TIMES SERIES ANALYSIS OF ANNUAL FLOW VOLUMES OF TEN SRI LANKA RIVERS

R. Srikanthan

Introduction

The design and operation of a water resource system inevitably depends on the prediction of the future inflows over the economic life of the system. Reliance only on historical records has been criticized for the following three reasons. Firstly, the historic record is not likely to be repeated in the future. Secondly, it provides only one sample of possible futures and is insufficient to measure uncertainty in system response that is assignable to variability in hydrology. Thirdly, existing records may be of insufficient length to provide even one sample over the economic life of the system. As a consequence, considerable attention has been given to the development of methods for generating pseudo random sequences of synthetic streamflows on a digital computer where the generated sequences have similar statistical properties as those estimated from the historical record. These synthetic sequences have become an important tool to the planner of a water resources system because when used in conjunction with computer simulation, they allow him to evaluate proposed system designs more thoroughly than was possible with previously available methods. In order to generate synthetic streamflows, the time series structure of the streamflow series should be analyzed first to obtain a suitable model. In this paper, annual streamflow volumes of ten Sri Lankan rivers are analyzed to indicate the probable form of the models.

Analysis of Data

The names and gauging stations of the ten rivers used in the study are given in Table 2. The annual flow volumes of each river is plotted to find out one or more of the following characteristics :

(1) seasonality
(2) trends in the mean or variance of the series
(3) persistence and long-term cycles
(4) extreme values

A visual inspection of these plots did not reveal any of the above characteristics. Definition and removal of trend require careful study since the usual methods of fitting polynomial functions or moving average schemes [1] with or without weighting factors may produce a change in the short period oscillations of the series. An appropriate statistical test is Kendall's rank correlation procedure [2].

Tests of Dependence

The annual flow series are tested for short-term dependence using the following six tests :

1. Autocorrelation Test

Short term dependence is usually measured by the magnitude of the low order autocorrelation coefficients. The autocorrelation coefficient, r_k, of lag k is estimated using [3]

$$r_k = \frac{\sum_{i=1}^{n-k} (X_i-\bar{X})(X_{i+k}-\bar{X})}{\sum_{i=1}^{n} (X_i-\bar{X})^2} \qquad (1)$$

where x_i = annual flow volume at time i,
n = sample size, and

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The lag one autocorrelation, r_1, is calculated from Eq. (1) and checked whether or not it is significantly different from zero. The distribution of r_1 can be taken to be approximately normal with mean zero and variance 1/n for moderate to large values of n.

2. Fisz Test

In the Fisz test [4], the annual flow volumes are replaced by O if $x_i < \tilde{x}$ (median) and by 1 if $x_i \geq \tilde{x}$. If the original series has been generated by a purely random process, then, m, the number of times O is followed by 1 or 1 is followed by O is approximately normally distributed with

$$\text{mean} = (n-1)/2 \text{ and} \qquad (2)$$
$$\text{variance} = (n-1)/4$$

If m is not significantly different from the expected value, then the hypothesis that the series results from a random process is accepted.

R. Sri Kanthan graduated from the Faculty of Engineering, Peradeniya in 1971 (Civil Eng.) obtained M. Eng. degree (in coastal Engineering) from Asian Institute of Technology, Bangkok in 1976 and Ph. D. degree (Hydrology) from Monash University, Australia in 1979.

Presently working as a lecturer in the Department of Civil Engineering, University of Peradeniya, Sri Lanka.

3. Kendall's Test

Kendall's test [1] is based on the number of turning points, p, which is approximately normally distributed with

$$\text{mean} = 2(n-2)/3 \text{ and} \quad (3)$$
$$\text{variance} = (16n-29)/90$$

4. Length of Runs Test

In Gold's[5] length of runs test, a run length l is defined by a set of l consecutive flows either above or below the median. If $m'(l)$ denotes the total number of runs above and below the median of length l, then for a purely random process the expected value of $m'(l)$ is given by

$$E\left[m'(l)\right] = (n+3-l)/2^{l+1} \quad (4)$$

and the statistic

$$\sum_{l=1}^{l'} \frac{\left\{m'(l)-E\left[m'(l)\right]\right\}}{E\left[m'(l)\right]}$$

is distributed as Chi square with l'-1 degrees of freedom where l' is the maximum run length in the sequence.

5. Rank Difference Test

In the rank difference test [6] the flow volumes are replaced by their relative ranks (R_i) with the lowest being denoted by rank $1(R_1)$. The U statistic is calculated by

$$U = \sum_{i=2}^{n}\left[R_i - R_{i-1}\right] \quad (5)$$

where U = sum of the absolute rank-difference ($R_i - R_{i-1}$) between successive items in a single sequence.

