STATISTICAL MODELS AND METHODS FOR RIVERS IN THE SOUTHWEST

by

Sidney Yakowitz

ABSTRACT

Riverflow modeling is believed useful for purposes of decision making with respect to reservoir control, irrigation planning, and flood forecasting and design of structures to contain floods. This author holds the view that present riverflow models in vogue are unsatisfactory because, for one thing, sample simulations according to these models do not resemble observed southwestern river records. The purpose of this paper is to outline a general Markov model which assumes only that rivers have a finite memory. We show how to calibrate the model from river records and then present evidence to support our contention that some success has been realized in mimicking typical flows by our simulation procedure.

REVIEW OF STATUS OF STREAMFLOW MODELING

The objective in streamflow modeling is as follows. A sequence $\{X_i\}_{i=1}^{n}$ of measured daily flows of a particular river at a particular gauging station is presented. The hydrologist is required to provide some model and calibrate this model by statistical analysis of the record $\{X_i\}$. This model should be successful in duplicating certain salient features of the recorded streamflow. Of course, the choice of features to be preserved in the simulated record depends on the decision problem for which the river flow analysis is ultimately to be used.

In Figure 1, we have presented a connected plot of a USGS record of a representative three years of flow at a certain station on the Cheyenne River. This sporadic behavior of no flow interrupted by bursts of flow is characteristic of many rivers in the southwest. Currently, the most popular streamflow models discussed in the literature are autoregressive moving average (ARMA) models. Such models are elaborately discussed in the monograph by Fiering (1967) and also in the papers by Jackson (1975a and 1975b), Weiss (1977), Kottegoda (1972), Ven Te Chow and Kareliotis (1970), and Mejia and Rodrigues-Iturbe (1974), to mention but a few of the many works on this subject.

An ARMA model for streamflow $\{X_i\}$ is characterized by the stochastic difference equation

$$X_{i+1} = \sum_{i=0}^{N} a_i X_{n-i} + \sum_{j=0}^{M} b_j U_{n-j} \qquad (1)$$

where the U_i's are independent, identically distributed random variables. (Sometimes the U_i's have different distributions according to the season in which day j falls, but this adaptation is not essential to the comments to follow.)

There are several drawbacks to using the ARMA model for modeling streamflows in the southwest, as this investigator has discussed in Yakowitz (1972) and (1973). First, if the noise input $\{U_i\}$ is assumed Gaussian (or even nonzero with probability 1), either the flows will never be identically zero, or else negative flows will be possible. On the other hand, if the U_i's have high probability of being zero, so that decays to zero flow are possible, then for any stable linear difference equation of the form (1), the decay will be exponential; that is,

