REGIONAL STOCHASTIC GENERATION OF STREAMFLOW USING AN ARIMA (1,0,1) PROCESS AND DISAGGREGATION

U. S. Geological Survey
Water-Resources Investigations 79-3

Prepared in cooperation with the
Susquehanna River Basin Commission

BIBLIOGRAPHIC DATA SHEET	1. Report No.	2.	3. Recipient's Accession No.

4. Title and Subtitle
REGIONAL STOCHASTIC GENERATION OF STREAMFLOWS USING AN ARIMA (1,0,1) PROCESS AND DISAGGREGATION

5. Report Date
May 1979

6.

7. Author(s)
Jeffrey T. Armbruster

8. Performing Organization Rept. No. USGS/WRI-79-3

9. Performing Organization Name and Address
U.S. Geological Survey, Water Resources Division
Post Office Box 1107
Harrisburg, Pennsylvania 17108

10. Project/Task/Work Unit No.

11. Contract/Grant No.

12. Sponsoring Organization Name and Address
U.S. Geological Survey, Water Resources Division
Post Office Box 1107
Harrisburg, Pennsylvania 17108

13. Type of Report & Period Covered
Final

14.

15. Supplementary Notes
Prepared in cooperation with the Susquehanna River Basin Commission

16. Abstracts An ARIMA (1,0,1) model is used to generate annual flow sequences at three sites in the Juniata River basin, Pennsylvania. The study was designed to analyze low-flow frequency characteristics of a basin. The model preserves the mean, variance, and cross-correlations of the observed station data. In addition, it has a desirable blend of both high and low frequency characteristics and therefore is capable of preserving the Hurst coefficient, h. The generated annual flows are disaggregated into monthly sequences using a modification of the Valencia-Schaake model. The low-flow frequency and flow duration characteristics of the generated monthly flows, with length equal to the historical data, compare favorably with the historical data. Once the models are calibrated and verified, 100-year sequences are generated and analyzed for their low flow characteristics. One-, three-, and six-month low-flow frequencies at recurrence intervals greated than 20 years are generally found to be lower than flow computed from the historical flows. Procedures are presented for application of the models presented here to ungaged sites.

17. Key Words and Document Analysis. 17a. Descriptors

*Stochastic processes, *synthetic hydrology, *model studies, Pennsylvania, correlation, persistence, flow characteristics

17b. Identifiers/Open-Ended Terms

Autoregressive integrated moving average (ARIMA), disaggregation

17c. COSATI Field/Group

18. Availability Statement	19. Security Class (This Report)	21. No. of Pages
No restriction on distribution	UNCLASSIFIED	54
	20. Security Class (This Page) UNCLASSIFIED	22. Price

REGIONAL STOCHASTIC GENERATION OF STREAMFLOWS USING AN ARIMA (1,0,1) PROCESS AND DISAGGREGATION

By Jeffrey T. Armbruster

U.S. GEOLOGICAL SURVEY

Water-Resources Investigations 79-3

Prepared in cooperation with the

Susquehanna River Basin Commission

May 1979

UNITED STATES DEPARTMENT OF THE INTERIOR

CECIL D. ANDRUS, Secretary

GEOLOGICAL SURVEY

H. William Menard, Director

For additional information write to:

U.S. Geological Survey
P.O. Box 1107
Harrisburg, Pennsylvania 17108

CONTENTS

	Page
Abstract	1
Introduction	1
Acknowledgments	4
Statistics of observed flows	5
Hydrologic persistence and the Hurst coefficient	8
Description and calibration of the ARIMA (1,0,1) model	10
Estimation of model parameters	14
Generation of annual flows	17
Description and calibration of the disaggregation model	20
Description of the Valencia-Schaake model	20
Modifications to the Valencia-Schaake model	22
Wrap-around correlation	22
Transformation of data	22
Other modifications to the Valencia-Schaake model	23
Disaggregation of seasonal and monthly flows from generated annual flows	24
Flow synthesis	24
Generation of long-term annual flows	32
Generation of long-term seasonal and monthly flows	32
Regional estimates of streamflow statistics	44
Mean and standard duration	44
Cross correlation	44
Estimation of diagonal elements	44
Estimation of off-diagonal elements	49
Lag-zero cross correlation	49
Lag-one and lag-two cross correlations	51
Summary	51
References	53

ILLUSTRATIONS

Figure	1.--Map of the Juniata River basin, Pennsylvania	3
	2.--Comparison of the 1-month duration observed and five typical generated 30-year sequences low-flow frequency curves at site 1.	27
	3.--Comparison of the 1-month duration observed and five typical generated 30-year sequences low-flow frequency curves at site 2.	27
	4.--Comparison of the 1-month duration observed and five typical generated 30-year sequences low-flow frequency curves at site 3.	28

Page

Figure 5.--Comparison of the 3-month duration observed and five typical generated 30-year sequences low-flow frequency curves at site 1.---------- 28
6.--Comparison of the 3-month duration observed and five typical generated 30-year sequences low-flow frequency curves at site 2.---------- 29
7.--Comparison of the 3-month duration observed and five typical generated 30-year sequences low-flow frequency curves at site 3.---------- 29
8.--Comparison of the 6-month duration observed and five typical generated 30-year sequences low-flow frequency curves at site 1.---------- 30
9.--Comparison of the 6-month duration observed and five typical generated 30-year sequences low-flow frequency curves at site 2.---------- 30
10.--Comparison of the 6-month duration observed and five typical generated 30-year sequences low-flow frequency curves at site 3.---------- 31
11.--Comparison of observed and five typical generated 30-year sequences flow duration curves at site 1.- 34
12.--Comparison of observed and five typical generated 30-year sequences flow duration curves at site 2.- 34
13.--Comparison of observed and five typical generated 30-year sequences flow duration curves at site 3.- 35
14.--Comparison of the low-flow frequency curves for the 1-month duration 30-year observed flows and five typical 100-year generated flow sequences at site 1.---------- 39
15.--Comparison of the low-flow frequency curves for the 1-month duration 30-year observed flows and five typical 100-year generated flow sequences at site 2.---------- 39
16.--Comparison of the low-flow frequency curves for the 1-month duration 30-year observed flows and five typical 100-year generated flow sequences at site 3.---------- 40
17.--Comparison of the low-flow frequency curves for the 3-month duration 30-year observed flows and five typical 100-year generated flow sequences at site 1.---------- 40
18.--Comparison of the low-flow frequency curves for the 3-month duration 30-year observed flows and five typical 100-year generated flow sequences at site 2.---------- 41

Page

Figure 19.--Comparison of the low-flow frequency curves for the 3-month duration 30-year observed flows and five typical 100-year generated flow sequences at site 3.---------- 41
20.--Comparison of the low-flow frequency curves for the 6-month duration 30-year observed flows and five typical 100-year generated flow sequences at site 1.---------- 42
21.--Comparison of the low-flow frequency curves for the 6-month duration 30-year observed flows and five typical 100-year generated flow sequences at site 2.---------- 42
22.--Comparison of the low-flow frequency curves for the 6-month duration 30-year observed flows and five typical 100-year generated flow sequences at site 3.---------- 43
23.--Comparison of observed and five typical generated 30-year sequences flow duration curves at site 1.---------- 45
24.--Comparison of observed and five typical generated 30-year sequences flow duration curves at site 2.---------- 46
25.--Comparison of observed and five typical generated 30-year sequences flow duration curves at site 3.---------- 46
26.--Relation between lag-one cross correlation and distance between centroids of overlapping basins---------- 50

TABLES

Table 1.--Gaging stations used in this study---------- 2
2.--Lag-zero correlation matrix for nine stations in the Juniata River basin---------- 6
3.--Lag-one correlation matrix for nine stations in the Juniata River basin---------- 7
4.--Lag-two correlation matrix for nine stations in the Juniata River basin---------- 7

TABLES--Continued

		Page
Table 5.	--Estimates of the Hurst coefficient from historical data	9
6.	--Mean and standard deviation of twenty generated 30-year sequences	18
7.	--Estimated lag-zero, -one, and -two cross correlation matrices for one generated 30-year sequence	19
8.	--Summary of Hurst coefficients computed for the 30-year historical sequence and the twenty 30-year generated sequence	19
9.	--Comparison of historical seasonal means and standard deviations with the average means and standard deviations of twenty generated 30-year sequences	25
10.	--Comparison of historical monthly means and standard deviations with the average means and standard deviations of twenty generated 30-year sequences	26
11.	--Summary of means and standard deviations of 100-year generated annual flow sequences	33
12.	--Summary of values of the Hurst coefficient estimated from ten sequences of generated flows, 100 years long each	36
13.	--Summary of generated seasonal means and standard deviations of ten 100-year sequences	37
14.	--Summary of generated monthly means and standard deviations of ten generated 100-year sequences	38
15.	--Summary of regression equations for mean of annual and monthly flows in the Juniata River basin	47
16.	--Summary of regression equations for standard deviations of annual and monthly flows in the Juniata River basin	48

REGIONAL STOCHASTIC GENERATION OF STREAMFLOWS USING AN ARIMA (1,0,1) PROCESS AND DISAGGREGATION

By Jeffrey T. Armbruster

ABSTRACT

An ARIMA (1,0,1) model was calibrated and used to generate long annual flow sequences at three sites in the Juniata River basin, Pennsylvania. The model preserves the mean, variance, and cross correlations of the observed station data. In addition, it has a desirable blend of both high and low frequency characteristics and therefore is capable of preserving the Hurst coefficient, h.

The generated annual flows are disaggregated into monthly sequences using a modification of the Valencia-Schaake model. The low-flow frequency and flow duration characteristics of the generated monthly flows, with length equal to the historical data, compare favorably with the historical data.

Once the models were verified, 100-year sequences were generated and analyzed for their low flow characteristics. One-, three- and six-month low-flow frequencies at recurrence intervals greater than 10 years are generally found to be lower than flow computed from the historical flows.

A method is proposed for synthesizing flows at ungaged sites.

INTRODUCTION

The Susquehanna River Basin Commission (SRBC) has the responsibility to develop and implement a water-supply program as part of its overall objective to manage the water resources of the basin. Three critical needs of the SRBC are to define long-duration, low-flow frequency characteristics, to predict any impending water shortage so that emergency water conservation measures can be implemented, and to evaluate the effectiveness of water-supply storage projects. Statistical inferences could be made about all three of these items based on data generated from a properly designed simulation model.

In 1975, a 2-year cooperative program by the U.S. Geological Survey and the SRBC was begun to explore the use of a stochastic streamflow-generating model to satisfy these needs. Although a single model for the entire Susquehanna River Basin was desired, it was felt that due to constraints of present day computers, and inexperience with large scale models of this type subbasin models may be equally useful. The Juniata River basin, a subbasin in the southeast part of the Susquehanna River basin was selected for a pilot project.

The study involves several major parts and is outlined by the steps discussed in the remainder of this section. Initially, the mean and standard deviation of observed annual data are calculated. Because this step is elementary, no further discussion of it will be presented. The statistics of the observed flows as discussed here and below were used in the model and will be described later. Adequacy of model output is based on a comparison of observed to generated statistics.

The next step examines the correlation structure of the annual flows at each station in the basin. The nine regular gaging stations listed in table 1 and shown on figure 1 were used. A concurrent 30-year period of data, 1945-74, was available for this analysis. The structure of these correlations must resemble the correlation structure of the model before it can be used to simulate annual flows. For many years autoregressive models have been used to generate synthetic streamflow sequences. However, these models have been criticized during the past 10 years because they do not preserve the Hurst coefficient and do not generally provide flows more extreme than those in the observed sequence. For these reasons the model being applied here is a first-order autoregressive, zero-order integrated, first-order moving-average process, referred to hereafter as an ARIMA (1,0,1) process (Box and Jenkins, 1970).

Table 1.--Gaging stations used in this study

Site No.	USGS station no.	Name	Period of record
1	01556000	Frankstown Br. Juniata R. at Williamsburg	1917-74
2	01557500	Bald Eagle Creek at Tyrone	1945-74
3	01558000	Little Juniata R. at Spruce Creek	1939-74
4	01559000	Juniata R at Huntingdon	1942-74
5	01560000	Dunning Cr. at Belden	1940-74
6	01562000	Raystown Br. Juniata R at Saxton	1912-74
7	01563500	Juniata R. at Mapleton Depot	1938-74
8	01564500	Aughwick Cr nr Three Springs	1939-74
9	01567000	Juniata R. at Newport	1900-74

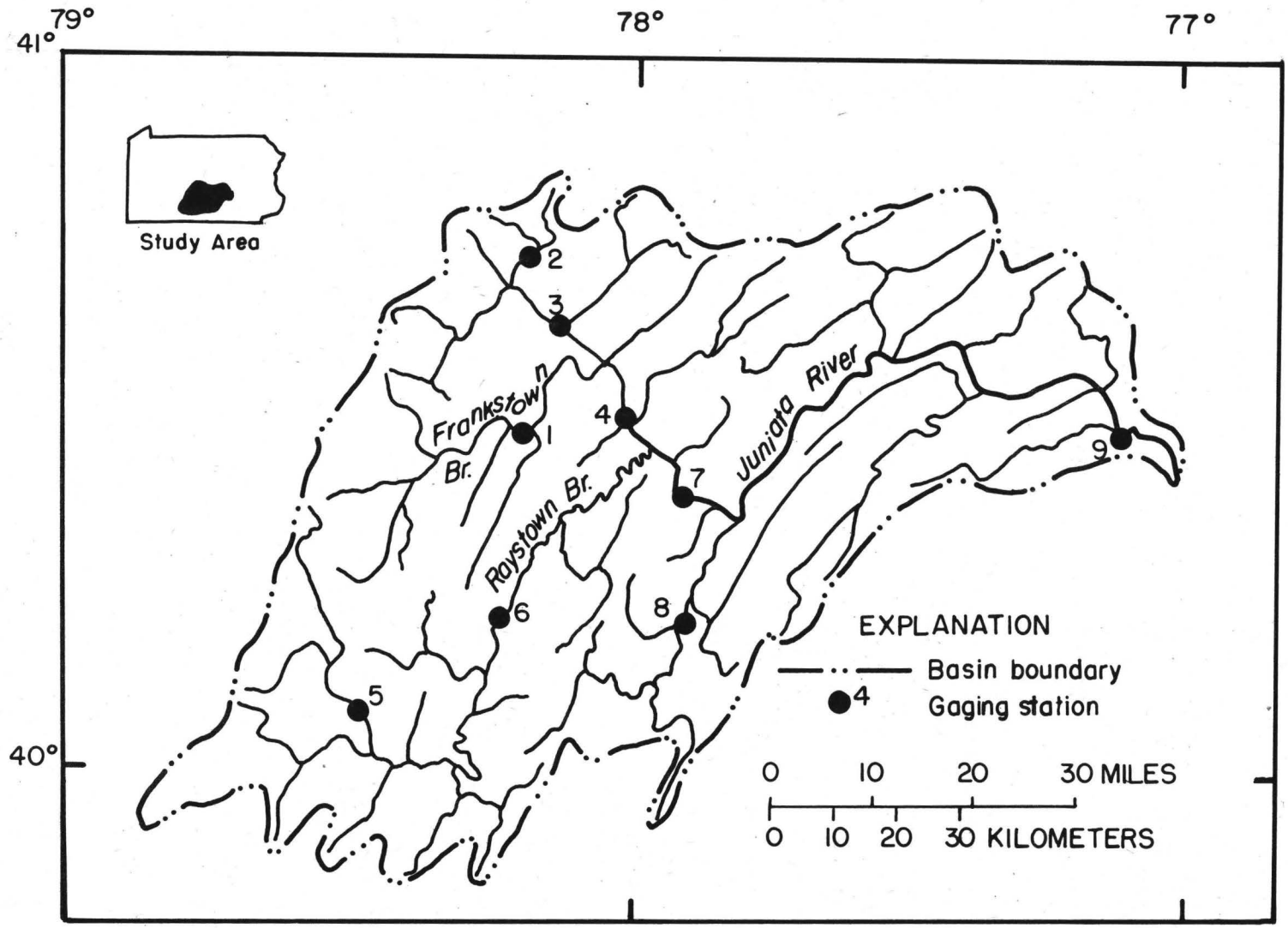

Figure 1.--Map of the Juniata basin, Pennsylvania.

