Forecasting of seasonal streamflows with Box-Jenkins models

HUYNH NGOC PHIEN AND **CHUNG-TUNG TWU**
Asian Institute of Technology, P.O. Box 2754, Bangkok 10501, Thailand

Abstract

This paper deals with the application of the Box-Jenkins models in forecasting seasonal (monthly) streamflows of the Mekong River. Although seasonal models are readily available and have been extensively used in the literature, standardization is introduced to remove the seasonality of monthly streamflows whereby the resulting series may be represented by non-seasonal models which are much easier to identify, require less computational efforts and are simpler in the calculation of forecasts. Moreover, the posterior possibility criterion is used in model selection instead of the modified portmanteau statistic, which is found to be indecisive in such situations.

Introduction

Planning and control of water resources systems have become increasingly important in recent years. At the same time, attempts have also been made to increase the efficiency of existing reservoirs, especially those with many purposes. All of them require the use of inflow forecasts. If the reservoir under consideration is mainly for flood control, short-term forecasts would be sufficient. In this case, except for the period with floods, the forecasting is quite simple because the inflows during those periods without floods do not change much from day to day or hour to hour. However, when one intends to make use of the water in the existing reservoir for planning work such as the planning of cropping systems, the estimation of the irrigated areas of some crops in the coming year, long-term forecasts are needed.

Forecasting of seasonal streamflows is a particular case of long-term forecasting problem. In this case, forecasts with *lead time* ranging from one month to a year may be required.

In a recent survey, Phien (1983) summarized various techniques which are likely to be useful in seasonal streamflow forecasting. They include regression analysis, conceptual models, group method of data handling and what was termed as time series analysis, in which only streamflow data are employed. In this study, an attempt is made to evaluate the ability of the Box-Jenkins method which is commonly believed to perform very well in seasonal data. Moreover, besides following the seasonal models developed by Box and Jenkins (1976), an effort is also made to use the nonseasonal models after a suitable removal of the seasonalities. Comparison of their performance can then be made and conclusions are drawn.

Box-Jenkins models

Let $x_1, x_2, \ldots, x_N$ be a discrete time series measured at equal time intervals. For example, if $F_{i,j}$ denotes the streamflow in month j (j = 1, ..., 12) of year i, then such a time series can be obtained by setting

$$x_t = F_{i,j}\ ,\ t = 12(i-1)+j \qquad (1)$$

where the time interval is equal to one month. The most general form of a time series known as the seasonal autoregressive integrated moving average (ARIMA) model can be written as (see Box and Jenkins, 1976):

$$\phi(B)\,\Phi(B^s)\,\{[(1-B)^d(1-B^s)^D x_t^\lambda] - \mu\} = \theta(B)\,\Theta(B^s)\,a_t \qquad (2)$$

where

x_t^λ = Box-Cox transformed value of x_t
t = discrete time
s = seasonal length, s = 12 for monthly flows
B = backward shift operator defined by
$Bx_t^\lambda = x_{t-1}^\lambda$
$B^s x_t^\lambda = x_{t-s}^\lambda$
μ = mean level of the time series, usually taken as the average of v_t where
$$v_t = (1-B)^d(1-B^s)^D x_t^\lambda \qquad (3)$$
(so that $w_t = v_t - \mu$ has a zero mean)
a_t = normally independently distributed white noise with mean 0 and variance σ_a^2, commonly denoted as NID $(0, \sigma_a^2)$
$\phi(B)$ = $1 - \phi_1 B - \ldots - \phi_p B^p$: seasonal autoregressive (AR) operator
$(1-B)^d$ = ∇^d : nonseasonal differencing operator of order d to produce nonseasonal stationarity
$\Phi(B^s)$ = $1 - \Phi_1 B^s - \ldots - \Phi_P B^{Ps}$: seasonal AR operator of order P
$(1-B^s)^D$ = seasonal differencing operator of order D to produce seasonal stationarity
$\theta(B)$ = $1 - \theta_1 B - \ldots - \theta_q B^q$: nonseasonal moving average (MA) operator
$\Theta(B^s)$ = $1 - \Theta_Q B^{Qs}$: seasonal MA operator

Usually, both d and D take the values 0, 1 or 2. For the stationarity of the series w_t, the roots of $\phi(B)$ and $\Phi(B^s)$ must lie outside the unit circle. Likewise the roots of $\phi(B)$ and $\Phi(B^s)$ must lie outside the unit circle for the invertibility of the process. The above general form is commonly denoted as $(p,d,q) \times (P,D,Q)_s$. The parameters $\phi_1, \ldots, \phi_p$ $(\Phi_1, \ldots, \Phi_P)$ and $\theta_1, \ldots, \theta_q$ $(\Theta_1, \ldots, \Theta_Q)$ are respectively the nonseasonal (seasonal) autoregressive and moving average parameters of the model. When (P,D,Q) = (0,0,0), the model is nonseasonal and is denoted by (p,d,q).

Model building

The model expressed by eq. 2 is very general. In order to arrive at an appropriate selection, Box and Jenkins (1976) suggested a three-stage procedure consisting of identification, estimation and diagnostic checking.

Identification:

In the identification stage, a candidate model is selected. This means that a set of values of p, d, q, P, D and Q is chosen. Hipel *et al.* (1977) discussed the use of autocorrelation function (ACF), partial autocorrelation function (PACF), inverse autocorrelation function (IACF) and inverse partial autocorrelation function (IPACF) for model identification. Interested readers should refer to their work for more elaborated information. Recent development includes the S-array (Gray *et al.*, 1978) and the generalized partial autocorrelation function (Woodward and Gray, 1981) as additional tools for model identification. Different tools can suggest different candidates. Fortunately, in modelling monthly streamflows, the work of McKerchar and Delleur (1974), Tao and Delleur (1976), Hipel *et al.* (1977) and McLeod *et al.* (1977) has provided some typical guidances, and the selection of a suitable candidate is not too difficult. Moreover, as will be seen later, the selected model at this stage is just only tentative, one should not place too much emphasis on such selection. For simplicity, the ACF and PACF would be sufficient.

