Autocorrelation of Rainfall and Streamflow Minimums

GEOLOGICAL SURVEY PROFESSIONAL PAPER 434-B

Matalas—AUTOCORRELATION OF RAINFALL AND STREAMFLOW MINIMUMS—Geological Survey Professional Paper 434-B

Autocorrelation of Rainfall and Streamflow Minimums

By NICHOLAS C. MATALAS

STATISTICAL STUDIES IN HYDROLOGY

GEOLOGICAL SURVEY PROFESSIONAL PAPER 434-B

UNITED STATES GOVERNMENT PRINTING OFFICE, WASHINGTON : 1963

UNITED STATES DEPARTMENT OF THE INTERIOR

STEWART L. UDALL, *Secretary*

GEOLOGICAL SURVEY

William T. Pecora, *Director*

First printing 1963

Second printing 1967

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C. 20402 - Price 20 cents (paper cover)

CONTENTS

	Page
Abstract	B1
Introduction	1
Characteristics of time series	1
Effect of nonrandomness in hydrologic studies	2
Methods for investigating nonrandomness	2
Trend analysis	2
Measure of nonrandomness	3
Generating processes	3
Methods for investigating nonrandomness—Continued	
Correlogram analysis	B4
Investigation of nonrandomness in hydrologic time series	5
Serial correlation	5
Generating process	6
Effect of serial correlation on the estimate of the variance	6
Discussion and conclusions	7
References	7

ILLUSTRATIONS

	Page
FIGURES 1–6. Correlograms for annual minimum rainfall.	
1. Boston, Mass	B8
2. Providence, R.I	8
3. Charleston, S.C	8
4. Washington, D.C	8
5. Baltimore, Md	8
6. Philadelphia, Pa	8
7, 8. Correlograms for annual minimum 7-day discharge.	
FIGURE 7. South Fork Holston River at Bluff City, Tenn	B9
8. James River at Buchanan, Va	9
9. Correlogram for annual minimum 1-day discharge for Middle Branch Westfield River near Goss Heights, Mass	9
10. Expected value of the variance for a first-order Markov process	9

TABLES

	Page
TABLE 1. Rainfall stations used in analyzing annual minimum monthly rainfall	B9
2. Gaging stations used in analyzing annual minimum discharge	9
3, 4. Significance of first-order serial-correlation coefficients.	
3. Rainfall data	10
TABLE 4. Discharge data	B10
5, 6. Serial-correlation coefficients for orders 1 through 19.	
5. Rainfall data	10
6. Discharge data	10
7. Chi-square goodness of fit of first-order Markov process to discharge correlograms	10

AUTOCORRELATION OF RAINFALL AND STREAMFLOW MINIMUMS

By Nicholas C. Matalas

ABSTRACT

Hydrologic time series of annual minimum mean monthly rainfall and annual minimum 1-day and 7-day discharge, considered as drought indices, were used to study the distribution of droughts with respect to time. The rainfall data were found to be nearly random. The discharge data, however, were found to be nonrandomly distributed in time and generated by a first-order Markov process. The expected value of the variance for a time series generated by a first-order Markov process was compared with the expected value of the variance for a random time series. This comparison showed that the expected value of the variance for a nonrandom time series converged to the population variance with an increase in sample size at a slower rate than for a random time series.

INTRODUCTION

The object of the investigation was to determine if droughts are randomly or nonrandomly distributed in time and to modify tests of reliability to account for the effect of nonrandomness. For these purposes, the hydrologic time series used in the investigations were annual minimum monthly rainfall at 18 rainfall stations and annual minimum 1-day and 7-day discharge at 9 gaging stations.

A time series is defined as a sequence formed by the values of a variable at increasing points in time. A time series may be composed of the sum of two components: a random element and a nonrandom element. If the values of the time series are not independent of each other, the nonrandom element exists, and the values are said to be serially dependent.

The nonrandom element in a time series may be due to a trend or an oscillation about the trend or both. In a given time series, the nonrandom element need not be due to both of these causes. In order to analyze properly a time series, the random and nonrandom elements must be isolated and studied separately. Trend must be eliminated from the nonrandom element before studying the oscillatory character of a time series. The oscillatory movement in a trend-free time series may be due to 1 of 3 schemes: (a) moving averages, (b) sums of cyclic components, and (c) autoregression.

CHARACTERISTICS OF TIME SERIES

Time series may be classified as either stationary or nonstationary. Assume that a time series is divided into several segments and that the data within each segment are characterized by statistical parameters such as the mean and variance. If the expected values of these parameters are the same for each segment, the time series is said to be stationary. If the expected values are not the same for all segments, the time series is nonstationary. In stationary time series, absolute time is not important, and the series maybe assumed to have started somewhere in the infinite past. However, in nonstationary time series, absolute time must be considered since the series cannot be assumed to have begun prior to the time of the initial observation.

Time series may be considered as composed of the sum of a random element and a nonrandom element. The nonrandom element consists of a trend and an oscillation about the trend. Trend is usually thought of as a smooth motion of the time series over a long period of time. For any given time series, the sequence of values will follow an oscillatory pattern. If this pattern indicates a more or less steady rise or fall, the pattern is defined as a trend. No matter what the length of a time series, one can never state with certainty that an apparent trend is not part of a slow oscillation, unless the series ends.

An oscillatory movement is often confused with a cyclical movement. In a cyclical time series, the maximum and minimum values occur at equal intervals of time with constant amplitude. The random component, if present, tends to distort this pattern. In an oscillatory time series, the amplitude and the interval of time between maximum and minimum values are distributed about mean values. A cyclical time series is oscillatory, but an oscillatory time series is not necessarily cyclical.

In order to analyze properly a time series, it is necessary to separate the random and the nonrandom components. Trend must be eliminated from the nonrandom component before studying the oscillatory character of the time series.

EFFECT OF NONRANDOMNESS IN HYDROLOGIC STUDIES

The design of dams and water-storage reservoirs is based on the expected hydrologic conditions many years in the future, depending upon the life expectancy of the hydraulic structures. To obtain long-range estimates, the frequency with which hydrologic events occur must be known, if it is assumed that the future will follow the pattern shown by the past. In a nonrandom time series, there is a tendency for low values to follow low values and for high values to follow high values. In the case of a water-supply reservoir, the reservoir is filled during a wet season to meet the anticipated demand during a dry season. If droughts are nonrandomly distributed in time, then to overcome the tendency for low values to follow low values, the reservoir capacity must be increased to provide sufficient water to meet the need for more than one dry season.