If the ARIMA (1,0,1) process can be verified as a proper representation of the observed annual streamflows, model parameters, for use in generating long-term annual flows, are estimated using short-term observed streamflow records.

The third major step disaggregates or separates annual flows into seasonal flows and seasonal flows into monthly flows. During the disaggregation process, the correlation structure of the observed seasonal and monthly flows must be maintained in the simulated flows. The model used to carry out disaggregation was developed by Valencia and Schaake (1972, 1973).

The fourth major part of the study compares the simulated flow sequences and estimates of their statistics to similar values of the historic sequences. Generated low-flow frequency and flow duration curves are also compared to historical data. The reliability of generated flows as a basis for estimating population (rather than sample) statistics depends heavily on the assumptions that observed flows are a representative temporal and spatial sample and that the statistical distribution of flows is adequately described.

The final part of the study regionalizes or generalizes the above procedure for use at ungaged sites. The statistics of annual and monthly flows were related to basin parameters using standard linear regression techniques. A method is also developed for estimating the cross-correlation structure of annual flows between and among ungaged sites.

ACKNOWLEDGMENTS

The author thanks Dr. P. Enda O'Connell of the Institute of Hydrology, Wallingford, England, for providing the computer subroutines used to estimate the parameters of the ARIMA (1,0,1) multisite model and the subroutine used to generate the annual flows.

Special thanks and appreciation are extended to Dr. John C. Schaake, National Oceanic and Atmospheric Administration, National Weather Service, for providing the subroutines used in the disaggregation process. Dr. Schaake was also very instrumental in helping to solve the seasonal and monthly flow skewness problem.

Finally, the author also thanks Marshall E. Moss, U.S. Geological Survey, Reston, Virginia, for his assistance and encouragement.

STATISTICS OF OBSERVED FLOWS

In generating sequences of synthetic streamflows at multiple sites, the interrelation between flows within each sequence and the interrelation between flows at all pairs of sites should be maintained. In very simple terms the correlation between two sets of paired observations is a measure of the linear interrelation between the two sets. If X and Y are two series of observations, the correlation between the two series is described by the correlation matrix $\hat{\rho}$. The diagonal elements are called autocorrelations and the off-diagonals are called cross correlations. Cross correlations describe how X relates to Y, $\hat{\rho}_{xy}$, and how Y relates to X, $\hat{\rho}_{yx}$. Autocorrelations however, describe how X and Y relate to themselves, $\hat{\rho}_{xx}$ and $\hat{\rho}_{yy}$. Thus the correlation matrix for two sites is

$$\hat{\rho} = \begin{vmatrix} \hat{\rho}_{xx} & \hat{\rho}_{xy} \\ \hat{\rho}_{yx} & \hat{\rho}_{yy} \end{vmatrix}$$

When the X's and Y's are observations at the same time step, the diagonals or autocorrelations are always equal to one. Similarly $\hat{\rho}_{xy}$ is always equal to $\hat{\rho}_{yx}$. The same theory can be used to expand this definition to n sites. Thus for the case of nine sites, the lag-zero or paired observations at the same time step, correlation matrix is 9 x 9.

If one set of observations or time series is offset by one or more time steps from the other, $\hat{\rho}_{xy}$ is no longer equal to $\hat{\rho}_{yx}$. Similarly for the diagonal elements, when series X_t is offset by k time steps or lags, X_t correlated with X_{t+k} is no longer equal to unity. Thus, in general, the correlation matrix is estimated by

$$\hat{\rho}_{xy(k)} = \frac{\sum_{t=1}^{n-k} X_t \sum_{t=1}^{n-k} Y_{t+k} - \frac{1}{n-k} \sum_{t=1}^{n-k} X_t \sum_{t=1}^{n-k} Y_{t+k}}{\sqrt{\left(\sum_{t=1}^{n-k} X_t^2 - \frac{1}{n-k}\left(\sum_{t=1}^{n-k} X_t\right)^2\right)\left(\sum_{t=1}^{n-k} Y_{t+k}^2 - \frac{1}{n-k}\left(\sum_{t=1}^{n-k} Y_{t+k}\right)^2\right)}} \quad \text{for } k=0,1,2\ldots,m \quad (1)$$

Where $\hat{\rho}_{xy(k)}$ is an estimate of the correlation between X and Y, k time steps apart. If applied to streamflow records, X_t is the flow at site X at time t, Y_{t+k} is the flow at site Y at time t+k, n is the total length of each series, and m is the maximum number of time lags to be considered where $m < n-k$. For the particular case of autocorrelations, the Y_{t+k}'s are simply replaced by X_{t+k}'s.

Thus,

$$\hat{\rho}_{xy(k)} = \frac{\sum_{t=1}^{n-k} X_t X_{t+k} - \frac{1}{n-k}\left(\sum_{t=1}^{n-k} X_t\right)\left(\sum_{t=1}^{n-k} X_{t+k}\right)}{\sqrt{\left(\sum_{t=1}^{n-k} X_t^2 - \frac{1}{n-k}\left(\sum_{t=1}^{n-k} X_t\right)^2\right)\left(\sum_{t=1}^{n-k} X_{t+k}^2 - \frac{1}{n-k}\left(\sum_{t=1}^{n-k} X_{t+k}\right)^2\right)}} \quad \text{for } k=0,1,2\ldots,m \quad (2)$$

Tables 2-4 show the lag-zero, lag-one, and lag-two sample correlation matrices computed for the nine streamflow records used in this study.

Table 2.--Lag-zero correlation matrix for nine stations in the Juniata River basin

Site	1	2	3	4	5	6	7	8	9
1	1.000	0.864	0.955	0.982	0.960	0.942	0.970	0.919	0.963
2	.864	1.000	.942	.905	.786	.760	.858	.803	.864
3	.955	.942	1.000	.983	.921	.900	.964	.922	.962
4	.982	.905	.983	1.000	.947	.937	.975	.939	.983
5	.960	.786	.921	.947	1.000	.977	.960	.945	.952
6	.942	.760	.900	.937	.977	1.000	.967	.968	.959
7	.970	.858	.964	.975	.960	.967	1.000	.968	.985
8	.919	.803	.922	.939	.945	.968	.968	1.000	.971
9	.963	.864	.962	.983	.952	.959	.985	.971	1.000

Table 3.--Lag-one correlation matrixfor nine stations in the Juniata River basin

Site	1	2	3	4	5	6	7	8	9
1	0.370	0.356	0.344	0.421	0.310	0.297	0.309	0.303	0.409
2	.319	.380	.298	.374	.244	.278	.284	.303	.391
3	.337	.333	.299	.382	.279	.290	.277	.300	.388
4	.332	.331	.302	.385	.275	.274	.269	.287	.380
5	.349	.332	.335	.403	.301	.277	.287	.290	.387
6	.330	.312	.314	.379	.281	.256	.267	.272	.358
7	.349	.344	.322	.394	.291	.284	.285	.285	.381
8	.306	.293	.281	.355	.274	.269	.248	.273	.344
9	.309	.303	.287	.362	.266	.262	.251	.269	.358

Table 4.--Lag-two correlation matrix for nine stations in the Juniata River basin

Site	1	2	3	4	5	6	7	8	9
1	0.104	0.096	0.109	0.173	0.098	0.117	0.070	0.132	0.155
2	.053	.068	.048	.110	.048	.095	.057	.111	.115
3	.099	.072	.098	.166	.113	.141	.085	.160	.161
4	.105	.083	.106	.172	.107	.132	.081	.153	.157
5	.107	.085	.117	.186	.108	.148	.085	.173	.178
6	.187	.147	.198	.264	.177	.216	.163	.241	.246
7	.163	.126	.167	.234	.157	.185	.134	.208	.216
8	.179	.143	.190	.252	.156	.192	.155	.232	.231
9	.150	.120	.159	.218	.144	.167	.121	.196	.197

HYDROLOGIC PERSISTENCE AND THE HURST COEFFICIENT

Hydrologic persistence is the tendency for high flows to follow high flows and low flows to follow low flows. It can be attributed to storage in the atmosphere, on the land surface, or underground (Wallis and Matalas, 1971a), or to persistence in the meteorologic processes which produce rainfall in a given area.

Persistence is often described by the structure of serial dependence, or correlation of a streamflow sequence. A picture of this dependence is given by a correlogram, ρ_k versus k, where ρ_k is the autocorrelation at k time steps apart. Problems arise in the use of correlograms, however, because large sampling errors often exist in estimating ρ_k from a small sample.

Another measure of persistence was included in studies of many natural phenomena by H. E. Hurst (1951). Hurst's analysis of nearly 900 natural sequences showed that the range of cumulative departures (R) from the mean (X) for a sequence of N observations took the form

$$R/S \approx N^h$$

where S is the standard deviation of the sample and h is a coefficient. Hurst estimated h by equating R/S to $(N/2)^h$, computing h for each sample. Thus if K is an estimate of h, then

$$K = \log (R/S)/\log (N/2) \qquad (3)$$

Hurst found that for nearly 900 sequences, the mean and standard deviation of K were 0.73 and 0.08, respectively. For purely random processes, Hurst (1951) and Feller (1951) independently showed that h = 1/2.

The tendency for streamflows and other natural time series to have values of h between 1/2 and 1, has become known as the Hurst phenomenon (Matalas, 1971; Wallis and Matalas, 1970, 1971a, 1971b; O'Connell, 1971).

Several estimators of the Hurst coefficient have been proposed in addition to Hurst's K described above. Wallis and Matalas (1971b) presented a method to estimate another index of the Hurst phenomena described by Mandelbrot and Wallis (1969) as H, and referred to as the G Hurst procedure. The method, as programmed by Slack (personal communication, 1977), is as follows.

The estimate H is the slope of the linear least squares fit of the relation between the log of average R/S and log n for replicate sets of subsequences of various lengths n. A sequence of flows of length n is divided into a set of m_k nonoverlapping subsequences of a specified length n_k. The ratio R/S is calculated for each subsequence, averaged, then regressed against the log n_k. The procedure is repeated for various values of n_k. H is the slope of this line.

Wallis and Matalas (1971b) found that in general K>H, but that the variance of K is smaller than that for H. They also found that, although H and K are both biased estimators of the asymptotic value of h, the bias for K is larger than for H. Thus a trade-off must be made between bias and variance. Both estimates have been used here and presented in table 5 for the historical data. H and K have been computed using the period of record at each site as well as the period of concurrent record 1945-74.

Table 5.--Estimates of the Hurst coefficient from historical data

	H - Hurst		K - Hurst	
Site Number	Period of Record	1945-74	Period of Record	1945-74
1	0.4378	0.8944	0.6033	0.7997
2	.9451	.9451	.8116	.8116
3	.6285	.8036	.6818	.7963
4	.8744	.8229	.7854	.8120
5	.7179	.7524	.7336	.7991
6	.5172	.6843	.7389	.7768
7	.6459	.8041	.6939	.7810
8	.6355	.7099	.7037	.7679
9	.5457	.7706	.7427	.8205

The autocorrelation structure of Markov or autoregressive models fails to preserve the Hurst phenomenon. The autocorrelation function for a first-order Markov process is

$$\rho_k = (\rho_1)^k$$

where k is the number of time lags, ρ_1 is the lag 1 autocorrelation, and ρ_k is the lag k autocorrelation. It can be readily seen that ρ_k approaches zero quickly, thus the "effective memory" is short (O'Connell, 1971).

In a subsequent section it will be shown that the autocorrelation of the ARIMA (1,0,1) model does not suffer from this drawback if its model parameters, ϕ and θ, are sufficiently large.

DESCRIPTION AND CALIBRATION OF THE ARIMA (1,0,1) MODEL

An autoregressive-integrated-moving average, ARIMA, process is a powerful, but general, family of models proposed by Box and Jenkins (1968,1970). It is "capable of describing virtually any form of stationary or nonstationary behavior in time series" (O'Connell, 1971).

O'Connell (1974) presents a thorough development of the ARIMA process. The particular model that he proposes for use in generating annual streamflow sequences is the ARIMA (1,0,1) process. The single site model is defined as

$$X_t - \phi_1 X_{t-1} = \varepsilon_t - \theta_1 \varepsilon_{t-1} \qquad (4)$$

where X_t and X_{t-1} are flows at times t and t_1, ε_t and ε_{t-1} are random variates at times t and t_1, and ϕ_1 and θ_1 are model parameters. The model has an autocorrelation function defined by

$$\rho_1 = \frac{(1-\phi\theta)(\phi-\theta)}{1+\theta^2 - 2\phi\theta} \qquad \text{for } k = 1$$

$$\rho_k = \phi\rho_{k-1} \qquad \text{for } k \geq 2$$

where ϕ and θ are the model parameters describing the strength of the autoregressive and moving average aspects of the model, respectively (Box and Jenkins, 1970). The absolute value of the autocorrelation decays exponentially from ρ_1 onward. If both ϕ and θ are positive and $\phi > \theta$, ρ_1 and ρ_k are always positive. For a fixed θ, the autocorrelation dies out more slowly as θ is increased. As a result, the model has a desirable blend of high and low frequency characteristics important for use in applications to streamflow generation (O'Connell, 1971). O'Connell (1974) also presented a more complex, multisite ARIMA (1,0,1) model. A brief outline of that model is presented below.

The process is formulated as

$$\underline{x}(t) - \underline{A}\underline{x}(t-1) = \underline{B}\varepsilon(t) - \underline{C}\varepsilon(t-1) \qquad (5)$$

where $\underline{x}(t)$ and $\underline{x}(t-1)$ are m x 1 matrices. Their elements are $x_i(t) = (X_i(t) - \mu_i)$ and $x_i(t-1) = (X_i(t-1) - \mu_i)$ respectively, for i = 1,2,...., m, where m is the number of sites, and $X_i(t)$ and $X_i(t-1)$ are the annual flows at site i and times t and t-1, respectively. $\underline{A}$, $\underline{B}$, and $\underline{C}$ are (m x m) matrices of coefficients, or model parameters, and $\varepsilon(t)$ and $\varepsilon(t-1)$ are vectors of independent random variables at times t and (t-1) respectively. To define matrices $\underline{A}$, $\underline{B}$, and $\underline{C}$, matrices $\underline{M}_0$, $\underline{M}_1$, and $\underline{M}_2$ are required. $\underline{M}_0$ $\underline{M}_1$, and $\underline{M}_2$ are, respectively, the lag-zero, lag-one, and lag-two covariance matrices. If, however, $x_i(t)$ is redefined as

$$x_i(t) = \frac{X_i(t) - \mu_i}{\sigma_i} \qquad (6)$$

where μ_i and σ_i are the mean and standard deviation of the X_i's, respectively, then $\underline{M}_0$, $\underline{M}_1$, and $\underline{M}_2$ become the lag-zero, lag-one, and lag-two cross correlation matrices.

Thus,

$$\underline{M}_1 = \begin{vmatrix} \rho_{11}(1) & \rho_{12}(1) & \cdots\cdots & \rho_{1_m}(1) \\ \rho_{21}(1) & \rho_{22}(1) & \cdots\cdots & \rho_{2_m}(1) \\ \vdots & & & \\ \rho_{m_1}(1) & \rho_{m_2}(1) & \cdots\cdots & \rho_{mm}(1) \end{vmatrix}$$

where $\rho_{11}(1)$, $\rho_{22}(1)$,.... $\rho_{mm}(1)$ are the lag-one autocorrelations at sites 1, 2, ...,m and $\rho_{ij}(1)$ and $\rho_{ji}(1)$ are the lag-one cross correlations between sites i and j, and j and i, respectively.

