèles choisis ont été validés pour la signification de la moyenne résiduelle, la signification des périodicités dans les résidus et la signification de la corrélation dans les résidus. Les modèles choisis basés sur le critére du MP pour la production synthétique de séries de débits de trois mois des rivières Kaveri, Hemavathi et Malaprabha sont respectivement AR(4), ARMA(2,1) et ARMA(3,1). Pour les séries décadaires de la rivière Malaprabha, le modèle AR(4) a été choisi. Le modèle AR(1) a conduit à la valeur minimale de moyenne quadratique dans tous les cas étudiés et on le recommande en prévoyant les débits avec un pas de temps en avance.

*now with the Department of Civil Engineering, Indian Institute of Technology, Bombay 400 076, India.

Open for discussion until 1 February 1991

INTRODUCTION

The development and use of stochastic models of hydrological phenomena play an important role in water resources engineering, including their use to forecast river flows. The choice of the right model for a given hydrological series is an important aspect of the modelling process. The statistical models that are best suited for three South Indian rivers, viz. the Cauvery, Malaprabha and Hemavathy, are investigated herein. Many models of the ARMA (Auto-Regressive Moving Average) family are considered in this study, and for each river a model is selected for the representation of data and for one step ahead forecasting. As demonstrated below, the best models for these two needs are often not the same. The ARIMA (Auto-Regressive Integrated Moving Average) models are deliberately excluded from the study, as differencing the series (which is an essential feature of such models) causes the variance to increase continuously and hence such models cannot be used for the simulation of data (Kashyap & Rao, 1976). They may, however, be used for one step ahead forecasting, where they may perform as well as the ARMA models.

Box & Jenkins (1970) give a method to estimate the orders of the AR and MA terms of a model based on autocorrelations and partial autocorrelations. Procedures for estimating these orders from the given data based on testing residuals, given in Kashyap & Rao (1976), are used in the present study.

A popular decision rule for comparing models in the time series literature is the Akaike Information Criterion (AIC) (Akaike, 1974). However, investigations, both theoretical (Kashyap, 1980) as well as numerical, have indicated flaws in the AIC rule. Firstly, the AIC has no optimal property, i.e. it does not minimize the average value of any criterion function. Secondly, the AIC rule is not consistent, i.e. the probability that the decision rule will choose a wrong model does not go to zero even when the number of observations tends to infinity (Shibata, 1976).

Rao *et al.* (1982) have given a rule that is consistent. They have generalized the technique used by Kashyap (1977) in which a minimum probability of error rule was developed for comparing generalized AR models of different orders.

In some time series applications, the given data are often transformed by a non-linear transformation such as a Box-Cox transformation (Box & Cox, 1964). By employing the methods of Granger & Newbold (1976) it is now possible to obtain Minimum Mean Square Error (MMSE) forecasts of the original series when the data have been changed by a non-linear transformation.

Several investigators have used Bayes decision theory for choosing the model type and order (Valdes *et al.*, 1979; Schwartz, 1978). Akaike (1979) interprets the AIC criterion as a Bayes rule.

In constructing an appropriate model for a given streamflow series, the following procedure is usually followed: (a) the selection of the appropriate type of model among AR, MA, ARMA, ARIMA and seasonal ARIMA models; (b) the choice of orders for the selected model; (c) the estimation of the parameters in the model using the given streamflow series; and (d) validation of the model by residual testing and by simulation. This procedure is applied to identify models for forecasting and synthetic generation of four streamflow series from three rivers in Karnataka State, India. The three rivers considered for the study are the Cauvery, Hemavathy & Malaprabha. In the case of the Malaprabha, both monthly and ten-daily series are studied, whereas only monthly series are considered for the other two rivers. Table 1 gives the summary details of the data used for the four series.

In the following paragraphs, the selection of models of the ARMA family with the two criteria, MLE and MMSE, is discussed.

Table 1 Data used for the study

Stream & site name	*Period for which data are available*	*Type of data*
Cauvery at Krishna Raja Sagara Reservoir	*June 1934 - May 1974 (40 years)*	*Monthly*
Hemavathy at Akkihebbal	*June 1916 - May 1974 (58 years)*	*Monthly*
Malaprabha at Manoli	*June 1950 - May 1985 (35 years)*	*Monthly*
Malaprabha at Manoli	*June 1950 - May 1985 (35 years)*	*Ten-daily*

MODEL DESCRIPTION

The models belonging to the ARMA family may be written as:

$$y(t) = \Sigma_{j=1}^{m_1} \phi_j y(t - j) + \Sigma_{j=1}^{m_2} \theta_j w(t - j) + C + w(t) \qquad (1)$$

where $\{y(t), t=1,2,.....\}$ is the series being modelled;

m_1 is the number of AR parameters;

ϕ_j is the jth AR parameter;

m_2 is the number of MA parameters;

θ_j is the jth MA parameter;

C is a constant; and

$\{w(t), t=1,2,.....\}$ is the residual series.

The important assumptions involved in such models are that $[w(t)]$ has zero mean with terms which are uncorrelated and form an independently identically distributed random variable.

