Arab Gulf J. scient. Res. **4** (2), pp. 761-774 (1986)

The Generation of Missing River Flow Data

Adnan A. Al-Samawi

Civil Engg. Department,
University of Basrah, Basrah, Iraq.

ABSTRACT. A time series model is presented for use in generating missing river flow data. A 14-year record of average monthly flows of the river Khabur at Zakho (Iraq) is used as a case-study. The time series model of monthly river flows consists of a deterministic component and a stochastic component. The deterministic component is represented by a periodic model, while the stochastic component is fitted to a third-order-autoregressive AR(3) model using the Box-Jenkins approach of time series analysis. The model explains more than 87% of the total variance of the original series. The model is then used in generating a sequence of missing data.

Engineers and hydrologists who take up the task of analysing flows in Iraqi rivers for the purposes of design and planning are often confronted with the problem of working with records having sequences of missing data. One of such rivers is the river Khabur which originates in Turkey and flows into Iraqi territory for a total distance of 160 km before it finally discharges into the river Tigris. The infilling of missing values in hydrological data involves the use of statistical procedures of data generation. In the United States, Beard (1965) pioneered the practical use of several data generation methods. One of such methods is the use of univariate linear stochastic models. The objective of this paper is to investigate the applicability of such stochastic models to Iraqi rivers with the river Khabur at Zakho as a case-study.

Theory

The following theoretical derivations are mainly based on the work of Maidment and Parzen (1984). The time series model may be summarized as follows: the mean monthly river flow F(t) is expressed as the sum of a long-memory or deterministic component $F_l(t)$ and a short-memory or stochastic component $F_s(t)$:

$$F(t) = F_l(t) + F_s(t) \qquad t = 1,2,\ldots.T \tag{1}$$

The concept of long-memory and short-memory time series models do not have standard definitions. But short-memory time series is defined as being stationary with an invertible infinite autoregressive scheme representation. A long-memory time series is one with more or less regular periodic variations, whose pattern is approximately preserved year after year.

In this paper, the original time series of monthly river flow $F_a(t)$ is written for clarification of expression at this point as $F_a(m,y)$ to denote the river flow in month m of year y. Long-memory components of time series are modelled by deterministic models employing the Fourier-fitted means method.

Deseasonalizing

To find the Fourier-fitted means, the arithmetic monthly means of river flows,

$$\bar{F}_a(m) = \frac{1}{\gamma} \sum_{y=1}^{\gamma} F_a(m,y) \tag{2}$$

where m=1,2, ... 12 and m=1 corresponds to January, 2 to February;...

y = 1,2, ... γ and

γ = Total years considered

are represented as Fourier series:

$$\bar{F}_a(m) = \bar{F}_a + \sum_{k=1}^{6} \left[a_k \, cos \, \frac{2\pi k}{12} \, m + b_k \, sin \, \frac{2\pi k}{12} \, m \right] \tag{3}$$

where $\bar{F}_a$ = grand mean of the series:

$$\text{and } k = \begin{cases} 1,2, \ldots. \dfrac{m}{2} & \text{for even values of m} \\ 1,2, \ldots. \dfrac{(m-1)}{2} & \text{for odd values of m} \end{cases}$$

The coefficients a_k and b_k are computed by using the discrete Fourier transform of $\bar{F}_a(m)$ (Kottegoda 1980). The cycle corresponding to k=1 has a 12-month period; the cycles for k=2, 3....6 are harmonics of periods $\frac{12}{k}$ months; they allow seasonal cycles of river flow, which is periodic but not directly sinusoidal to be modelled. An analysis of variance is then carried out to identify the significant and insignificant cycles.

Those with insignificant amplitudes have their coefficients (ak, b_k) set to zero. The significant cycles are used in equation (3) to form the Fourier-fitted means $\bar{F}_a(m)$, and the deseasonalized series $F_b(m,y)$ is found as:

$$F_b(m,y) = F_a(m,y) - \bar{F}_a(m) \qquad (4)$$

The series $F_b(m,y)$ may be checked for stationarity in the variance. Stationarity in the covariance (without periodic autocorrelation) is also assumed. The series $F_b(m,y)$ is, therefore, taken to be a second-order stationary time series; it is now more appropriate to use the time t in months from the beginning of the series as the time index, noting that $t=12\,(y-1) + m$. Therefore, the series $F_b(m,y)$ is written as $F_b(t)$.