Equation (5) can be rewritten as

$$\underline{x}(t) = \underline{A}\underline{x}(t-1) + \underline{B}\varepsilon(t) - \underline{C}\varepsilon(t-1) \qquad (7)$$

By postmultiplying equation 7 by $\underline{x}(t)^T$ and taking expected values we get

$$E[\underline{x}(t)\underline{x}(t)^T] = \underline{A}E[\underline{x}(t-1)\underline{x}(t)^T] + \underline{B}E[\underline{\varepsilon}(t)\underline{x}(t)^T] - \underline{C}E[\underline{\varepsilon}(t-1)\underline{x}(t)^T] \qquad (8)$$

From the definition of cross correlations

$$\underline{M}_0 = E[\underline{x}(t)\underline{x}(t)^T] \qquad (9)$$

$$\underline{M}_1 = E[\underline{x}(t)\underline{x}(t-1)^T] \qquad (10)$$

$$\underline{M}_2^T = E[\underline{x}(t)\underline{x}(t-2)^T] \qquad (11)$$

The transpose of equations 10 and 11 are

$$\underline{M}_1^T = E[\underline{x}(t-1)\underline{x}(t)^T] \tag{12}$$

$$\underline{M}_2^T = E[\underline{x}(t-2)\underline{x}(t)^T] \tag{13}$$

The terms on the right hand side of equation 8 can be rewritten individually as (O'Connell, 1974)

$$\underline{A}E[\underline{x}(t-1)\underline{x}(t)^T] = \underline{A}\underline{M}_1^T$$

$$\underline{B}E[\underline{\varepsilon}(t)\underline{x}(t)^T] = \underline{B}\underline{B}^T$$

$$\underline{C}E[\underline{\varepsilon}(t-1)\underline{x}(t)^T] = \underline{C}\underline{B}^T\underline{A}^T - \underline{C}\underline{C}^T$$

Equation 8 can now be rewritten as

$$\underline{M}_0 = \underline{A}\underline{M}_1^T + \underline{B}\underline{B}^T - \underline{C}\underline{B}^T\underline{A}^T + \underline{C}\underline{C}^T$$

or

$$\underline{B}\underline{B}^T + \underline{C}\underline{C}^T = \underline{M}_0 - \underline{A}\underline{M}_1^T + \underline{C}\underline{B}^T\underline{A}^T \tag{14}$$

By postmultiplying equation 7 by $\underline{x}(t-1)^T$ it can be shown that

$$\underline{M}_1 = \underline{A}\underline{M}_0 - \underline{C}\underline{B}^T$$

or

$$\underline{C}\underline{B}^T = \underline{A}\underline{M}_0 - \underline{M}_1 \tag{15}$$

Similarly, if equation 7 is postmultiplied by $\underline{x}(t-2)^T$

$$\underline{M}_2 = \underline{A}\underline{M}_1$$

Thus

$$\underline{A} = \underline{M}_2\underline{M}_1^{-1} \tag{16}$$

Now, substituting for $\underline{A}$ and CB^T in equation 14

$$\underline{B}\underline{B}^T + \underline{C}\underline{C}^T = \underline{M}_0 - \underline{M}_2\underline{M}_1^{-1}\underline{M}_1^T + \underline{M}_2\underline{M}_1^{-1}\underline{M}_0\underline{M}_1^{T-1}\underline{M}_2^T - \underline{M}_1\underline{M}_1^T\underline{M}_2^T \tag{17}$$

The entire right side of equation 17 can be compiled into a symmetric matrix which will be called $\underline{S}$. Thus

$$\underline{S} = \underline{B}\underline{B}^T + \underline{C}\underline{C}^T \tag{18}$$

Rewriting equation 15 yields

$$\underline{T} = \underline{C}\underline{B}^T \tag{19}$$

where $\underline{T}$ is a nonsymmetric matrix in terms of $\underline{M}_0$, $\underline{M}_1$, and $\underline{M}_2$. Because $\underline{M}_0$, $\underline{M}_1$, and $\underline{M}_2$ do not provide sufficient conditions to define $\underline{B}$ and $\underline{C}$, a lower triangular form of either $\underline{B}$ or $\underline{C}$ may be assumed in order to preserve $\underline{M}_0$, $\underline{M}_1$, and $\underline{M}_2$ as the correlation matrices.

To obtain real-valued coefficients for $\underline{B}$ and $\underline{C}$, the matrices $\underline{S}$ and $\underline{T}$ must satisfy certain conditions. These conditions can be specified from equations derived from equations 18 and 19 as

$$(\underline{B} + \underline{C})\ (\underline{B} + \underline{C})^T = \underline{B}\underline{B}^T + \underline{C}\underline{C}^T + \underline{C}\underline{B}^T + \underline{B}\underline{C}^T$$

$$= \underline{S} + \underline{T} + \underline{T}^T \qquad (20)$$

and

$$(\underline{B} - \underline{C})(\underline{B} - \underline{C})^T = \underline{S} - \underline{T} - \underline{T}^T \qquad (21)$$

According to O'Connell (1974), matrices $(\underline{B} + \underline{C})$ and $(\underline{B} - \underline{C})$ are real valued if matrices $(\underline{S} + \underline{T} + \underline{T}^T)$ and $(\underline{S} - \underline{T} - \underline{T}^T)$ are positive semidefinite, which they are.

An iterative solution of equations 18 and 19 is necessary because no analytical solution has been found that permits preservation of the correlation matrices (O'Connell, 1974).

If equation 19 is rewritten as

$$\underline{C} = \underline{T}(\underline{B}^T)^{-1} \qquad (22)$$

then

$$\underline{C}^T = \underline{B}^{-1}\underline{B}^T$$

When these are substituted into equation 18 and terms rearranged

$$\underline{B}\underline{B}^T + \underline{T}(\underline{B}\underline{B}^T)^{-1}\underline{T}^T = \underline{S} \qquad (23)$$

By letting $\underline{B}\underline{B}^T = \underline{U}$, an iterative solution for $\underline{U}$ can be developed by

$$\underline{U}_j = \underline{S} - \underline{T}\underline{U}_{j-1}^{-1}\underline{T}^T \qquad (24)$$

where $\underline{U}_j$ is the value of $\underline{U}$ on the jth iteration. To start the iterative process, it is convenient to let $\underline{U}$ be the identity matrix. Thus, each iteration, in theory, provides a closer approximation to the solution, $\underline{B}$. O'Connell (1974) states that in certain cases $\underline{U}$ may not converge even if both $(\underline{S} + \underline{T} + \underline{T}^T)$ and $(\underline{S} - \underline{T} - \underline{T}^T)$ are positive semidefinite.

Once a solution is obtained for $\underline{B}$, $\underline{C}$ is determined by using equation 22.

With the assumptions made above, $\underline{A}$ will be a diagonal matrix, $\underline{B}$ will be a lower triangular matrix, and $\underline{C}$ will be a full (m x m) matrix. In addition, $\underline{M}_0$ and $\underline{M}_1$ will be preserved, as will the diagonal elements of the $\underline{M}_2$ matrix.

Estimation of Model Parameters

The model parameters, matrices $\underline{A}$, $\underline{B}$, and $\underline{C}$, are estimated using the formulation in the previous section. Computer subroutines for the iterative estimation procedure were developed by O'Connell (personal communication, 1977). To simplify model development, only three of the nine sites were used. The first three gaging stations listed in table 1 were arbitrarily selected. Parameter estimation using the iterative procedure resulted in the following $\underline{A}$, $\underline{B}$, and $\underline{C}$ matrices:

$$\underline{A} = \begin{bmatrix} .2806 & 0.0 & 0.0 \\ 0.0 & .1783 & 0.0 \\ 0.0 & 0.0 & .3280 \end{bmatrix}$$

$$\underline{B} = \begin{bmatrix} .8969 & 0.0 & 0.0 \\ .6981 & .4041 & 0.0 \\ .8437 & .2316 & .1422 \end{bmatrix}$$

and

$$\underline{C} = \begin{bmatrix} -.0991 & -.1108 & .2371 \\ -.1841 & -.1807 & .4708 \\ -.0264 & -.0140 & .3804 \end{bmatrix}$$

A check was made to determine the correlation matrices preserved using these parameters. They are

$$\underline{M}_0 = \begin{bmatrix} 1.0123 & .8870 & .9720 \\ .8870 & 1.0438 & .9729 \\ .9720 & .9729 & 1.0226 \end{bmatrix}$$

$$\underline{M}_1 = \begin{bmatrix} .3730 & .3629 & .3483 \\ .3232 & .3876 & .3037 \\ .3424 & .3432 & .3068 \end{bmatrix}$$

$$\underline{M}_2 = \begin{bmatrix} .1047 & .1018 & .0978 \\ .0576 & .0691 & .0541 \\ .1123 & .1126 & .1006 \end{bmatrix}$$

One feature of the preserved correlations that appears to violate the numerical constraints placed on values of a correlation coefficient is the diagonal elements of the $\underline{M}_0$ matrix. Correlation coefficients cannot, by definition, be outside the range -1.0 to +1.0. The diagonal elements of $\underline{M}_0$, although greater than 1.0, simply reflect numerical rounding errors in converting the $\underline{A}$, $\underline{B}$, and $\underline{C}$ matrices back to $\underline{M}_0$, $\underline{M}_1$, and $\underline{M}_2$.

It should be noted that the $\underline{M}_0$ and $\underline{M}_1$ matrices and the diagonal of the $\underline{M}_0$ matrix are preserved quite well (see tables 2-4), but biased upward a small amount. The off-diagonal elements of $\underline{M}_2$, although not specifically preserved by the parameter estimation technique, do resemble the off-diagonal elements of the historical $\underline{M}_2$ matrix.

When parameter estimation was extended to the nine-site problem, two types of serious problems were encountered. Each type, by itself, prevents estimation of a valid set of model parameters.

The first type of problem was encountered when all nine sites were included in the parameter estimation. Matrix $\underline{A}$ was estimated using equation 16. Diagonal elements of a lower triangular form of $\underline{B}$, b_{ii}, were calculated using:

$$b_{ii} = \sqrt{\underline{u}_{ii}} \qquad \text{for } i = 1 \qquad (25)$$

$$b_{ii} = \sqrt{\underline{u}_{ii} - \sum_{j=1}^{i} b_{ij}} \qquad \text{for } i = 2,3,\ldots,m \qquad (26)$$

where the b's and u's are elements of matrices $\underline{B}$ and $\underline{U}$. The computer subroutines used in the calculations set negative values of b_{ii}^2, equal to zero to prevent taking the square root of a negative number. The result of a zero on the diagonal of $\underline{B}$ is that all elements in the respective column are then zero. During the calculation of $\underline{B}$ from $\underline{U}$, the squares of several diagonal elements of matrix $\underline{B}$ were negative, thus calculations of subsequent diagonal and offdiagonal elements of $\underline{B}$ were contaminated as were the elements of the $\underline{C}$ matrix.

Examination of the $\underline{U}$ matrix, used in solving for $\underline{B}$ in the 9-site case, showed that a diagonal element much smaller than off-diagonal elements in the corresponding row was the cause of the diagonal element of $\underline{B}$ being equated to zero. Because elements of $\underline{U}$ are a function of $\underline{M}_0$, $\underline{M}_1$, and $\underline{M}_2$, an abnormally small element in $\underline{U}$ is caused by the estimated cross correlations. Although it is difficult to precisely trace an element of $\underline{U}$ back to $\underline{M}_0$, $\underline{M}_1$, or $\underline{M}_2$, highly correlated historical data (see tables 2-4) may be the cause of the problem, especially when all the cross correlations are nearly equal. Use of double precision calculations in the above analyses did not materially change the results.

The second type of problem is common to many iterative solution techniques, namely, under certain conditions, the solution does not converge. Even though both matrices $(\underline{S} + \underline{T} + \underline{T}^T)$ and $(\underline{S} - \underline{T} - \underline{T}^T)$ are positive semidefinite, the solution for $\underline{U}$ can continue to oscillate from iteration to iteration. O'Connell (1974) states that when convergence cannot be obtained for $\underline{U}$ in equation 24, a damping coefficient λ can be inserted such that

$$\underline{U}_j = \underline{S} - \lambda \underline{T} \underline{U}_{j-1}^{-1} \underline{T}^T \qquad (27)$$

where $0.0 < \lambda < 1.0$. He points out, however, that the equations being solved are now

$$\underline{B}\underline{B}^T + \lambda \underline{C}\underline{C}^T = \underline{S} \qquad (28)$$

and

$$\underline{C}\underline{B}^T = \underline{T} \qquad (29)$$

Although $\lambda < 1.0$ permits convergence, the equations being solved are not exactly the same as equations 18 and 19.

The ARIMA (1,0,1) is a conditional model. Slack (1973) states that "a conditional model presents not only an operational problem by occasionally rejecting historical sequences, but also a philosophical dilemma by occasionally rejecting its own produce." The inability to compute a set of model parameters from observed data, that by design preserves correlation structure is what Slack (1972, 1973) calls self-denial. The problems found here are self-denial if the ARIMA (1,0,1) model truly represents the annual streamflow processes. Thus a solution with $\lambda \neq 1.0$ may be possible if a solution cannot be obtained with $\lambda = 1.0$. Several numerical experiments were run using $\lambda = 0.90$. Some solutions were affected only slightly. In many cases, however, using $\lambda = 0.90$ produced $\underline{A}$, $\underline{B}$, and $\underline{C}$ matrices that preserved correlation matrices that had elements greater than 1.2 or 1.3. These elements were not the result of minor roundoff errors, but were due to incorrect elements in the parameter matrices, which in turn were caused by $\lambda \neq 1.0$.

The experiments led to the conclusion that λ can take on values less than 1.0 and provide a usable solution. If a solution does result, however, all resulting parameters must be carefully checked. The solutions should not be used if the check indicates a violation of correlation constraints.

Because parameter estimates for the 9-site problem could not be calculated, annual streamflow generation was carried out for the 3-site case discussed earlier. Three-site generation is sufficient to proceed to the subsequent phase of this project.

Generation of Annual Flows

Ten sequences of 30 years of annual flows were generated using equation 5 with appropriate values of matrices A, B, and C presented earlier, normally distributed random numbers with zero mean and unit variance, and means and standard deviations of observed flows. These 30-year sequences were generated for ultimate use in the disaggregation process, which will be discussed later. It was felt that if the characteristics of the generated 30-year sequences resembled the 30-year historical flows used here, then the model could be used to generate long-term annual flow sequences. The mean and standard deviation of each sequence were computed and are presented in table 6. For even relatively short sequences, mean and standard deviation are preserved. Preservation of the mean and standard deviation of observed flows requires that the mean of standard deviations of the generated sequences, divided by the square root of the generated sequence length, be equal to the standard deviation of the means of the generated sequences. Using site 1 data from table 6 as an example, the mean and standard deviation would be preserved if $103/\sqrt{30}$ equalled 18.8. Some simple calculations show that this condition is not strictly met for the data presented in table 6. The distribution of the means is skewed. This situation arises because of the short generated sequence length, 30 years, and the small number of generated sequences, 20. Examination of the generated 100-year sequences described in a subsequent section shows that the statistics are well preserved, as expected.

The lag-zero, lag-one, and lag-two cross correlations matrices were computed for each generated 30-year sequence. Because each sequence has three 3 x 3 correlation matrices, only one set is presented in table 7. These values should be compared to those tables 2-4.

Estimates of the Hurst coefficient, for both the 30-year historical and for 20 generated 30-year sequences were computed using K and H (G-procedure). The summary in table 8 shows, as expected from the results presented by Wallis and Matalas (1970), that estimates of h using K produces values generally lower in variance than H. For these relatively short sequences, however, H was generally higher than K, in contrast to the results obtained by Wallis and Matalas (1971b).

Table 6.--Mean and standard deviation of twenty generated 30-year sequences.

	Site					
	1		2		3	
Sequence	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Historical	400	108	75	20	375	86
1	353	114	67	22	336	90
2	416	87	76	21	386	81
3	393	116	75	24	372	97
4	412	130	74	22	379	99
5	342	117	69	19	332	88
6	387	81	73	19	363	68
7	387	89	74	19	366	78
8	336	95	61	18	323	73
9	414	100	79	19	390	80
10	424	85	79	20	392	80
11	456	91	85	18	419	75
12	382	93	75	18	364	74
13	442	99	80	19	404	79
14	359	128	65	24	340	102
15	313	102	60	20	308	85
16	364	110	70	19	348	84
17	423	108	81	18	398	82
18	391	116	75	22	369	93
19	412	101	75	17	383	76
20	423	95	79	18	391	75
Summary						
Mean	391	103	74	20	368	83
Standard deviation	38	14	7	2	29	9

Table 7.--Estimated lag-zero, -one, and -two cross-correlation matrices for one generated 30-year sequence.