Hurst (1950), in his study of long-term storage capacities of reservoirs, concluded that annual total discharge is nonrandomly distributed in time. This study was based on an analysis of the cumulative sum of departures of annual total discharges from the mean annual total discharge. The absolute value of the difference between the maximum and the minimum of the cumulative sums was defined as the range. To test for nonrandomness, Hurst determined the expected values of the range for various natural phenomena, and compared these values with the expected value of the range for a random normally distributed time series.

For a random normally distributed time series, Hurst defined the expected value of the range, R, as

$$R=1.25\sigma\sqrt{N}, \tag{1}$$

where σ is the standard deviation and N is the sample size. Feller (1951) showed that the variance of the range is

$$V(R)=0.074\sigma^2N. \tag{2}$$

Feller also pointed out that the asymptotic probability distribution of the range is independent of the form of the probability distribution of the variate values. Equations 1 and 2 are valid for large N. Anis and Lloyd (1953) defined the expected value of the range for any value of N as

$$R=\sqrt{2/\pi}\sigma \sum_{i=1}^{N-1} i^{-1}. \tag{3}$$

Hurst investigated 75 different phenomena consisting of 690 sets of records and found the range to be of the form

$$R=\sigma\left(\frac{N}{2}\right)^k. \tag{4}$$

The value of k varied slightly with different phenomena. For the 690 sets of records, the mean and standard deviation of k were 0.729 and 0.091, respectively.

By this analysis, Hurst concluded that estimates of the standard deviation based on N observations are more variable for a nonrandom time series than for a random time series. In a continuation of this study, Hurst (1956) pointed out that statistical averages based on 30 or 40 observations from a nonrandom time series do not yield close approximations to the true averages.

The difference between the ranges defined by equations 1 and 4 may be due to the data not being randomly distributed in time. This, however, requires that the nonrandomness be of a very special type. Moran (1959) suggests that the experimental series used by Hurst are, as a result of nonrandomness, of insufficient length for the asymptotic formula (eq. 1) to become valid. Hurst (1957) constructed a probability model which tends to illustrate this point.

METHODS FOR INVESTIGATING NONRANDOMNESS

TREND ANALYSIS

Trend is thought of as a smooth motion of the time series over a long period of time. If the oscillatory pattern of a time series indicates a more or less steady rise or fall, the pattern is defined as a trend. To study the oscillatory characteristics of a time series, the trend, if present, should be removed. Various methods of removing trend are available. All the methods, however, are not fully understood as to how they affect the time series. The most commonly used method is that of moving averages.

Assume that the observations $x_1, x_2, \ldots, x_N$ are taken at equal intervals of time. The method of moving averages consists of determining overlapping means of m successive weighted values. An example of moving averages for $m=3$ is

$$\left.\begin{aligned} y_2 &= (b_1x_1+b_2x_2+b_3x_3)/3 \\ y_3 &= (b_1x_2+b_2x_3+b_3x_4)/3 \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ y_{N-1} &= (b_1x_{N-2}+b_2x_{N-1}+b_3x_N)/3 \end{aligned}\right\} \tag{5}$$

The weights of the moving average, b_1, b_2, and b_3, are such that their sum equals 3. In general, for moving

$$r_k=\frac{\frac{1}{N-k}\sum_{i=1}^{N-k}x_ix_{i+k}-\frac{1}{(N-k)^2}\left(\sum_{i=1}^{N-k}x_i\right)\left(\sum_{i=1}^{N-k}x_{i+k}\right)}{\left[\frac{1}{N-k}\sum_{i=1}^{N-k}x_i^2-\frac{1}{(N-k)^2}\left(\sum_{i=1}^{N-k}x_i\right)^2\right]^{1/2}\left[\frac{1}{N-k}\sum_{i=1}^{N-k}x_{i+k}^2-\frac{1}{(N-k)^2}\left(\sum_{i=1}^{N-k}x_{i+k}\right)^2\right]^{1/2}}, \tag{6}$$

averages of m, the sum of the weights equals m. The weights may be positive or positive and negative. A simple moving average refers to the case where each of the weights equals one. Although a simple moving average tends to smooth out the data, it does not preserve the main features of the time series as well as a weighted moving average.

Generally, even a smooth trend obtained by the method of moving averages cannot be represented conveniently by a mathematical equation. The simplest mathematical expression is a straight line. However, a time series is apt to be such that a single linear trend can not be used throughout the time of observation. In such cases, it is possible to approximate the trend by using linear trends for portions of the time series.

After the trend line has been established, the trend can be removed from the data in one of several ways. One way is to take as a new variable the deviations about the trend line. These deviations must comprise a stationary time series. In some cases, the deviations may not comprise a stationary time series, but the deviations divided by their corresponding trend values may comprise a stationary time series.

If the method of moving averages is used to determine the trend in a time series which has an oscillatory movement about a trend, then a long-period oscillation tends to be included as part of the trend. Oscillations which are comparable in period to the length of the moving average, m, or shorter are damped out. The moving average also introduces an oscillatory movement into the random component of the time series. These consequences of the moving-average method are referred to as the Slutzky-Yule effect (Slutzky, 1937; Yule, 1921). Because of the Slutzky-Yule effect, care must be taken in discussing the oscillatory character of a time series if its trend has been removed by the moving-average method.

MEASURE OF NONRANDOMNESS

A time series is said to be randomly distributed if each event is independent of all preceding and following events. In analyzing sunspot data for periodicities, Yule (1927) found that the scatter diagrams for events k time units apart were linear. As a measure of linear dependence, Yule proposed equations 6 and 7.

In equation 6, x_i and x_{i+k} are the events at times i and $i+k$, respectively, N is the number of events forming the time series, and r_k is the kth-order serial-correlation coefficient.

where

$$\left.\begin{array}{l} r_o=1 \\ r_k=r_{-k} \end{array}\right\}. \tag{7}$$

Under the assumption that x_i varies linearly with x_{i+k}, r_k serves as a measure of linear dependence. If a time series is random, $r_k=0$ for all values of $k\geq 1$. However, for a finite sample, computed values of r_k may differ from zero because of sampling errors. For hydrologic time series, N is small, so that the sampling errors may be quite large. Thus the values of r_k must be tested to determine if they are significantly different from zero.

Anderson (1942) developed a test of significance based on a circular definition of the serial-correlation coefficients which supposes that the last event is followed by the first event so that

$$r_k=\frac{\frac{1}{N}\sum_{i=1}^{N}x_ix_{i+k}-\frac{1}{N^2}\left(\sum_{i=1}^{N}x_i\right)^2}{\frac{1}{N}\sum_{i=1}^{N}x_i^2-\frac{1}{N^2}\left(\sum_{i=1}^{N}x_i\right)^2}. \tag{8}$$

For N large and k small, the values of r_k given by equation 6 are nearly equal to those given by equation 8.