Autoregression

The time series $F_b(t)$ is a short-memory process. The stochastic model which represents $F_b(t)$ accounts for the autocorrelation of $F_b(t)$ by fitting the autoregression model

$$F_b(t) = \sum_{j=1}^{p} \phi_j\, F_b(t-j) + F_c(t) \qquad (5)$$

$$t=p+1,\ p+2,\ \ldots.\ T$$

Where ϕ_j for ($j=1,2,\ \ldots.,\ p$) are the model coefficients and $F_c(t)$ is the residual of regression and is the final residual error of the model. The coefficients ϕ_j are found by using the Yule-Walker equations (Box and Jenkins 1976). The optimal model order p is determined by autocorrelation-function and partial autocorrelation function analysis.

Residual Analysis

The residual series $F_c(t)$ is subjected to further analysis including autocorrelation coefficients and Chi-squared goodness-of-fit tests to check for its independence and normality. If the tentative model is to be adopted, then the residual series $F_c(t)$ must be independent and identically distributed in time (Box and Jenkins 1976). However normality of the residual series $F_c(t)$ is needed to define the confidence limits of the forecasted values.

Application Procedure

Data and Procedure

The data used in this study were the average monthly flows of the river Khabur at Zakho (Directorate General of Irrigation 1976).

A complete set of average monthly flows for the 15 year period 1959-1973 were available for the analysis. The first 14 years of average monthly flows were used to build the time series model. The model then was used in an ex post forecast to generate the monthly flows of the 15th year.

These forecasted monthly flows were then compared with the observed flows using the average absolute relative error (AARE) and the percentage explained variance as tools to evaluate the adequacy of the model.

The mean and standard deviations of the monthly flow data were 68.14 m^3/sec and 67.48 m^3/sec respectively.

This resulted in a rather high coefficient of variation of 0.99, showing that the river was subjected to an enormous range of monthly flows.

Time Series Behaviour

The average monthly flows of the river Khabur (1959-1974) are shown in Figure 1. The series exhibits a regular seasonal cycle, with maximum monthly flow usually occuring in May and averaging 9.5 times the minimum flow, which usually occurs in September. The series is fairly homogeneous from 1959 and up to 1962. However, the average monthly flows for the remaining years show irregular and abnormal sequences. These irregularities in the record were mentioned in the source of the data (Directorate General of Irrigation 1976), and are due to the fact that the automatic water stage recorder was damaged and ceased to function after 1962. Subsequent readings were taken from an auxiliary staff gauge which was situated downstream. Two points in the data were identified as outliers or freak points: one corresponded to the flow in April 1963 and the other to that in December 1968. These points, which proved to be outliers in later analysis, were smoothed according to the statistical procedure given by Montgomery and Johnson (1976).

Transformation

Due to the presence of irregular fluctuation in the data, the series was divided to five equal parts each containing 36 consecutive observations. The range and the mean of each part were calculated and then plotted as shown in Figure 2. According to Box and Cox (1964), the shape of the graph indicated that a logarithmic transformation of the data was needed to stabilize the variance and to make multiplicative effects additive. This indicated that the different components of the natural hydrological process, F(t), acted in a multiplicative manner, *i.e.*,

$$f(t) = f_I(t). \qquad f_s(t) \tag{6}$$

However, the logarithmic transformation of the series turned it into the form of equation (1), *i.e.*:

$$F(t) = \ell n\ f(t), \quad F_1(t) = \ell n\ f_1(t), \text{ and}$$
$$F_s(t) = \ell n\ f_s(t).$$

The transformed series F(t) was found to be fairly homogeneous and hence amenable to subsequent analysis.

