A Correlation Procedure for Augmenting Hydrologic Data

GEOLOGICAL SURVEY PROFESSIONAL PAPER 434-E

U.S. DEPARTMENT OF THE INTERIOR

March 3, 1849

A Correlation Procedure for Augmenting Hydrologic Data

By NICHOLAS C. MATALAS *and* BARBARA JACOBS

STATISTICAL STUDIES IN HYDROLOGY

GEOLOGICAL SURVEY PROFESSIONAL PAPER 434-E

UNITED STATES GOVERNMENT PRINTING OFFICE, WASHINGTON : 1964

UNITED STATES DEPARTMENT OF THE INTERIOR

STEWART L. UDALL, *Secretary*

GEOLOGICAL SURVEY

Thomas B. Nolan, *Director*

The U.S. Geological Survey Library has cataloged this publication as follows:

Matalas, Nicholas Constantinos, 1930–

A correlation procedure for augmenting hydrologic data, by Nicholas C. Matalas and Barbara Jacobs. Washington, U.S. Govt. Print. Off., 1964.

iii, 7 p. tables. 30 cm. (U.S. Geological Survey. Professional Paper 434–E)

Statistical studies in hydrology.

Bibliography: p. 6.

1. Hydrology - Stat. I. Reiher, Barbara S. (Jacobs) 1938– joint author. II. Title. (Series)

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C. 20402 - Price 15 cents (paper cover)

CONTENTS

TABLES

STATISTICAL STUDIES IN HYDROLOGY

A CORRELATION PROCEDURE FOR AUGMENTING HYDROLOGIC DATA

By NICHOLAS C. MATALAS and BARBARA JACOBS

ABSTRACT

A linear regression for a short and long sequence of hydrologic events is used to lengthen the short sequence. The lengthened sequence consists of the original observations and regressed values plus noise, where the noise is a random variable with zero mean and variance proportional to the variance of the observations for the short sequence about the line of regression. Estimates of the mean and variance for the lengthened sequence are shown to be unbiased. If the correlation coefficient, which measures the strength of the linear regression, exceeds about 0.5, then the estimates of the mean and variance based on the lengthened sequence are better estimators of the population values of the mean and variance than the estimates based on the observations for the short sequence. If noise is not added to the regressed values, the correlation coefficient must exceed about 0.8 to obtain improvement in the estimates by use of correlated values.

INTRODUCTION

In statistical studies of hydrology, the assumption is made that a sequence of a finite number of observed events represents a random sample from an infinite population of such events, where the outcome of each event is governed by some probability distribution. Any change in the hydrologic regime with which a given sequence is associated is reflected in a change of the probability distribution.

Although various parameters may be used to describe a probability distribution, the mean, variance, and skewness are three parameters which provide information about the most useful properties of any probability distribution. The mean is a measure of central tendency, a value about which the events tend to cluster, whereas the variance measures the dispersion or average spread of the events about the mean. The skewness is a measure of the asymmetry of the distribution of the events about the mean. These three parameters cannot define a probability distribution uniquely, except in special cases; nevertheless, they do provide characteristics which may be used to describe and compare various hydrologic phenomena. The population values of these parameters are generally unknown, but the values may be estimated from a sequence of observations. How reliable these estimates are depends primarily upon the period or length of the sequence—the total number of observations. If the estimates are unbiased, their reliability increases with an increase of the sequence length.

To increase a sequence length by making additional observations in time in order to obtain more reliable estimates of the mean, variance, and skewness is not always operationally or economically feasible. If not, recourse must be had to other procedures for increasing a sequence length. One such procedure, which is the topic of discussion in this paper, is to utilize the relations among hydrologic phenomena.

A relation among the concurrent observations for a short and a long sequence can be used to obtain estimates of the nonobserved events for the short sequence which correspond to the observed events for the nonconcurrent portion of the long sequence. In this manner the short sequence is lengthened. The observed and estimated events for the lengthened sequence can be used to obtain estimates of the mean, variance, and skewness. Whether or not the reliability of these estimates is greater than that for the estimates of these parameters based only on the observations depends mainly upon the strength of the relation between the concurrent observations for the short and long sequences.

A mathematical evaluation of this procedure has been made by several investigators under the assumptions that (1) the events are independently distributed in time, (2) the concurrent events for two sequences have a joint normal distribution, (3) the relation between the concurrent events is defined by a linear regression, and (4) no changes occur in the hydrologic regimes with which the sequences are associated. The strength of the linear regression is measured by the product-moment correlation coefficient. Under the assumption of normality, only the mean and variance need to be considered because these two parameters uniquely define a normal probability distribution. For this distribution, the skewness is zero.