By choosing different values of m_1 and m_2 different models of the ARMA family can be generated. The simplest model belonging to this family would be the AR(1) model, for which $m_1 = 1$ and $m_2 = 0$, i.e.:

$$y(t) = \phi_1 y(t - 1) + C + w(t) \qquad (2)$$

ARMA models may be used with different transformations of the original (observed) series (Granger & Newbold, 1976; Granger & Anderson, 1978). Commonly used transformations are the logarithm transform (Box & Jenkins, 1970) and the square root transform (McLeod *et al.*, 1977). These transformations decide the class to which the model belongs. The observed series and the standardized series also constitute important classes. A standardized series $\{x_t\}$, in this context, is defined as the series $\{y_t\}$:

$$y_t = \frac{x_t - \bar{x}_i}{S_i}$$

in which $\bar{x}_i$ is the estimate of the mean streamflow of the period i (month or ten days) to which t belongs and S_i is the estimate of the standard deviation of the streamflows of the period i. Standardization ensures the removal of periodicities inherent in the process. In the present work, only the standardized series are considered for the selection of models. It may be more useful to study different classes of models and select the best model for each of the classes and compare the performances of these models before a model is finally selected.

Both contiguous and non-contiguous models are studied. The non-contiguous models account for the most significant periodicities without considering the intermediate terms which may be insignificant. For example, a non-contiguous AR(3) model with significant periodicities at first, fourth and twelfth lags would be,

$$y(t) = \phi_1 \, y(t-1) + \phi_4 \, y(t-4) + \phi_{12} \, y(t-12) + C + w(t) \qquad (3)$$

The moving average terms are similarly considered. The obvious advantage of non-contiguous models is the reduction in the number of parameters to be estimated while accounting for the significant periodicities. Exactly which terms to include in the non-contiguous models would have to be decided based on the spectral analysis of the series under consideration. In the present case, the term corresponding to the twelfth lag was included for all the monthly series and that at 36th lag for the ten day series of the Malaprabha river was considered. In the latter case, a year consists of exactly 36 periods. Thus a non-contiguous ARMA(3,3) model for the monthly series will be:

$$\begin{aligned} y(t) = {} & \phi_1 \, y(t-1) + \phi_2 \, y(t-2) + \phi_{12} \, y(t-12) + \\ & \theta_1 \, w(t-1) + \theta_2 \, w(t-2) + \theta_{12} w(t-12) + C + \mathrm{w}(t) \end{aligned} \qquad (4)$$

In the following paragraphs, the model selection based on the Maximum Likelihood Estimate (MLE) and the Mean Square Error (MSE) criteria is discussed and the results are presented for the three rivers chosen for the study. The validation tests carried out on the selected models are subsequently presented.

MODEL SELECTION

The problem of model selection is an important one in time series analysis as there are infinitely many possible models and the choice of a wrong model may result in a costly decision. Out of these possible models however, only a few need to be considered for modelling a given streamflow sequence. AR parameters of up to order 6 and MA parameters of up to order 2 would, in general, serve the purpose. In this study, therefore, only the following models are investigated: AR(1), AR(2), ..., AR(6), ARMA(1,1), ARMA(2,1), ARMA(3,1), ARMA(1,2) and ARMA(2,2).

A model may be selected as the best among those investigated by using the following two criteria: Maximum Likelihood rule (ML) and Mean Square Error (MSE). Many other criteria are available and these two are representative of those available. Both these methods are used for the selection of the best model for each of the three rivers considered. The two criteria used for the model selection are discussed below.

Maximum likelihood rule

Selection of a model by this criterion involves evaluating a likelihood value for each of the candidate models and choosing the model which gives the highest value. The general form of the log-likelihood function for the *i*th model for a Gaussian process is (Kashyap & Rao, 1976):

$$L_i = \ln\, [p(z, \hat{\phi}_i)] - n_i \tag{5}$$

where L_i is the likelihood value;
p is the probability density function;
z is the vector of the historical series;
$\hat{\phi}_i$ is the vector of the parameters and residual variance,
$(\theta_1, \theta_2, ..., \phi_1, \phi_2, ..., \rho_i)$;
ρ_i is the residual variance; and
n_i is the number of parameters.

An immediate observation of the likelihood function (equation (5)) is that, in general, as the number of parameters, n_i, increases, the likelihood value decreases. Thus it is to be expected that the ML rule selects models with a small number of parameters. This is the principle of parsimony propounded by Box & Jenkins (1970). A particular likelihood function within this general framework is (Kashyap & Rao, 1976):

$$L_i = \left[\frac{-N}{2} \ln \hat{\rho}_i - n_i\right] + \left[\frac{-N}{2} \ln 2\pi + \frac{m_1}{2} \ln \left[\frac{\rho_i}{\rho_y}\right] - \frac{N}{2}\right] + \tfrac{1}{2}\left[m_1 - \frac{\Sigma_{t=1}^{m_1} y^2(t)}{\rho_y}\right] \tag{6}$$

where ρ_y is the variance of $\{y(t)\}$.