Lag-Zero

Site Number	1	2	3
1	1.000	.879	.962
2	.879	1.000	.934
3	.962	.934	1.000

Lag-One

Site Number	1	2	3
1	.454	.241	.363
2	.182	.649	.101
3	.506	.573	.409

Lag-Two

Site Number	1	2	3
1	.076	-.153	.127
2	-.338	.566	-.371
3	.051	-.148	.077

Table 8.--Summary of Hurst coefficients computed for the 30-year historical sequence and the twenty 30-year generated sequences.

Sequence	Site 1		Site 2		Site 3	
	H	K	H	K	H	K
Historical	0.8944	0.7997	0.9451	0.8116	0.8036	0.7963
Mean of generated	.7752	.7474	.7211	.7274	.7851	.7273
Standard deviation of generated	.2845	.0748	.3104	.1049	.2603	.0862

DESCRIPTION AND CALIBRATION OF THE DISAGGREGATION MODEL

Because the generated monthly series were the level of flows needed here, a technique was required to obtain monthly flows from the annual flows generated using the ARIMA (1,0,1) process. Several such models, referred to as disaggregation models or processes, were examined. Included among these were the models proposed by Harms and Campbell (1967), Tao and Delleur (1976), Young and Jettmar (1976), and Valencia and Schaake (1972, 1973). The Valencia - Schaake model was selected for use because of its ease of application, flexibility, and its overall capability. The notation used throughout this section is the same as the notation used by Valencia and Schaake (1972, 1973) and should not be confused with the ARIMA (1,0,1) model notation.

Description of the Valencia-Schaake Model

The model has a simple form,

$$\underline{Y} = \underline{A}\underline{X} + \underline{W} \tag{30}$$

where $\underline{Y}$ is an (n x 1) vector of correlated random variables, $\underline{X}$ is an (m x 1) vector of correlated random variables, A is an (n x m) matrix of coefficients, and W is an (n x 1) vector of correlated random variables independent of X. Let $\underline{X}$ be a vector of annual flows, and Y be a vector of seasonal flows, where the x_i's and y_i's are transformed so as to be normally distributed and have zero mean if the data are skewed. The model can be rewritten in a form equivalent to the model presented by Matalas (1967), if $Y = Q_{t+1}$ and $X = Q_t$

$$\underline{Q}_{t+1} = \underline{A}\underline{Q}_t + \underline{B}\underline{V}_{t+1}$$

where

$$\underline{B}\underline{V}_{t+1} = \underline{W}$$

Rewriting equation 30 for simpler use

$$\underline{Y}_t = \underline{A}\underline{X}_t + \underline{B}\underline{V}_{t+1} \tag{31}$$

where $\underline{Y}_t$ is an (n x 1) vector of seasonal flows, $\underline{X}_t$ is an (m x 1) vector of annual flows, m is the number of sites being considered, n equals 4m, $\underline{V}_t$ is a vector of independently distributed standard normal deviates and $\underline{A}$ and $\underline{B}$ are coefficient matrices. In like fashion, $\underline{Y}_t$ can be a vector of monthly flows and $\underline{X}_t$ a vector of season flows. Vector and matrix sizes would be adjusted accordingly.

The coefficient matrices are based on historical data, and are computed in such a manner that the generated flows $\underline{Y}_t$ resemble the historical values of $\underline{Y}$ according to some prescribed resemblance criteria. The criterion generally specified is preservation of mean, standard deviation, and cross correlation.

Valencia and Schaake (1972, 1973) showed that the above properties of the historical flows can be preserved by specifying that

$$\underline{A} = \underline{S}_{yx}\,\underline{S}_{xx}^{-1} \tag{32}$$

$$\underline{B}\underline{B}^T = \underline{S}_{yy} - \underline{S}_{yx}\,\underline{S}_{xx}^{-1}\,\underline{S}_{xy} \tag{33}$$

where $\underline{S}_{xx}$, $\underline{S}_{xy}$, $\underline{S}_{yx}$, and $\underline{S}_{yy}$ are matrices equal, respectively to $E[XX^T]$, $E[XY^T]$, $E[YX^T]$, and $E[YY^T]$, where X and Y have zero mean. Note that mean and standard deviation are preserved here regardless of the underlying multivariate distribution of $\underline{X}$ and $\underline{Y}$. If r observations of $\underline{X}$ and $\underline{Y}$ are considered, and without loss of generality consider that $\underline{X}$ and $\underline{Y}$ have zero mean and unit variance, then

$$\underline{S}_{xx} = \frac{1}{r}\,\underline{X}\underline{X}^T$$

$$\underline{S}_{xy} = \frac{1}{r}\,\underline{X}\underline{Y}^T$$

$$\underline{S}_{yx} = \frac{1}{r}\,\underline{Y}\underline{X}^T$$

$$\underline{S}_{yy} = \frac{1}{r}\,\underline{Y}\underline{Y}^T$$

Equations 32 and 33 remain valid using these new definitions. $\underline{X}$ and $\underline{Y}$ are now (m x r) and (n x r) matrices.

The model thus described also insures that the four seasons sum to the annual flow from which they were computed or the three months sum to the season from which they were computed. This two-phase disaggregation process is used to reduce computer storage requirements and subsequently increase the ease of use (Valencia and Schaake, 1973).

Under the structure presented here, the model is unable to correlate the last season of one year with the first season of the next year and the last month of one season with the first month of next season. Linking to the past at different levels of disaggregation through correlation is referred to here as "wrap-around" correlation.

Modifications to the Valencia-Schaake Model

Wrap-around correlation

Mejia and Rousselle (1976) presented a modification to the Valencia-Schaake model that overcomes the wrap-around correlation problem. Their method adds the product of two matrices to the right side of equation 29. One is a matrix of coefficients and the other is a matrix whose elements are made up of the last elements at the particular disaggregation level. For example, when X_t is the annual flow at year t, and the Y_t's are seasonal flows for year t, the added matrix element is the last season of year t-1.

A simplified computational procedure was developed in conjunction with Schaake (personal communication, 1977). The technique augments each row of the $\underline{X}$ and $\underline{A}$ matrices. Thus when the elements of $\underline{X}$ represent annual flows, the last season of r + 1 year is inserted in $\underline{X}$ as an additional column of row r. By so doing, one additional column is added to each row of $\underline{X}$ for each site used. When $\underline{X}$ is the matrix of seasonal values, it is augmented by the flows for the last month of the previous season, again lengthening each row by one element for each site. These additions to the $\underline{X}$ matrix, at either level of disaggregation, also require corresponding expansions in the coefficient matrices. For example, with $\underline{X}$ equal to a matrix of annual flows in the original version of the Valencia-Schaake model, $\underline{A}$ is a (4m x m) matrix and $\underline{B}$ is a (4m x 4m) matrix. In the modified version proposed here, $\underline{A}$ becomes a (4m x 2m) matrix and $\underline{B}$ remains the same. When $\underline{X}$ equals seasonal flows in the original version, $\underline{A}$ is a (3m x m x 4) matrix and $\underline{B}$ is a (3m x 3m x 4) matrix. Under the new scheme, $\underline{A}$ takes the dimensions (3m x 2m x 4) and $\underline{B}$ remains the same. The final results are identical to the Mejia and Rouselle version, but are simpler to compute.

Transformation of data

One of the assumptions for use of the Valencia-Schaake model was that all data were transformed to be normally distributed with zero mean. A further condition was that the data also have unit variance. The transform required to obtain zero mean and unit standard deviation is

$$\hat{X} = \frac{X - \mu}{\sigma} \qquad (34)$$

where $\hat{X}$ is the transformed value of X, and μ and σ are the mean and standard deviation, respectively, of X. If the cumulative probability distribution, when plotted on normal probability paper, is nonlinear then a data transformation is required. Two transformations commonly used are logarithmic and square-root transforms. Both have the desirable property that the reverse transform always results in a positive number.

Cumulative probability plots of annual data used in this study showed that the annual data reasonably could be assummed to be normally distributed. The seasonal and monthly data, however, were both found to be highly skewed. Neither logarithmic nor square-root transforms were, by themselves, able to remove the skew. The transform required to linearize the cumulative probability distribution is

$$\hat{X} = \sqrt{\ln X} \qquad (35)$$

where $\hat{X}$ is the transformed X. This transformation is obviously more severe than those popularly used. However, such a transformation must be used to approach normality and to satisfy the restrictions of the disaggregation model. Data thus transformed were then standardized using equation 34. Estimates of the parameter matrices $\underline{A}$ and $\underline{B}$ in equation 31 can now be computed using the transformed data. The cross-correlation structure of flows generated from transformed data is altered and may not be preserved.

Other modifications to the Valencia-Schaake Model

Because the same transformation may not normalize the data at all levels of disaggregation, the Valencia-Shaake model was also modified to permit a different transform at each level.

Whenever data are transformed, streamflow generation occurs in the transformed space. However it is no longer assured that seasonal flows will sum to the annual flow from which they were disaggregated or that monthly flows will add to the seasonal flow from which they were computed. The respective sums will be properly preserved in the transformed space, but probably lost in the reverse transformation. This is true simply because the transforms are nonlinear.

Loss of this continuity may be practically or esthetically displeasing. As a result, an option has been added to the Valencia-Schaake model that forces a real number flow balance after the reverse transform by distributing flows on a percentage basis. In the streamflow generation performed as part of this study, however the mass differences of the flows when transformed back to real numbers were found negligible without balancing; thus balancing was not used.

Use of the balancing option could also, under adverse conditions, cause a contamination of the cross-correlation structure of the generated flows. It should therefore be used with caution.

Disaggregation of Seasonal and Monthly Flows from Generated Annual Flows

Using the modified Valencia-Schaake model just described, twenty 30-year sequences of seasonal and monthly flows were disaggregated from the generated annual flows. The length of the sequences matched those of the historical flows. Tables 9 and 10 are summaries of the means and standard deviations of the generated seasonal and monthly sequences, respectively. The statistics, of course, vary from sequence to sequence; however, the averages are very close to those of the historical sequences. In long sequences, to be discussed later, the statistics are even more closely in agreement with the historical values. Cross correlations estimated from generated monthly sequences resemble historical estimates only in the transformed space because the flows are generated in the transformed space. Because of the complexity of the transform, the correlation matrices of untransformed flows do not compare favorably to estimates based on historical flows.

The disaggregated seasonal and monthly flows were also analyzed for their low-flow frequency and flow-duration characteristics. Characteristics of the generated flows were then compared to the characteristics of the historical data. Figures 2-10 show the annual low-flow frequency curves for 1-, 3-, and 6-month durations at each of the three sites. Each figure shows data from the first five sequences generated along with comparable data from the historical sequences. If curves were drawn through the data, little difference would be found at the low probabilities. The curves represent multiple 30-year sequences, with the observed sequence being only one, equally likely sequence. Differences are therefore likely and desirable.

Figures 11-13 are flow-duration curves of the observed and five typical generated sequences of monthly flows. The length of the sequences in all cases is 30 years. Interpretation of the differences between observed and simulated sequences is the same as above. If curves were drawn through plots of the various sequences, little difference would be seen.

FLOW SYNTHESIS

Because the twenty generated sequences of flows discussed above, show close resemblence in the mean, standard deviation, cross correlation (in the transformed space), and low-flow frequency and flow-duration characteristics, the model used is assumed to be an adequate representation of the underlying generating process. The model can now be used to generate long sequences of monthly flows that may be assumed to yield reasonable estimates of low probability (high recurrence interval), low-flow frequency characteristics.

Table 9.--Comparison of historical seasonal means and standard deviations with the average means and standard deviations of twenty generated 30-year sequences.

Season		Site 1		Site 2		Site 3	
		Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
		Historical					
1	(Fall)	267	190	50	35	253	166
2		638	175	116	34	553	137
3		543	217	111	34	537	175
4		155	95	24	20	153	70
		Generated/Disaggregated					
1	(Fall)	208[1/] (32)[2/]	143[3/] (50)[4/]	55(13)	54(22)	200(33)	123(38)
2		770(54)	225(38)	129(8)	37(6)	625(38)	160(27)
3		567(135)	222(53)	110(9)	30(5)	558(47)	149(24)
4		164(25)	85(24)	28(8)	25(14)	171(20)	71(15)

1/ The mean of the distribution of twenty 30-year mean fall flows.

2/ The standard deviation of the distribution twenty 30-year mean fall flows.

3/ The mean of the distribution of twenty 30-year standard deviations of fall flows.

4/ The standard deviation of the distribution of twenty 30-year standard deviations of fall flows.

Table 10.--Comparison of historical monthly means and standard deviations with the average means and standard deviations of twenty generated 30-year sequences.

Month		Site 1 Mean	Site 1 Standard deviation	Site 2 Mean	Site 2 Standard deviation	Site 3 Mean	Site 3 Standard deviation
				Historical			
10	(October)	145	108	24	24	152	102
11		265	223	50	46	258	215
12		391	310	74	54	362	254
1		451	274	83	54	296	229
2		565	304	97	48	474	211
3		900	322	169	69	790	294
4		764	354	149	62	705	277
5		525	261	114	51	540	212
6		338	359	70	74	366	376
7		183	129	31	30	185	118
8		152	130	22	23	144	74
9		131	124	19	20	131	76
				Generated			
10	(October)	124[1/](14)[2/]	60[3/](15)[4/]	21(4)	17(6)	124(13)	60(14)
11		171(27)	122(40)	45(11)	46(18)	174(27)	112(35)
12		310(65)	284(120)	85(26)	101(54)	291(54)	228(79)
1		559(108)	385(110)	101(17)	73(19)	460(60)	272(68)
2		687(119)	415(115)	104(19)	63(18)	520(95)	260(66)
3		1090(83)	398(78)	185(14)	75(17)	898(64)	326(68)
4		1067(88)	511(129)	188(10)	81(16)	894(44)	345(62)
5		492(70)	240(59)	99(13)	49(12)	522(58)	203(39)
6		282(53)	198(55)	51(9)	40(11)	288(45)	179(41)
7		178(35)	116(37)	31(13)	33(30)	189(32)	104(31)
8		184(23)	84(20)	30(6)	26(10)	182(16)	70(14)
9		124(15)	74(21)	23(5)	20(7)	145(13)	64(13)

1/ The mean of the distribution twenty 20 30-year mean October flows.

2/ The standard deviation of the distribution twenty 30-year mean October flows.

3/ The mean of the distribution of twenty 30-year standard deviations of October flows.

4/ The standard deviation of the distribution of twenty 30-year standard deviations of October flows.

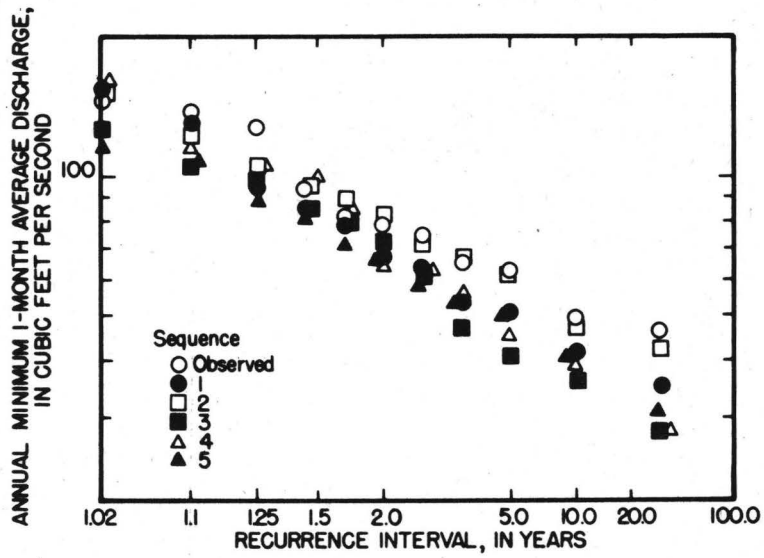

Figure 2.--Comparison of the 1-month duration observed and five typical generated 30-year sequences low-flow frequency curves at site 1.