For a random normal time series, Anderson showed that r_1 is approximately normally distributed with mean $[-1/(N-1)]$ and variance $(N-2)/(N-1)^2$. Thus a computed value of r_1 may be tested for significance by

$$\hat{r}_1=\frac{-1\pm t_\alpha\sqrt{N-2}}{N-1}, \tag{9}$$

where t_α is the standard normal variate corresponding to a probability level α. At a given probability level, if $r_1>\hat{r}_1$, then r_1 is considered to be significantly different from zero.

GENERATING PROCESSES

A generating process refers to the manner by which the causal forces act to produce a time series. Some processes can be expressed mathematically, in which case it is possible to determine directly the various characteristics of the time series. Often a time series is approximated by a certain process. The choice of this process is based upon how well the mathematical structure of the process conforms to the physical characteristics underlying the time series. The processes which have been studied extensively and which have found wide application in practice are the moving average, the sum of harmonics, and the autoregression.

The moving average process may be expressed as

$$x_i=b_o+b_1y_i+b_2y_{i-1}+\ \ldots\ +b_my_{i-(m-1)}, \qquad (10)$$

where y is a random variable and m is the extent of the moving average. Equation 10 may be taken as a model representing the relation between annual runoff, x, and annual effective precipitation (Folse, 1929). The effective precipitation prior to a given time interval is referred to as antecedent effective precipitation. Since the contribution of effective precipitation to the runoff converges to zero very rapidly with an increase in antecedent time, the effective precipitation for only a finite period of antecedency affects the runoff. This finite period of antecedency, which is defined as the carryover period, is a function of the water-retardation characteristics of the river basin and the distribution of the effective precipitation with respect to time. Wold (1933) shows that the serial-correlation coefficients for a series generated by a moving average process are given by

$$r_k=\frac{\sum_{i=0}^{m-1} b_ib_{i+k}}{\sum_{i=0}^{m} b_i^2};\ 0\leq k\leq m-1, \qquad (11)$$

$$r_k=0;\ k\geq m.$$

Many hydrologic studies involve the use of time series which exhibit a cyclic movement; that is, sequences of daily or monthly discharges. The generating process for a cyclic time series may be represented by

$$x_i=A\sin\theta i+y_i, \qquad (12)$$

where A and θ are the amplitude and period, respectively, of the cyclic movement, and y is a random component. The random component, y, tends to distort the amplitude and the period. If the random component is large, A and θ may be badly distorted. For a cyclic time series, the serial-correlation coefficients are given by

$$r_k=\frac{A^2}{2\sigma_x^2}\cos\theta k, \qquad (13)$$

where σ_x^2 is the variance of the x's.

Equation 12 is a special case where only one harmonic, namely θ, is involved. Some hydrologists argue that there are hidden harmonics, or periodicities, in hydrologic data. These periodicities are called hidden because, if there are a large number of different periods, the series is very erratic—seemingly random.

In the discussion of the moving-average process as a physical model in hydrology, only a finite period of antecedency was considered. If the antecedent period is considered to be infinite, then a value of x at time i is a function of all previous values of x. This condition may be expressed as an autoregression process by defining x_i in terms of previous values of x and a random component. An example of such a process is

$$x_{i+2}=ax_{i+1}+bx_i+\epsilon_{i+2}, \qquad (14)$$

where a and b are constants and ϵ is the random component.

Kendall (1951) shows that by letting $p=\sqrt{\ }$ and $\cos\theta=-a/(2\sqrt{b})$, the serial-correlation coefficients for the autoregression process given by equation 13 are defined as

$$r_k=p^k\frac{\sin(k\theta+\psi)}{\sin\psi}, \qquad (15)$$

where

$$\tan\psi=\frac{1+p^2}{1-p^2}\tan\theta. \qquad (16)$$

It is assumed that $\sqrt{b}$ is taken with a positive sign and that $4b>a^2$. Moreover, it is assumed that $\sqrt{b}$ is not greater than unity.

A special type of autoregression is given by

$$x_{i+1}=r_1x_1+\epsilon_{i+1}, \qquad (17)$$

where r_1 is the first-order serial-correlation coefficient. This process is often referred to as the first-order Markov process. The serial-correlation coefficients are given by

$$r_k=r_1^k. \qquad (18)$$

CORRELOGRAM ANALYSIS

A graph with the serial-correlation coefficients plotted as ordinates and their respective orders as abscissas is called a correlogram. To reveal the features of the correlogram, the plotted points are joined each to the next by a straight line.

For a moving-average process, the values of r_k are given by equation 11. Since $r_k=0$ for $k\geq m$, the correlogram may oscillate, depending upon the values of b, but the correlogram will vanish for $k\geq m$. The values of r_k for the harmonic process are given by equation 13. For this process, the correlogram will oscillate but never vanish. The oscillation of the correlogram is strictly cyclic with amplitude $A^2/2\sigma_x^2$ and period θ. Note that the period of the correlogram is identical to the period of the time series. Equation 15 indicates that the correlogram of the autoregression process given by equation 14 will oscillate, be damped, but will not vanish. If the autoregression process is given by equation 17, then it is seen by equation 18 that the form of the correlogram will be "exponential." Thus for infinitely long sequences, the correlogram provides a theoretical basis for distinguishing among the three types of oscillatory time series.

In practice, N is small so that observed correlograms show less damping than theoretical correlograms because the observed serial-correlation coefficients are inflated by sampling errors. Thus one cannot determine the generating process by simply observing the correlogram. The goodness of fit of a given generating process fitted to an observed correlogram must be determined. Tests of goodness of fit which have been developed are based on N being very large.

In this study only the autoregression process is considered. For this generating process, a chi-square test developed by Quenouille (1949) can be used to test the goodness of fit. If the autoregression model is defined by equation 17, the value of chi-square, χ^2_k, associated with the kth-order serial-correlation coefficient is given by

$$\chi^2_k=\frac{(N-k)R^2_k}{(1-r^2_1)^2}, \qquad (19)$$

where

$$R_k=r_k-2r_1r_{k-1}+r^2_1r_{k-2}. \qquad (20)$$

The sum of the values of chi-square forms the basis for testing the goodness of fit. If

$$\sum_{k=1}^{m}\chi^2_k>\chi^2_m(\alpha), \qquad (21)$$

where $\chi^2_m(\alpha)$ is the value of chi-square at the probability-level α with m degrees of freedom (m is the highest order for which a serial-correlation coefficient is determined), then the first-order Markov process is rejected.