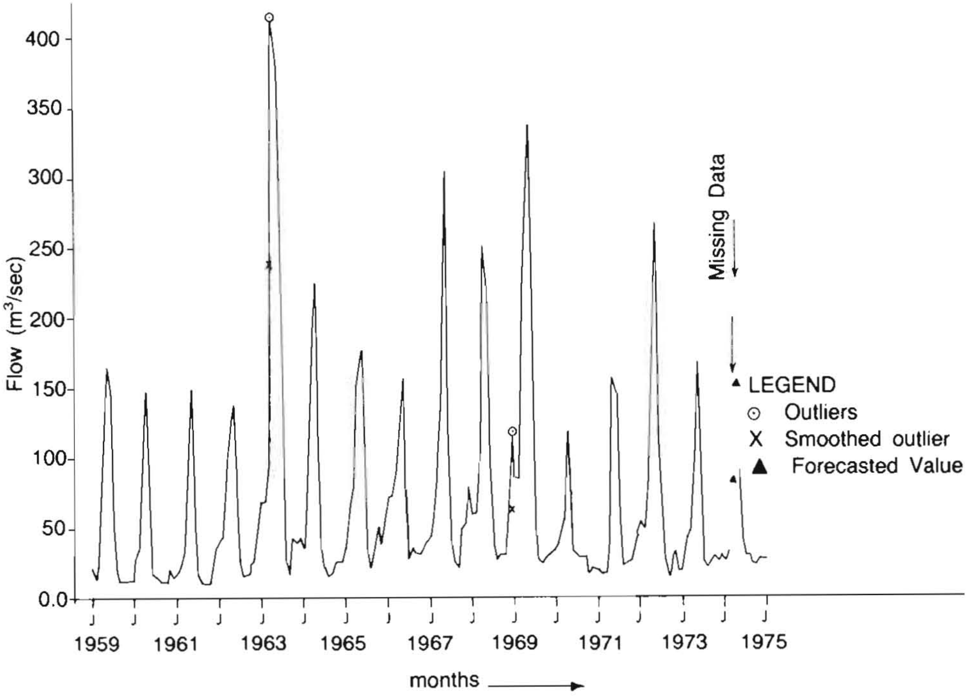

Fig. 1. Monthly Flows of the River Khabour at Zakho

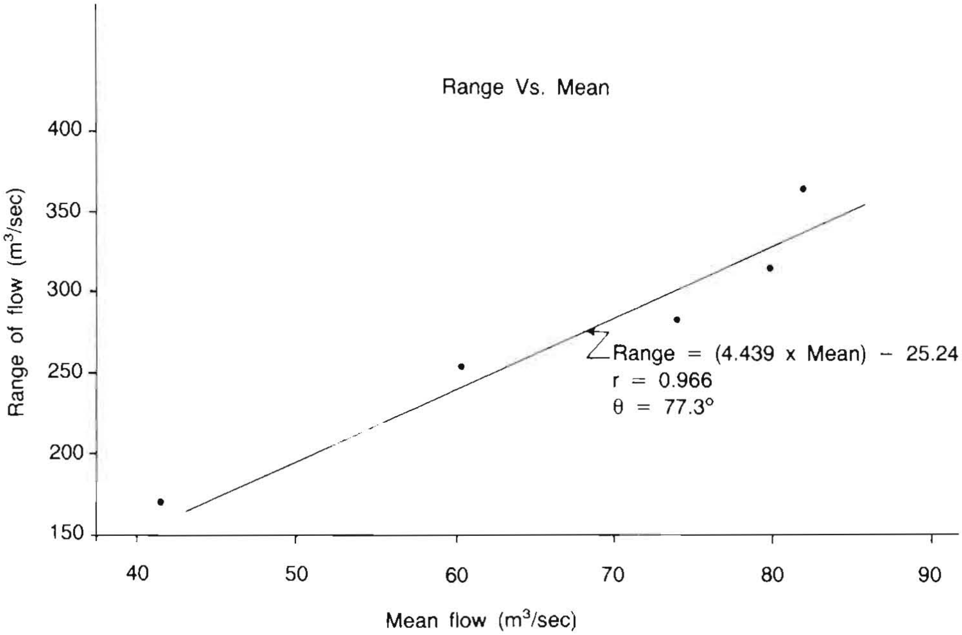

Fig. 2. Monthly Flow of Khabour River at Zakho

Seasonality

The autocorrelation function of the transformed series, $F_a(m,y)$, was computed and plotted in Figure 3. The seasonal nature of the series and the predominance of the fundamental harmonic were clear. The series $F_a(m,y)$, represented as a Fourier series by equation (3), was found to have a significant 12-month cycle (k=1) and a 6-month harmonic (k=2) by the analysis of variance techniques.