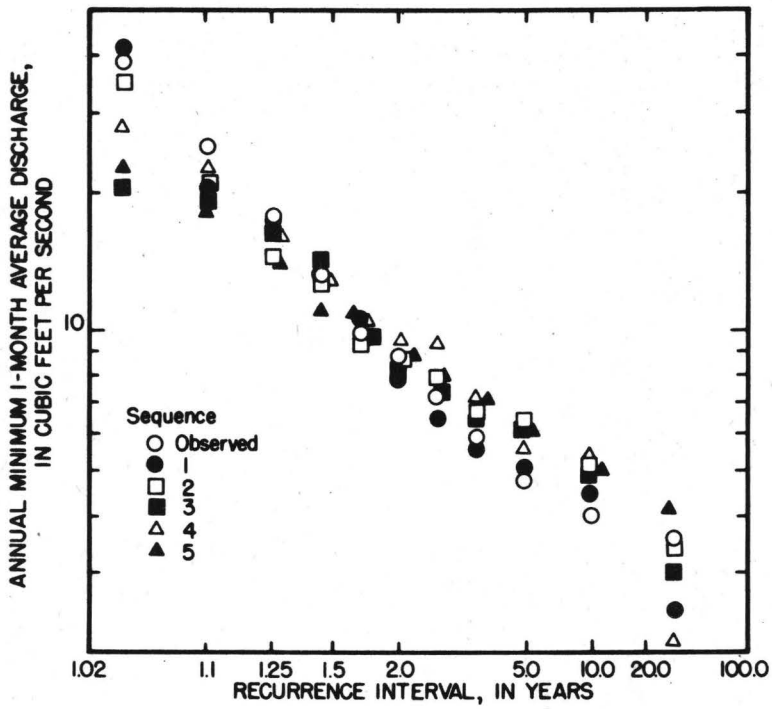

Figure 3.--Comparison of the 1-month duration observed and five typical generated 30-year sequences low-flow frequency curves at site 2.

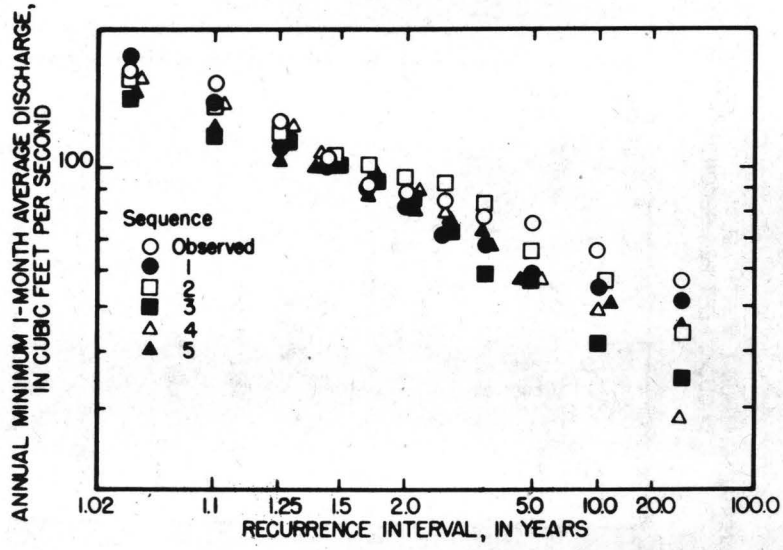

Figure 4.--Comparison on the 1-month duration observed and five typical generated 30-year sequences low-flow frequency curves at site 3.

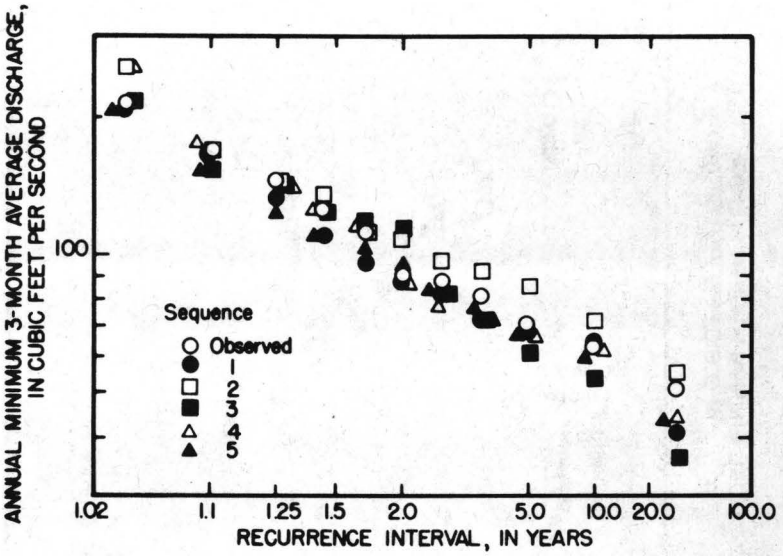

Figure 5.--Comparison of the 3-month duration observed and five typical generated 30-year sequences low-flow frequency curves at site 1.

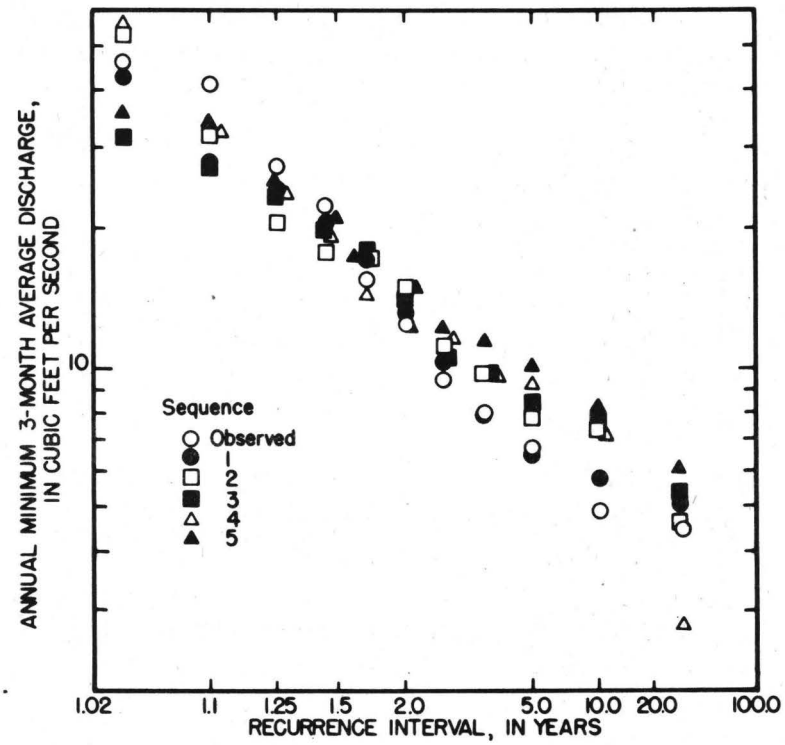

Figure 6.--Comparison of the 3-month duration observed and five typical generated 30-year sequences low-flow frequency curves at site 2.

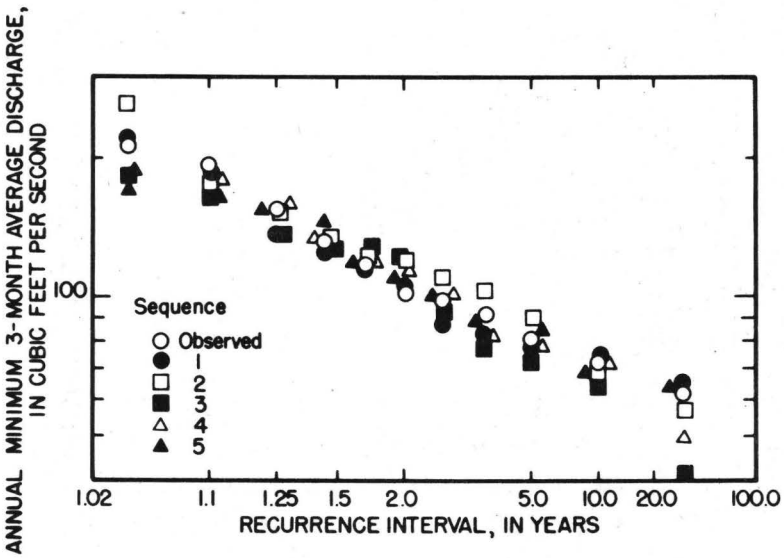

Figure 7.--Comparison of the 3-month duration observed and five typical generated 30-year sequences low-flow frequency curves at site 3.

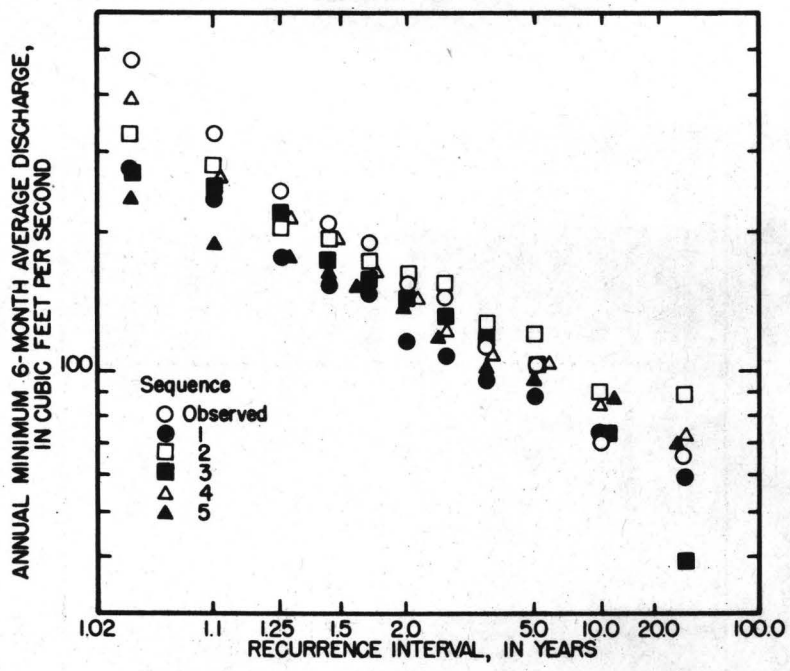

Figure 8.--Comparison of the 6-month duration observed and five typical generated 30-year sequences low-flow frequency curves at site 1.

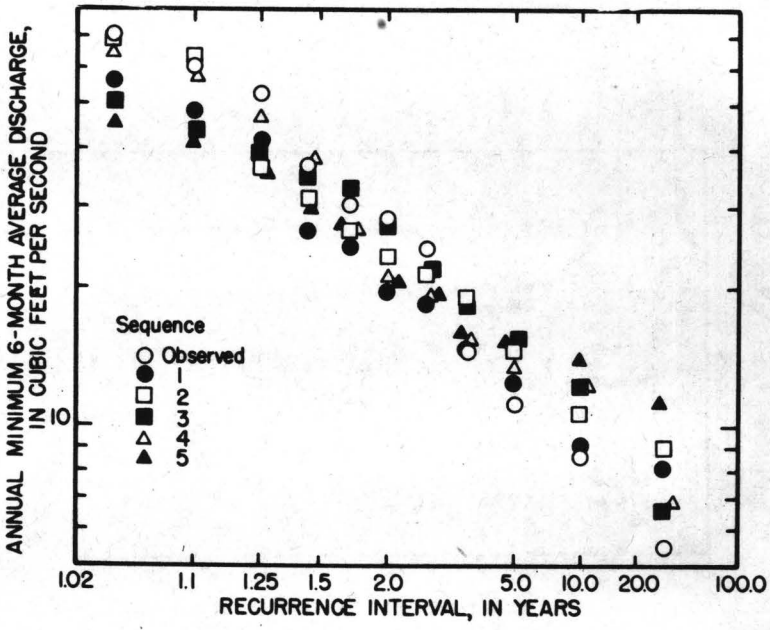

Figure 9.--Comparison of the 6-month duration observed and five typical generated 30-year sequences low-flow frequency curves at site 2.

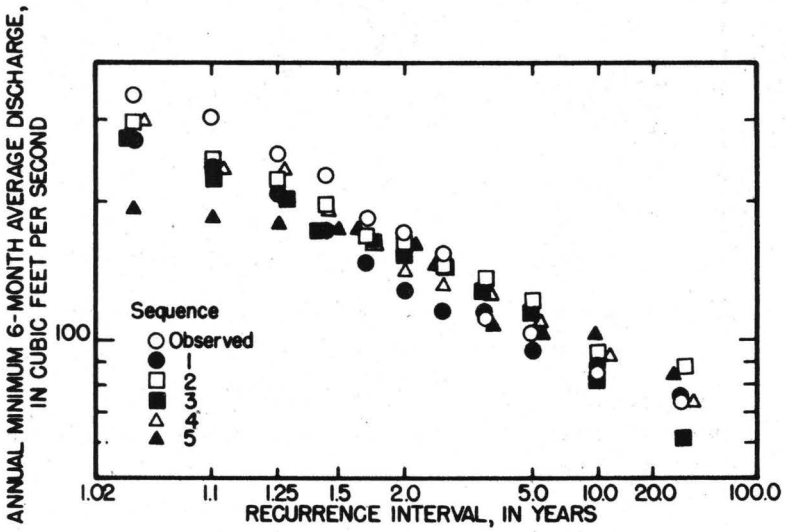

Figure 10.--Comparison of the 6-month duration observed and five typical generated 30-year sequences low-flow frequency curves at site 3.

Generation of Long-Term Annual Flows

Ten 100-year sequences of annual flows were generated using the ARIMA model at each of the three sites being analyzed. A summary of the results is presented in table 11. The average means and standard deviations of the 100-year generated sequences are closer to the historical values than the 30-year sequences, a fact that is not surprising because generated statistics often approach the historical statistics as the generated sequence increases. Although these statistics agree very closely with historical statistics, individual values more extreme than those in the historical sequence can be and sometimes are generated. Estimates of the Hurst coefficient for the generated sequences, presented in table 12, also behaved in a predictable fashion. K-Hurst estimates are generally higher than H-Hurst estimates and generally have a lower variance. Both H and K estimates based on the 100-year sequences are probably much closer, numerically, to the asymtotic values than the estimates computed from either observed or simulated 30-year sequences. Both estimation techniques improve significantly as the sequence length increases. Note that estimates of the historcial values of the Hurst coefficient, using either estimation procedure, do not fall within one standard deviation of the mean of estimates based on the simulated flows. This is probably caused by transient behavior, because the ARIMA model tends asymtotically to a value of 0.5 for h. The generated 30-year sequences (table 8) should produce higher values than the 100-year sequences. H and K for the 100-year sequences, however, are still significantly different from 0.5, a feature not attainable with a Markov model.

Generation of Long-Term Seasonal and Monthly Flows

The annual flows generated in the previous step were disaggregated using the modified Valencia-Schaake procedure. Tables 13 and 14 summarize the generated data. For comparison with historical statistics see tables 9 and 10. In much the same fashion as the statistics of the generated annual flows, the generated seasonal and monthly statistics are closer to their historical values than the generated 30-year sequences statistics were. Figures 14-22 compare low-flow frequency characteristics of the observed and five 100-year generated sequences. Many of the generated low probability (recurrence intervals of 10 years or more) low flows are less than the estimates based on observed flows. Some are about the same as historical estimates, and a few are greater. The generated sequences should have values with this mix. If all values were consistently higher or lower, one might conclude that either the transformations used in the model were incorrect or that the actual sample was not representative at those probabilities. The data generated here show that the model occasionally generates flows more extreme than those of the historical sequence. In contrast, if the observed sequence contains a value much lower than the generated sequences, the observed data may contain a rare extreme with a low probability of recurrence. Because the flow characteristics of the generated flows generally bracket those of the observed flows, the model is probably a valid representation of the underlying generating process. Long duration ($\geq$ 1 month) low-flow frequency characteristics of the generated sequences are probably more reliable than those estimated using the procedures in Armbruster (1976 b).

