Journal of Geophysical Research Volume 67, No. 9 August 1962

Information Content of the Mean

N. C. Matalas and W. B. Langbein

U. S. Geological Survey, Washington, D. C.

Abstract. The amount of information given by a statistical estimate is defined as the reciprocal of the variance of the estimate. In this study, a random series is adopted as the standard of information content. The relative information content of the mean for various types of hydrologic series is defined as the ratio of the variance of the mean for a random series to the variance of the mean for given types of series. Beyond the well-known fact that the sequential correlation in hydrologic series reduces the effective length of the series, the sequential correlation tends to impair the effectiveness of cross-correlations between series. There are limits, therefore, upon increasing the information content of a mean by utilizing the information contained in a related series. Moreover, cross-correlation tends to decrease the effective length of series for computing regional means.

Introduction

Although hydrologic records are carefully kept with due attention to accuracy and the needs of consistency over the years, records are either short, scant, or both, and so nearly every water development requires some estimates or extensions of available data. Comparisons between records, or averaging in time or over space are also frequently required. The reliability of these extensions or of the means of time and space have therefore often been subjects of study, and this paper extends the range of inquiry in a field of practical interest.

A characteristic of hydrologic phenomena is that their observation under a given set of conditions does not always yield the same observed outcome. However, the observed outcomes vary in such a way that there is statistical regularity. If the observed outcomes are arranged in order of observation, the ordered set is referred to as a series; if the observations are arranged in order of time, the ordered set of observed outcomes is called a time series.

For a stationary time series, there is a mean μ about which the observed outcomes fluctuate. This mean value is unknown and must be estimated from the observations. The estimate of the mean, $\bar{x}$, for a series consisting of N observations is given by

$$\bar{x} = \sum_{i=1}^{N} x_i/N \tag{1}$$

where x_i is the ith observed outcome.

It is intuitive to consider that a long series conveys more information about the mean than a short series. If the observed outcomes are mutually independent, the series is said to be random. For a random series, the variance of the estimated mean is σ^2/N where σ^2 is the population variance. Since there are only N observations, σ^2 is unknown, but its unbiased estimate, s^2, is given by

$$s^2 = \sum_{i=1}^{N} (x_i - \bar{x})^2/(N - 1) \tag{2}$$

A random series is adopted as the standard of information content. For other models of series having N observations, if the variance of the estimated mean is greater than σ^2/N, the series conveys less information about the mean than a random series of the same length. If the variance of the estimated mean is less than σ^2/N, the series conveys more information about the mean than a random series. Thus the relative information content may be defined as

$$I = (\sigma^2/N)/\sigma_m^2 \tag{3}$$

where σ_m^2 is the variance of the estimated mean for a particular series. If σ_m^2 is less than σ^2/N, I is greater than unity, and if σ_m^2 is greater than σ^2/N, I is less than unity. Since the variance is always a positive value, I can vary from zero to infinity.

Effect of Nonrandomness on Information Content

A time series may or may not consist of observed outcomes which are independent of one another. Streamflow is a hydrologic variable whose observations, equally spaced in time, are not necessarily randomly distributed through time. If the streamflow is regulated by storage in the river basin, there is carry-over of flow from one time interval to the next. Hence the streamflow values are correlated with one another, and the time series is nonrandom.

Spectrum analysis of several records of streamflow led *Julian* [1961, p. 78] to the conclusion that 'the type of non-randomness exhibited by the stream-runoff data was that of an autoregressive or moving average nature, with the stream runoff amounts exhibiting various degrees of persistence from water year to water year.' Julian added that the records were not sufficiently long to permit the erection of a statistical model more refined than a simple Markov chain. *Brooks* [1948] adopted a similar statistical model and so shall we, leaving a fuller explanation to our posterity.

The simple Markov model is

$$x_{i+1} = \tau_1 x_i + \epsilon_{i+1} \tag{4}$$

where x_i and x_{i+1} are the variate values at times i and $(i + 1)$, respectively, ϵ_{i+1} is a random component independent of x_{i+1} and all preceding x's and τ_1 is the first-order autocorrelation coefficient, which is a measure of the correlation between values separated by one time unit. If μ and σ^2 denote the mean and the variance of x, then the mean and the variance of ϵ are denoted by $(1 - \tau_1)\ \mu$ and $(1 - \tau_1^2)\ \sigma^2$.