The estimated $\hat{F}_a(m)$ of monthly flow by this Fourier series was,

$$\hat{F}_a(m) = 3.84\text{-}0.491 \cos\left(\frac{\pi m}{6}\right) + 0.808 \sin\left(\frac{\pi m}{6}\right) + 0.159 \cos\left(\frac{\pi m}{3}\right) - 0.424 \sin\left(\frac{\pi m}{3}\right) \quad (7)$$

The de-seasonalized time series or short-memory series, $F_b(m,y)$, was produced by substituting $\hat{F}_a(m,y)$ from equation (7) into equation (4).

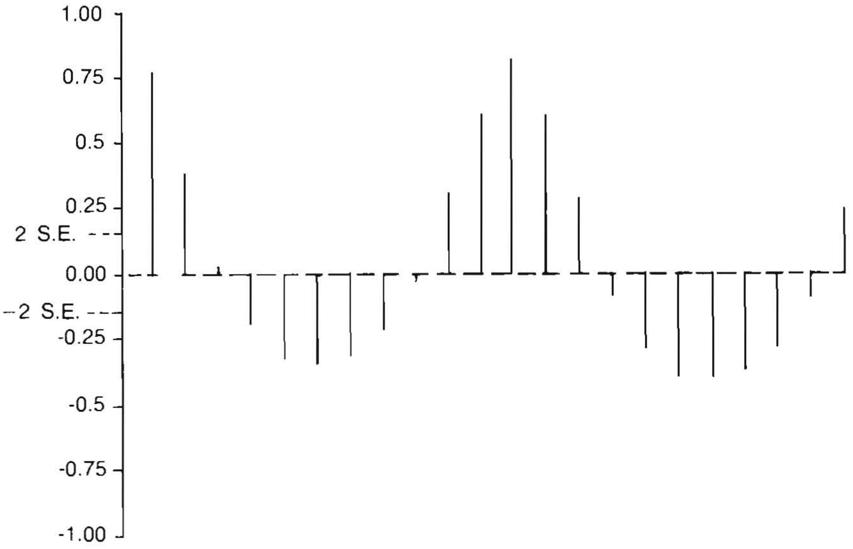

Fig. 3. ACF of Transformed Series $F_a(m,y)$

The short-memory series, $F_b(m,y)$, oscillated about a zero origin in an arbitrary manner without any seasonal variation.

Autoregression

The short-memory series, $F_b(m,y)$, was properly indexed by time alone as $F_b(t)$ $(t=1,2, \ldots.\ 168)$ for ease in subsequent analysis.

The autocorrelation function of the series $F_b(t)$ was computed and plotted in Figure 4. The autocorrelation coefficients decayed down well inside the limits of ± twice the standard error in about 10% of the number of points of the series (17 lags) and stayed there, indicating the stationarity of the series and the autoregressive nature of the structure of the series. The partial autocorrelation function of the series $F_b(t)$ was computed to identify the order of the autoregression (p) to be included in the stochastic model (see Figure 5). They truncated after lag 3, confirming the autoregressive nature of the series and giving the order of autoregression as three, *i.e.* AR(3).