H. A. Thomas, Jr. (written commun. 1956), showed that the lengthened sequence yields an unbiased estimate of the mean, and if the product-moment correlation coefficient exceeds $1/\sqrt{N_1-2}$ where N_1 is the length of the concurrent period, the reliability of this estimate of the mean is greater than that based only on the observations for the short sequence. These results also were obtained by W. G. Cochran (1953) in an earlier and independent study of a double sampling problem. J. R. Rosenblatt (1959) showed that the lengthened sequence yields a biased estimate of the variance, and that the reliability of this estimate is greater than that based only on the observations for the short sequence if the product-moment correlation coefficient exceeds about 0.8. M. B. Fiering (1963) obtained somewhat similar results for the estimates of the mean and variance when more than one long sequence is related to a short sequence by a multiple linear regression.

In the studies cited above, the estimated events in the lengthened sequence were regression estimates, that is, estimates which correspond to values on the line of regression. These estimates tend to yield a smaller variance than would the real observations. In order to "preserve" the variance inherent with the observations, a random component must be added to the regression estimates. This component, often referred to as noise, is normally distributed with zero mean and variance proportional to the variance of the observations for the short sequence about the line of regression.

This paper reports the results of an investigation made to determine the reliability of estimates of the mean and variance computed from a lengthened sequence when noise is added to the regression values. Noise is shown to have no effect on the reliability of the estimate of the mean. However, the addition of noise is shown to lead to an unbiased estimate of the variance for the lengthened sequence. The reliability of this estimate is greater than that when no noise is added to the regression values.

STATISTICAL MODEL

The long and short sequences for a pair of hydrologic phenomena are denoted by x and y. In general, the two phenomena need not be the same. If, for example, y denotes streamflow, x may denote streamflow, precipitation, temperature, or a geochronologic phenomena such as tree-ring widths. The observed events for the long and short sequences are represented as

$$x_1, \ldots, x_{N_1}, x_{N_1+1}, \ldots, x_{N_1+N_2},$$

$$y_1, \ldots, y_{N_1},$$

where N_1 is the length of the short sequence and (N_1+N_2) is the length of the long sequence. For this representation of the two sequences, N_1 also denotes the concurrent period of observation. In practice, the N_1 observations for the short sequence need not correspond to the first N_1 observations of the long sequence, nor need the N_1 concurrent observations of x and y occur consecutively. However, there is no loss of generality if the two sequences are represented as above.

The concurrent observations x and y are assumed to have a joint normal probability distribution with parameters μ_x, μ_y, σ_x^2, σ_y^2, and $\rho=\beta\sigma_x/\sigma_y$, where μ_x and σ_x^2 denote the population values of the mean and variance, respectively, for x, and μ_y and σ_y^2, the population mean and variance, respectively, for y. The parameter ρ is the product-moment correlation coefficient, and β is the population value of the slope for the linear regression of y on x.

For the short sequence, the estimates of the mean and variance are given by

$$\bar{y}_1=\frac{1}{N_1}\sum_{i=1}^{N_1} y_i, \tag{1}$$

$$s_{y_1}^2=\frac{1}{(N_1-1)}\sum_{i=1}^{N_1} (y_i-\bar{y}_1)^2, \tag{2}$$

respectively. Similarly for the long sequence

$$\bar{x}=\frac{1}{(N_1+N_2)}\sum_{k=1}^{N_1+N_2} x_k=\frac{N_1\bar{x}_1+N_2\bar{x}_2}{N_1+N_2}, \tag{3}$$

$$\begin{aligned} s_x^2&=\frac{1}{(N_1+N_2-1)}\sum_{k=1}^{N_1+N_2} (x_k-\bar{x})^2 \\ &=\frac{1}{(N_1+N_2-1)}\left[(N_1-1)s_{x_1}^2+(N_2-1)s_{x_2}^2\right. \\ &\qquad\left.+\frac{2N_1N_2}{(N_1+N_2)}(\bar{x}_1-\bar{x}_2)^2\right], \end{aligned} \tag{4}$$

where

$$\bar{x}_1=\frac{1}{N_1}\sum_{i=1}^{N_1} x_i, \tag{5}$$

$$\bar{x}_2=\frac{1}{N_2}\sum_{j=N_1+1}^{N_1+N_2} x_j, \tag{6}$$

$$s_{x_1}^2=\frac{1}{(N_1-1)}\sum_{i=1}^{N_1} (x_1-\bar{x}_1)^2, \tag{7}$$

$$s_{x_2}^2=\frac{1}{(N_2-1)}\sum_{j=N_1+1}^{N_1+N_2} (x_j-\bar{x}_2)^2. \tag{8}$$

In equations 1 through 8, the subscripts 1 and 2 indicate that the estimates are based on the observational periods N_1 and N_2, respectively. No numerical subscript is used with those estimates based on the period (N_1+N_2).