Usually $m_1 << N$, in which case equation (6) may be written as (Kashyap & Rao, 1976):

$$L_i = \frac{-N}{2}\left[\ln \hat{\rho}_i\right] - n_i \qquad (7)$$

Equation (7) is the likelihood function that is used to select the model for a given site. The twelve models mentioned above are the candidate models. Table 2 gives the likelihood values for these twelve models for each of the three streams for both contiguous and non-contiguous models. It is apparent that, for a given model, the value of the likelihood function differs significantly from one site to another. This large difference may be due to the difference in the lengths of the data available as the variance of the residuals (or the logarithm of it) is not likely to cause the observed magnitude of change in the values of the likelihood function. The relative values of the likelihood function for different models when applied to a given site, rather than those for a given model for different sites, is of interest. In Table 2, the values shown with an asterisk superscript are the maximum values in their respective rows. For the Cauvery river at the KRS reservoir site, the model corresponding to the maximum likelihood value is AR(4). This is in

Table 2 *Maximum likelihood values*

Site	*Contiguous ARMA*										
	(1,0)	*(2,0)*	*(3,0)*	*(4,0)*	*(5,0)*	*(6,0)*	*(1,1)*	*(1,2)*	*(2,1)*	*(2,2)*	*(3,1)*
KRS Reservoir (Monthly)	29.33	28.91	28.96	31.63*	30.71	29.90	30.58	29.83	29.83	28.80	29.45
Hemavathy (Monthly)	22.53	22.55	22.64	22.94	22.47	21.51	23.38	24.96*	24.48	23.94	22.37
Malaprabha (Monthly)	0.588	0.830	-0.16	-0.86	-0.68	-0.63	0.660	-0.07	-0.74	-1.12	-1.19*
Malaprabha (Ten-daily)	59.71	60.26	59.87	61.97*	60.97	60.80	60.66	61.52	58.94	58.64	60.91
Site	*Non-contiguous ARMA*										
	(2,0)	*(3,0)*	*(4,0)*	*(5,0)*	*(6,0)*	*(7,0)*	*(2,2)*	*(2,3)*	*(3,2)*	*(3,3)*	*(4,2)*
KRS Reservoir[a] *(Monthly)*	28.52	28.12	28.21	30.85*	29.94	29.12	29.81	28.82	28.48	28.06	28.65
Hemavathy[a] *(Monthly)*	22.52	22.49	22.57	23.09	22.76	21.77	23.37	24.81*	24.33	22.87	22.85
Malaprabha[a] *(Monthly)*	-0.37	-0.17	-1.16	-1.86	-1.66	-1.63	-0.33	-1.07	-1.74	-2.11	-2.34*
Malaprabha[b] *(Ten-daily)*	58.80	59.33	58.94	61.13*	60.14	59.98	59.73	60.64	58.03	57.67	60.06

a - AR & MA parameters at 12th lag (ref. equation (4))
b - AR & MA parameters at 36th lag (ref. equation (4))
* *indicates the selected model.*

accordance with observation of the spectral analysis of the series. For the Hemavathy river at Akkihebbal, the model selected is ARMA (1,2). For the monthly series of the Malaprabha river at Manoli, ARMA(3,1) is chosen, whereas for the ten-day series of the same river the AR(4) model is selected.

For a given series, the choice of a contiguous model or a non-contiguous model is decided by the relative likelihood values for the two models. Thus for the monthly series of the Cauvery and Hemavathy, contiguous models are adequate whereas for the Malaprabha a non-contiguous ARMA(4,2) model is the best one among those considered. Table 3 gives the parameters estimated for the models selected to represent the four series. The values indicated within the brackets are the standard error values associated with the parameters. For a parameter to be significant, its absolute value must be larger than the standard error.

Table 3 Parameter values and their standard error for models selected on ML rule basis

Site	*Model selected*	*Parameters with their standard error in brackets*
KRS Reservoir (Monthly)	*ARMA(4,0)*	ϕ_1 = *0.2137 (0.0644)* ; ϕ_2 = *0.0398 (0.0659)* ϕ_3 = *0.0540 (0.0661)* ; ϕ_4 = *0.1762 (0.0652)* C = *-0.0157 (0.1070)*
Hemavathy (Monthly)	*ARMA(1,2)*	ϕ_1 = *0.9015 (0.0605)* ; θ_1 = *0.5404 (0.0838)* θ_2 = *0.1737 (0.0659)* ; C = *-0.1706 (0.1418)*
Malaprabha (Monthly)	*ARMA(3,1)*	ϕ_1 = *0.9402 (0.0650)* ; ϕ_2 = *-0.0830 (0.0612)* ϕ_3 = *-0.0897 (0.0560)* ; θ_1 = *0.6059 (0.1756)* C = *0.1425 (0.1124)*
Malaprabha (Ten-daily)	*ARMA(4,0)*	ϕ_1 = *0.4852 (0.0410)* ; ϕ_2 = *0.0441 (0.0456)* ϕ_3 = *-0.0051 (0.0457)* ; ϕ_4 = *0.1034 (0.0416)* C = *0.1243 (0.0974)*

Models such as those in Table 3 are often used for the synthetic generation of data. Sequences generated by such models are used for the design of reservoirs. Such simulated sequences would obviously be different from one model to another. Designs based on such sequences would thus depend on the right choice of model. The maximum likelihood estimate criterion is suited for the selection of a model for simulation purpose. For short-term forecasting, such as one step ahead forecasting, the mean square error (MSE) criterion may be more useful (Kashyap & Rao, 1976). Selection of a model based on an MSE criterion is known as the prediction approach, and is discussed below.