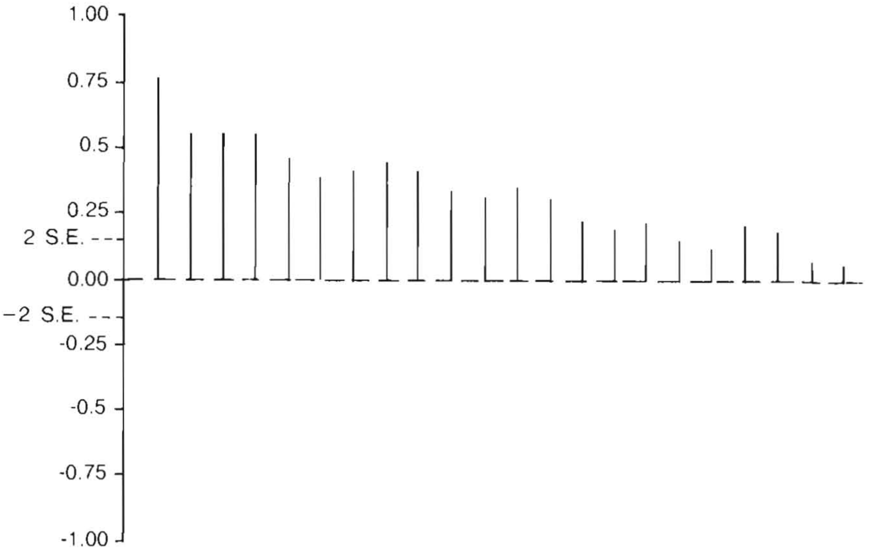

Fig. 4. ACF of Series $F_b(m,y)$

The stationarity constraints on the autoregressive coefficients, as set by Box and Jenkins (1976), were checked and found to be satisfactory.

The following model was constructed for the short-memory series:

$$\hat{F}_b(t) = 0.831 \quad F_b(t-1) - 0.334 \quad F_b(t-2) + 0.335 \quad F_b(t-3) \tag{8}$$

for t=4,5, ... 168, and where $\hat{F}_b(t)$ is the autoregressive model estimate of $F_b(t)$. The residual of this regression, $F_c(t) = F_b(t) - \hat{F}_b(t)$, had a zero mean and a variance of 0.079.

The autocorrelation function of the residual series, $F_c(t)$, was computed and plotted in Figure 6. All autocorrelation coefficients for lags > zero, fell well inside the ± 2 S.E. (standard errors of the estimated coefficients) limits except the coefficient at lag 17 which fell just outside the limits, although it was inside the 99% confidence limits (± 0.1999). The residual series, $F_c(t)$, was considered to be independent and identically distributed in time. The series $F_c(t)$ was further subjected to the test of normality.