If (4) represents the model for a given series, according to *Brooks and Carruthers* [1953], the estimate of the mean has variance

$$\sigma_m^2 = \frac{\sigma^2}{N}\left[\frac{1+\tau_1}{1-\tau_1} - \frac{2}{N}\,\frac{\tau_1(1-\tau_1^N)}{(1-\tau_1)^2}\right] \tag{5}$$

For nonrandom hydrologic time series, τ_1 is positive. The quantity in brackets in (5) is positive and greater than unity for $\tau_1 > 0$. Thus the information content given by

$$I = \left[\frac{1+\tau_1}{1-\tau_1} - \frac{2}{N}\,\frac{\tau_1(1-\tau_1^N)}{(1-\tau_1)^2}\right]^{-1} \tag{6}$$

is less than unity for $\tau_1 > 0$.

In an autocorrelated series, each observation repeats part of the information contained in previous observations. Therefore, N observations of a nonrandom time series having $\tau_1 > 0$ give as much information about the mean as some lesser number, N, of observations in random time series. This lesser number of observations is called the effective number of observations and is defined as

$$N' = NI \tag{7}$$

For the simple Markov model,

$$N' = N\left[\frac{1+\tau_1}{1-\tau_1} - \frac{2}{N}\,\frac{\tau_1(1-\tau_1^N)}{(1-\tau_1)^2}\right]^{-1} \tag{8}$$

If $\tau_1 = 0$, then $N' = N$; if $\tau_1 > 0$, then $N' < N$. Equation 8 is expressed graphically in Figure 1. As an example, a 50-year record for which

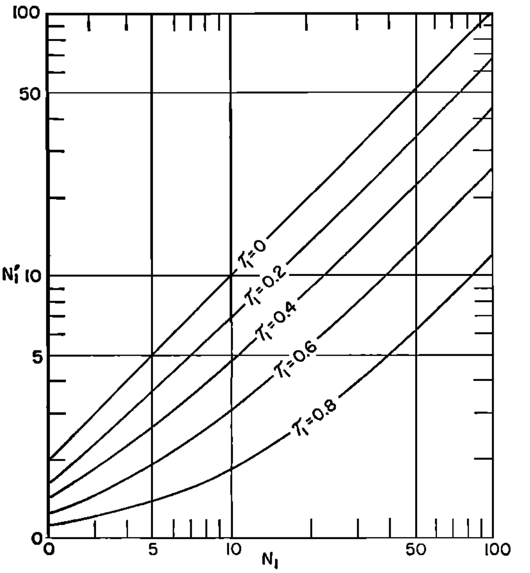

Fig. 1. Relation between the effective number of observations, N_1', and the number of observations, N_1, for given values of autocorrelation, τ_1.

$\tau_1 = 0.20$, a value characteristic of hydrologic series [*Yevdjevich*, 1961], contains only as much information about the mean as a 33-year record in which the observations are independent.

Information Inferred by Cross-Correlation

Cross-correlation of random series. Let us assume that two hydrologic variables are both functions of a common set of factors and are therefore related. For example, the streamflow for two rivers in a given region is more or less affected by the same occurrence of rainfall over the region. Under this assumption, if simultaneous observations of the two variables are made, the observed outcomes will be correlated. Owing to this cross-correlation, the observed outcomes for one variable repeat some part of the information contained in the observed outcomes for the other variable.

On the basis of N_1 simultaneous observations of the two variables, the information content of the mean for either or both series can be determined. If N_2 additional observations are made of one of the variables, the information content of the mean for this variable may be determined on the basis of $N_1 + N_2 = N$ observations. For the other variable for which only N_1 observations are available, additional information about its mean can be inferred by virtue of the cross-correlation of the two variables.

Let y and x denote the random variables for which there are N_1 and $N_1 + N_2$ observations, respectively. If the simultaneously observed outcomes are linearly related, the simple regression of y on x is

$$y = a + b(x - \bar{x}_1) \tag{9}$$

The regression constants, a and b, are computed by the usual formulas derived from the theory of least squares. In (9), $\bar{x}_1$ is the estimate of the mean of x based on the N_1 observations. From (9), N_2 estimates of y corresponding to the N_2 additional observations of x may be obtained. If $\bar{y}_1$ and $\bar{y}_2$ denote the estimates of the mean of y based on N_1 observations and N_2 regression estimates, respectively, then a weighted estimate of the mean of y is given by