For large n, U is normally distributed with

$$\text{mean} = (n+1)(n-1)/3 \text{ and} \quad (6)$$
$$\text{variance} = (n-2)(n+1)(4n-7)/90$$

6. Cumulative Periodogram Test

The periodogram of a time series is defined as [7]

$$I(f_j) = \frac{2}{n}\left[\left\{\sum_{i=1}^{n} x_i \cos 2\pi i f_j\right\}^2 + \left\{\sum_{i=1}^{n} x_i \sin 2\pi i f_j\right\}^2\right] \quad (7)$$

where $f_j = \frac{j}{n}$ is the frequency, $j = 1,2, \ldots \left\|\frac{n-1}{2}\right\|$ in which $\left\|\frac{n-1}{2}\right\|$ stands for the integer value of $\frac{n-1}{2}$.

Thenormalised cumulative periodogram is obtained from

$$C(f_j) = \frac{\sum_{i=1}^{j} I(f_i)}{n S^2} \quad (8)$$

where S^2 = estimate of the variance of x_i. For a white noise series, the plot of $C(f_j)$ against f_j would be scattered about a straight line joining points (O,O) and (0.5,1). The approximate confidence limit lines for a truely random series are drawn at distances

$$\pm K_\alpha \Big/ \left\{\left\|\frac{n-1}{2}\right\|\right\}^{\frac{1}{2}}$$

above and below thetheoretical line. Values of K_α are given below :

α	0.01	0.05	0.	0.25
K_α	1.63	1.36	1.22	1.02

Table 1. Values of K_α (7)

If the plotted points lie within the confidence limit lines, then the series is assumed to be a random series. Otherwise, it is considered as dependent or correlated.

The above six tests have been applied to the annual flow volumes for the ten rivers and the results are set out in Table 2.

For annual flow volumes of a river to be considered non-random, assuming at least 3 tests should indicate such an effect, then the analysis shows that only one river (Maha Oya) is non-random.

Annual Time Series Models

Annual time series models can be broadly classified into two groups, namely, short-memory and long-memory models. A long memory model uses the Hurst coefficient in addition to the lower order moments and its application is straight forward once a decision is made on the value of Hurst coefficient. Consequently, only the fitting of short-memory models such as autoregressive (AR), moving average (MA) or autoregressive moving average (ARMA) is considered.

An autoregressive model of order p is defined as

TABLE 2

RESULTS OF TESTS FOR DEPENDENCE OF ANNUAL FLOW VOLUMES

No.	*River*	*Station*	*Length of Record (yrs)*	*r1*	*Fisz*	*Kendall*	*Gold*	*Meacham*	*Cumulative Periodogram*
1	Kelani Ganga	Glencourse	23	ns	ns	ns	ns	ns	ns
2	Kalu Ganga	Putupaula	31	ns	ns	ns	ns	ns	ns
3	Nilwala Ganga	Bopagoda	34	ns	ns	ns	ns	ns	ns
4	Mahaweli Ganga	Manampitiya	33	ns	ns	ns	ns	ns	10
5	Amban Ganga	Elahara	33	ns	ns	ns	ns	ns	ns
6	Maha Oya	Alawwa	27	ns	5	ns	10	5	ns
7	Menik Ganga	Kataragama	30	ns	ns	ns	ns	ns	ns
8	Kirindi Oya	Lunuganwehera	30	ns	ns	ns	ns	ns	ns
9	Yan Oya	Pangurgaswewa	28	ns	ns	ns	1	10	ns
10	Malwattu Oya	Kappchchi	29	ns	ns	ns	ns	ns	ns

ns — not significant. The numerical numbers indicate the level of significance.

$$x_t = \sum_{i=1}^{p} \phi_i x_{t-i} + a_t \tag{9}$$

where ϕ_i = i^{th} autoregressive parameter and
a_t = random noise component

A moving average model of order q is defined as

$$x_t = a_t - \sum_{i=1}^{q} \theta_i a_{t-i} \tag{10}$$

where θ_i = i th moving average operator and
a_t = random noise component

An autoregressive moving average process of order (p,q) is defined as

$$x_t = \sum_{i=1}^{p} x_{t-i} + a_t - \sum_{i=1}^{q} \theta_i a_{t-i} \tag{11}$$

where ϕ_i, θi and a_t are defined as before.

One of the outcomes of a time series analysis of data is the determination of the structure of the process giving rise to the time series. An initial step in this is to decide whether the flows are generated by a white noise drocess. The tests in the previous section are a basis of this step along with the autocorrelation function (acf) (Eq. 1) and partial autocorrelation function (pacf).

The pacf is a more sensitive criterion for determining the order of an AR process than acf. The pacf analysis is carried out by computing for each value of k (the order of the process), the estimate ϕ_{kk} of the last coefficient ϕ_k in the fitted AR model of the form

$$x_t = \sum_{i=1}^{k} \phi_i x_{t-i} + a_t \tag{12}$$

The plot of ϕ_{kk} versus k is called the pacf. For large n, ϕ_{kk} is approximately normally distributed with mean zero and variance 1/n.

Thus, based on acf, pacf and dependence tests the annual flow series are classified into two groups :

(1) those which do not have any significant dependence, acf and pacf.

(2) those which have significant dependence, acf and pacf for some lags.