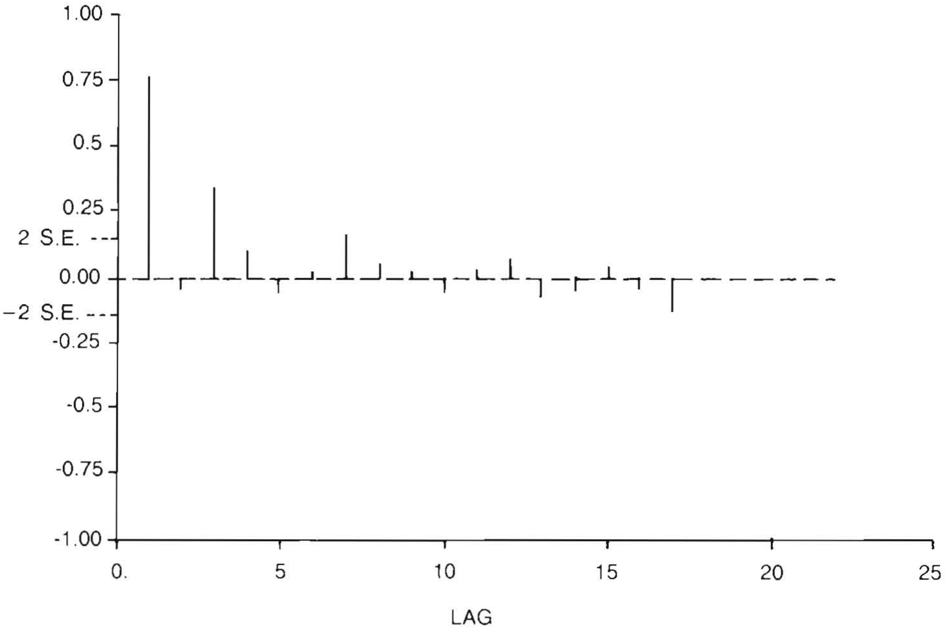

Fig. 5. PACF of Series $F_b(m.y)$

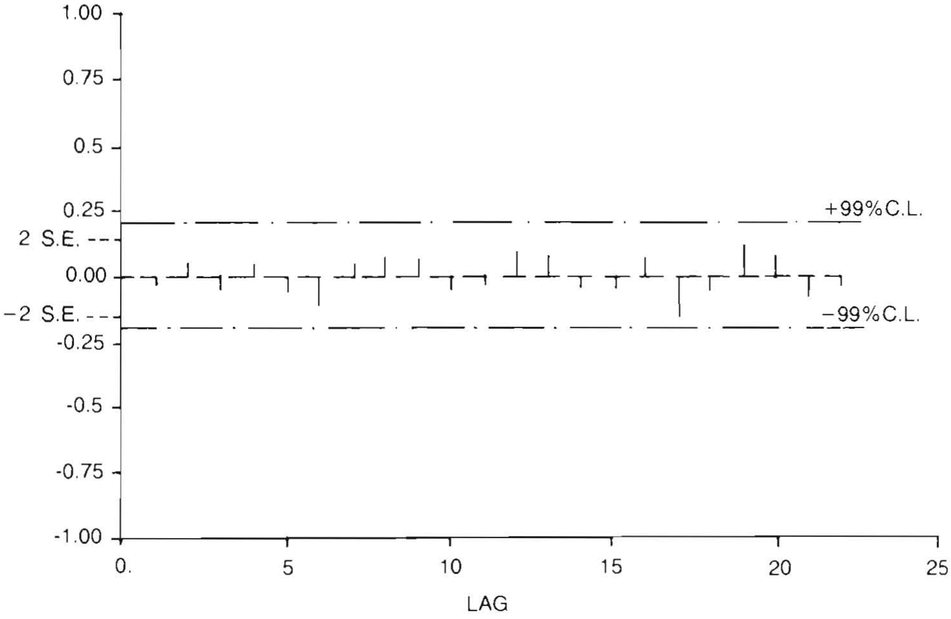

Fig. 6. ACF of Residual Series $F_c(m.y)$

A Chi-squared goodnes-of-fit test was carried out and the calculated chi-squared value, for 3 degrees of freedom, was found to be (χ^2=5.3). This compared favourably with the critical, $\chi^2_{0.05}(3)=7.8$, which meant the acceptance of the null hypothesis (assumption of normality distribution of the series $F_c(t)$ at the 5% level of significance). Hence the series $F_c(t)$ was considered to be a white noise series.

Ex-post Forecast

An ex-post forecst is critical if a proper evaluation of the time series model is to be made. The final adopted model consisted, as shown, of a deterministic component plus a stochastic component. The components were represented by equations (7) and (8) respectively.

The adopted model was used to forecast F(t) for t=169 to 180, *i.e.*, monthly flows for 1973. The average absolute relative error, (AARE), which is defined as,

$$\text{AARE} = \frac{1}{n}\,\Sigma\left(\left|\frac{A_t-P_t}{A_t}\right|\right) \tag{9}$$

where, A_t = actual monthly flows
P_t = forecasted monthly flows
n = number of forecasted flows

was used to evaluate the accuracy of the forecast. The percentage AARE for the (1973) year was found to be 6.5 for the logarithmically transformed series.

The actual and forecast monthly flows for 1973 are plotted in Figure 7 in both transformed and untransformed forms. The model explained 87.6% of the total variance of the raw data.

The adopted model was then used to generate a sequence of missing data. This sequence consisted of two points corresponding to March and April 1974. These values were found to be 86.4 and 155.5 m^3/sec respectively. The values are shown in Figure 1 of the original time series.