Prediction approach (MSE criterion)

The procedure involved in this approach is quite simple and can be summarized as follows: (a) estimate the parameters of different models using a portion, usually half, of the available data; (b) forecast the second half of the series one step ahead by using the candidate models; (c) estimate the MSE corresponding to each model; and (d) select the model that results in the least value of the MSE. The one step ahead forecast, $\hat{y}[(t + 1)/t]$, for ARMA (m_1, m_2) is given by:

$$\hat{y}[(t + 1)/t] = \sum_{j=1}^{m_1} \phi_j \; y(t - j) + \sum_{j=1}^{m_2} \phi_j w(t - j) + C \tag{8}$$

$\hat{y}[(t + 1)/t]$ represents the forecast streamflow for the time, $t + 1$, given the streamflow up to and including the time, t. The one step ahead forecast error is given by:

$$e(t + 1) = y(t + 1) - \hat{y}[(t + 1)/t] \tag{9}$$

When the series consists of N observations, the first $N/2$ observations are used for the parameter estimation of the candidate models. The streamflows from $N/2 + 1$ to N are forecast by using these models and their errors calculated. The MSE for a model is then given by:

$$\text{MSE} = \frac{\sum_{i=N/2+1}^{N} e(i)^2}{N/2} \tag{10}$$

Table 4 gives MSE values for contiguous as well as non-contiguous models for all the series considered. For all the cases the simplest model, AR(1), results in the least value of the MSE, underlining the fact that for one step ahead forecasting quite often the simplest model is sufficient. Also, in the present case, as the number of parameters increases, the MSE increases, which is an interesting result contrary to the common belief that models with larger numbers of parameters give better forecasts. For all the four series of the streamflows considered, the AR(1) model is strongly recommended for use in forecasting the series one step ahead. However it should be noted that this is not a general conclusion and for other time series a similar analysis has to be carried out separately to decide the model that suits the particular sequence the best. Table 5 gives the estimated parameters, with their standard error in brackets, for the models selected on the MSE criteria.

The exercise so far has been to identify a model both for simulation and for forecasting. Before the model is used, however, it has to be validated. The major assumptions that have gone into the construction of the model must be checked for their validity in the selected model. The following section discusses the validation tests carried out in this study and the results.

VALIDATION TESTS

The following tests are carried out to examine whether the following

Table 4 Mean square error values

Site	*Contiguous ARMA*										
	(1,0)	*(2,0)*	*(3,0)*	*(4,0)*	*(5,0)*	*(6,0)*	*(1,1)*	*(1,2)*	*(2,1)*	*(2,2)*	*(3,1)*
KRS Reservoir (Monthly)	0.97*	1.92	2.87	3.82	4.78	5.74	2.49	2.17	3.44	4.29	1.89
Hemavathy (Monthly)	0.78*	1.54	2.31	3.08	3.85	4.62	0.98	0.81	0.80	1.35	2.41
Malaprabha (Monthly)	0.77*	1.54	2.31	3.08	3.85	4.62	0.98	0.81	0.80	1.35	2.41
Malaprabha (Ten-daily)	0.62*	1.24	1.85	2.47	3.09	3.72	0.77	1.02	1.07	2.44	1.62

Site	*Non-contiguous ARMA*										
	(2,0)	*(3,0)*	*(4,0)*	*(5,0)*	*(6,0)*	*(7,0)*	*(2,2)*	*(2,3)*	*(3,2)*	*(3,3)*	*(4,2)*
KRS Reservoir[a] *(Monthly)*	0.96*	1.89	2.84	3.79	4.74	5.70	2.42	1.99	2.52	1.15	1.71
Hemavathy[a] *(Monthly)*	0.76*	1.51	2.25	3.00	3.75	4.49	1.04	1.52	8.68	0.83	1.39
Malaprabha[a] *(Monthly)*	0.77*	1.54	2.31	3.07	3.85	4.62	0.98	0.81	0.80	1.24	3.76
Malaprabha[b] *(Ten-daily)*	0.62*	1.24	1.85	2.47	3.09	3.72	0.68	0.88	1.13	2.04	1.65

a - AR & MA parameters at 12th lag (ref. equation (4))
b - AR & MA parameters at 36th lag (ref. equation (4))
* *indicates the selected model.*

Table 5 Parameter values and their standard error for models selected on MSE basis

Site	*Model selected*	*Parameters with their standard error in brackets*
KRS Reservoir (Monthly)	*ARMA(1,0)*	ϕ_1 = 0.2557 (0.0627) ; C = -0.009 (0.0765)
Hemavathy (Monthly)	*ARMA(1,0)*	ϕ_1 = 0.4204 (0.0502) ; C = -0.1816 (0.0864)
Malaprabha (Monthly)	*ARMA(1,0)*	ϕ_1 = 0.3742 (0.0654) ; C = 0.1404 (0.1110)
Malaprabha (Ten-daily)	*ARMA(1,0)*	ϕ_1 = 0.5311 (0.0348) ; C = -0.1189 (0.0780)

assumptions used in building the model are in fact valid for the model selected:

(a) the residual series $\{w(t)\}$ has zero mean;
(b) no significant periodicity is present in the residual series; and
(c) the residual series is uncorrelated.

The residual series is constructed from equation (1) as follows:

$$w(t) = y(t) - \Sigma_{j=1}^{m_1} \phi_j y(t-j) - \Sigma_{j=1}^{m_2} \theta_j w(t-j) + C \tag{11}$$

Fig. 1 shows the histogram of residuals for the inflows into the KRS reservoir resulting from an AR(4) model. The histogram is skewed to the right, suggesting that a skewed distribution such as the log-Pearson type III should be used in a simulation. A similar skewness is also seen in the histograms of the residuals from the other models selected.