Table 11.--Summary of means and standard deviations of 100-year generated annual flow sequences

	Site 1		Site 2		Site 3	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Historical	400	108	75	20	375	86
1	408	102	77	18	383	79
2	408	103	75	21	379	85
3	390	105	71	20	364	84
4	410	109	76	20	382	86
5	390	111	74	18	368	85
6	393	118	74	19	370	90
7	396	99	76	16	374	76
8	400	97	76	19	376	82
9	399	107	76	20	375	86
10	390	109	78	19	373	85
Summary						
Mean	398	106	75	19	374	84
Standard deviation	8	6	2	1	6	4

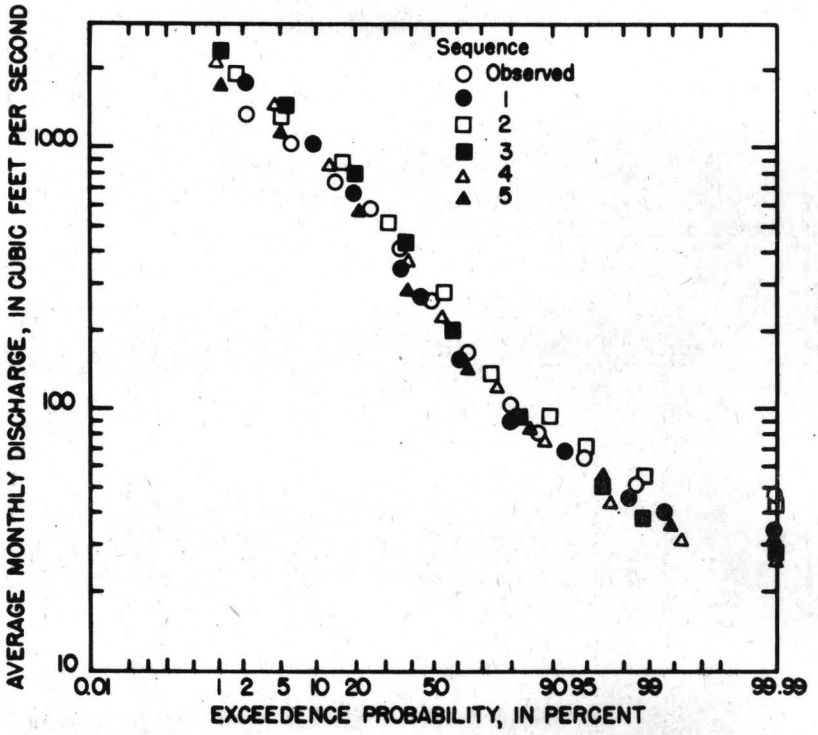

Figure 11.--Comparison of observed and five typical generated 30-year sequences flow duration curves at site 1.

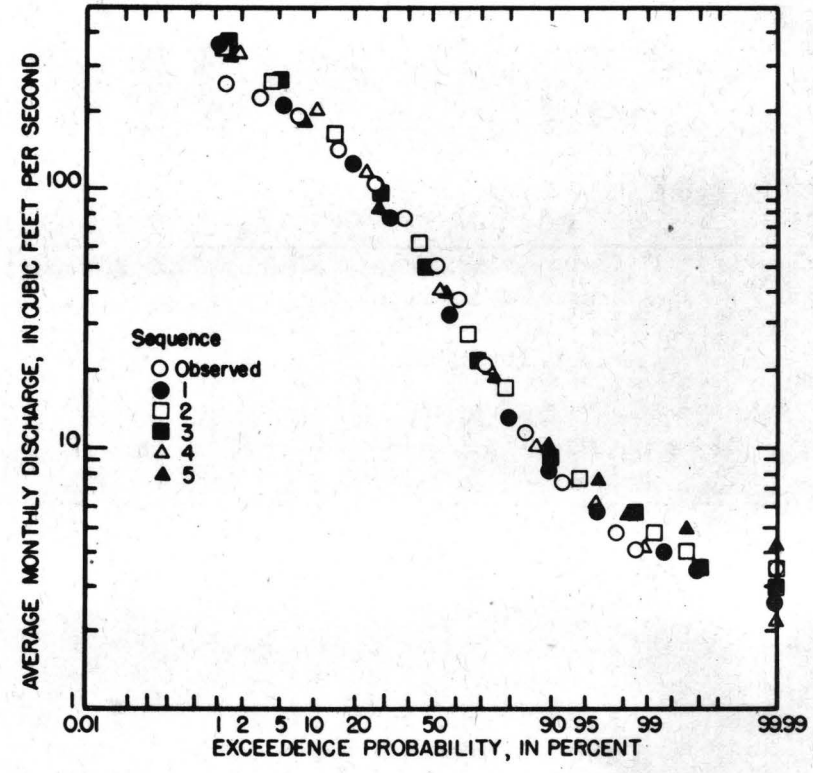

Figure 12.--Comparison of observed and five typical generated 30-year sequences flow duration curves at site 2.

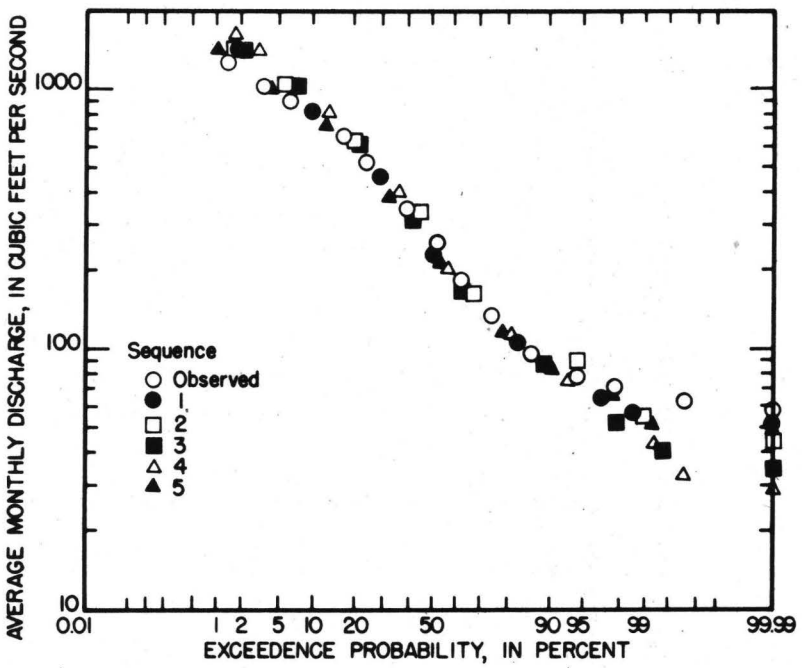

Figure 13.--Comparison of observed and five typical generated 30-year sequences flow duration curves at site 3.

Table 12.--Summary of values of the Hurst coefficient estimated from ten sequences of generated flows, 100 years long each

	Site 1		Site 2		Site 3	
	H	K	H	K	H	K
Historical	.8944	.7997	.9451	.8116	.8036	.7963
Generated 1	.4032	.5614	.5830	.6324	.4694	.5855
2	.5375	.6913	.4950	.7020	.5236	.6856
3	.5689	.6059	.6454	.6310	.5611	.6009
4	.7081	.7705	.6757	.7728	.6985	.7734
5	.5720	.6676	.6359	.6640	.5570	.6436
6	.6881	.7198	.7678	.7124	.7062	.7233
7	.6589	.7044	.6200	.6890	.6482	.6834
8	.6238	.6730	.4963	.6574	.5914	.6630
9	.7592	.7984	.6830	.7959	.7342	.7946
10	.7851	.7155	.6259	.7712	.7324	.7319
Summary						
Mean	.6305	.6908	.6228	.7028	.6222	.6885
Standard deviation	.1146	.0702	.0830	.0598	.0944	.0688

Table 13.--Summary of generated seasonal means and standard deviations of ten 100-year sequences.

Season	Site 1		Site 2		Site 3	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
1 (Fall)	207[1/](13)[2/]	144[3/](27)[4/]	52(5)	51(13)	200(11)	124(19)
2	782(20)	236(22)	131(3)	38(3)	634(12)	168(14)
3	608(21)	234(25)	112(3)	32(3)	563(14)	161(16)
4	161(4)	87(9)	26(6)	26(9)	170(4)	70(7)

1/ The mean of the distribution of ten 100-year mean fall flows.

2/ The standard deviation of the distribution ten 100-year mean fall flows.

3/ The mean of the distribution of ten 100-year standard deviations of fall flows.

4/ The standard deviation of the distribution of ten 100-year standard deviations of fall flows.

Table 14.--Summary of generated monthly means and standard deviations of ten generated 100-year sequences.

Month	Site 1		Site 2		Site 3	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
10 (October)	124[1/](7)[2/]	63[3/](14)[4/]	21(2)	24(14)	125(7)	63(13)
11	170(10)	128(24)	43(4)	47(14)	172(7)	115(15)
12	307(25)	289(47)	80(11)	96(29)	287(22)	228(39)
1	584(54)	388(81)	106(11)	85(19)	473(40)	290(58)
2	683(46)	490(11)	103(6)	75(22)	530(25)	288(56)
3	1116(51)	418(53)	190(8)	79(10)	919(36)	343(44)
4	1041(68)	458(64)	185(8)	79(16)	880(41)	321(47)
5	502(18)	260(18)	100(4)	51(5)	527(17)	221(18)
6	299(16)	240(42)	55(5)	50(11)	304(18)	216(36)
7	175(7)	126(23)	29(2)	30(6)	186(6)	102(10)
8	179(6)	86(12)	31(3)	29(9)	173(23)	67(8)
9	123(4)	74(7)	24(4)	27(10)	146(5)	66(7)

1/ The mean of the distribution of ten 100-year mean October flows.

2/ The standard deviation of the distribution ten 100-year mean October flows.

3/ The mean of the distribution of ten 100-year standard deviations of October flows.

4/ The standard deviation of the distribution of ten 100-year standard deviations of October flows.

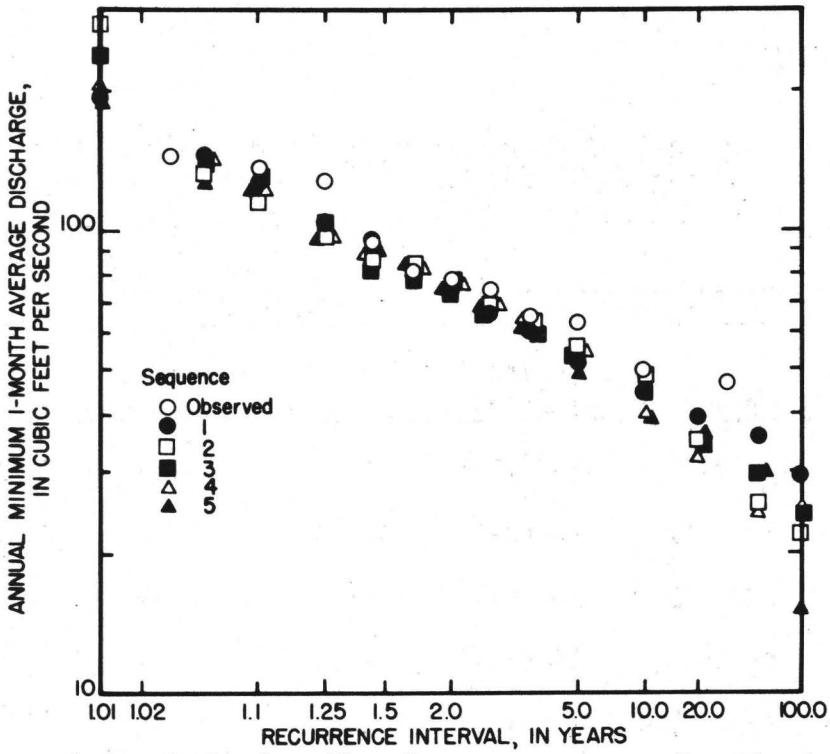

Figure 14.--Comparison of the low-flow frequency curves for the 1-month duration 30-year observed flows and five typical 100-year generated flow sequences at site 1.

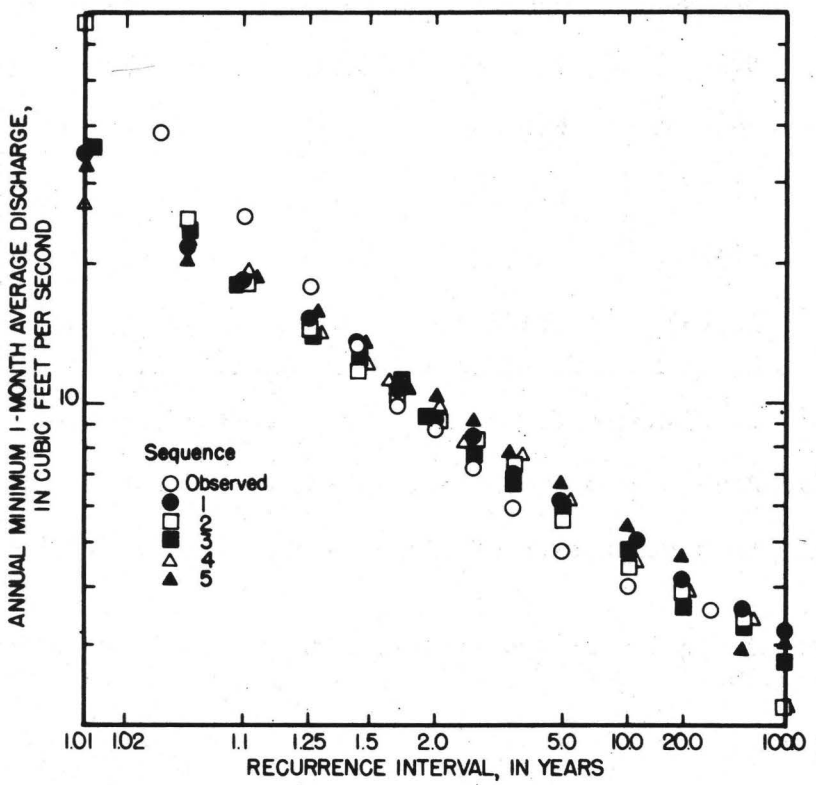

Figure 15.--Comparison of the low-flow frequency curves for the 1-month duration 30-year observed flows and five typical 100-year generated flow sequences at site 2.

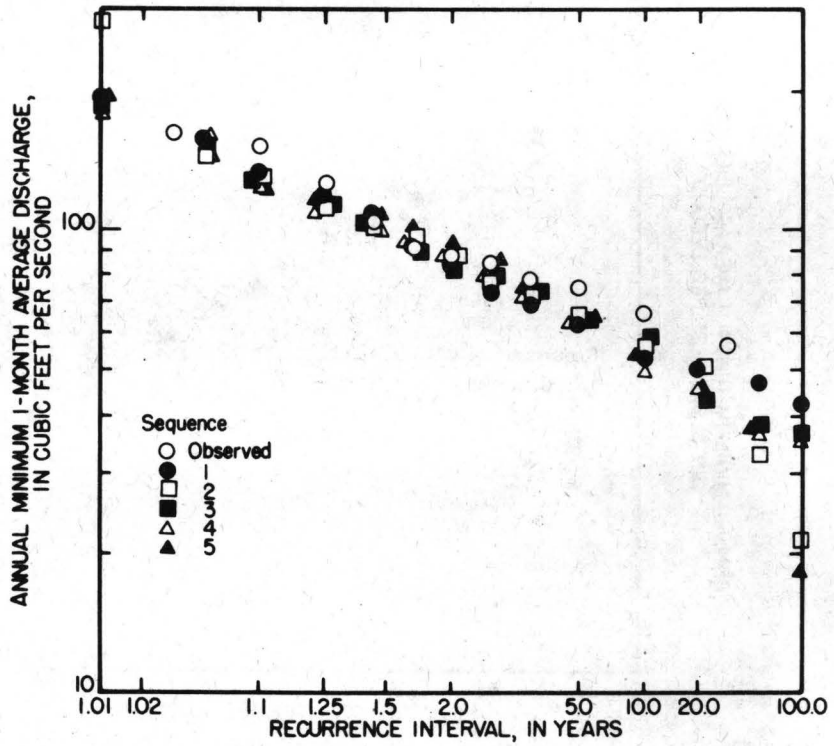

Figure 16.--Comparison of the low-flow frequency curves for the 1-month duration 30-year observed flows and five typical 100-year generated flow sequences at site 3.