Because each of the above estimates converges to its respective population value as the observational periods tend to infinity, the estimates are unbiased. Therefore, the expectations of the estimates equal their respective population values (Wilks, 1962). If the symbol E is used to denote the expectation, then

$$E(\bar{y}_1)=\mu_y, \qquad (9)$$

$$E(s^2_{y_1})=\sigma^2_y, \qquad (10)$$

$$E(\bar{x}_1)=E(\bar{x}_2)=E(\bar{x})=\mu_x, \qquad (11)$$

$$E(s^2_{x_1})=E(s^2_{x_2})=E(s^2_x)=\sigma^2_x. \qquad (12)$$

A measure of the reliability of an estimate is given by its variance. From the assumption of normality, the variance, denoted by Var, for each of the estimates (Wilks, 1962) is

$$\text{Var }(\bar{y}_1)=\sigma^2_y/N_1, \qquad (13)$$

$$\text{Var }(s^2_{y_1})=2\sigma^4_y/(N_1-1), \qquad (14)$$

$$\text{Var }(\bar{x}_1)=\sigma^2_x/N_1, \qquad (15)$$

$$\text{Var }(\bar{x}_2)=\sigma^2_x/N_2, \qquad (16)$$

$$\text{Var }(\bar{x})=\sigma^2_x/(N_1+N_2), \qquad (17)$$

$$\text{Var }(s^2_{x_1})=2\sigma^4_x/(N_1-1), \qquad (18)$$

$$\text{Var }(s^2_{x_2})=2\sigma^4_x/(N_2-1), \qquad (19)$$

$$\text{Var }(s^2_x)=2\sigma^4_x/(N_1+N_2-1). \qquad (20)$$

The variance of each estimate varies inversely with the period of observation. Therefore, as the period of observation increases, the variance of an estimate decreases, and consequently, the reliability of the estmate increases.

The true regression of y on x is defined as

$$Y_i=\mu_y+\beta(x_i-\mu_x), \qquad (21)$$

where Y_i denotes the regression value for y for a given value x_i. The y_i are normally and independently distributed about the regression values with variance $(1-\rho^2)\sigma^2_y$. Therefore, the y_i may be expressed as

$$y_i=\mu_y+\beta(x_i-\mu_x)+\sqrt{1-\rho^2}\sigma_y\epsilon_i, \qquad (22)$$

where ϵ is a random normal variable with zero mean and unit variance, and the term $\sqrt{1-\rho^2}\sigma_y\epsilon_i$ represents noise.

From the method of least squares, the estimate of the true regression of y on x is given by

$$\hat{Y}_i=\bar{y}_1+b(x_i-\bar{x}_1), \qquad (23)$$

where $\hat{Y}_i$ is the estimate of Y_i and b, the estimate of β, is given by

$$b=\sum_{i=1}^{N_1} y_i(x_i-\bar{x}_1)/\sum_{i=1}^{N_1} (x_i-\bar{x}_1)^2. \qquad (24)$$

If the right-hand side of equation 22 is substituted for y_i in equation 24, b may be expressed as

$$b=\beta+\sqrt{1-\rho^2}\sigma_y \sum_{i=1}^{N_1} \epsilon_i(x_i-\bar{x}_1)/\sum_{i=1}^{N_1} (x_i-x_1)^2. \qquad (25)$$

If the noise term is multiplied by $\alpha\theta$ and the population parameters are replaced by their sample estimates, equation 22 may be expressed as

$$y_i=\bar{y}_1+b(x_i-\bar{x}_1)+\alpha\theta\sqrt{1-r^2}s_{y_1}e_i, \qquad (26)$$

where e is a random normal variable with zero mean and unit variance, α is a constant which will be defined later, and r, the estimate of ρ, is given by

$$r=bs_{x_1}/s_{y_1}. \qquad (27)$$

The parameter θ is introduced to facilitate the comparison of the case when noise is added with the case when noise is not added. If noise is added, $\theta=1$, and if noise is not added, $\theta=0$. In equation 26, if i is replaced by j, then N_2 values of x outside the concurrent period may be used to obtain estimates of $y_{N_1+1}, \ldots, y_{N_1+N_2}$. These estimates, denoted by $\hat{y}_{N_1+1}, \ldots, \hat{y}_{N_1+N_2}$, may be pooled with the observations $y_1, \ldots, y_{N_1}$, to form the lengthened sequence

$$y_1, \ldots, y_{N_1}, \hat{y}_{N_1+1}, \ldots, \hat{y}_{N_1+N_2}.$$

For the lengthened sequence, the estimate of the mean is

$$\bar{y}=\frac{N_1\bar{y}_1+N_2\bar{y}_2}{N_1+N_2}=\bar{y}_1+\frac{N_2}{(N_1+N_2)}\,b(\bar{x}_2-\bar{x}_1)$$
$$+\frac{N_2\alpha\theta}{(N_1+N_2)}\sqrt{1-r^2}s_{y_1}\bar{e}_2, \qquad (28)$$

where $\bar{y}_1$ is defined by equation 1 and

$$\bar{y}_2=\frac{1}{N_2}\sum_{j=N_1+1}^{N_1+N_2}\hat{y}_j=\bar{y}_1+b(\bar{x}_2-\bar{x}_1)+\alpha\theta\sqrt{1-r^2}s_{y_1}\bar{e}_2. \qquad (29)$$