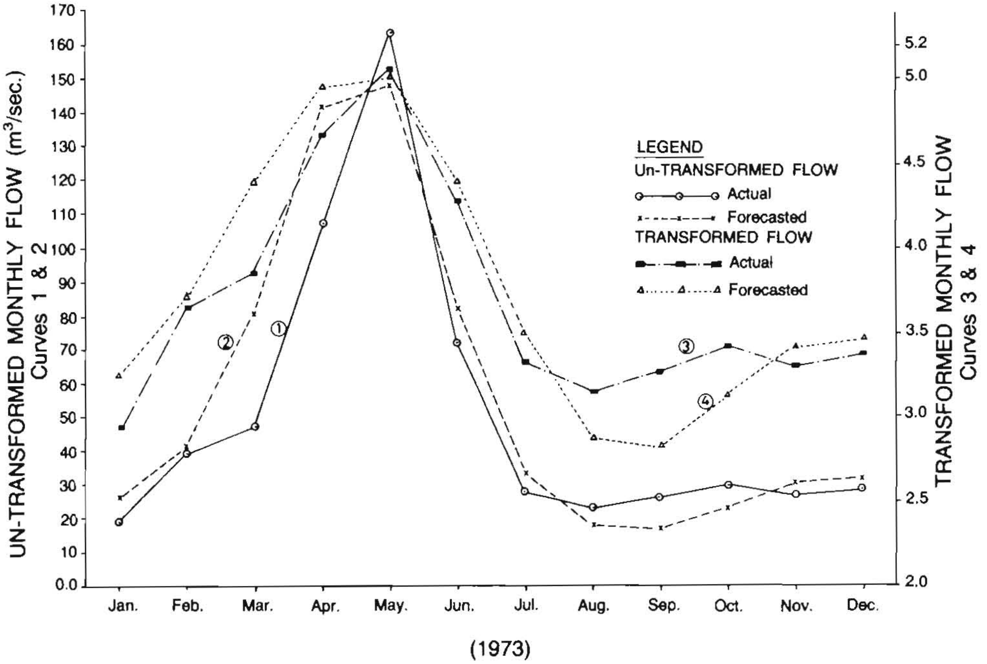

Fig. 7. One-Year-Ahead Forecast

Discussion

Several trials were attempted to model the monthly flow data without transformation, but every trial ended in failure of the tentative model during the diagnostic checking stage of model construction. All tentative models failed to render the residual series, $F_c(t)$, independent and identically distributed in time and none passed the normality test.

The condition that the residual series, $F_c(t)$, is to be independent and identically distributed in time is necessary (Box and Jenkins 1976), while normality is needed to obtain probability limits to the forecast.

Therefore logarithmic transformations were used to satisfy the above condition on the residual series, $F_c(t)$. Logarithmic transformations of hydrologic data, such as river flows, are common. Such transformations were used by many investigators, for instance, by Beard (1965), by Gupta and Fordham (1972), and by Codner and McMahon (1973). Chow (1954) argued that any hydrological event is the result of the joint action of many hydrological and geological factors which can be expressed in the mathematical form of equation (6). The deterministic component of the model was represented by a Fourier series model which contained 2 harmonics, the fundamental (first) and the second harmonics. The contribution ratio of the fundamental harmonic was about 56.8%, while that of the second harmonic was only 13.4%. The significance of the first and second harmonics was in full agreement with Kottegoda (1980), who stated that "In practice, periodicties can be represented by one or two harmonic in monthly series". Kottegoda (1980) fitted a harmonic model to monthly river flows of the Teme at Tenbury Wells (England) which contained only the first harmonic.

The stochastic component of the model consisted of a third-order-autoregressive model, AR(3).

The autoregressive nature of the stochastic component of river flows is well documented in water resources literature: for instance, Roesner and Yevjevich (1966) applied a first-order autoregressive model, AR(1), to monthly river flows data, while a second-order-model, AR(2), was suggested by Quimpo (1968) for daily river flows, and Kotegoda (1972) used a fourth-order-model, AR(4), for generating 5-day river flow data. Earlier, Thomas and Fiering (1962) introduced their famous model in the field of synthetic hydrology. Their model described monthly river flows by an autoregressive model of order one, AR(1). Furthermore, Kottegoda (1980) stated that a linear autoregressive model is a first approximation to many natural processes and that it is particularly applicable to flow in a river which is supplemented by time-dependent components of ground water, surface run-off, catchment retention and the like. Finally, O'Donnell *et al.* (1972) discussed the practicalities of using an autoregressive log-normal daily flow model for the Dart in southern England and the Vardar in Yugoslavia.

Summary and Conclusions

The results of the study of the river Khabur flows at Zakho indicate the following:

(1) The process of generating monthly flows of the river Khabur is adequately represented by a multiplicative model. The model is then linearised by a logarithmic transformation of the data.

(2) The time series model representing monthly flows of the river Khabur at Zakho consists of two components, a deterministic and a stochastic one.

(3) The deterministic component accounts for seasonality only. Seasonality is represented by a Fourier series model, which contains two harmonics. Only the fundamental (12-month cycle) and the second (6-month cycle) harmonics are significant.

(4) The stochastic component is adequately represented by an autoregressive model of order three, AR(3).