INVESTIGATION OF NONRANDOMNESS IN HYDROLOGIC TIME SERIES

SERIAL CORRELATION

If a time series is randomly distributed, the population values of the serial-correlation coefficients, called autocorrelation coefficients, between events for all orders are zero. For such a time series, the serial-correlation coefficients are not significantly different from zero at a probability-level α.

To investigate the distribution of droughts with respect to time, the annual minimum mean monthly rainfall and the annual minimum 1-day and 7-day discharge were considered as indices of droughts. The rainfall and discharge stations are listed in tables 1 and 2, respectively. A preliminary investigation for trend was made. Only the rainfall data for Charleston, S.C., strongly indicated a trend. However, the apparent trend was not removed.

By using equation 6, the first-order serial-correlation coefficients for the rainfall and discharge data were determined and are given in tables 3 and 4, respectively. The first-order serial-correlation coefficient given in tables 3 and 4 range from −0.296 to 0.300 for rainfall data and from −0.085 to 0.385 for discharge data. Although the discharge data exhibit higher first-order serial-correlation coefficients than the rainfall data, the variability of the first-order serial-correlation coefficients is greater for the rainfall data than for the discharge data.

These values of r_1 were tested for significance at the 90-percent and 95-percent levels by equation 9. An inspection of tables 3 and 4 shows that at the 90-percent level of significance, the data for 6 rainfall and 4 gaging stations exhibit significant values of r_1. At the 95-percent level, the data for 4 rainfall and 3 gaging stations yield significant values of r_1. Thus at the usual levels of significance employed in hydrologic studies, approximately ¼ to ⅓ of the stations for both phenomena possessed data yielding significant first-order serial-correlation coefficients.

The data which exhibited nonsignificant values of r_1 are not necessarily random in time. Serial-correlation coefficients of order higher than one, if significant, would indicate a lack of randomness. To determine if the data were random in time, the serial-correlation coefficients for orders 1 through 19 were considered for series having both significant and nonsignificant values of r_1. The data for 6 rainfall and 3 gaging stations were used. The values of r_k for $k=1, 2, \ldots, 19$ for the rainfall and discharge data are given in tables 5 and 6. Equation 9 was used as an approximate test of significance. The letter *a* denotes significance at the 90-percent level.

If serial-correlation coefficients were determined for a random time series, an occasional significant value would be expected to occur by chance. With respect to the 6 rainfall records, 114 serial-correlation coefficients were determined. If the data were randomly distributed and if the data for different stations were uncorrelated, 12 of these values would be expected to be significant at the 90-percent level. However, at the 90-percent level, 22 values were found to be significant. For the 3 discharge records, 57 serial-correlation coefficients were determined. If the data were random, 6 values would be expected to be significant by chance. Actually 10 values were found to be significant.

These results seem to indicate that the data are nonrandom. However, this investigation is handicapped by many factors. The number of stations used is very small, and the data for different stations, particularly the rainfall stations, are probably correlated. Also, most of the significant serial correlations for the rainfall data belonged to the two records having the largest values of r_1. One of these records was the Charleston, S.C., rainfall which apparently is not trend free. If these factors are taken into consideration, the rainfall data can be regarded as random in time. For the rainfall data to be nonrandom, there would need to be some form of atmospheric storage.

The discharge data may be nonrandomly distributed in time, with the nonrandomness being due to storage within the river basins. Note that the values of r_1 given in table 4 are positive except for the Pemigewasset River at Plymouth, N.H. Also, the significant serial correlations, given in table 6, are mainly those of low order.

GENERATING PROCESS

In figures 1 through 9, correlograms are shown for the data considered in the investigation of the serial-correlation coefficients for orders 1 through 19. A visual inspection of these correlograms shows no clear indication of a moving-average, cyclic, or autoregression process. By considering the significance of the serial-correlation coefficients, the rainfall correlograms may be considered as indicative of random processes where the nonzero values of the serial-correlation coefficients are due to sampling errors. For the discharge data, the low-order serial correlations were significant. This suggests that the first-order Markov process may approximate the generating process.

The first-order Markov process fitted to the discharge correlograms was tested for goodness of fit by Quenouille's chi-square at the 90-percent level. The values of R_k, χ_k^2, and $\Sigma\chi_k^2$ are given in table 7. The value of chi-square at the 95-percent level for $m=19$ degrees of freedom is 27.2. Since this value of chi-square is less than $\Sigma\chi_k^2$ for the discharge data for each of the three streams, the hypothesis that the generating process is a first-order Markov process is not rejected.

EFFECT OF SERIAL CORRELATION ON THE ESTIMATE OF THE VARIANCE

If a sequence of values is nonrandomly distributed in time, then each value repeats some of the information contained in previous values. Thus a longer sequence of nonrandom events is needed in order to arrive at a reliable estimate of the variance than in the case of a sequence of random events.

For a sequence of N' events, taken from a random time series having mean $\mu=0$ and variance σ^2, the estimate of the variance is given by

$$S^2=\frac{\sum_{i=1}^{N'} x_i^2}{N'}-\bar{x}^2, \qquad (22)$$

where $\bar{x}$ is the estimate of the mean defined by $\bar{x}=\sum_{i=1}^{N'} x_{i/N'}$. It follows that equation 22 can be written as

$$S^2=\frac{\sum_{i=1}^{N'} x_i^2}{N'}-\frac{\sum_{i=1}^{N'} x_i^2}{(N')^2}-2\,\frac{\sum_{i\neq j}\sum x_i x_j}{(N')^2}. \qquad (23)$$

The expectation of x_i is $E(x_i)=0$ and the expectation of x_i^2 is $E(x_i^2)=\sigma^2$. Since x is randomly distributed, $E(x_i x_j)=E(x_i)E(x_j)=0$. Thus the expectation of S^2 is

$$E(S^2)=\frac{N'-1}{N'}\,\sigma^2. \qquad (24)$$

Let ρ_k denote the population kth-order serial-correlation coefficient whose estimate is given by r_k. For a sequence of N events, taken from a nonrandom time series generated by a first-order Markov process having mean $\mu=0$ and variance σ^2, the estimate of the variance is given by equation 23. For this generating process, $E(x_i x_j)=\sigma^2\rho_{|i-j|}=\sigma^2\rho_{|k|}=\sigma^2\rho_1^{|k|}$. Also $E(x_i)=0$ and $E(x_i^2)=\sigma^2$. Hence

$$E(S^2)=\left[1-\frac{(1-\rho_1^2)}{N(1-\rho_1)^2}+\frac{2\rho_1(1-\rho_1^N)}{N^2(1-\rho_1)^2}\right]\sigma^2. \qquad (25)$$

If $\rho_1=0$ and $N=N'$, equation 25 reduces to equation 22. Assuming that $N=N'$, it can be shown that the quantity in the brackets in equation 25 is smaller than $(N'-1)/N'$. Hence N' must be larger than N in order to obtain as reliable an estimate of the variance for a nonrandom time series as for a random time series. By equating equations 23 and 25, it is possible to determine N' as a function of N and ρ_1. If N is the actual length of the time series, N' can be considered as the effective length of the time series. Thus if the variance is estimated from N events, the estimate is only as reliable as that estimated by a lesser number, N', of random events.