Also, for the lengthened sequence, the estimate of the variance is given by

$$
\begin{aligned}
s_y^2 &= \frac{1}{(N_1+N_2-1)}\left[\sum_{i=1}^{N_1}(y_i-\overline{y})^2+\sum_{j=N_1+1}^{N_1+N_2}(\hat{y}_j-\overline{y})^2\right], \\
&= \frac{1}{(N_1+N_2-1)}\Bigg[(N_1-1)s_{y_1}^2+(N_2-1)b^2s_{x_2}^2 \\
&\quad +\frac{N_1N_2}{(N_1+N_2)}b^2(\overline{x}_2-\overline{x}_1)^2+\frac{N_1N_2}{(N_1+N_2)}\alpha^2\theta^2(1-r^2)s_{y_1}^2\overline{e}_2^2 \\
&\quad +\frac{2N_1N_2}{(N_1+N_2)}\alpha\theta\sqrt{1-r^2}s_{y_1}(\overline{x}_2-\overline{x}_1\ )\overline{e} \\
&\quad +2(N_2-1)\alpha\theta b\sqrt{1-r^2}s_{y_1}s_{xe_2}+(N_2-1)\alpha^2\theta(1-r^2)s_{y_1}^2s_{e_2}^2\Bigg],
\end{aligned} \tag{30}
$$

where $\overline{x}_1$, $\overline{x}_2$, $s_{y_1}^2$, $s_{x_2}^2$, b, r, and $\overline{y}$ are defined by equations 5, 6, 2, 8, 24, 27, and 28, respectively, and

$$\overline{e}_2=\frac{1}{N_2}\sum_{j=N_1+1}^{N_1+N_2}e_j, \tag{31}$$

$$s_{e_2}^2=\frac{1}{(N_2-1)}\sum_{j=N_1+1}^{N_1+N_2}(e_j-\overline{e}_2)^2, \tag{32}$$

$$s_{xe_2}=\frac{1}{(N_2-1)}\sum_{j=N_1+1}^{N_1+N_2}(x_j-\overline{x}_2)(e_j-\overline{e}_2). \tag{33}$$

The use of equations 28 and 30 necessitates that noise actually be added to the regression estimates. Truly random numbers do not exist, and therefore recourse must be had either to generating pseudorandom numbers or utilizing such numbers that have been tabulated and are to be found in the statistical literature. Regardless of the questions which can be raised as to the goodness of the pseudorandom numbers, their use is not too appealing. Independent studies of the same sequences of x and y by several investigators lead to different values of $\overline{y}$ and s_y^2, because the same sequence of pseudorandom numbers is unlikely to be used by the investigators. Some of these values are apt to appear absurd, even though the values are within the limits expected by chance.

To overcome these objections to the use of equations 28 and 30, the following approximations are offered. On the basis of the observed sequences $x_1, \ldots, x_{N_1}$, and $y_1, \ldots, y_{N_1}$, the best estimates of $\overline{e}_2$, $s_{e_2}^2$, and s_{xe_2} are 0, 1, and 0, respectively. Thus, if $\overline{e}_2$, $s_{e_2}^2$, and s_{xe_2} are replaced by their respective estimates, equations 28 and 30 reduce to

$$\overline{y}=\overline{y}_1+\frac{N_2}{(N_1+N_2)}b(\overline{x}_2-\overline{x}_1), \tag{34}$$

$$
\begin{aligned}
s_y^2 &= \frac{1}{(N_1+N_2-1)}\Bigg[(N_1-1)s_y^2+(N_2-1)b^2s_{x_2}^2 \\
&\quad +(N_2-1)\alpha^2\theta^2(1-r^2)s_{y_1}^2+\frac{N_1N_2}{(N_1+N_2)}b^2(\overline{x}_2-\overline{x}_1)^2\Bigg]
\end{aligned} \tag{35}
$$

respectively. Both equations 34 and 35 are expressed in terms that can be determined directly from the observations.

Equation 34, which gives an approximate estimate of the population mean from the lengthened sequence, does not contain any terms that reflect the "addition" of noise. Actually equation 34 gives the same estimate of the mean for the lengthened sequence as when noise is not added to the regression values. This is seen to be the case if θ is set equal to zero, in which case equation 28 reduces to equation 34.