(5) The time series model can be confidently used to generate monthly river flows for one year ahead. The model explains 87.6% of the total variance of monthly river flows. The model can also be used to generate a sequence of missing data.

(6) Similarly structured models could be build for other rivers which have incomplete flow records using the mathematical approach described in this paper.

References

Beard, L.R. (1965) Use of interrelated records to simulate streamflow, *J. Hydraul. Div., ASCE.* **91**(5): 13-32.

Box, G.E.P. and **Cox, D.R.** (1964) An analysis of transformations, *J. Roy. Statist. Soc. B,* **26:** 211-52.

Box, G.E.P. and **Jenkins, G.M.** (1976) Time Series Analysis: Forecasting and Control, Revised ed., Holden-Day, San Francisco, Calif. 575p.

Chow, V.T. (1954) The log-probability law and its engineering applications, *Proc. ASCE.* **80:** Sept. no. 536.

Gonder, G.P. and **McMahon, T.A.** (1973) log-normal streamflow generation models re-examined, *J. Hydraul, Div., ASCE.* **99:** 1421-1431.

Directorate General of Irrigation (1976) *Discharges For Selected Gauging Stations in Iraq 1959-1975,* Ministry of Irrigation, Iraq, 374p.

Gupta, V.L. and **Fordham, J.W.** (1972) Streamflow synthesis - A case study, *J. Hydraul. Div., ASCE.* **98**(6): 1049-1055.

Kottegoda, N.T. (1972) Stochastic five daily stream flow model, *J. Hydraul. Div., ASCE.* **98**(9): 1469-1485.

Kottegoda, N.T. (1980) *Stochastic Water Resources Technology,* First Pub., The Macmillan Press Ltd., London, 384p.

Maidment, D.R. and **Parzen, E.** (1984) Cascade model of monthly municipal water use, *Water Resources Research,* **20:** 15-23.

Montgomery, D.C. and **Johnson, L.A.** (1976) *Forecasting and Time Series Analysis,* McGraw-Hill Co., N.Y., 304p.

O'Donnell, T., Hall, M.J. and **O'Connel. P.E.** (1972) Some application of stochastic hydrological, models, *in:* **Biswas, A.K. (ed)** *Modelling of Water Resources Systems,* Harvest House, Montreal, pp. 250-262.

Quimpo, R.G. (1968) Stochatic analysis of daily riverflows, *J. Hydraul. Div., ASCE.* **94**(1): 43-57.

Roesner, L.A. and **Yevjevich, V.M.** (1966) *Mathematical models for time series of monthly precipitation and monthly runoff,* Colo. St. Univ., Fort Collins, Hydrol. Papers, **No.** 15.

Thomas, H.A. and **Fiering, M.B.** (1962) Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation *in:* **Maass, A. (ed)** *Design of Water Resources Systems,* Harvard Univ., Press, Camb., Mass., pp. 459-493.

(Received 20/05/1985;
in revised form 05/11/1985)

توليد القيم المفقودة في تصريفات مياه الأنهر

عدنان عباس السماوي

قسم الهندسة المدنية ـ جامعة البصرة ـ العراق

يتناول هذا البحث مشكلة القيم المفقودة من سجلات تصريفات المياه لبعض الأنهر العراقية، ويقدم أحد الحلول الإحصائية لها.

يتمثل الحل بنموذج مسلسل زمني مؤلف من حد حتمي وحد ستوكاستيكي. لقد تم بناء نموذج مسلسل زمني للتصريفات الشهرية لمياه نهر الخابور عند مدينة زاخو. تم تمثيل الحد الحتمي بنموذج سلسلة فورير (Fourier-Series)، وأما الحد الستوكاستيكي فمثل بنموذج الإنحدار الذاتي (Autoregressive) من المرتبة الثالثة، AR (3)، وباستعمال أسلوب بوكس وجنكنز (Box and Jenkins Approach) لتحليل السلاسل الزمنية. فسر النموذج أكثر من 87% من مجموع التباينات الكلية عندما استعمل في عملية تنبؤ لمدة سنة واحدة، ثم استعمل النموذج للغرض الرئيسي وهو توليد قيم تصريفات شهرية مفقودة.