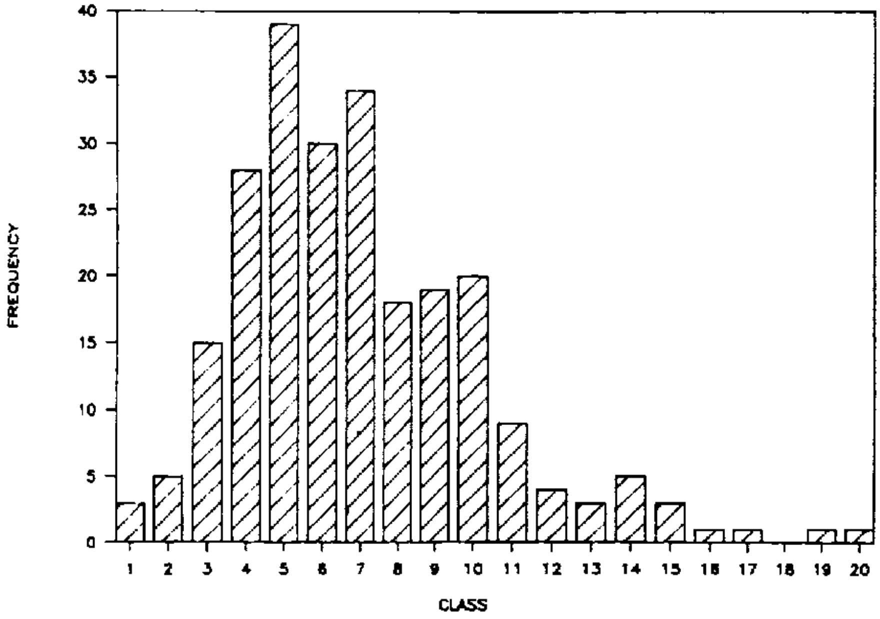

Fig. 1 Histogram of residuals.

All the validation tests are carried out on the residual series only. The tests are summarized briefly in the following paragraphs.

Test 1 Significance of the residual mean

The purpose of this test is to examine the validity of the assumption that the series $\{w(t)\}$ has zero mean. For this purpose a statistic, $\eta(w)$, is defined as:

$$\eta(w) = N^{1/2}\, \bar{w} / \hat{\rho}^{1/2} \tag{12}$$

where $\bar{w}$ is the estimate of the residual mean; and

$\hat{\rho}$ is the estimate of the residual variance.

The statistic, $\eta(w)$, is approximately distributed as $t(\alpha, N-1)$, where α is the significance level at which the test is being carried out. If the value of $\eta(w) \leq t(\alpha, N-1)$, then the mean of the residual series is not significantly different from zero and hence the series passes this test. Table 6 gives the values of the statistic, $\eta(w)$, and $t(\alpha, N-1)$ for all the models selected for the different sites. At the 95% significance level, it is observed that the residual series passes the test in all the cases. This must be true when the models are fitted to the standardized series.

Table 6 Results of tests 1 and 2 for KRS data

Model	*Test 1*		*Test 2*				
			η value for the periodicity				
	η	$t_{0.95}(239)$	*One*	*Two*	*Three*	*Four*	$F_{0.95}(2, 238)$
ARMA (1,0)	*0.002*	*1.645*	*0.527*	*1.092*	*0.364*	*0.065*	*3.00*
ARMA (2,0)	*0.006*	*1.645*	*1.027*	*2.458*	*0.813*	*0.129*	*3.00*
ARMA (3,0)	*0.008*	*1.645*	*1.705*	*4.319*	*1.096*	*0.160*	*3.00*
ARMA (4,0)	*0.025*	*1.645*	*3.228*	*6.078*	*0.948*	*0.277*	*3.00*
ARMA (5,0)	*0.023*	*1.645*	*3.769*	*7.805*	*1.149*	*0.345*	*3.00*
ARMA (6,0)	*0.018*	*1.645*	*4.190*	*10.130*	*1.262*	*0.441*	*3.00*
ARMA (1,1)	*0.033*	*1.645*	*4.737*	*10.090*	*2.668*	*0.392*	*3.00*
ARMA (1,2)	*0.104*	*1.645*	*6.786*	*10.670*	*2.621*	*0.372*	*3.00*
ARMA (2,1)	*0.106*	*1.645*	*7.704*	*12.120*	*2.976*	*0.422*	*3.00*
ARMA (2,2)	*0.028*	*1.645*	*6.857*	*13.220*	*3.718*	*0.597*	*3.00*