$$X_n \approx \lambda^{n-N} X_N$$

where λ is the largest (in modulus) root of the polynomial associated with the homogeneous version of the difference equation (1) and N is the last day of positive input U_i. But examination of records shows that southwestern rivers tend to die away faster
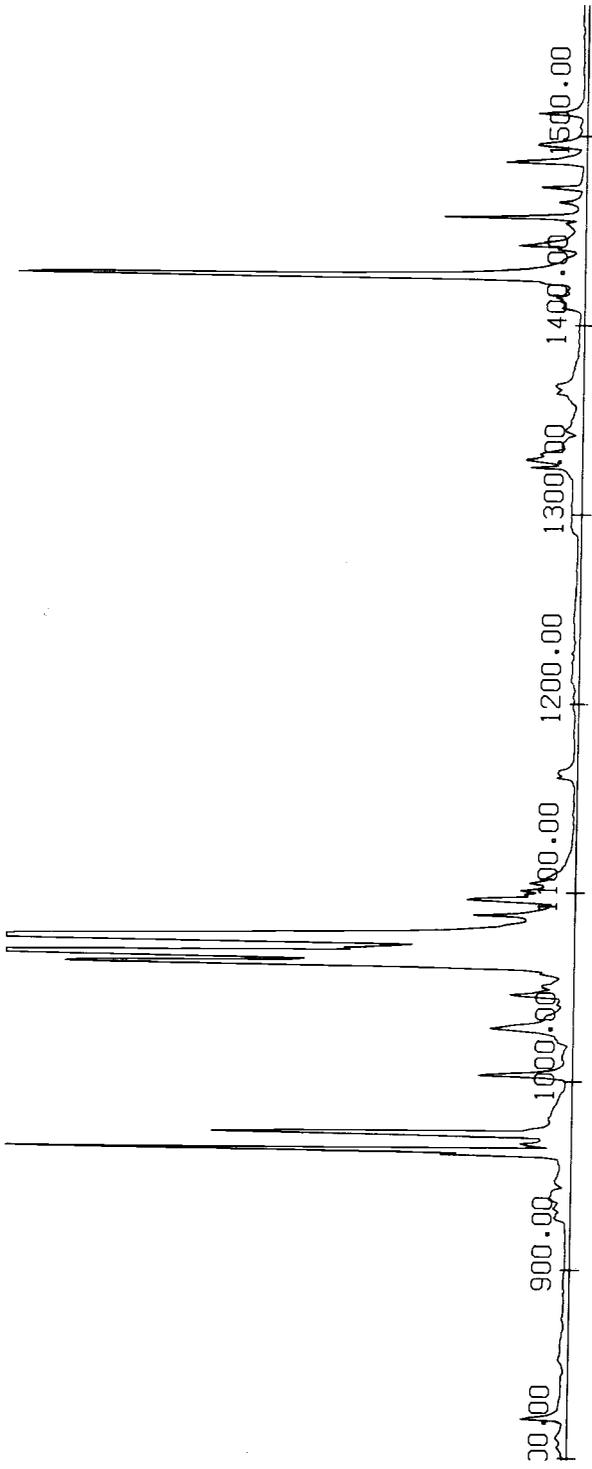
Figure 1. Record of the Cheyenne River

than exponentially during dry periods. Secondly, according to the ARMA models (1), if dry periods are possible, the episodes of flow in the model must be evenly distributed along the time axis (because the U_j's must be independent) whereas characteristically, southwestern rivers exhibit bursts of activity associated with several consecutive days of precipitation.

Another point is that there is no physical reason to think that streamflow satisfies the linear difference equation (1), since channel flow equations themselves are nonlinear differential equations and thus lead to nonlinear difference equations. We mention in passing that the broken line process (e.g., Mandelbrot, 1972) and the Markov-ARMA model of Jackson (1975a,b) are also subject to the above-listed limitations.

As a consequence of these ideas, this investigator has been casting around for alternative models which are more inclusive and yet computationally manageable. One type of model which circumvented the above drawbacks was described in Yakowitz (1973), but it must be admitted that statistical analysis required by this model was somewhat sophisticated and this model also had some arbitrary elements. In the next section, we describe what we feel is a promising and general approach toward overcoming the limitations inherent in current models.

A GENERAL STREAMFLOW MODEL AND A STATISTICAL PROCEDURE

We make the assumption that the streamflow process $\{X_i\}$ is a Markov chain of known order r. That is, for every integer n and for every vector $\underline{X}_n = (X_n,\ldots,X_{n-r+1})$, and every event A,

$$P[X_{n+1} \in A | X_j, j \leq n] = P[X_{n+1} \in A | \underline{X}_n]. \tag{2}$$

It may be demonstrated that the ARMA process (1) is a particular type of Markov chain of order r < M + N, but it is evident that there are many Markov chains which are not representable as ARMA processes. The papers by Denny, Kisiel, and Yakowitz (1974), Yakowitz and Denny (1973), and Yakowitz (1976) are addressed to the problem of inferring the Markov order from streamflow records. The initial state $\underline{X}_r$ having been specified, the probabilistic behavior of a Markov chain is completely determined by the transition probability function $F(y|\underline{X})$ where