Estimation:

Once a tentative model has been selected, its parameters are estimated. The most desirable estimates are provided by the method of maximum likelihood which, however, due to computational difficulties, has become available recently (Ansley, 1979). The likelihood function takes the form

$$\ell(\beta, \sigma_a^2 \mid W) \,\alpha\, (\sigma_a^2)^{-n/2} g(\beta) \exp[-\tfrac{1}{2} S(\beta, W)/\sigma_a^2] \quad (4)$$

where

σ_a^2 = variance of the white noise
n = N − d − sD
W = vector with components $W_1, W_2, \ldots, W_n$
β = vector of the unknown parameters
$g(\beta)$ = a function of the parameters but not of the w_i.

An exact expression for the likelihood was given by Newbold (1974) but the alternative form of Ansley (1979) is more suitable for computation. If $\hat{\beta}$ denotes the maximum estimator of β, the maximum likelihood estimator of σ_a^2 is obtained as

$$\hat{\sigma}_a^2 = S(\hat{\beta}, W)/n \quad (5)$$

Using the form provided by Ansley and the general procedure described by Box and Jenkins (1976), the parameters of the tentatively selected model can be estimated.

Diagnostic checking:

At this stage, one needs to check if the selected model is adequate. There are several tools which may be utilized for this purpose. However, the modified portmanteau statistic (Ljung and Box, 1978) is commonly used. This statistic is computed as

$$U = n(n+2) \sum_{k=1}^{L} r_k^2(\hat{a})/(n-k) \quad (6)$$

where $r_k(\hat{a})$ is the autocorrelation of the residuals, and
L has a value from 15 to 25 for nonseasonal models ($L \leq n/4$), and a value of 4s for seasonal time series. This statistic will asymptotically follow a chi-square distribution with $\nu = L - p - q - P - Q$ degrees of freedom. With a specified significance level α, the adequacy of the selected model is rejected if the computed value of U is greater than $\chi_\alpha^2(\nu)$. It should be noted that the modified portmanteau statistic is recommended in preference to the original portmanteau test of Box and Pierce (1970), because the former approximates the asymptotic distribution much better for moderate values of n.

In recent years, the Akaike information criterion (AIC), proposed by Akaike (1974), has been extensively used. This criterion incorporates the parsimony criterion suggested by Box and Jenkins (i.e. to use a model with as few parameters as possible) by penalizing the inclusion of a large number of parameters. The simplified and most commonly used form of the AIC is as follows:

$$\text{AIC} = n \log \hat{\sigma}_a^2 + 2m \quad (7)$$

where log denotes the natural logarithm as commonly used in mathematics nowaday and m is the number of parameters involved in the model. When there are several competing models to choose from, select the one that gives the minimum value of the AIC. Even if it is commonly used, when viewed as an estimator of the model order, the AIC was found to be inconsistent (Kashyap, 1980) in the sense that it does not pick up the true model. Another selection criterion, known as the posterior possibility criterion (PPC), was developed independently by Kashyap (1977) and Schwarz (1978). The PPC can be expressed as follows:

$$\text{PPC} = n \log \hat{\sigma}_a^2 + m \log n \quad (8)$$

According to the this criterion, one selects the model which minimizes PPC. It is known that the PPC gives a consistent rule for selecting the true model. With the use of the AIC or PPC, one can avoid the requirement of choosing the significance level in hypothesis testing. Cline (1981) suggested that m in eqs. 7 and 8 be computed as follows:

$$m = p + q + P + Q + pp + qQ + 1 \quad (9)$$

In this study, the PPC is used. However, the AIC and modified portmanteau statistic are also computed in order to support the selection made by the PPC.

Forecasting

In this study, Box-Jenkins models are used to forecast monthly streamflows. Among the three alternative forms provided by Box and Jenkins (1976), the model expressed as a difference equation and as an infinite sum of the white noise is useful in computing the forecasts and forecast errors, respectively.

Let h denote the forecast lead time. The minimum mean squared error (MMSE) predictor $\hat{x}_t(h)$ of w_{t+h} made at t is obtained by taking the conditional expectation of w_{t+h} at time t. Using the square brackets to signify conditional expectation, one can write

$$\begin{array}{lll} [w_{t-j}] = w_{t-j} & j = 0,1,2,\ldots \\ [w_{t+j}] = \hat{w}_t(j) & j = 1,2,\ldots \\ [w_{t-j}] = w_{t-j} - \hat{z}_{t-j-1}(1) & j = 0,1,2,\ldots \\ [a_{t+j}] = 0 & j = 1,2,\ldots \end{array} \quad (10)$$

In other words, the w_{t-j} ($j = 0, 1, 2, \ldots$) which have already happened at instant t are left unchanged; the w_{t+j} ($j = 1, 2, \ldots$) which have not happened, are replaced by their forecasts $\hat{w}_t(j)$; the a_{t-j} ($j = 0, 1, \ldots$) which have happened, are computed as the forecast error $w_{t-j} - \hat{w}_{t-j-1}(1)$, while the a_{t+j} ($j = 1, 2, \ldots$)

which have not yet happened, are replaced by zeroes. With this rule and using the difference equation form, the forecasts can easily be computed.