Test 2 Significance of the periodicities

For the model to be applicable, the residual series, $\{w(t)\}$, must not have any significant periodicity in it. The following test is carried out to ensure that this is, in fact, true. This test is conducted for different periodicities, and the significance of each of the periodicities is tested. A statistic, $\eta(w)$, is defined as:

$$\eta(w) = \frac{\gamma^2(N-2)}{4\,\hat{\rho}_1} \tag{13}$$

where $\gamma^2 = \hat{\alpha}^2 + \hat{\beta}^2$;

$$\hat{\rho}_1 = 1/N \left\{\Sigma_{t=1}^{N} \left[w(t) - \hat{\alpha}\cos(W_1 t) - \hat{\beta}\sin(W_1 t)\right]^2\right\};$$

$$\hat{\alpha} = 2/N \;\; \Sigma_{t=1}^{N} w(t)\cos(W_1 t);$$

$$\hat{\beta} = 2/N \;\; \Sigma_{t=1}^{N} w(t)\sin(W_1 t); \text{ and}$$

$2\pi/W_1$ is the periodicity for which the test is being carried out.
The statistic, $\eta(w)$, is distributed approximately as $F_\alpha(2, N-2)$, α being the significance level. The periodicity corresponding to W_1 is not significant if:

$$\eta(w) \leqslant F_\alpha(2, N-2)$$

This test was carried out on the residual series resulting from each of the four streamflow series considered. Table 6 presents the values of $\eta(w)$ for the different periodicities tested. All the periodicities tested were found to be insignificant for the models selected and thus the models passed the test.

Another test carried out for the significance of periodicities is the cumulative periodogram test, also known as Bartlett's test (Bartlett, 1946). Unlike the previous test which has to be carried out for one periodicity at a time, this test is conducted to detect the first significant periodicity in the series. If a significant periodicity is observed, the next significant periodicity will be detected by carrying out the test on the series from which the first periodicity is removed, and so on. The test is briefly explained below:

$$\text{Define } \gamma_k^2 = \left[2/N \ \Sigma_{t=1}^{N} \ w(t) \cos\,(W_k t)\right]^2 + \left[2/N \ \Sigma_{t=1}^{N} \ w(t) \sin\,(W_k t)\right]^2$$

$$k = 1,2,....,N/2 \tag{14}$$

$$\text{Compute } g_k = \frac{\Sigma_{j=1}^{k} \gamma^2{}_j}{\Sigma_{k=1}^{N/2} \gamma_k^2} \tag{15}$$

It is noted that $0 \leqslant g_k \leqslant 1$. The plot of g_k versus k is known as the cumulative periodogram. On the cumulative periodogram two confidence limits are drawn. These are given by $\pm\lambda/(N/2)^{1/2}$. The value of λ prescribed (Kashyap & Rao, 1976) is 1.35 for 95% confidence and 1.65 for 99% confidence. If all the values of g_k lie within the significance band, then there is no significant periodicity present in the series. When one of the g_k values lies outside the significance band (the subsequent values will also lie outside the band), the periodicity corresponding to that value of g_k is significant.

Fig. 2 shows the cumulative periodogram of the residual series resulting from the AR(4) model applied to the KRS inflows. All the values of g_k lie within the significance band, thus confirming the result of the earlier test that no significant periodicity is present in the residual series. The same result is also observed in the case of the models selected for each of the other three streamflow series considered. To contrast the cumulative periodogram, shown in Fig. 2, the cumulative periodogram of the original series (without standardizing) is shown in Fig. 3. It is seen from Fig. 3 that, corresponding to $k = 40$, the periodicity is significant. This value of k corresponds to a periodicity of 12 months.

Of the two tests mentioned in this section, the latter (Bartlett's test) is more convenient computationally. These two tests are carried out both on the original series and on the residual series. In the cases studied, the conclusions drawn from the two tests do not differ from each other at any time. However, the cumulative periodogram test is preferred because of its ability to test all the periodicities at a time.

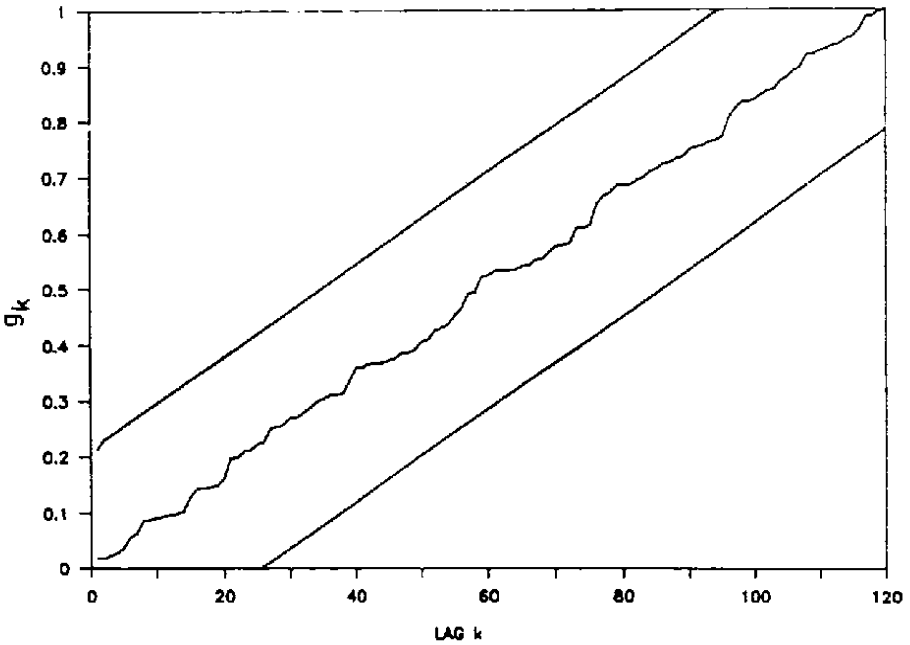

Fig. 2 Cumulative periodogram for residuals.