$$F(y|\underline{X}) = P[X_{n+1} \leq y | X_n = \underline{X}].$$

In principle, the task of statistically inferring $F(y|\underline{X})$ from the observed streamflow record $\{X_i\}$ would seem impossible because there is an uncountable infinitude of distribution functions F to approximate, one for each possible r-tuple $\underline{X}$. But we have discovered a rule which gives a good approximation for modest size samples and, as proven in Yakowitz (1977), has the property that as the record length n increases, provided only that $F(y|\underline{X})$ is continuous in $\underline{X}$, if $F_n(y|\underline{X})$ is constructed according to the algorithm given below, for every y and every possible*r-tuple $\underline{X}$,

$$F_n(y|\underline{X}) \xrightarrow{n} F(y|\underline{X}).$$

ALGORITHM FOR APPROXIMATION OF TRANSITION FUNCTION

Assume the record length n and the record $\{X_i\}_{i=1}^n$ is specified.

STEP 1

Choose some representative set $\mathcal{L} = \{\underline{Y}_1,\ldots,\underline{Y}_M\}$ of vectors (r-tuples). They should be chosen so that $M \approx \sqrt{n}$, and so that they are placed as well as possible to lie "close" to r-tuples characteristic of the record. We used the K-means algorithm for cluster analysis to choose these vectors. (See Hartigan, Clustering Algorithms, 1975, for details.)

STEP 2

For each j, $1 \leq j \leq M$, partition the flows into sets S_j, where

$$S_j = \{X_{i+1}: ||\underline{X}_i - \underline{Y}_j|| \leq ||\underline{X}_i - \underline{Y}_k||, 1 \leq k \leq M\}$$

In words, S_j is composed of the flows that come after the states which are closer to $\underline{Y}_j$ than any of the other representative vectors

---

*Excludes vectors having negative components and components greater than the capacity of the river banks, etc.

149

STEP 3

Use any standard method of statistics to infer a cumulative distribution function $F_n(y|\underline{Y}^j)$ from the numbers in S_j, $1 \leq j \leq M$.

STEP 4

For any x and y, we approximate $F(y|\underline{X})$ by $F_n(y|\underline{Y}^*)$, where $\underline{Y}^*$ is defined to be the representative vector in $\mathcal{U}$ which is closest to $\underline{X}$.

It would seem reasonable that under the continuity condition for $F(y|\underline{X})$, as the size of the sample increases and consequently, the set $\mathcal{U}$ of representative vectors become more numerous in the set of all possible vectors, that the asserted convergence takes place. This mathematical fact is proven in Yakowitz (1977). The author is preparing a paper for a hydrology journal which describes this algorithm in greater detail and discusses hydrological ramifications and simulation results more fully.

In Figure 2, we present the simulated graph obtained by applying 30 years of daily record (divided into two annual seasons) into the transition function algorithm just described. The plots were obtained by computer simulation of the Markov chain using the inferred transition function $F_n(y|\underline{X})$.

REFERENCES CITED

Chow, Ven Te and S. Kareliotis. 1970. Analysis of stochastic hydrologic systems. Water Resources Research 6:1569-1582.

Denny, J., C. Kisiel and S. Yakowitz. 1974. Procedures for determining the order of dependence in streamflow records. Water Resources Research 10(5):947-954.

Fiering, M. 1967. Streamflow Synthesis. Harvard University Press, Cambridge, Massachusetts.

Hartigan, J. 1975. Clustering Algorithms. J. Wiley, New York.

Jackson, B. 1975a. The use of streamflow models in planning. Water Resources Research 11:54-63.

Jackson, B. 1975b. Markov mixture models for drought lengths. Water Resources Research 11:64-74.

Kottegoda, N. 1972. Five stochastic daily streamflow models. Journal of the Hydraulics Division, American Society of Civil Engineers HY9:1469-1485.

Mandelbrot, B. 1972. Broken line processes derived as an approximation to fractional noise. Water Resources Research 8:1354-1356.

Mejia, J. and I. Rodrigues-Iturbe. 1974. Correlation links between normal and log normal processes. Water Resources Research 10:689-693.