Now, writing the ARIMA model in an infinite order moving average form:

$$w_{t+h} = (a_{t+h} + \psi_1 a_{t+h-1} + \ldots + \psi_{h-1} a_{t+1}) + (\psi_h a_t + \psi_{h+1} a_{t-1} + \ldots) \quad (11)$$

where the weights ψ can be computed by rewriting eq. 2 and by equating the coefficients of a_{t+h}:

$$w_{t+h} = \{\phi(B)\Phi(B^s)\}^{-1}\theta(B)\Theta(B^s)a_{t+h} \quad (12)$$

The forecast error at lead time h is then given by

$$\begin{aligned} e_t(h) &= w_{t+h} - \hat{w}_t(h) \\ &= a_{t+h} + \psi_1 a_{t+h-1} + \ldots + \psi_{h-1} a_{t+1} \end{aligned} \quad (13)$$

so that the MMSE is

$$V(h) = \sigma_a^2 (1 + \sum_{i=1}^{h-1} \psi_i^2) \quad (14)$$

In practice, $\hat{\sigma}_a^2$ is replaced by $\hat{\sigma}_a^2$ and the ψ_i are computed from the estimates of the model parameters. The approximate $1-\varepsilon$ probability limits $w_{t+h}(-)$ and $w_{t+h}(+)$ for w_{t+h} will be given by

$$w_{t+h}(\pm) = \hat{w}_t(h) \pm z_{\varepsilon/2} [1 + \sum_{i=1}^{h-1} \psi_i^2]^{1/2} \hat{\sigma}_a \quad (15)$$

where $z_{\varepsilon/2}$ is the normal deviate exceeded by a portion $\varepsilon/2$. In this study, ε is taken equal to 5%, thus the 95% confidence limits are obtained by taking $z_{\varepsilon/2} = 1{,}96$.

Standardization

Phien and Balasuriya (1982) recommended the use of standardization in order to render seasonal data nonseasonal. Let $\bar{F}_j$ and S_j denote respectively the mean and standard deviation of flow in month j (j = 1, ..., 12):

$$\begin{aligned} \bar{F}_j &= T^{-1} \sum_{i=1}^{T} F_{i,j} \\ S_j &= [(T-1)^{-1} \sum_{i=1}^{T} (F_{i,j} - \bar{F}_j)^2]^{1/2} \end{aligned} \quad (16)$$

with T being the length of record employed in years, then in many cases,

$$x_t = (F_{i,j} - \bar{F}_j)/S_j \ , \ t = 12(i-1)+j \quad (17)$$

may be conveniently represented by nonseasonal models. Doing so introduces 24 parameters ($\bar{F}_1, \ldots, \bar{F}_{12}$; $S_1, \ldots, S_{12}$) to the parameters of the ARIMA model, but the resulting series $\{x_t\}$ is much easier to model. The forecasts of $F_{i+1,j}$, made at $F_{i,12}$, for example, can be obtained as follows:

$$\hat{F}_{i+1,j} = \hat{F}_{1,12}(j) = \bar{F}_j + S_j \hat{x}_t(j) \quad (18)$$

where $t = 12(i-1) + 12 = 12i$.
Likewise, the $1-\varepsilon$ probability limits for $F_{i+1,j}$ are obtained by replacing $\hat{x}_t(j)$ in eq. 18 by $x_{t+j}(-)$ and $x_{t+j}(+)$, which are computed according to eq. 15.

Data for the study

The Mekong River

The Mekong is an international river that starts at an elevation of about 5 000 m in the snow-covered mountain ranges of Tang Ku La on the great Tibet Plateau in Southwestern China. It flows generally southward for about 1 600 km between the mountain ranges of Yunnan province (China) and enters the peninsula at the common border of China, Burma and Laos. From this border, the river continues to flow to the south for an additional length of 2 400 km and discharges into the South China Sea. The lower section of the river forms part of the boundaries between Burma and Laos, between Thailand and Laos, flows across Kampuchea and the South of Vietnam. The Lower Mekong Basin is shown in Fig. 1 with the streamflow stations where data are employed. These stations are listed in Table 1 and the data were taken from the Mekong Secretariat whose main responsibilities are to develop the water resources of the Lower Mekong Basin. For each station, the last year of record is used for comparison between actual and forecast values, while all the data prior to that year are used in model building.

Preliminary data analysis

In this stage, the mean and standard deviation of the flow in each month are computed. They are collected in Table 2 for all the five stations employed. From these computed values, the seasonalities of monthly flows are obvious. In order to check the normality of monthly flows, the skewness (Cs) and excess (Ex) coefficients are computed. If, for a month,

$$\begin{aligned} |Cs| &< 1{,}96\,(6/T)^{1/2} \\ |Ex| &< 1{,}96\,(24/T)^{1/2} \end{aligned} \quad (19)$$

where T is the length of record (in years), then the corresponding monthly flow can be said to be normally distributed at the 5% significance level. Inspection of the computed values of Cs and Ex reveals that the streamflows in some months are not normally distributed. However, according to the Box-Jenkins method, monthly flows are not to be cut according to each month; instead, they are arranged as shown in eq. 1 to form the time series of x_t, which is found to be approximately normally distributed. So no transformation is needed for the data sets employed.

Data analysis

Use of raw data

First of all, monthly data are subject to Box-Jenkins modelling without standardization. For illustrative purposes, station 4 (Pakse) is used.

Identification

As mentioned before, the ACF and PACF would be sufficient at this stage. They are recommended and used because of their simple computation. The ACF (autocorrelation function) of monthly streamflows at station 4 for the priod 1945-1979 in shown in Fig. 2, from which the 12-month periodicity is clearly seen. Since the estimated autocorrelations at lags that are integer multiples of seasonal length s = 12 do not die out rapidly, seasonal differencing is needed to produce stationarity. Failure of other autocorrela-