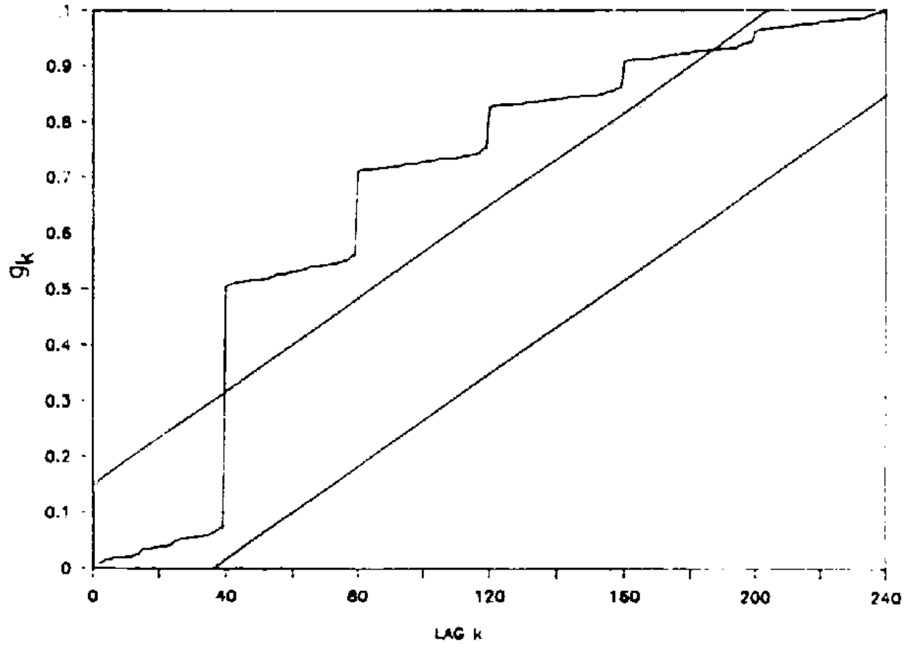

Fig. 3 Cumulative periodogram for monthly data.

Test 3 White noise test

An important assumption in the models studied is that the residual series, $\{w(t)\}$, is a white noise sequence (or that the series is uncorrelated). In this section the residuals are tested for absence of correlation. Two tests are carried out for this purpose and the results are compared.

Whittle's test This test (Whittle, 1952) involves the construction of the covariance matrix. The covariance R_k at lag k of the series, $\{w(t)\}$, is estimated by:

$$R_k = 1/(N-k)\ \Sigma_{j=k+1}^{N}\ w(j)\ w(j-k) \qquad (16)$$

$$k = 0,1,2,\ldots\ldots,k_{\max}$$

The value of $k_{\max}$ is normally chosen as 15% of the sample size., i.e. $k_{\max} = 0.15N$. The covariance matrix, Γ_{n1}, is then constructed as:

$$\Gamma_{n1} = \begin{bmatrix} R_0 & R_1 & R_2 & \ldots\ldots & R_{k_{\max}} \\ R_1 & R_0 & R_1 & \ldots\ldots & R_{k_{\max}-1} \\ R_2 & R_1 & R_0 & \ldots\ldots & R_{k_{\max}-2} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ R_{k_{\max}} & R_{k_{\max}-1} & \ldots & & R_0 \end{bmatrix} \qquad (17)$$

This is a square symmetric matrix of size $n1 = k_{\max}$.

A statistic, $\eta(w)$, is defined as:

$$\eta(w) = (N/n1 - 1)\ (\hat{\rho}_0/\hat{\rho}_1 - 1) \qquad (18)$$

where $\hat{\rho}_0$ is the lag zero correlation coefficient (= 1) and

$$\hat{\rho}_1 = \frac{\det \Gamma_{n1}}{\det \Gamma_{n1-1}}$$

The matrix Γ_{n1-1} is constructed by eliminating the last row and the last column from the matrix Γ_{n1}. The statistic, $\eta(w)$, defined by equation (18) is distributed approximately as $F_\alpha(n1, N-n1)$. If $\eta(w) \leqslant F_\alpha(n1, N-n1)$ then the residual series is uncorrelated. This test was carried out on the residual series resulting from the different models considered for all the four streamflow series. Table 7 gives the values of the statistic, $\eta(w)$, for the models applied to the KRS inflows. From Table 7 it is seen that when $n1 = 25$, ARMA(1,2) and ARMA(2,1) do not pass the test. In all other cases the test is successful. The result from the other sites indicates that the models selected for each series also passed the test.

Portmanteau test This test also uses the covariance, R_k, defined earlier. The statistic, $\eta(w)$, is defined as:

$$\eta(w) = (N - n1)\ \Sigma_{k=1}^{n1}\ (R_k/R_0)^2 \qquad (19)$$

Table 7 Whittle's test for KRS data (N = 240)

$F_{0.95}(n1, N-n1)$	$n1 = 73$ 1.29	$n1 = 49$ 1.39	$n1 = 25$ 1.52
Model	η	η	η
ARMA (1,0)	0.642	0.917	0.891
ARMA (2,0)	0.628	0.898	0.861
ARMA (3,0)	0.606	0.868	0.791
ARMA (4,0)	0.528	0.743	0.516
ARMA (5,0)	0.526	0.739	0.516
ARMA (6,0)	0.522	0.728	0.493
ARMA (1,1)	0.595	0.854	0.755*
ARMA (1,2)	0.851	1.256	1.581*
ARMA (2,1)	0.851	1.256	1.581
ARMA (2,2)	0.589	0.845	0.737

* *indicates that the model does not pass the test.*

This is distributed approximately as $\chi^2_\alpha(n1)$. If $\eta(w) \leqslant \chi^2_\alpha(n1)$, then the series is uncorrelated at the significance level, α. The value of $n1$ is normally chosen as $0.15N$. However the test was carried out for different values of $n1$. Table 8 gives the results of this test. It is seen that the residuals of all models except ARMA(1,2) and ARMA(2,1) pass the test.