In equation 35, which gives an approximate estimate of the variance for the lengthened sequence, the "addition" of noise is reflected by the third term in brackets.

RELIABILITY OF ESTIMATES

To judge the reliability of an estimate, two properties, namely the expected value and the variance of the estimate, must be considered. The estimate of the mean for the lengthened sequence, given by equation 34, is the same as that considered by Thomas and Cochran. Therefore, the discussion of the reliability of this estimate is given without the derivation of the expected value and variance of the estimate. Because the derivation of the expected value and variance of s_y^2 is rather long, the derivation is omitted from this section and outlined in the appendix on page E6.

The estimate $\overline{y}$ is an unbiased estimator of μ_y, so that

$$E(\overline{y})=\mu_y. \tag{36}$$

The variance of $\overline{y}$ is given by

$$\operatorname{Var}(\overline{y})=\frac{\sigma_y^2}{N_1}\left\{1-\frac{N_2}{(N_1+N_2)}\left[\rho^2-\frac{(1-\rho^2)}{(N_1-3)}\right]\right\}, \tag{37}$$

where the term σ_y^2/N_1 denotes the variance of $\overline{y}_1$. For $\overline{y}$ to be a better estimate of μ_y than $\overline{y}_1$, Var $(\overline{y})<$ Var $(\overline{y}_1)=\sigma_y^2/N_1$. This is the case if the term in the brackets in equation 37 is negative, that is, if

$$|\rho|>1/\sqrt{N_1-2}. \tag{38}$$

Equation 38 shows that to improve the estimate of the mean by the use of a linear regression relation, the strength of the regression, measured by ρ, must exceed a critical minimum value which depends only upon N_1. If inequality (38) is not satisfied, then Var $(\overline{y})>$ Var $(\overline{y}_1)$, in which case $\overline{y}_1$ is a better estimate of μ_y than $\overline{y}$ In this case, the linear regression relation serves no practical purpose. Equation 38 shows that the critical minimum value of ρ cannot be determined for $N_1 \leq 2$. Table 1 gives the minimum critical value of ρ for various values of $N_1 \geq 10$.

TABLE 1.—*Critical minimum values of ρ for the mean*

N_1:	10	15	20	25	30
ρ:	0. 35	0. 28	0. 24	0. 21	0. 19

The expected value of s_y^2, the derivation of which is outlined in the appendix, is

$$E(s_y^2)=\sigma_y^2 \left\{1-\frac{(1-\rho^2)}{(N_1+N_2-1)}\left[\frac{N_2(N_1-4)}{(N_1-3)} -\frac{(N_1-2)(N_2-1)}{(N_1-1)}\alpha^2\theta^2\right]\right\}. \qquad (39)$$

If

$$\alpha^2=\frac{N_2(N_1-4)(N_1-1)}{(N_2-1)(N_1-3)(N_1-2)},$$

and $\theta=1$, then $E(s_y^2)=\sigma_y^2$, so that s_y^2 is an unbiased estimator of σ_y^2. If $\theta=0$, equation 39 reduces to the biased estimator of σ_y^2 derived by Rosenblatt.

The variance of s_y^2 is

$$\text{Var}\ (s_y^2)=\frac{2\sigma_y^4}{(N_1-1)}+\frac{N_2\sigma_y^4}{(N_1+N_2-1)^2}[A\rho^4+B\rho^2+C], \qquad (41)$$

where

$$A=\left[\frac{(N_2+2)(N_1-6)(N_1-8)}{(N_1-3)(N_1-5)}-\frac{8(N_1-4)}{(N_1-3)} -\frac{2N_2(N_1-4)^2}{(N_1-3)^2}\theta^2+\frac{N_1N_2(N_1-4)^2}{(N_1-3)^2(N_1-2)}\theta^4\right]\rho^4, \qquad (42)$$

$$B=\left[\frac{6(N_2+2)(N_1-6)}{(N_1-3)(N_1-5)}+\frac{2(N_1^2-N_1-14)}{(N_1-3)} -\frac{2(N_1+N_2-1)(N_1-4)}{(N_1-3)}(1-\theta^2) +\frac{2N_2(N_1-4)(N_1-5)}{(N_1-3)^2}\theta^2-\frac{2(N_1-4)(N_1+3)}{(N_1-3)}\theta^2 -\frac{2N_1N_2(N_1-4)^2}{(N_1-3)^2(N_1-2)}\theta^4\right]\rho^2, \qquad (43)$$

$$C=\left[\frac{2(N_1+1)}{(N_1-3)}+\frac{3(N_2+2)}{(N_1-3)(N_1-5)}-\frac{(N_1+1)(2N_1+N_2-2)}{(N_1-1)} +\frac{2(N_1+N_2-1)(N_1-4)}{(N_1-3)}(1-\theta^2)+\frac{2N_2(N_1-4)}{(N_1-3)^2}\theta^2 +\frac{2(N_1-4)(N_1+1)}{(N_1-3)}\theta^2+\frac{N_1N_2(N_1-4)^2}{(N_1-3)^2(N_1-2)}\theta^4\right], \qquad (44)$$