A graphical procedure facilitates the determination of N'. In figure 10, a family of curves is shown for $E(S^2)/\sigma^2$ versus N as a function of ρ_1. As N tends to infinity, $E(S^2)/\sigma^2$ tends to unity for all values of ρ_1. The larger ρ_1 is, the slower is the rate of convergence. For a given sequence, N is known and ρ_1 can be estimated by r_1. Thus, starting with the value of N on the abscissa, a vertical line is drawn upward to the curve corresponding to ρ_1. From this point of intersection, a horizontal line is drawn to the left until it intersects the curve for $\rho_1=0$, and then a vertical line is drawn downward to the abscissa scale to determine N'. An example of this graphical procedure for determining N' is shown in figure 10. In this example $N=30$ and $\rho_1=0.4$. The graphical solution gives $N'=13$.

DISCUSSION AND CONCLUSIONS

Time series of annual minimum monthly rainfall and annual minimum 1-day and 7-day discharge, considered as drought indices, were used to study the distribution of droughts with respect to time. An investigation for trends in the data indicated that the time series were trend free. The one exception was the rainfall data for Charleston, S.C. However, no attempt was made to remove the trend from this set of data.

The first-order serial-correlation coefficient, r_1, was determined for each rainfall and discharge series. The values of r_1 ranged from -0.296 to 0.300 for the rainfall data and from -0.085 to 0.385 for the discharge data. The discharge data exhibited higher but less variable values of r_1 than the rainfall data. At the 90-percent level of significance, 6 and 4 values of r_1 for the rainfall and discharge data, respectively, were found to be significant. At the 95-percent level, the rainfall data gave 4 significant values of r_1 and the discharge data gave 3 significant values of r_1. Thus at the usual levels of significance employed in hydrologic studies, approximately ¼ to ⅓ of the time series for both phenomena possessed significant values of r_1.

Data exhibiting nonsignificant values of r_1 are not necessarily random since serial-correlation coefficients of order greater than one, if significant, would indicate a lack of randomness. Serial-correlation coefficients for orders 1 through 19 were determined for the data for 6 rainfall and 3 gaging stations. A total of 114 and 57 serial-correlation coefficients was determined for the rainfall and discharge data, respectively. If these series were random, then 12 and 6 serial-correlation coefficients for rainfall and discharge data, respectively, would be expected to be significant at the 90-percent level. However, 22 and 10 serial-correlation coefficients for the rainfall and discharge data, respectively, were found to be significant.

This analysis seems to indicate that the data for both phenomena are nonrandomly distributed. Before drawing such a conclusion, several factors affecting the analysis must be considered. The number of records used was very small, and the data for different stations, particularly the rainfall data, were probably correlated. Over half the significant serial correlations for the rainfall data belonged to the two records having the largest values of r_1. One of these records was the Charleston, S.C., rainfall which apparently is not trend free. Thus the rainfall data are probably randomly distributed. If they were nonrandomly distributed, the nonrandomness would need to be attributed to some form of atmospheric storage.

The discharge records are not correlated. All the values of r_1 except one are positive, and the low-order serial correlations tend to be significant. Thus the discharge data may be regarded as nonrandom, with the nonrandomness being attributed to storage within the river basin.

The first-order Markov process was found to approximate the generating process of the discharge data. By assuming a time series to be generated by this process, the expectation of the estimate of the variance based on N observations was given and compared with the expectation of the estimate of the variance based on N' observations taken from a random time series. Because of the serial correlation, the estimate of the variance for a nonrandom time series based on N observations is only as reliable as the estimate of the variance for a random time series based on N' observations where N' is less than N. N', defined as the effective number of observations, is a function of N and the first-order autocorrelation coefficient, ρ_1. A graphical procedure (fig. 10) is given for determining N' for various values of N and ρ_1.

REFERENCES

Anderson, R. L., 1942, Distribution of the serial correlation coefficient: Annals Math. Statistics, v. 13, p. 1–13.

Anis, A. A., and Lloyd, E. H., 1953, On the range of partial sums of a finite number of independent normal variates: Biometrika, v. 40, p. 35–42.

Feller, William, 1951, The asymptotic distribution of the range of sums of independent random variables: Annals Math. Statistics, v. 22, p. 427.

Folse, J. A., 1929, A new method of estimating streamflow: Carnegie Inst. Washington Pub. 400.

Hurst, H. E., 1950, Long-term storage capacity of reservoirs: Proc. Am. Soc. Civil Engineers, Separate 11.

——— 1956, Methods of using long-term storage in reservoirs: Inst. Civil Engineers, Rept. 6059.

——— 1957, A suggested statistical model of some time series which occur in nature: Nature, v. 180, p. 494.

Kendall, M. G., 1951, The advanced theory of statistics—Part 2: New York, N. Y., Hafner Press, p. 406.

Moran, P. A. P., 1959, The theory of storage: London, England, Methuen and Co. Ltd., p. 66.

Quenouille, M. H., 1949, A large sample test for the goodness of fit of autoregressive schemes: Royal Statistical Soc. Jour., v. 110, p. 123.

Slutzky, E., 1937, The summation of random causes as the source of cyclic processes: Econometrika, v. 84, p. 105.

Wold, Herman, 1954, A study in the analysis of stationary time series: Uppsala, Sweden, Almqvist and Wiksells, p. 122.

Yule, G. U., 1921, On the time series problem: Royal Statistical Soc. Jour., v. 89, p. 497.

——— 1927, On a method of investigating periodicities in disturbed series with special reference to Wolfer's sunspot numbers: Philos. Trans., v. A226, p. 267.

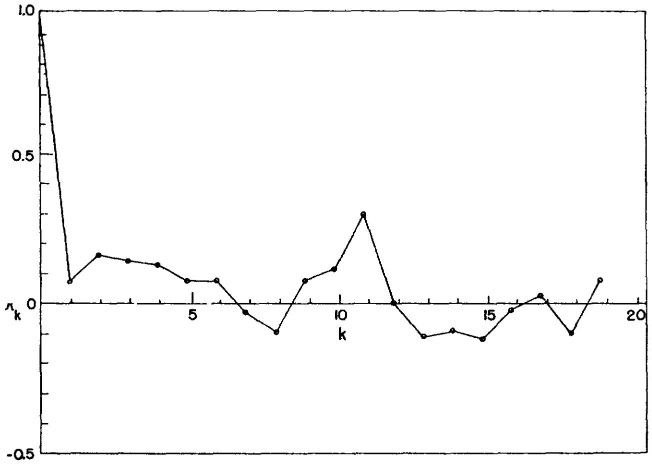

FIGURE 1.—Correlogram for annual minimum monthly rainfall at Boston, Mass.