Kashyap & Rao (1976) have proved that the portmanteau test is uniformly inferior to Whittle's test and recommended the latter for application.

Table 8 Portmanteau test for KRS data (N = 240)

$\chi^2_{0.95}(k_{max})$ *Model*	$k_{max} = 48$ 65.0	$k_{max} = 36$ 50.8	$k_{max} = 24$ 36.4	$k_{max} = 12$ 21.0
ARMA (1,0)	31.44	33.41	23.02	14.80
ARMA (2,0)	32.03	34.03	24.47	15.17
ARMA (3,0)	30.17	32.05	21.61	13.12
ARMA (4,0)	20.22	21.49	11.85	4.31
ARMA (5,0)	19.84	21.08	11.75	4.14
ARMA (6,0)	19.64	20.87	11.48	3.79
ARMA (1,1)	29.89	31.76*	22.24*	12.76*
ARMA (1,2)	55.88	59.38*	48.37*	39.85*
ARMA (2,1)	55.88	59.38	48.37	38.85
ARMA (2,2)	28.62	30.41	20.39	11.25

* *indicates that the model does not pass the test.*

CONCLUSIONS

Streamflow sequences for three south Indian rivers, viz. the Cauvery, Hemavathy and Malaprabha, are modelled. Ten candidate models of the

ARMA family are studied and the best model for each of the four streamflow series is selected (three monthly and one ten-day series). The best models resulting from the maximum likelihood criterion for the three monthly series of the Cauvery, Hemavathy and Malaprabha are respectively AR(4), ARMA(2,1) and ARMA(3,1). For the ten-day series of the Malaprabha, the model selected is AR(4). Selection of models based on a minimum mean square error criterion results in an AR(1) model for all the four streamflow series considered. The selected models are validated by tests on residuals for the significance of residual mean, the significance of periodicities (Bartlett's test) and the significance of correlations (Whittle's test and Portmanteau test). These tests revealed that the models selected by the two criteria pass all the tests and hence these models are recommended for use in practice for the three rivers.

Acknowledgements This work originated from a series of lectures on Time Series Analysis delivered by Dr. A. Ramchandra Rao, Professor, School of Civil Engineering, Purdue University, currently Visiting Professor, Indian Institute of Science, Bangalore. The authors wish to thank him for his encouragement and guidance.

REFERENCES

Akaike, H. (1979) Bayesian extension of the minimum AIC procedure. *Biometrika*. 66, 237-242.

Akaike, H. (1974) A new look at statistical model identification. *IEEE Trans. Automatic Control* AC-19, 716-722.

Bartlett, M. S. (1946) On the theoretical specification of sampling properties of autocorrelated time series. *J. Roy. Statist. Soc.* B8, 27.

Box, G. E. P. & Cox, D.R. (1964) An analysis of transformations. *J. Roy. Statist. Soc.* B26, 211.

Box, G. E. P. & Jenkins, G. M. (1970) Time Series Analysis. Holden Day, San Francisco, California, USA.

Granger, C. W. J. & Anderson, A. G. (1978) Non-linear time series modeling. In: *Applied Time Series Analysis* (ed. D. F. Findely), Academic Press, New York, USA.

Granger, C. W. J. & Newbold, P. (1976) Forecasting transformed series. *J. Roy. Statist. Soc.* B38 (2), 189-203.

Kashyap, R. L. (1977) Bayesian comparison of dynamic models. *IEEE Trans. Automatic Control* AC-22, 715-727.

Kashyap, R. L. (1980) Inconsistency of AIC rule for estimating the order of auto regressive models. *IEEE Trans. Automatic Control*. AC-25, 996-998.

Kashyap, R. L. & Rao, A. R. (1976) *Dynamic Stochastic Models from Empirical Data*. Academic Press, New York, USA.

McLeod, A. I., Hipel, K. W. & Lennox, W. C. (1977) Advances in Box-Jenkins modeling. 2. Applications. *Wat. Resour. Res.* 13 (3), 577-586.

Rao, A. R., Kashyap, R. L. & Liang-Tsi Mao (1982) Optimal choice of type-and order of river flow time series models. *Wat. Resour. Res.* 18 (4), 1097-1109.

Shibata, R. (1976) Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* 63, 117-126.

Schwartz, G. (1978) Estimating the dimensions of a model. *Ann. Statist.* 6 (2), 461-464.

Whittle, P. (1952) Tests of fit in time series. *Biometrika* 39, 309-318.

Valdes, J. B., Vicens, G. J. & Rodriguez-Iturbe, I. (1979) Choosing among alternative hydrologic regression models. *Wat. Resour. Res.* 15 (2), 347-358.

Received 23 August 1989; accepted 20 December 1989