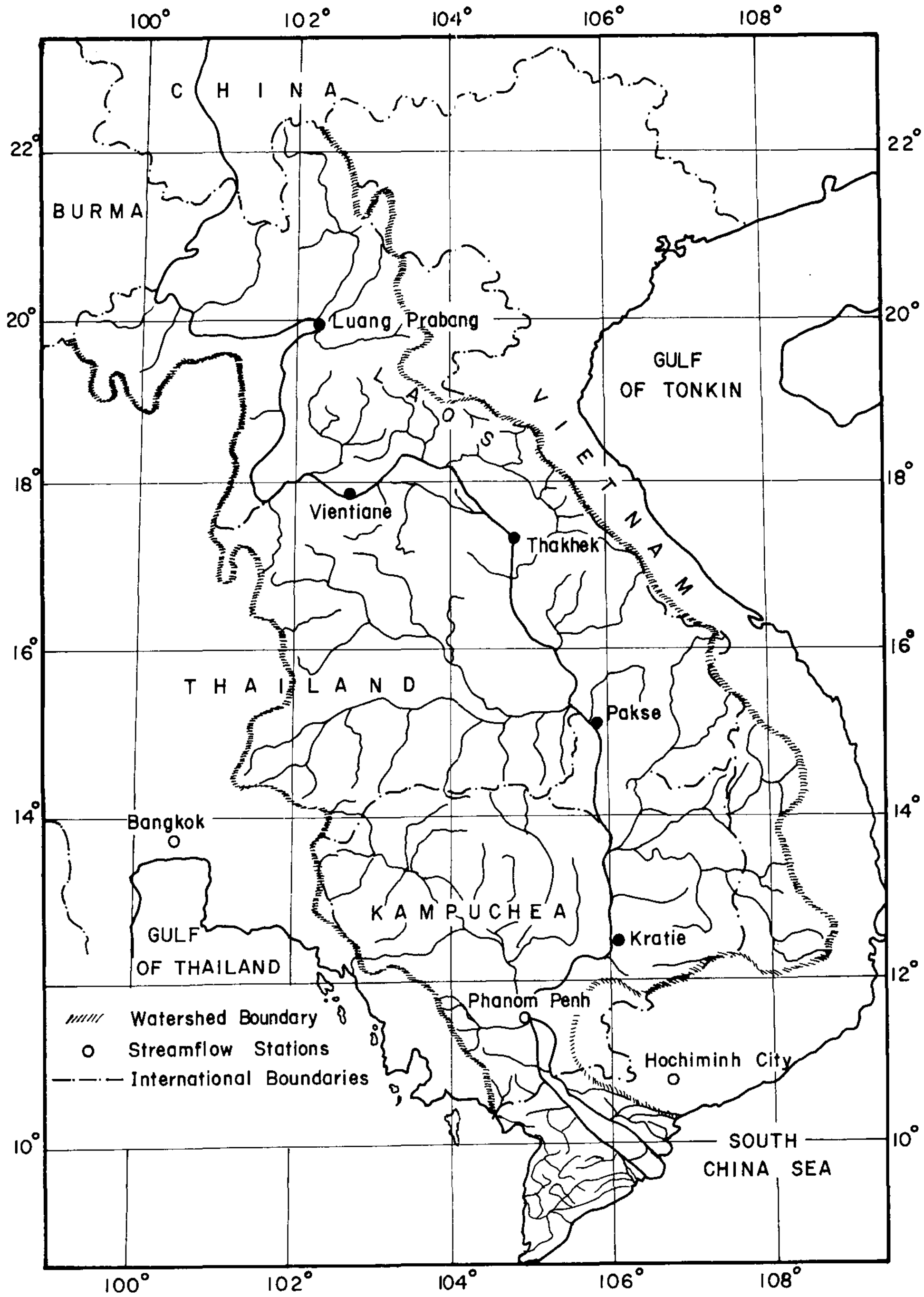

Figure 1
The Lower Mekong Basin with streamflow stations where data are employed.

TABLE 1
LIST OF STREAMFLOW STATIONS

Station	Location	Drainage Area (km²)	Period of Record Used
1	Luang Prabang	268 000	1950 – 1980
2	Vientiane	299 000	1946 – 1980
3	Thakhek	373 000	1946 – 1980
4	Pakse	545 000	1945 – 1980
5	Kratie	646 000	1945 – 1968(*)

(*) Data at this station were not available after 1968.

TABLE 2
MEAN AND STANDARD DEVIATIONS OF MONTHLY STREAMFLOWS (m^3/s) AT FIVE STATIONS UNDER CONSIDERATION

Station		Month 1	2	3	4	5	6	7	8	9	10	11	12
(1) Luang Prabang	(i)	1 624	1 258	1 038	1 063	1 436	3 010	6 151	10 370	9 074	5 329	3 529	2 272
	(ii)	308	202	158	140	408	882	1 645	2 387	2 177	1 233	1 065	608
(2) Vientiane	(i)	1 755	1 391	1 164	1 169	1 652	3 431	6 973	11 902	11 631	6 831	4 095	1 509
	(ii)	312	202	267	285	427	968	1 946	2 584	2 774	1 514	1 157	594
(3) Thakhek	(i)	2 319	1 812	1 475	1 422	2 300	6 856	13 130	20 032	20 349	11 114	5 694	3 350
	(ii)	343	244	193	214	639	2 137	3 259	3 825	4 123	2 521	1 223	576
(4) Pakse	(i)	2 800	2 145	1 740	1 681	2 847	9 034	16 940	26 676	28 659	16 780	8 432	4 333
	(ii)	485	337	231	260	833	2 973	4 503	5 548	5 411	3 968	1 703	729
(5) Kratie	(i)	3 621	2 640	2 088	1 974	3 684	11 181	21 297	33 553	40 136	24 638	11 848	5 944
	(ii)	575	395	278	246	1 188	3 545	5 466	6 076	5 776	5 716	2 435	1 116

NOTES: MONTH 1 = January, . . . , MONTH 12 = December
(i) Mean and (ii) standard deviation