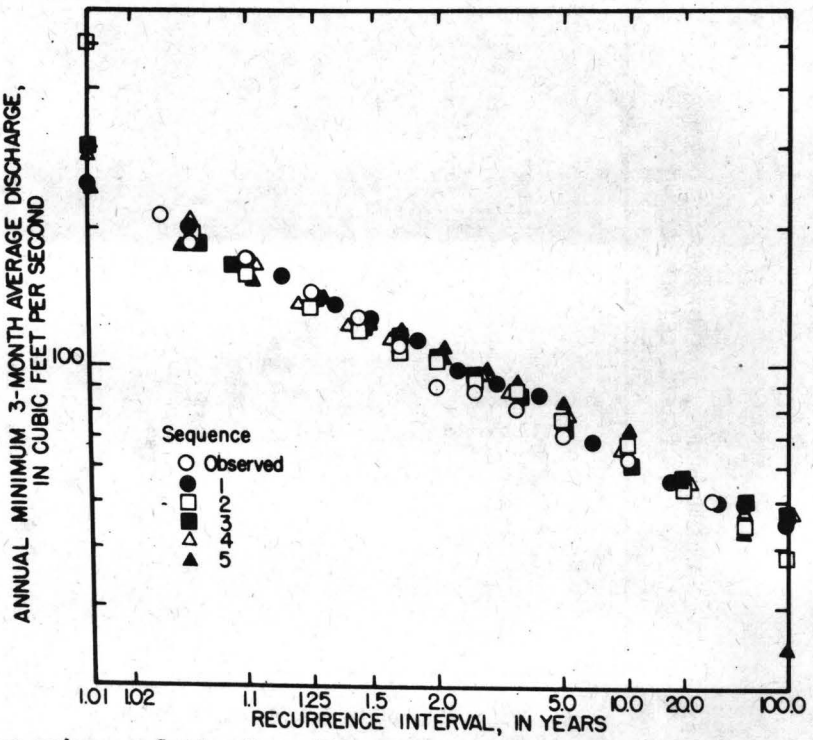

Figure 17.--Comparison of the low-flow frequency curves for the 3-month duration 30-year observed flows and five typical 100-year generated flow sequences at site 1.

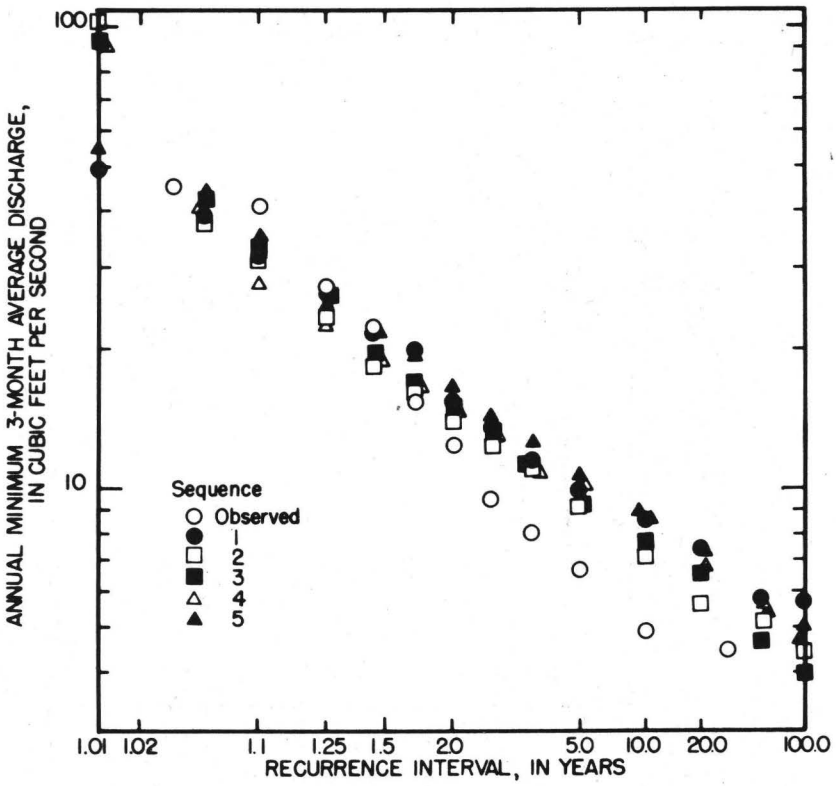

Figure 18.--Comparison of the low-flow frequency curves for the 3-month duration 30-year observed flows and five typical 100-year generated flow sequences at site 2.

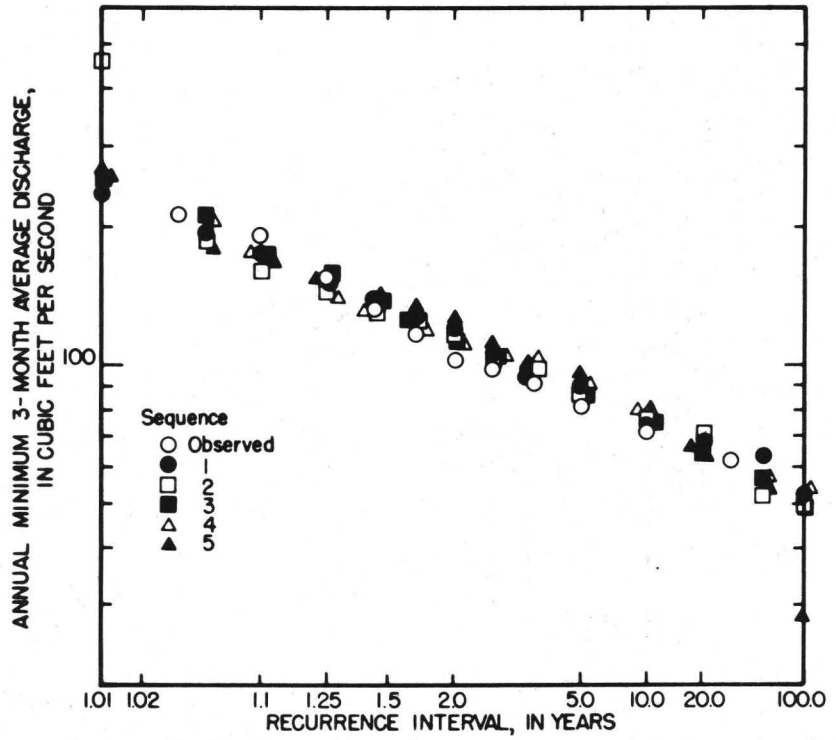

Figure 19.--Comparison of the low-flow frequency curves for the 3-month duration 30-year observed flows and five typical 100-year generated flow sequences at site 3.

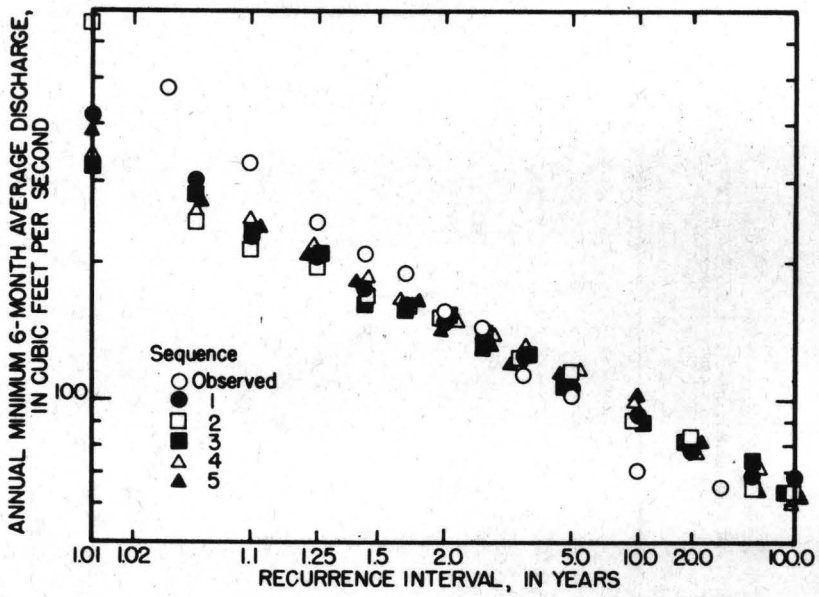

Figure 20.--Comparison of the low-flow frequency curves for the 6-month duration 30-year observed flows and five typical 100-year generated flow sequences at site 1.

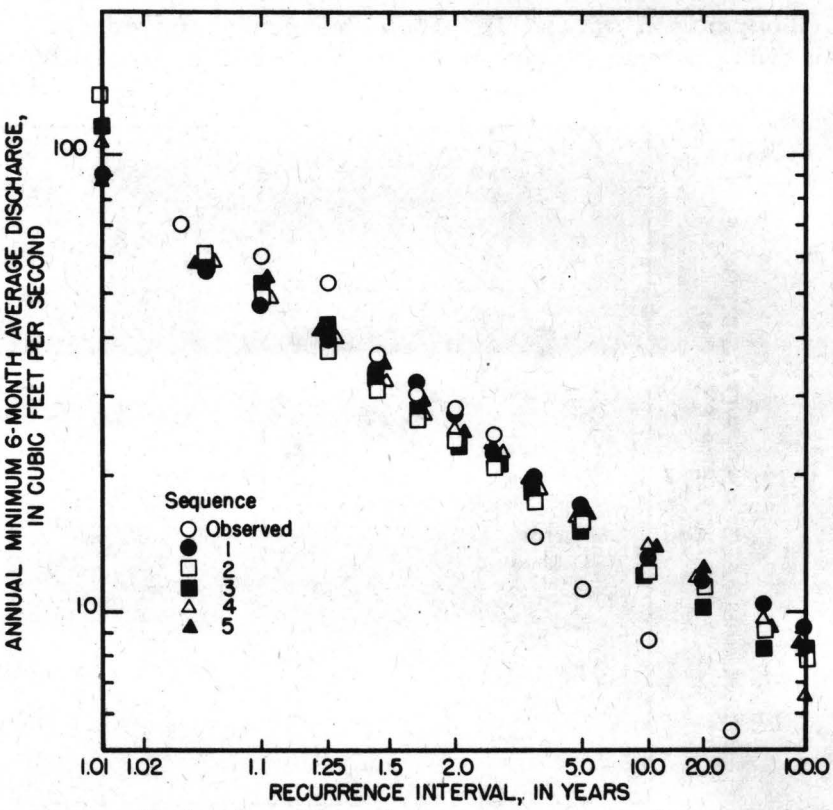

Figure 21.--Comparison of the low-flow frequency curves for the 6-month duration 30-year observed flows and five typical 100-year generated flow sequences at site 2.

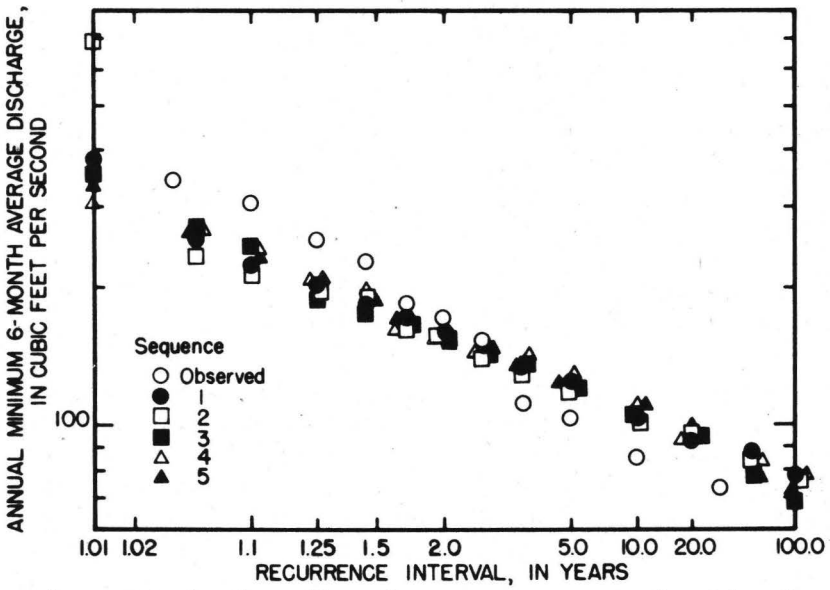

Figure 22.--Comparison of the low-flow frequency curves for the 6-month duration 30-year observed flows and five typical 100-year generated flow sequences at site 3.

Figures 23-25 further verify the ability of the model to generate flows that seem to be equally likely realizations of the same generating process that produced the observed flows.

REGIONAL ESTIMATES OF STREAMFLOW STATISTICS

Mean and Standard Duration

Streamflow statistics such as mean annual and mean monthly flows and standard deviations of annual and monthly flows are required on a regional basis before stochastic models can be applied to ungaged sites. Using techniques described by Benson and Matalas (1967), regression equations were developed for estimating the above statistics. The general form of the regression model is

$$\log Y = \log c + b_1 \log X_1 + + b_2 \log X_2 + \ldots.$$

where Y is the statistic being estimated, X's are basin parameters, b's are regression coefficients, and c is a regression constant.

Many different basin characteristics, such as drainage area, precipitation, channel slope and length, percent of basin covered by forest, and index of relative infiltration (Armbruster, 1976a) were used in the regression analyses, but drainage area was generally the only characteristic determined to be statistically significant. All nine sites listed in table 1 were used. Results of the regional analyses are presented in tables 15 and 16. For almost all the regressions presented here the standard errors of estimate of the equations are smaller than those presented by Page (1970).

Cross Correlation

The problem of estimating cross-correlation matrices, when an ungaged site is one or more elements, was divided into two parts -- first, estimate the diagonal elements of the matrix and second, estimate the off-diagonal elements.

Estimation of diagonal elements

The lag-one and lag-two autocorrelations at the gaged sites were computed using equation 2. These values were then related to physical and climatic characteristics of their respective drainage basins, using standard linear-regression techniques. The resulting relation for the lag-one autocorrelation is

$$\log \hat{\rho}(1) = 1.222 + 0.7140 \log (P-30) \qquad (36)$$

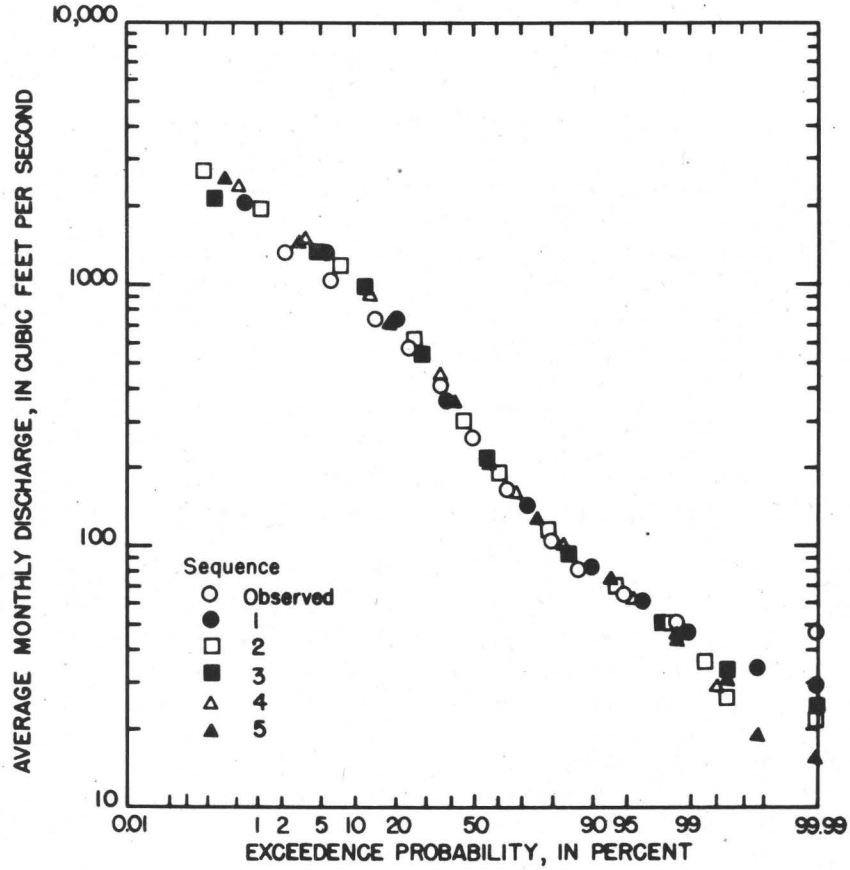

Figure 23.--Comparison of observed and five typical generated 30-year sequences flow duration curves at site 1.