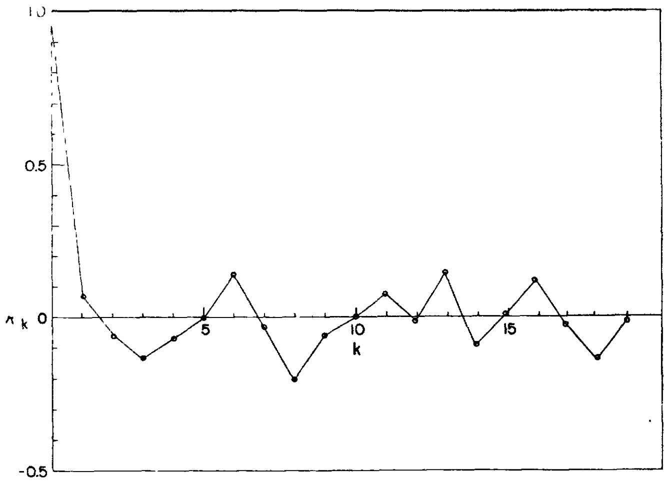

FIGURE 2.—Correlogram for annual minimum monthly rainfall at Providence, R.I.

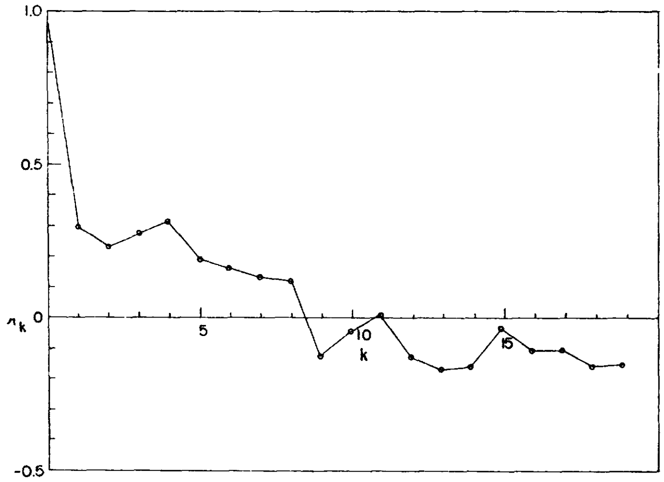

FIGURE 3.—Correlogram for annual minimum monthly rainfall at Charleston, S.C.

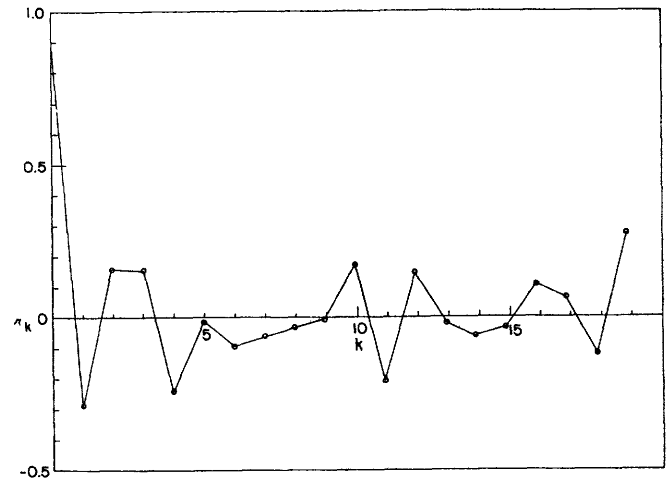

FIGURE 4.—Correlogram for annual minimum monthly rainfall at Washington, D.C.

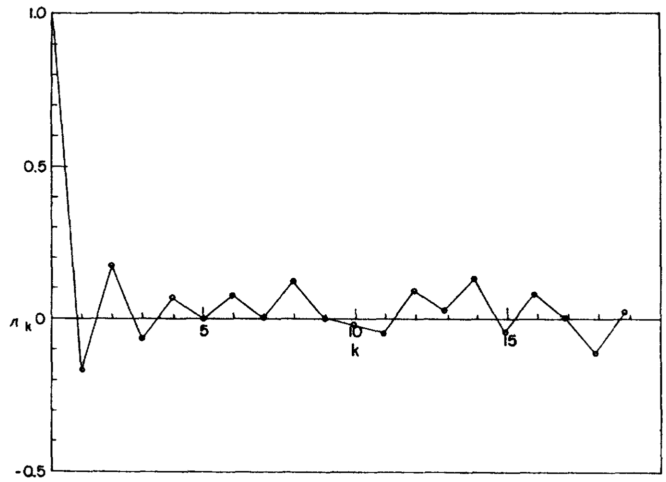

FIGURE 5.—Correlogram for annual minimum monthly rainfall at Baltimore, Md.

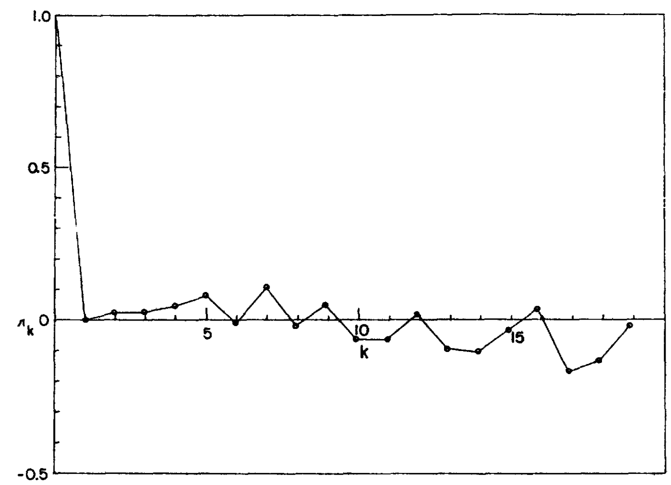

FIGURE 6.—Correlogram for annual minimum monthly rainfall at Philadelphia, Pa.