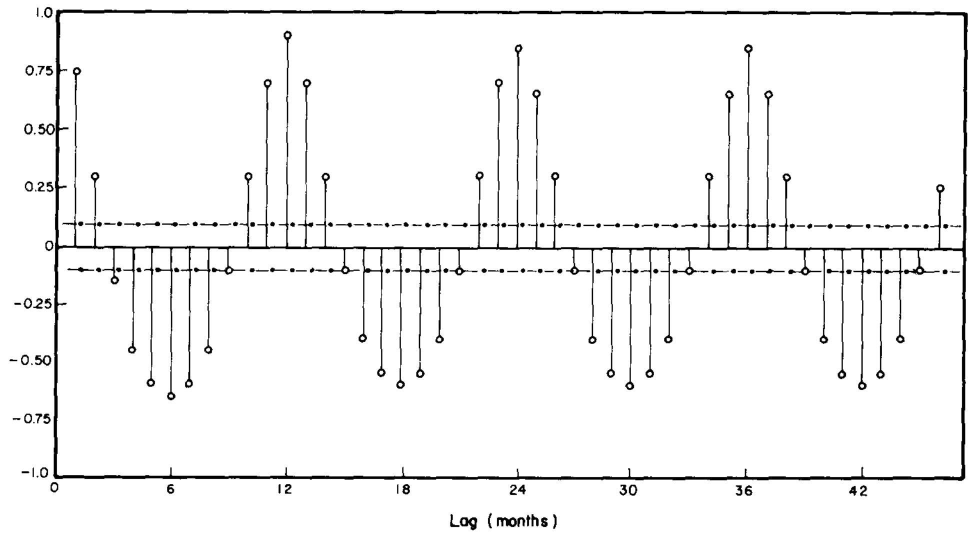

Figure 2
The autocorrelation function of monthly flows at Pakse

TABLE 3
STANDARD DEVIATION OF THE DIFFERENCED SEQUENCE

Order (d, D)	Standard Deviation (m^3/s)
(0,0)	9 818,2
(1,0)	6 824,4
(2,0)	7 360,6
(0,1)	4 193,2
(1,1)	4 861,1
(2,1)	7 918,5
(0,2)	7 315,9
(1,2)	8 746,3
(2,2)	14 422,0

tions to dying out may imply that nonseasonal differencing is also required. Table 3 shows the standard deviation of the process after various differencing orders. It is seen that $d = 0$ and $D = 1$ are desirable to yield the smallest value of the standard deviation. However, $d = 1$ and $D = 1$ are also suspected because the standard deviation of the resulting sequence is also quite close to the previous case. After differencing, the ACF and PACF of the series for $d = 0$ and $D = 1$ are shown in Figs. 3 and 4, respectively. From these figures and those with $d = D = 1$, the following models are tentatively selected:

(I) $(1,0,0) \times (1,1,1)_{12}$ (II) $(1,0,1) \times (1,1,1)_{12}$
(III) $(1,1,1) \times (1,1,1)_{12}$

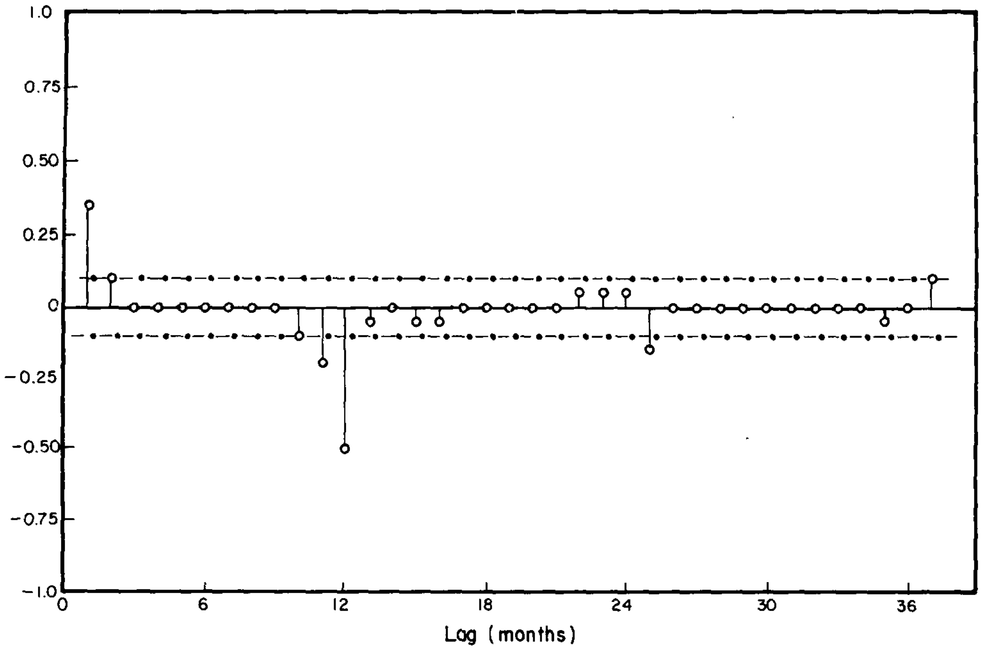

Figure 3
The autocorrelation function of differenced monthly flows (d = 0, D = 1) at Pakse