Weiss, G. 1977. Shot noise models for the generation of synthetic streamflow data. Water Resources Research 13:101-108.

Yakowitz, S. 1972. A statistical model for daily streamflow records with application to the Rillito River. Proceedings, International Symposium on Uncertainties in Hydrologic and Water Resource Systems, University of Arizona, Tucson.

Yakowitz, S. 1973. A stochastic model for daily river flows in an arid region. Water Resources Research 9(5):1271-1285.

Yakowitz, S. 1976. Small sample hypothesis tests of Markov order, with application to simulated and hydrologic chains. Journal of the American Statistical Association 71(353):132-136.

Yakowitz, S. 1977. Nonparametric inference of Markov transition functions. Submitted for publication.

Yakowitz, S. and J. Denny. 1973. On the statistics of hydrologic time series. Proceedings, 17th Annual Meeting of the Arizona Academy of Sciences, Tucson, Arizona 3:146-163.
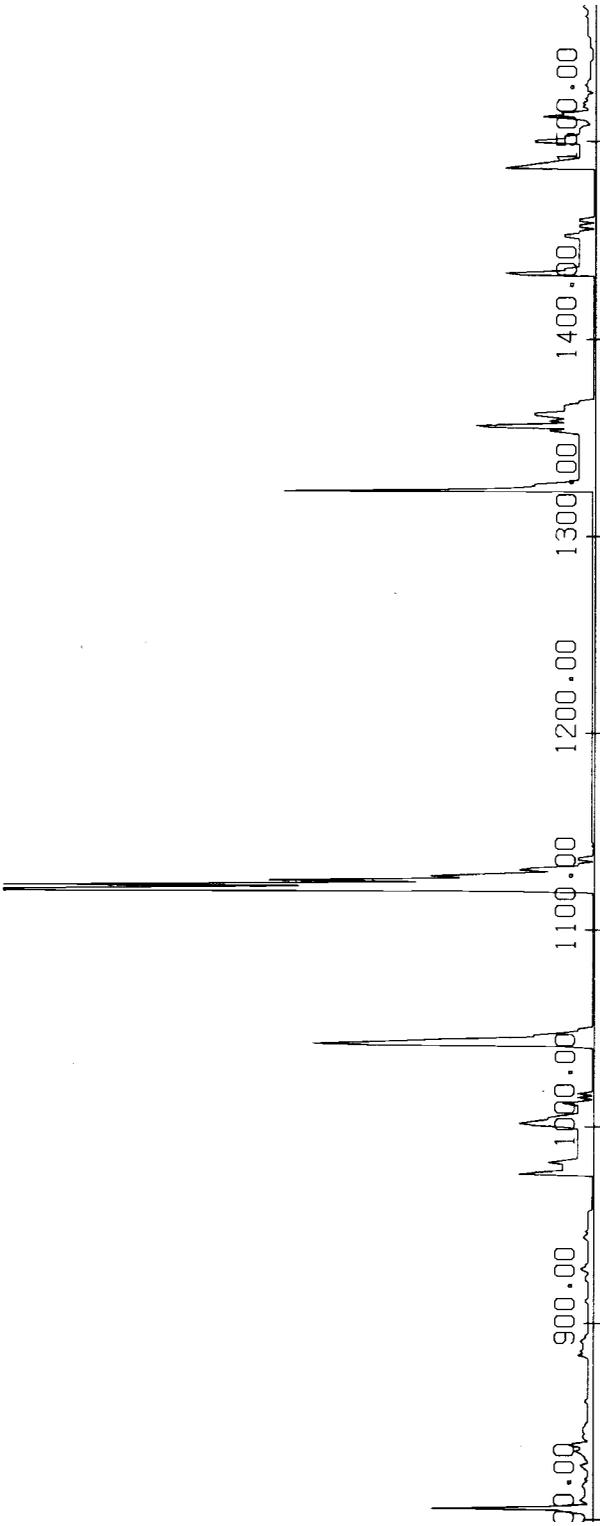
Figure 2. Simulated Record of the Cheyenne River