In equation 41, the term $2\sigma_y^4/(N_1-1)$ denotes the variance of $s_{y_1}^2$. If the second term on the right-hand side of equation 41 is negative, then Var $(s_y^2)<$Var $(s_{y_1}^2)=2\sigma_y^4/(N_1-1)$, so that s_y^2 is a better estimator of σ_y^2 than $s_{y_1}^2$. In this case, the correlation coefficient, σ, must exceed a critical minimum value which is given by

$$|\rho|>\left[\frac{-B\pm\sqrt{B^2-4AC}}{2A}\right]^{\frac{1}{2}}, \qquad (45)$$

where A, B, and C are defined by equations 42, 43, and 44, respectively. If inequality (45) is not satisfied, then $s_{y_1}^2$ is a better estimator of σ_y^2 than s_y^2. If $\theta=0$, equation 41 and the expressions for A, B, and C defined by equations 42, 43, and 44, respectively, reduce, after a rearrangement of terms, to the expression for the mean square error of s_y^2 derived by Rosenblatt.

Tables 2 and 3 give the critical minimum values of ρ for various values of N_1 and N_2 for $\theta=1$ and $\theta=0$, respectively.

TABLE 2.—*Critical minimum values of ρ for the variance: θ=1*

N_2 \ N_1	10	15	20	25	30
10	0. 65	0. 54	0. 52	0. 42	0. 38
15	. 65	. 54	. 51	. 42	. 39
20	. 65	. 54	. 51	. 42	. 39
25	. 65	. 54	. 50	. 42	. 39
30	. 65	. 54	. 50	. 42	. 39

TABLE 3.—*Critical minimum values of ρ for the variance: θ=0*

N_2 \ N_1	10	15	20	25	30
10	0. 73	0. 63	0. 70	0. 76	0. 76
15	. 75	. 77	. 79	. 80	. 80
20	. 76	. 79	. 81	. 81	. 82
25	. 78	. 80	. 84	. 83	. 81
30	. 77	. 80	. 82	. 83	. 84

An inspection of these tables show that for given values of N_1 and N_2, ρ is less when noise is added, $\theta=1$, than when noise is not added, $\theta=0$. Because the critical minimum values of ρ, equation 45, cannot be defined for values of $N_1<6$, tables 2 and 3 are based on $N_1 \geq 1.0$

DISCUSSION

In practice, the correlation coefficient, ρ, must be estimated by r. For given values of N_1 and N_2, if r exceeds the appropriate critical minimum values tabulated in tables 1–3, then the mean and variance for the lengthened sequence may be determined from the approximate expressions given by equations 34 and 35, respectively. The use of equations 34 and 35 when r does not exceed the appropriately tabulated critical minimum values results in estimates which are less reliable than estimates based only upon the observations. However, caution must be exercised in the use of these tables, because r may not be a good estimate of ρ and because the hydrologic variables, x and y, may not satisfy the assumptions upon which the above developments are based.

The reliability of r as an estimate of ρ depends upon the length of the concurrent period, N_1, for the short and long sequences. At this point, all that can be said is that N_1 should be as large as possible. If ρ were known, then table 1 could be used if $N_1>3$ and tables 2 and 3 could be used if $N_1>5$. Because ρ is not known and must be estimated by r, N_1 should be much larger than 5.

Most hydrologic variables exhibit skewness and therefore they cannot be considered as normally distributed. Usually, simple transformations, such as logarithms and square roots, may be used to normalize the data. If normalizing transforms are used, the developments noted above pertain not to the variables but to the transformed variables. Transformations may also be needed to linearize the relation between x and y.

Hydrologic data exhibit a tendency for high values to follow high values and for low values to follow low values, so that the values are not independently distributed in time. This dependency, which is attributed to storage processes in the hydrologic regime with which a given sequence is associated, decreases rather rapidly with an increase in the time interval between observations. For example, a sequence of monthly events exhibits greater dependence among adjacent events than an annual sequence. Simple transformations of a variable are unlikely to reduce the dependence among variate values. Therefore, the use of the developments noted above should be restricted to annual sequences whose hydrologic regimes are such that storage or carryover factors are a minimum.

REFERENCES

Cochran, W. G., 1953, Sampling techniques: New York, John Wiley & Sons, Inc.