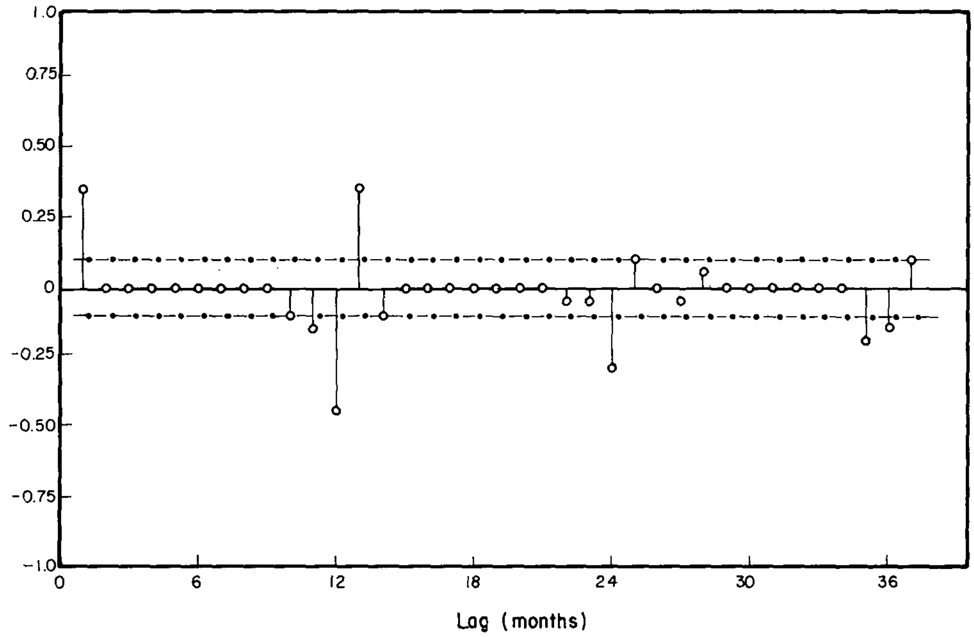

Figure 4
The partial autocorrelation function of differenced monthly flows (d = 0, D = 1) at Pakse

TABLE 4
PARAMETERS OF CANDIDATE MODELS FOR PAKSE

Model	ϕ	θ	Φ	Θ	σ_a^2
(I)	0,389 18	na	−0,165 62	0,856 25	0,622 17
(II)	0,341 49	−0,058 01	−0,168 75	0,855 25	0,628 19
(III)	0,404 66	0,839 06	−0,175 00	−0,186 10	0,625 68

na: not applicable

Estimation

At this stage, the parameters of all the possible models are estimated. A simple scheme for parameter estimation may be carried out using the Statistical Package of Social Science (SPSS) developed by Hull and Nie (1981). The values of the parameters so computed are shown in Table 4. Since the value of the nonseasonal moving average parameter θ is quite small, model II is almost the same as model I.

Diagnostic checking

Having computed the model parameters, the statistics U, AIC and PPC can readily be computed. Their values are collected in Table 5. Again model II gives nearly the same values for these statistics even though model I is the most suitable candidate for station 4 (Pakse):

$$(1-0{,}38918B)(1+0{,}16562B^{12})(1-B^{12})x_t=(1-0{,}85625B^{12})a_t \quad (20)$$

with $\hat{\sigma}_a^2 = 0{,}62217$.

Remarks

- The mean of $(1-B^{12})x_t$ may be neglected, i.e. $\mu = 0$.
- All three models under consideration are adequate according to the modified portmanteau statistic. The parsimony criterion (Box and Jenkins, 1976) selects model I also because it involves the least number of parameters. It should be noted that this criterion is difficult to apply in many situations where the simplicity of the candidate models is not obviously seen. In such situations, the AIC or PPC is more suitable.

TABLE 5
VALUES OF THE STATISTICS USED FOR DIAGNOSTIC CHECK (PAKSE)

Model	U(ν)	AIC	PPC
I	0,121 (45)	8 876,2	8 896,5
II	0,141 (44)	8 877,9	8 902,3
III	0,177 (44)	8 884,6	8 913,1

(ν): degree of freedom of the U-statistic

Forecasting for 1980

The model of eq. 20 is now used to produce forecasts for monthly streamflows in the year 1980, all made in December, 1979. The forecasts are shown together with the observed flows in Fig. 5 in which the lower 95% probability limits are plotted only when they are positive.

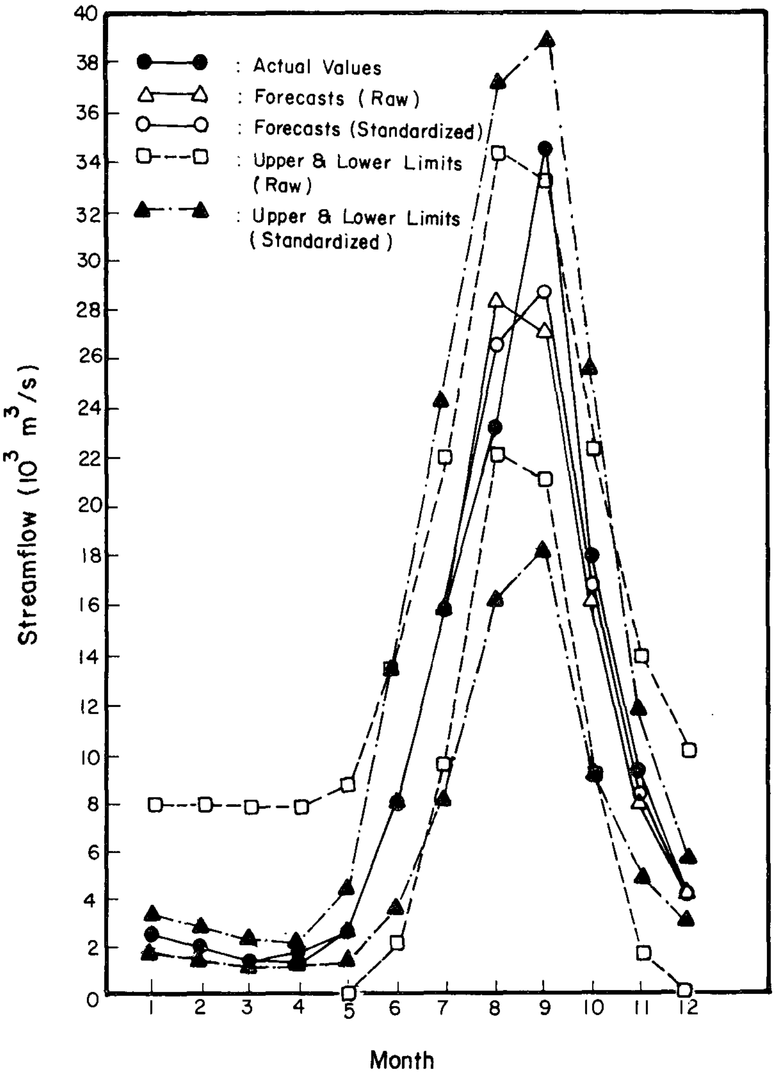

Figure 5
Comparison between forecast and actual monthly flows at Pakse

Remarks

- By the same procedure, the following $(1,0,0) \times (0,1,1)_{12}$ is reached for monthly streamflows at Vientiane (station 2):

$$(1-0{,}53416B)(1-B^{12})x_t = (1-0{,}95000B^{12})a_t \qquad (21)$$

with $\hat{\sigma}_a^2 = 0{,}50430$

- From eq. 14 and the fact that ψ_i becomes zero after a few values of i, the standard deviation of the forecast error becomes constant after some lead times.