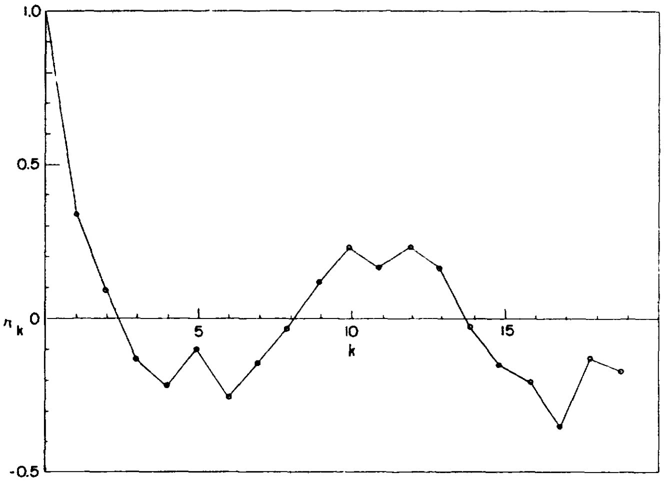

FIGURE 7.—Correlogram for annual minimum 7-day discharge for James River at Buchanan, Va.

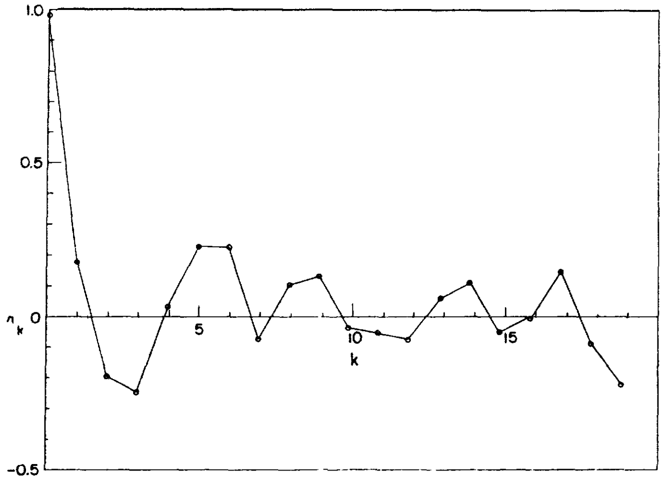

FIGURE 8.—Correlogram for annual minimum 7-day discharge for South Fork Holston River at Bluff City, Tenn.

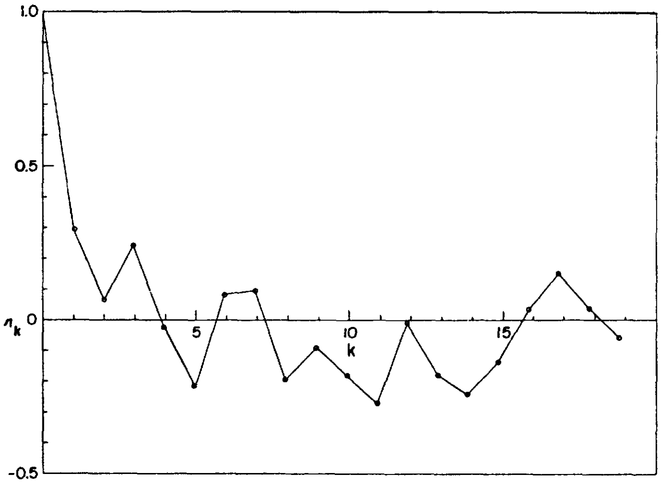

FIGURE 9.—Correlogram for annual minimum 1-day discharge for Middle Branch Westfield River near Goss Heights, Mass.

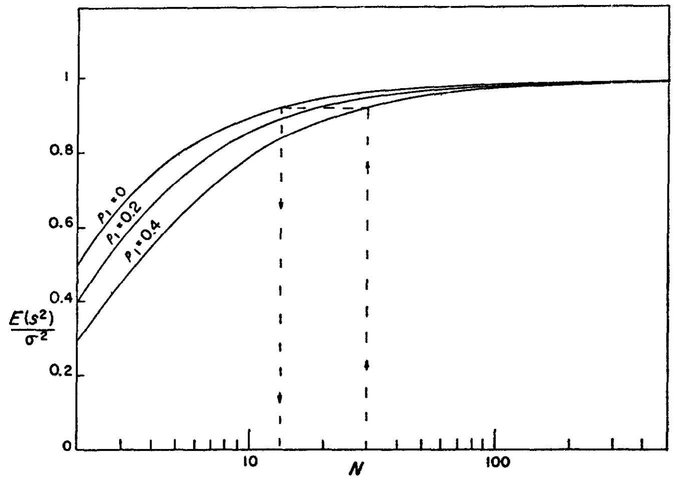

FIGURE 10.—Expected value of the variance for a first-order Markov process.

TABLE 1.—*Rainfall stations used in analyzing annual minimum monthly rainfall*

Rainfall station	Period of record	Length of record (years)	Rainfall station	Period of record	Length of record (years)
Amherst, Mass.	1836–1950	115	Albany, N.Y.	1826–1928	103
New Bedford, Mass.	1814–1950	137	Providence, R.I.	1832–1928	97
Lowell Locks and Canals, Lowell, Mass.	1855–1950	96	New York, N.Y.	1826–1943	118
Lowell, Mass.	1826–1925	100	Charleston, S.C.	1832–1953	122
Waltham, Mass.	1828–1930	102	St. Louis, Mo.	1837–1953	117
Cambridge, Mass.	1841–1921	81	Washington, D.C.	1852–1953	103
Boston, Mass.	1818–1913	96	Astoria, Ore.	1854–1953	100
Burlington, Vt.	1838–1928	91	New Burnswick, N.J.	1854–1953	100
			Baltimore, Md.	1818–1953	136
			Philadelphia, Pa.	1820–1953	134

TABLE 2.—*Gaging stations used in analyzing annual minimum discharge*

Gaging station	Period of record	Length of record (years)	Gaging station	Period of record	Length of record (years)
James River at Buchanan, Va.	1898–1954	57	Souhegan River at Merrimack, N.H.	1909–55	46
South Fork Holston River at Bluff City, Tenn.	1901–49	49	Susquehanna River at Conklin, N.Y.	1905–50	46
French Broad River at Asheville, N.C.	1903–53	50	Little Tennessee River at Judson, N.C.	1896–1941	46
Middle Branch of Westfield River at Goss Heights, Mass.	1910–55	45	Schroon River at River Bank, N.Y.	1908–53	46
Pemiguewasset River at Plymouth, N.H.	1903–55	52			

TABLE 3.—*Significance of first-order serial-correlation coefficients for rainfall*