Fiering, M. B., 1963, Use of correlation to improve estimates of the mean and variance: U.S. Geol. Survey Prof. Paper 434-C.

Rosenblatt, J. R., 1959, On the synthetic record problem: Natl. Bur. Standards Tech. Rept. 1.

Wilks, S. S., 1962, Mathematical statistics: New York, John Wiley & Sons, Inc.

APPENDIX

Expectation of s_y^2:

$$E(s_y^2)=E\left\{\frac{1}{(N_1+N_2-1)}\left[(N_1-1)s_{y_1}^2+(N_2-1)b^2s_{x_2}^2\right.\right.$$

$$\left.\left.+(N_2-1)\alpha^2\theta^2(1-r^2)s_{y_1}^2+\frac{N_1N_2}{(N_1+N_2)}\,b^2(\bar{x}_2-\bar{x}_1)^2\right]\right\}$$

$$=\frac{1}{(N_1+N_2-1)}\left[(N_1-1)E(s_{y_1}^2)\right.$$

$$+(N_2-1)E(b^2)E(s_{x_2}^2)+(N_2-1)\alpha^2\theta^2E(1-r^2)s_{y_1}^2$$

$$\left.+\frac{N_1N_2}{(N_1+N_2)}\,E(b^2)E(\bar{x}_2-\bar{x}_1)^2\right]$$

$$=\sigma_y^2\left\{1-\frac{(1-\rho^2)}{(N_1+N_2-1)}\left[\frac{N_2(N_1-4)}{(N_1-3)}\right.\right.$$

$$\left.\left.-\frac{(N_1-2)(N_2-1)}{(N_1-1)}\,\alpha^2\theta^2\right]\right\}, \qquad \text{(A-1)}$$

where

$$E(s_{y_1}^2)=\sigma_y^2, \qquad \text{(A-2)}$$

$$E(b^2)=\left[\rho^2+\frac{(1-\rho^2)}{(N_1-1)}\right]\frac{\sigma_y^2}{\sigma_x^2}, \qquad \text{(A-3)}$$

$$E(s_{x_2}^2)=\sigma_x^2, \qquad \text{(A-4)}$$

$$E(\bar{x}_2-\bar{x}_1)^2=\frac{(N_1+N_2)}{N_1N_2}\,\sigma_x^2, \qquad \text{(A-5)}$$

$$E(1-r^2)s_{y_1}^2=\frac{(N_1-2)}{(N_1-1)}\,(1-\rho^2)\,\frac{\sigma_y^2}{\sigma_x^2}. \qquad \text{(A-6)}$$

If $\theta=1$ and

$$\alpha^2=\frac{N_2(N_1-4)(N_1-1)}{(N_2-1)(N_1-3)(N_1-2)}, \qquad \text{(A-7)}$$

then equation A–1 reduces to

$$E(s_y^2)=\sigma_y^2. \qquad \text{(A-8)}$$

Variance of s_y^2:

$$\text{Var }(s_y^2)=E(s_y^2-\sigma_y^2)^2=E(s_y^2)^2-2\sigma_y^2E(s_y^2)+\sigma_y^4, \qquad \text{(A-9)}$$

$$
\begin{aligned}
E(s_y^2)^2=\frac{1}{(N_1+N_2-1)^2}\Big[&(N_1-1)^2E(s_{y_1}^2)^2\\
&+(N_2-1)^2E(b^4)E(s_{x_2}^2)^2\\
&+(N_2-1)^2\alpha^4\theta^4E[(1-r^2)^2(s_{y_1}^2)^2]\\
&+\frac{(N_1N_2)^2}{(N_1+N_2)^2}E(b^4)E(\bar{x}_2-x_1)^4\\
&+2(N_1-1)(N_2-1)E(b^2s_{y_1}^2)E(s_{x_2}^2)\\
&+2(N_1-1)(N_2-1)\alpha^2\theta^2E[(1-r^2)(s_{y_1}^2)^2]\\
&+\frac{2(N_1-1)N_1N_2}{(N_1+N_2)}E(b^2s_{y_1}^2)E(\bar{x}_2-\bar{x}_1)^2\\
&+2(N_2-1)^2\alpha^2\theta^2E[b^2(1-r^2)s_{y_1}^2]E(s_{x_2}^2)\\
&+\frac{2(N_2-1)N_1N_2}{(N_1+N_2)}E(b^4)E(\bar{x}_2-\bar{x}_1)^2E(s_{x_2}^2)\\
&+\frac{2(N_2-1)N_1N_2}{(N_1+N_2)}\alpha^2\theta^2E[b^2(1-r^2)s_{y_1}^2]E(\bar{x}_2-\bar{x}_1)^2\Big],
\end{aligned}
\tag{A-10}
$$