The annual flow series in the first group can be modelled using white noise with appropriate skewness. For the second group of annual flows, apporpriate models are identified using acf and pacf. The acf of an AR process of order p tails off while the pacf has a cut off after lag p. Similarly, the acf of a moving average (MA) process of order q has a cut off after lag q, while its pacf tails off. If both the acf and pacf tail off, a mixed process is suggested (7).

The inverse autocorrelation function (iacf) helps to confirm the models identified using acf and pacf and to obtain a parsimonious model. A parsimonious model is one which models a process adequately and has the least number of parameters. The iacf of a time-series is defined to be the autocorrelation function associated with the inverse of the spectral density of the series(8) An autoregressive estimate of the iacf, ri_k, is obtained from (8) as

$$ri_k = \frac{-\phi_k + \phi_1 \phi_{k+1} + \cdots + \phi_{p-k} \phi_p}{1 + \phi_1^2 + \cdots + \phi_p^2}$$

where p is the order of the fitted AR process, and ϕi the AR parameters. If the series is white noise, their ri_k is approximately normal with mean zero and variance 1/n. Also, a small magnitude of ri_k for any lag k indicates that the AR parameter ϕ_k should be set to zero.

TABLE 3

FORMS OF STOCHASTIC MODEL SUGGESTED FOR THE RIVERS

No.	*River*	*Model*
1	Kelani Ganga	WN*
2	Kalu Ganga	WN
3	Nilwala Ganga	AR (1)
4	Mahaweli Ganga	AR (3) with $\phi_2 = 0$
5	Amban Ganga	WN
6	Maha Oya	AR (3) with $\phi_2 = 0$
7	Menik Ganga	WN
8	Kirindi Oya	WN
9	Yan Oya	WN
10	Malwattu Oya	WN

*WN stands for white noise

Forms of stochastic model suggested for the rivers

The value of p may be chosen in practice by examining the acf and pacf. A large value of p (with respect to the number of observations) results in an estimate with large variance, while a small value of p causes bias in the estimate of iacf. Thus, in practice, it is worth while to calculate ri_k for a few values of p. If the ri_k resulting from a particular value of p is reliable, the values of ri_k should not change drastically when p is increased by a moderate amount.

Based on the above analysis, Table 3 indicates the form of the model suggested for each river.

Even though the dependence tests showed that only one river as non-random, Nilwala Ganga had fairly large iacf for lag one and Mahaweli Ganga for lags one and three. Consequently AR(1) and AR(3) models are suggested for Nilwala Ganga and Mahaweli Ganga respectively. Hence, seven out of ten rivers can be represented by a white noise model.

Conclusions

Dependence analysis carried out on ten annual stream flow series showed that only one of them (Maha Oya) cannot be considered as random. The procedure for identifying suitable models for these rivers are outlined and the appropriate stochastic models for these rivers are given.

Acknowledgements

The author wishes to acknowledge the assistance and cooperation given by Mr. K. B. E. Karunaratne (Director) and Mr. S. Sritharan (Irrigation Engineer) of the Irrigation Department in obtaining the data used in this study. The comments and suggestions made by Prof. T. D. M. A. Samuel (Engineering Mathematics Department, University of Peradeniya) on the manuscript are greatly appreciated.

References

1. Kendall, M. G. and Stuart, A., *The Advanced Theory of Statistics*, Vol. 3, Charles Griffin, 1976.
2. Kendall, M. G., *Rank Correlation Methods*, Griffin, London, 1948.
3. Jenkins, G. M. and Watts, D. G., *Spectral Analysis and Its Applications*, Holden-Day Inc., San Francisco, 1968.
4. Fisz, M., *Probability Theory and Mathematical Statistics*, John Wiley, N.Y., 1963.
5. Gold, E., *Notes on the Frequency of Occurrence of Sequence in a Series of Events of Two Types*, A. J. Roy. Met. Soc., Vol. 55, 1929, pp. 307-309.
6. Meacham, I., *Correlation in Sequential Data—Three Simple Indicators*, Civil Eng. Trans., I.E. Aust. Vol. CE10, No. 2, 1968, pp. 225-228.
7. Box, G. K. P. and Jenkins, G. M., *Time Series Analysis: Forecasting and Control*, Holden-Day Inc., San Francisco, 1970.
8. Cleveland, W. S., *The Inverse Autocorrelation of a Time Series and Their Applications*, Technometrics, Vol. 14, No. 2, 1972, pp. 277-298.

ශ්‍රී ලංකාවේ ගංගා දහයක වාර්ෂික ගැලීම් පරිමාවන්ගේ කාලශ්‍රේණි විශ්ලේෂණය

සාරාංශය

ශ්‍රී ලංකාවේ ගංගා දහයක වාර්ෂික ගැලීම් පරිමාවන්ගේ පරායත්තතාව සඳහා ප්‍රථමයෙන් විශ්ලේෂණය කරණු ලැබේ. ඉන් පසුව ගැලීම් ශ්‍රේණියේ ස්වයං සම්බන්ධතාය සහ අර්ධ ස්වයං සම්බන්ධනය භාවිතයෙන් එක් එක් ගංගාව සඳහා අදාල අනුමානික ආදර්ශය ලබාගනු ලැබේ.