Rainfall station	r_1	Confidence level			
		90-percent		95-percent	
Amherst, Mass	0.030	0.145	−0.152	0.174	−0.191
New Bedford, Mass	.070	.133	−.148	.160	−.175
Lowell Locks and Canals, Lowell, Mass	.060	.157	−.178	.190	−.211
Lowell, Mass	[1].166	.154	−.175	.186	−.206
Waltham, Mass	[2]−.213	.153	−.173	.184	−.204
Cambridge, Mass	.059	.171	−.196	.205	−.230
Boston, Mass	.071	.159	−.178	.190	−.211
Burlington, Vt	.089	.161	−.184	.195	−.217
Albany, N.Y	−.070	.152	−.172	.183	−.203
Providence, R.I	.070	.157	−.177	.189	−.210
New York, N.Y	−.139	.143	−.160	.171	−.188
Charleston, S.C	[2].300	.141	−.157	.168	−.185
St. Louis, Mo	[2]−.222	.143	−.161	.172	−.189
Washington, D.C	[2]−.296	.152	−.172	.183	−.203
Astoria, Oreg	.119	.154	−.175	.186	−.206
New Brunswick, N.J	−.046	.154	−.175	.186	−.206
Baltimore, Md	[1]−.171	.134	−.149	.160	−.175
Philadelphia, Pa	.000	.135	−.150	.162	−.177

[1] Significant at 90-percent level.
[2] Significant at 95-percent level.

TABLE 5.—*Serial-correlation coefficients for orders 1 through 19 for rainfall*

	Boston, Mass.	Providence, R.I.	Charleston, S.C.	Washington, D.C.	Baltimore, Md.	Philadelphia, Pa.
r_1	0.071	0.070	[1]0.300	[1]−0.296	[1]−0.171	0.000
r_2	[1].164	−.060	[1].238	[1].157	[1].180	.024
r_3	.142	−.138	[1].282	.150	−.067	.024
r_4	.130	−.072	[1].317	[1]−.249	.074	.049
r_5	.078	−.004	[1].196	−.013	.000	.087
r_6	.075	.140	[1].163	−.100	.086	−.003
r_7	−.031	−.031	.132	−.065	.004	.113
r_8	−.099	−.210	.120	−.034	.130	−.014
r_9	.078	−.063	−.129	−.008	.007	.054
r_{10}	.117	−.004	−.048	[1].175	−.011	−.059
r_{11}	[1].303	.077	.009	[1]−.217	−.043	−.059
r_{12}	.000	−.018	−.135	.147	.101	.028
r_{13}	−.117	.150	[1]−.175	−.021	.032	−.089
r_{14}	−.095	−.096	[1]−.163	−.066	[1].143	−.095
r_{15}	−.124	.007	−.037	−.037	−.039	−.023
r_{16}	−.023	.123	−.112	.105	.095	.044
r_{17}	.028	−.028	−.109	.062	.018	[1]−.162
r_{18}	−.107	.140	[1]−.164	−.213	−.106	−.127
r_{19}	.080	−.014	[1]−.159	[1].278	.035	−.003

[1] Significant at 90-percent level.

TABLE 7.—*Chi-square goodness of fit of first-order Markov process to discharge correlograms*

k	James River at Buchanan, Va.		South Fork Holston River at Bluff City, Tenn.		Middle Branch Westfield River near Goss Heights, Mass.	
	R_k	χ_k^2	R_k	χ_k^2	R_k	χ_k^2
1	−0.304	6.664	−0.172	1.521	−0.271	3.879
2	−.035	.085	−.232	2.692	−.027	.036
3	−.148	1.524	−.177	1.540	.232	2.718
4	−.120	.976	.121	.706	−.165	1.346
5	.038	.100	.218	2.225	−.184	1.621
6	−.218	3.107	.150	1.032	.210	2.069
7	.053	.180	−.146	.949	.030	.041
8	.018	.019	.142	.883	−.251	2.809
9	.129	1.033	.107	.486	.034	.051
10	.150	1.368	−.082	.283	−.147	.914
11	.024	.036	−.025	.025	−.169	1.174
12	.152	1.345	−.046	.086	.187	.742
13	.029	.052	.090	.311	−.203	1.585
14	−.119	.787	.096	.348	−.136	.694
15	−.111	.666	−.080	.228	−.011	.004
16	−.101	.533	.068	.165	.092	.293
17	−.229	2.705	.145	.714	.126	.535
18	.094	.447	−.134	.592	−.047	.071
19	−.122	.229	−.170	.928	−.011	.004
$\Sigma\chi_k^2$	------------	22.356	------------	15.714	------------	20.586

TABLE 4.—*Significance of first-order serial-correlation coefficients for discharge*

Gaging station	r_1	Confidence level			
		90-percent		95-percent	
James River at Buchanan, Va	[2]0.345	0.200	−0.236	0.241	−0.277
South Fork Holston River at Bluff City, Tenn	.178	.214	−.256	.258	−.300
French Broad River at Asheville, N.C	.154	.212	−.253	.259	−.300
Middle Branch Westfield River near Goss Heights, Mass	.297	.222	−.268	.268	−.314
Pemiguewasset River at Plymouth, N.H	−.082	.208	−.248	.253	−.292
Souhegan River at Merrimack, N.H	[2].385	.223	−.267	.267	−.311
Susquehanna River at Conklin, N.Y	[1].224	.223	−.267	.267	−.311
Little Tennessee River at Judson, N.C	.096	.223	−.267	.267	−.311
Schroon River at River Bank, N.Y	.059	.223	−.267	.267	−.311

[1] Significant at 90-percent level.
[2] Significant at 95-percent level.

TABLE 6.—*Serial-correlation coefficients for orders 1 through 19 for discharge*

	James River at Buchanan, Va.	South Fork Holston River at Bluff City, Tenn.	Middle Branch Westfield River near Goss Heights, Mass.
r_1	[1]0.345	0.178	[1]0.297
r_2	.084	−.200	.061
r_3	−.131	[1]−.254	[1].242
r_4	[1]−.220	.037	−.027
r_5	−.098	[1].239	[1]−.221
r_6	[1]−.259	[1].234	.081
r_7	−.114	−.070	.098
r_8	−.030	.110	−.200
r_9	.122	.148	−.093
r_{10}	[1].238	−.033	−.185
r_{11}	.174	−.041	−.271
r_{12}	[1].244	−.060	−.008
r_{13}	.177	.070	−.184
r_{14}	−.026	.123	−.245
r_{15}	−.130	−.038	−.140
r_{16}	−.201	.051	.030
r_{17}	[1]−.348	.164	.156
r_{18}	−.122	−.077	.093
r_{19}	−.165	−.203	−.061

[1] Significant at 90-percent level.

☆ U.S. GOVERNMENT PRINTING OFFICE: 1967 O—248-124