- The forecasts can easily be updated. However, updating is not undertaken in this study.

Use of standardized data

For seasonal data, it is quite difficult to arrive at a suitable model. Several other models have also been attempted for Pakse and Vientiane and only those which are good are reported in the previous section. Since the monthly flows are approximately normally distributed, the mean and standard deviation are sufficient to determine the marginal (normal) distribution in each month. Consequently, more emphasis is placed upon these two statistics. Romoval of seasonality is therefore achieved by standardization (eq. 17) and the resulting series x_t is to be represented by nonseasonal models, ARIMA (p,d,q).

For each station, several candidate models are tested. Table 6 shows the computed values of the statistics U, AIC and PPC for the models, which are most likely to be selected for each of the five stations. From this table, the following models are selected:

TABLE 6
COMPUTED VALUES OF U, AIC AND PPC FOR STANDARDIZED DATA

Model	U(ν)	AIC	PPC
	1. Luang Prabang		
(1,1,1)	0,043 58 (13)	−252,27	−240,61
(2,1,1)	0,042 53 (12)	−249,52	−233,98
(1,1,2)	0,046 43 (12)	−250,45	−234,91
	2. Vientiane		
(1,0,1)	0,111 23 (13)	−281,37	−269,16
(2,0,1)	0,231 56 (12)	−286,52	−235,97
(1,0,2)	0,259 68 (12)	−286,23	−234,91
	3. Thakhek		
(1,1,1)	0,046 17 (13)	−172,51	−160,39
(3,1,1)	0,036 92 (11)	−167,99	−147,79
	4. Pakse		
(1,1,1)	0,035 66 (13)	−204,81	−188,03
(1,1,3)	0,028 61 (11)	−204,53	−184,19
	5. Kratie		
(1,0,0)	0,033 46 (14)	−154,90	−147,66
(1,0,1)	0,100 01 (13)	−152,90	−142,04
(2,0,0)	0,099 19 (13)	−151,94	−141,08

(ν): degree of freedom of the U-statistic

1. Luang Prabang: ARIMA (1,1,1):
$$(1-0{,}50712B)(1-B)\,x_t = 1-0{,}91892B)\,a_t\ , \qquad \hat{\sigma}_a^2 = 0{,}51184 \qquad (22)$$

2. Vientiane: ARIMA (1,0,1):
$$(1+0{,}45994B)\,x_t = (1-0{,}88661B)\,a_t\ , \qquad \hat{\sigma}_a^2 = 0{,}50071 \qquad (23)$$

3. Thakhek: ARIMA (1,1,1):
$$(1-0{,}47134B)(1-B)\,x_t = (1-0{,}95408B)\,a_t\ , \qquad \hat{\sigma}_a^2 = 0{,}63995 \qquad (24)$$

4. Pakse: ARIMA (1,1,1):
$$(1-0{,}37964B)(1-B)\,x_t = (1-0{,}81210B)\,a_t\ , \qquad \hat{\sigma}_a^2 = 0{,}62077 \qquad (25)$$

5. Kratie: ARIMA (1,0,0):
$$(1-0{,}63207B)\,x_t = a_t\ , \qquad \hat{\sigma}_a^2 = 0{,}54575 \qquad (26)$$

Remarks

- For data at Vientiane, the AIC and PPC do not give the same selection. According to the AIC, MODEL (2,0,1) should be chosen while the PPC suggests model (1,0,1). As mentioned before, the AIC does not give consistent order (Kashyap, 1980), the model shown in eq. 23 is selected according to the PPC statistic.

- When d = 0, nonseasonal ARIMA models reduce to autoregressive moving average models which are commonly denoted by ARMA (p,q). For example, eq. 23 for standardized data at Vientiane represents an ARMA (1,1). When q = 0, ARMA models reduces to autoregressive models, denoted by AR (p), and when p = 0, they reduces to moving average models, denoted by MA (q). For example, standardized monthly flows at Kratie can be presented by an AR (1) model.

- The modified portmanteau statistic indicates that all the listed models (in Table 6) are acceptable. It does not help in the selection of the best model.

Comparison

In order to evaluate the performance of the two different approaches which are based respectively on raw and standardized data, the results obtained for Pakse and Vientiane are compared, using the mean squared error (MSE) defined as

$$\text{MSE} = \frac{1}{12}\sum_{j=1}^{12}(F_{i+1,j} - \hat{F}_{i+1,j})^2 \qquad (27)$$

where i + 1 denotes the year 1980, and $\hat{F}_{i+1,j}$ denotes the forecast of the streamflow $F_{i+1,j}$ in month j, made at the end of 1979, i.e.

$$\hat{F}_{i+1,j} = \hat{F}_{i,12}(j)\ , \quad j = 1, \ldots, 12\ .$$

However, the results shown in Table 7 are inconclusive. The MSE statistic indicates that standardization is better for Pakse while it is worse for Vientiane. Besides the MSE, one can use some other useful criteria. In this study, the mean absolute deviation (MAD), mean relative error (MRE) and maximum relative error (RE_{max}) are also employed. These are defined as follows:

TABLE 7
COMPARISON OF FORECASTS FOR THE YEAR 1980 AT PAKSE AND VIENTIANE

Station		Statistics			
		MSE	MAD	MRE	REmax (%)
Pakse	(i)	7 294 968	1 413,02	0,006	21,7
	(ii)	4 368 284	1 198,58	−0,004	17,4
Vientiane	(i)	892 067	759,75	0,002	51,2
	(ii)	2 126 586	104,75	0,120	33,6