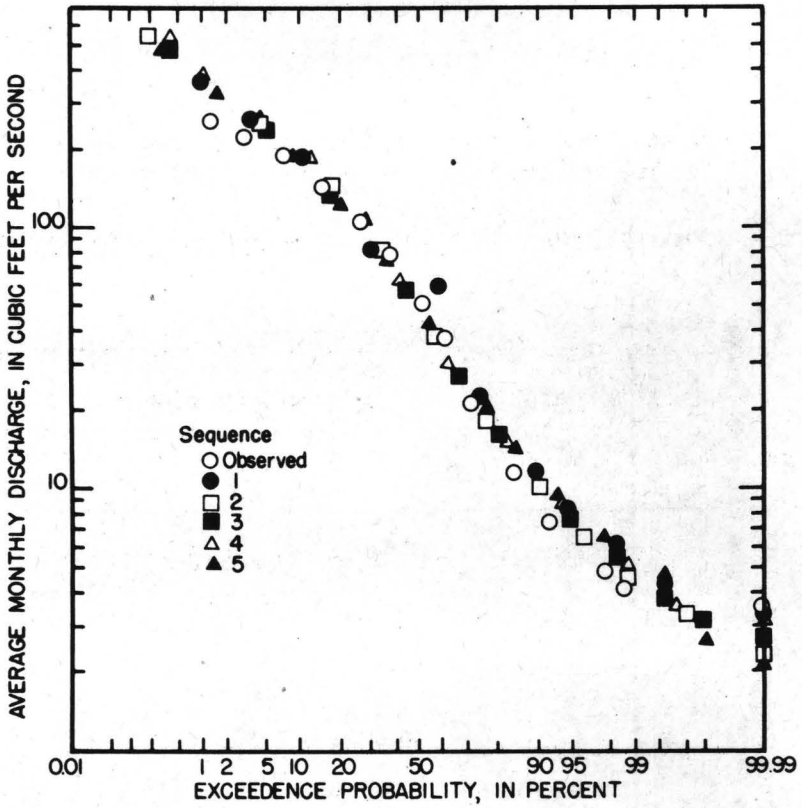

Figure 24.--Comparison of observed and five typical generated 30-year sequences flow duration curves at site 2.

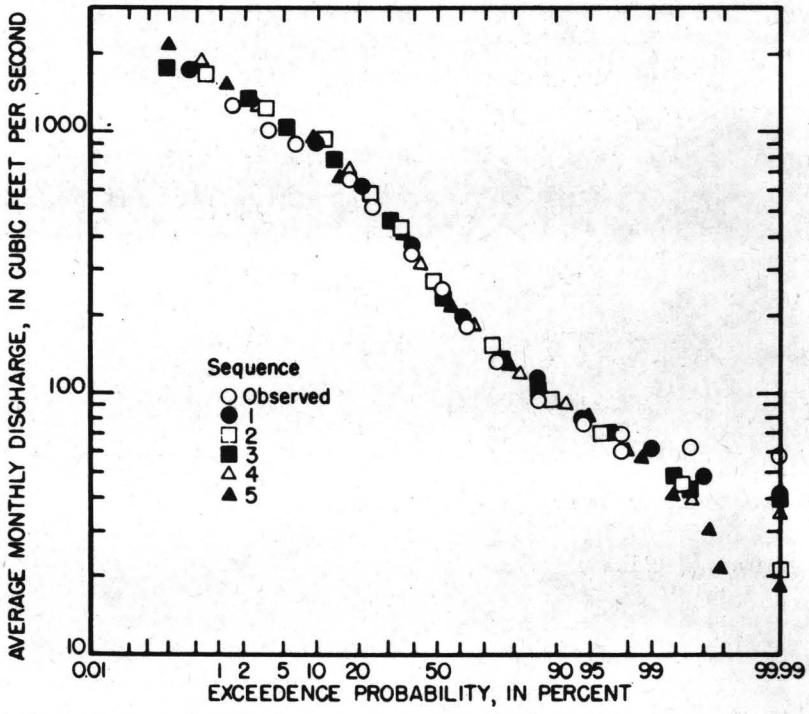

Figure 25.--Comparison of observed and five typical generated 30-year sequences flow duration curves at site 3.

Table 15.--Summary of regression equations for mean of annual and monthly flows in the Juniata River basin

[log Y = log c + b_1 log A, where A is drainage area (square miles)]

Streamflow characteristic Y	Regression constant log c	Regression coefficient b_1	Standard error of estimate percent
Q10*	-0.1945	0.9540	17.2
Q11	.1439	.9333	12.3
Q2	.3289	.9262	7.6
Q1	.3819	.9211	9.8
Q2	.4626	.9284	4.9
Q3	.6985	.9166	7.6
Q4	.6248	.9148	12.8
Q5	.4769	.9174	16.8
Q6	.2370	.9378	19.2
Q7	-.1706	.9674	28.3
Q8	-.1602	1.0108	25.7
Q9	-.3302	.9644	28.9
Q_A**	.3197	.9267	11.5

* Q10 = mean flow for month 10, October.

** Q_A = mean annual flow.

Table 16.--Summary of regression equations for standard deviations of annual and monthly flows in the Juniata River basin.

[$\log Y = \log c + b_1 \log A$, where A is drainage area (square miles)]

Streamflow characteristic Y	Regression constant log c	Regression coefficient b_1	Standard error of estimate percent
S10	0.0522	0.8742	14.5
S11	.1634	.9101	15.3
S22	.1979	.9393	6.4
S1	.2094	.9119	6.1
S2	.2012	.9227	10.2
S3	.3223	.8991	7.5
S4	.2866	.9181	4.9
S5	.2326	.8833	6.6
S6	.2944	.9405	17.0
S7	-.1266	.9073	21.2
S8	-.0238	.8207	21.1
S9	-.0525	.8341	11.5
SA**	-.2884	.9467	6.4

* S10 = standard deviation of flows for month 10, October.

** SA = standard deviation of annual flows.

where $\hat{\rho}$ (1) is an estimate of the lag-one autocorrelation coefficient and P is mean annual precipitation, in inches. The standard error of estimate of this relation is 10 percent. Use of the mean of the sample of lag-one autocorrelation has a standard deviation of 16 percent of the mean value. Thus equation 36 provides only a small improvement, over using the mean lag-one autocorrelation. The lag-two autocorrelation is estimated by

$$\log \hat{\rho}(2) = 0.8902 + 2.091 \log (P-30) + 0.6963 \log I \qquad (37)$$

where $\hat{\rho}$ (2) is an estimate of the lag-two autocorrelation and I is an index of relative infiltration described by Armbruster (1976a, 1976b). The standard error of estimate of this relation is 24 percent. The standard error associated with use of the mean diagonal element, by contrast, is 40 percent.

Estimation of Off-Diagonal Elements

Estimation of the off-diagonal elements or cross correlations, required a completely different approach than the one used for estimating the autocorrelations. An attempt was made to estimate cross correlations at ungaged sites using the parameters of the multisite ARIMA model (Moss, personal communication, 1977). Because the function was extremely complex in terms of the parameters, and because there is no simple way to estimate the ARIMA model parameters, an attempt was made to estimate cross correlations empirically, using known or easily obtainable information. The procedures used to estimate the lag-zero matrix will be discussed separately from the lag-one, and lag-two matrices because they are handled differently.

Lag-zero cross correlation

The interrelation between annual flows of pairs of sites was related to the distance between the centroids of the respective basins. In theory the distance between centroids of nonoverlapping basins is probably a better representation because it would assume that no drop of water is accounted for more than once. Practically, however, this approach did not yield useable results, probably because of basin shape, topography, and time sampling errors. However, it was found that the lag-zero cross correlations of annual flows were linearly related to the distance between the centroids of overlapping basins, as shown in figure 26.

The curve, fitted to the data using a linear regression model, is

$$\hat{\rho}_{ij} (0) = 1.02143 - .003786 D \qquad (38)$$

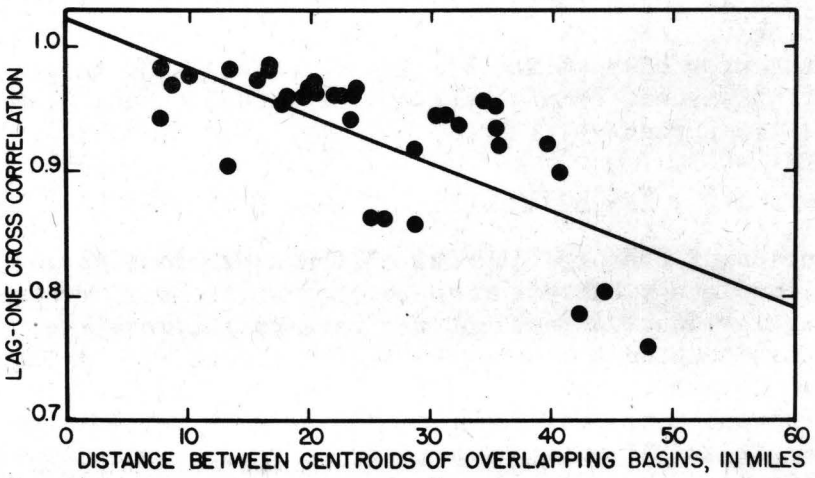

Figure 26.--Relation between lag-one cross correlation and distance between centroids of overlapping basins.

where $\hat{\rho}_{ij}(0)$ is an estimate of the lag-zero cross correlation between flows at sites i and j and D is the distance between centroids of basins i and j. The standard error of estimate of this relation is 0.0417, and the correlation coefficient is 0.704. Esthetically, equation 38 is flawed, as the intercept is greater than one -- and correlations can never exceed unity. Equation 38 should be used only for estimating lag-zero cross correlation between basins with centroids greater than about 5.6 miles apart.

Lag-one and lag-two cross correlations

Conceptually, lag-one and lag-two cross correlations should be a function of the respective individual autocorrelations. The exact form of the functions, however, is unknown. Several attempts were made to define the relations, but none could be found that provided any improvement over using the mean of the sample cross correlations. For the lag-one cross correlations, the standard deviation is about 14 percent of the mean, and, for the lag-two cross correlations, the standard deviation is about 35 percent. Therefore, when one or more ungaged sites are included as stations in a multisite problem, off-diagonal elements for the ungaged sites should be equated to 0.316 and 0.145 for lag-one and lag-two cross correlations, respectively.

SUMMARY

An ARIMA (1,0,1) model was calibrated and used to generate long (100-year) sequences of annual streamflows at multiple sites. The model generates annual flows that preserve the means, standard deviations, and low-lag cross correlations at each site, and yields estimates of the Hurst coefficient close to estimates based on observed data.

To obtain monthly flows from the generated annual sequences, the Valencia-Schaake disaggregation model was calibrated and used. Initially seasonal flows were disaggregated from annual flows, then monthly flows were disaggregated from the seasonal values. Several modifications were made to the Valencia-Schaake to improve its flexibility.

Finally a method was suggested for synthesizing flows at ungaged sites. Streamflow statistics can be estimated using easily measured basin characteristics.

There are limitations and constraints to the models used here. For example, probably the most serious computational constraint is the parameter estimation procedure used in calibrating the multisite ARIMA model. It is very sensitive to the cross correlation of observed flows among the sites being analyzed. Although an attempt was made to calibrate the model for a 9-site problem, the model could only be calibrated for 3-sites. The parameters are estimated using an iterative procedure that sometimes fails to converge.

As with other models of this type, the reliability of the generated flows is strongly dependent on the assumption that observed flows are a representative sample in both time and space.

The methods proposed here are valid only for streams with natural or unregulated flows.

REFERENCES

Armbruster, Jeffrey T., 1976a, An infiltration index useful in estimating low-flow characteristics of drainage basins: U.S. Geological Survey, Journal of Research, v. 4, no. 5, p. 533-538.

_____, 1976b, Technical manual for estimating low-flow frequency characteristics of streams in the Susquehanna River basin: U.S. Geological Survey Water Resources Investigation 76-51, 51 p.

Benson, M. A. and Matalas, N. C., 1967, Synthetic hydrology based on regional statistical parameters: Water Resources Research, v. 3, no. 4, p. 931-935.

Box, G. E. P. and Jenkins, G. M., 1968, Some recent advances in forecasting and control, I: Applied Statistics, v. 17, p. 41-109.

_____, 1970, Time series analysis, forecasting and control: Holden-Day, San Francisco, Calif., 553 p.

Feller, W., 1951, The asymtotic distribution of the range of sums of independent random variables: Annals Mathamatical Statistics, v. 22, p. 427-432.

Harms, A. A. and Campbell, T. H., 1967, An extension to the Thomas-Fiering model for the sequential generation of streamflow: Water Resources Reasearch, v. 3, no. 3, p. 653-661.

Hurst, H. E., 1951, Long-term storage capacities of reservoirs: Transactions of the American Society Civil Engineers, v. 116, p. 770-808.

Mandelbrot, B. B. and Wallis, J. R., 1969, Some long run properties of geophysical records: Water Resources Research, v. 5, no. 2, p. 321-340.

Matalas, N. C., 1967, Mathematical assessment of synthetic hydrology: Water Resources Research, v. 3, no. 4, p. 931-935.

_____, 1971, Stochastic aspects in hydrology: Proceedings of the international symposium in hydrology, Waterloo University.

Mejia, J. M., and Rousselle, J., 1976, Disaggregation models in hydrology revisited: Water Resources Research, v. 12, no. 2, p. 185-186.

O'Connell, P. E., 1971, A simple stochastic modelling of Hurst's law: Proceedings of the international symposium on mathematical models in hydrology, Warsaw, Poland, July, 1971.

_____, 1974, Stochastic modelling of hydrologic persistence in streamflow sequences: University of London, London, England, Ph. D. Thesis.

Page, L. V., 1970, A proposed streamflow data program for Pennsylvania: Pennsylvania Department of Forests and Waters Technical Bulletin 3, 56 p.

Slack, J. R., 1972, Bias, illusion and denial as data uncertainties: International symposium on uncertainties in hydrology and water resources systems, Tuscon, Arizona, Dec 11-14, 1972, p. 122-132.

_____, 1973, I would if I could (self-denial by conditional models): Water Resources Research, v. 9, no. 1, p. 247-249.

Tao, P. C. and Delleur, J. W., 1976, Seasonal and nonseasonal ARMA models in hydrology: American Society Civil Engineers Proceedings Journal Hydraulics Division, HY10, p. 1541-1559.

Valencia R. D. and Schaake, J. C., 1972, A disaggregation model for time series analysis and synthesis: Report 149, Ralph M. Parsons Laboratory for water resources and hydrodynamics, Massachusetts Institute of Technology, Cambridge, Massachusetts, 190 p.

_____, 1973, Disaggregation processes in stochastic hydrology: Water Resources Research, v. 9, no. 3, p. 580-585.

Wallis, J. R. and Matalas, N. C., 1970, Small sample properties of H and K--estimators of the Hurst coefficient h: Water Resources Research, v. 6, no. 6, p. 1583-1594.

_____, 1971a, Correlogram analysis revisited: Water Resources Research, v. 7, no. 6, p. 1448-1459.

_____, 1971b, In hydrology h is a household word: International symposium on mathematical models in hydrology, Warsaw, Poland, July, 1971.

Young, G. K., and Jettmar, R. U., 1976, Modeling monthly hydrologic persistence: Water Resources Research, v. 12, no. 5, p. 829-835.