$$E(s_{y_1}^2)^2=\frac{(N_1+1)}{(N_1-1)}\sigma_y^4, \tag{A-11}$$

$$E(s_{x_2}^2)=\frac{(N_2+1)}{(N_2-1)}\sigma_x^4, \tag{A-12}$$

$$E(b^4)=\left[\rho^4+\frac{6\rho^2(1-\rho^2)}{(N_1-3)}+\frac{3(1-\rho^2)^2}{(N_1-3)(N_1-5)}\right]\frac{\sigma_y^4}{\sigma_x^4}, \tag{A-13}$$

$$E(\bar{x}_2-\bar{x})^2=\frac{(N_1+N_2)}{N_1N_2}\sigma_x^2, \tag{A-14}$$

$$E(\bar{x}_2-\bar{x})^4=\frac{3(N_2+N_2)^2}{(N_1N_2)^2}\sigma_x^4, \tag{A-15}$$

$$E[(1-r^2)^2(s_{y_1}^2)^2]=\frac{N_1(N_1-2)}{(N_1-1)^2}(1-\rho^2)^2\sigma_y^4, \tag{A-16}$$

$$E(b^2s_{y_1}^2)=\left[\rho^4+\frac{(N_1+4)}{(N_1-1)}\rho^2(1-\rho^2)+\frac{(N_1+1)(1-\rho^2)^2}{(N_1-1)(N_1-3)}\right]\frac{\sigma_y^4}{\sigma_x^2}, \tag{A-17}$$

$$
\begin{aligned}
E[(1-r^2)(s_{y_1}^2)^2]=\Big[&\frac{(N_1+1)}{(N_1-1)}(1-\rho^4)-\frac{(N_1+4)}{(N-1)}\rho^2(1-\rho^2)\\
&-\frac{(N_1+1)}{(N_1-1)^2}(1-\rho^2)^2\Big]\sigma_y^4,
\end{aligned}
\tag{A-18}
$$

$$
\begin{aligned}
E[b^2(1-r^2)s_{y_1}^2]=\Big[&\frac{(N_1-2)}{(N_1-1)}\rho^2(1-\rho^2)\\
&+\frac{(N_1-2)}{(N_1-1)(N_1-3)}(1-\rho^2)^2\Big]\frac{\sigma_y^4}{\sigma_x^2}.
\end{aligned}
\tag{A-19}
$$

From equations A–1 and A–7 and equations A–10 through A–19, equation A–9 may be expressed as

$$\text{Var}\ (s_y^2)=\frac{2\sigma_y^4}{(N_1-1)}+\frac{N_2\sigma_y^4}{(N_1+N_2-1)^2}(A\rho^4+B\rho^2+C), \tag{A-20}$$

where

$$
\begin{aligned}
A=\Big[&\frac{(N_2+2)(N_1-6)(N_1-8)}{(N_1-3)(N_1-5)}-\frac{8(N_1-4)}{(N_1-3)}-\frac{2N_2(N_1-4)^2}{(N_1-3)^2}\\
&+\frac{4(N_1-4)}{(N_1-3)}\theta^2+\frac{N_1N_2(N_1-4)^2}{(N_1-3)^2(N_1-2)}\theta^4\Big]\rho^4,
\end{aligned}
\tag{A-21}
$$

$$
\begin{aligned}
B=\Big[&\frac{6(N_2+2)(N_1-6)}{(N_1-3)(N_1-5)}+\frac{2(N_1^2-N_1-14)}{(N_1-3)}\\
&-\frac{2(N_1+N_2-1)(N_1-4)}{(N_1-3)}(1-\theta^2)\\
&+\frac{2N_2(N_1-4)(N_1-5)}{(N_1-3)^2}\theta^2-\frac{2(N_1-4)(N_1+3)}{(N_1-3)}\theta^2\\
&-\frac{2N_1N_2(N_1-4)^2}{(N_1-3)^2(N_1-2)}\theta^4\Big]\rho^2
\end{aligned}
\tag{A-22}
$$

$$
\begin{aligned}
C=\Big[&\frac{2(N_1+1)}{(N_1-3)}+\frac{3(N_2+2)}{(N_1-3)(N_1-5)}-\frac{(N_1+1)(2N_1+N_2-2)}{(N_1-1)}\\
&+\frac{2(N_1+N_2-1)(N_1-4)}{(N_1-3)}(1-\theta^2)+\frac{2N_2(N_1-4)}{(N_1-3)^2}\theta^2\\
&+\frac{2(N_1-4)(N_1+1)}{(N_1-3)}\theta^2+\frac{N_1N_2(N_1-4)^2}{(N_1-3)^2(N_1-2)}\theta^4\Big].
\end{aligned}
\tag{A-23}
$$