Notes: (i) raw data
(ii) standardized data

$$MAD = \frac{1}{12}\sum_{j=1}^{12} | F_{i+1,j} - \hat{F}_{i+1,j} | \quad (28)$$

$$MRE = \frac{1}{12}\sum_{j=1}^{12} \frac{F_{i+1,j} - \hat{F}_{i+1,j}}{F_{i+1,j}} \quad (29)$$

$$RE_{max} = \max 100 \times | \frac{F_{i+1,j} - \hat{F}_{i+1,j}}{F_{i+1,j}} | (\%) \quad (30)$$

From their computed values collected in Table 7, the MAD and RE_{max} criteria consistently indicate that it would be better to standardize monthly streamflows before the Box-Jenkins models are attempted. Moreover, a comparison of the two approaches is also made by plotting forecast and actual flows on the same graph. Typical results are shown in Fig. 5 for Pakse. The performance of the two approaches is almost identical. However, the upper and lower probability limits provided by the case of standardized data seems to be more reasonable, because the seasonalities are clearly reflected. Moreover, it is observed that the actual flow in September 1980 was out of the probability limits computed with raw data. This, to some extent, indicates that they are not reliable. Finally, it is much simpler to work with standardized data: appropriate models are easier to identify, estimation of parameters requires less computational effort, and forecasts are easier to compute. Consequently, standardization is recommended as a tool ro remove the seasonality of monthly streamflows of the Mekong River so that nonseasonal Box-Jenkins models can be entertained.

Summary and Conclusions

In the present study, Box-Jenkins models were used to forecast monthly streamflows of the Mekong River. The basic procedure consisting of identification, estimation and diagnostic checking were briefly described. Instead of relying on the modified portmanteau statistic, which is very less sensitive and is not useful for model selection, the final models are selected according to the posterior possibility criterion (PPC) whose selection is more reliable than the Akaike information criterion (AIC). In fact, both raw and standardized data were attempted. In the first case, seasonal models were fitted and forecasting was made with the resulting model; in the second situation, the data were treated to be nonseasonal and ARIMA (p,d,q) models were employed. Although the performance of models obtained with raw and standardized data is almost identical, standardization is recommended because it leads to simpler identification, easier parameter estimation and simpler computation of forecasts. It was observed that the forecasts for the months with peak discharges are not very good. This indicates that other methods should be tried. At present an attempt is being made to evaluate the various possible techniques. Related results will soon be reported.

Acknowledgements

The authors wish to express their sincere thanks to the Mekong Secretariat for the provision of the data employed in this study, and to the Asian Institute of Technology for its general support and for use of its research facilities.

References

AKAIKE, H. (1974) A new look at statistical model identification. *IEEE Trans. Auto. Contr.* AC-19 716-722.

ANSLEY, C.F. (1979) An algorithm for the exact likelihood of a mixed autoregressive – moving average process. *Biometrika* **66**(1) 59-65.

BOX, G.E.P. and JENKINS, G.M. (1976) *Time Series Analysis: Forecasting and Control.* Revised Edition, Holden-Day, San Francisco.

BOX, G.E.P. and PIERCE, D.A. (1970) Distribution of residual autocorrelations in autoregressive integrated moving average time series models. *J. Am. Statist. Assoc.* **65** 1509-1526.

CLINE, T.B. (1981) Selecting seasonal streamflow models. *Water Resources Research* **17**(4) 975-984.

GRAY, H.L., KELLEY, G.D. and McINTIRE, D.D. (1978) A new approach to ARMA modelling. *Communications in Statistics* **B7** 1-77.

HIPEL, K.W., McLEOD, A.I. and LENNOX, W.C. (1977) Advances in Box-Jenkins Modelling, I: Model Construction. *Water Resources Research* **13**(3) 567-575.

HULL, C.H. and NIE, N.H. (1981) *SPSS update 7-9: New Procedures and Facilities for Releases 7-9.* McGraw-Hill, New York.

KASHYAP, R.L. (1977) A bayesian comparison of different classes of dynamic models using empirical data, *IEEE Trans. Autom. Contr.* AC-**22**(5) 715-727.

KASHYAP, R.L. (1980) Inconsistency of the AIC rule for estimating the order of autoregressive models. Technical report, Dep. of Electr. Eng., Purdue Univ., Lafayette.

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

McKERCHAR, A.I. and DELLEUR, J.W. (1974) Application of seasonal parametric linear stochastic models to monthly flow data. *Water Resources Research* **10**(2) 246-255.

McLEOD, A.I., HIPEL, K.W. and LENNOX, W.C. (1977) Advances in Box-Jenkins modelling. 2: Applications. *Water Resources Research* **13**(3) 577-586.

NEWBOLD, P. (1974) The exact likelihood function for a mixed autoregressive-moving average process. *Biometrika* **61** 423-426.

PHIEN, H.N. (1983) Introduction to seasonal streamflow forecasting techniques. Paper presented at the seminar on Seasonal Streamflow Forecasting, held in Vientiane, Laos, by Mekong Committee – World Meteorological Organization. (October 20-27, 1983).

PHIEN, H.N. and BALASURIYA, B.L.N. (1982) Some useful techniques for modelling seasonal data. *Proc. Int. AMSE Conference* Paris-Sud, **11** 52-56.

SCHWARZ, G. (1978) Estimating the dimension of a model. *Ann. Statist.* **6**(2) 461-468.

TAO, P.C. and DELLEUR, J.W. (1976) Seasonal and nonseasonal ARMA models in hydrology. *J. Hydraul. Div, ASCE* **102**(HY10) 1541-1559.

WOODWARD, W.A. and GRAY, H.L. (1981) On the relationship between the S-array and the Box-Jenkins method of ARMA model identification. *J. Am. Statist. Assoc.* **76**(375) 579-587.