$$\hat{\mu}_y = \frac{N_1\bar{y}_1 + N_2\bar{y}_2}{N_1 + N_2} \tag{10}$$

If the random variables x and y are both normally distributed, then, according to *Thomas* [1958], the variance of $\hat{\mu}_y$ is

$$\sigma_m^2 = \frac{\sigma_y^2}{N_1}\left\{1 - \frac{N_2}{(N_1 + N_2)} \cdot \left[\rho^2 - \frac{(1-\rho^2)}{(N_1 - 3)}\right]\right\} \tag{11}$$

where ρ is the coefficient of cross-correlation of x and y. The information content is given by

$$I = \left\{1 - \frac{N_2}{(N_1 + N_2)} \cdot \left[\rho^2 - \frac{(1-\rho^2)}{(N_1 - 3)}\right]\right\}^{-1} \tag{12}$$

For the cross-correlation to provide additional information ($I > 1$) about the mean of y, it is necessary that

$$\rho > \frac{1}{N_1 - 2} \tag{13}$$

If inequality 13 is not satisfied, the N_2 regression estimates will provide an unreliable estimate of the mean of y; consequently, I will be less than unity. In this case, additional information about the mean of y can be obtained only by increasing N, until inequality 13 is satisfied.

Cross-correlation of nonrandom series. Matalas and Rosenblatt [1962] have shown that if both series are normally distributed and represented by simple Markov models where τ_1 is the same for both series, the variance of $\hat{\mu}_y$ is

$$\begin{aligned}\sigma_m^2 = \frac{\sigma_y^2}{N_1}\Bigg\{&\left(\frac{1+\tau_1}{1-\tau_1} - \frac{2}{N_1}\frac{\tau_1}{(1-\tau_1)^2}\right) \\ &- \frac{N_2}{N}\left[\rho^2\left(\frac{1+\tau_1}{1-\tau_1} - \frac{2(N+N_1)\tau_1}{NN_1(1-\tau_1)^2}\right) \right. \\ &\left. - \frac{(1-\rho^2)(1+\tau_1^2)}{N_1(1-\tau_1)^2}\right] + \frac{N_2(1-\rho^2)}{NN_1} \\ &\cdot\left[\frac{3(1+\tau_1^4) + 2\tau_1(1-\tau_1)^2}{N_1(1-\tau_1)^3(1+\tau_1)} \right. \\ &- \frac{2\tau_1}{N_2(1-\tau_1^3)(1+\tau_1)} \\ &\left. + \frac{N_1(1+\tau_1^2)[1-(1-\tau_1)(1+\tau_1)^3]}{NN_2(1-\tau_1)^4(1+\tau_1)^2}\right]\Bigg\}\end{aligned} \tag{14}$$

where $N = N_1 + N_2$. Equation 14 is an approximation retaining terms up to order $1/N_1^3$. Also, (14) is based on the assumption that N_1 is sufficiently large and/or τ_1 is sufficiently small to cause τ_1^N to be negligible. Because of these approximations, (14) does not reduce to (11) when $\tau_1 = 0$. However, the agreement is good for $N_1 > 10$.

The information content can be derived by dividing σ_y^2/N_1 by σ_m^2, given by (14). Because of the complexity of the resulting expression for I, the effect of ρ and τ_1 on the information content cannot be readily determined. Some idea as to the influence of ρ and τ_1 on I is provided by determining I for given values of N_1 and N_2. Values of I for various values of ρ and r_1 are given in Table 1 for $N_1 = N_2 = 10$ and in Table 2 for $N_1 = N_2 = 20$. The values of I in Tables 1 and 2 indicate that the information content tends to decrease as r increases and to increase as ρ increases. Moreover, large values of ρ can offset the loss of information due to autocorrelation only if r_1 is small.

TABLE 1. Information Content of the Mean Relative to a Random Series: $N_1 = N_2 = 10$

ρ	r_1: 0	0.2	0.4	0.6	0.8
0	0.939	0.648	0.432	0.264	0.126
.1	0.944	0.652	0.434	0.266	0.126
.5	1.083	0.746	0.500	0.304	0.146
.9	1.646	1.126	0.746	0.457	0.230
1.0	2.000	1.362	0.900	0.552	0.286

TABLE 2. Information Content of the Mean Relative to a Random Series: $N_1 = N_2 = 20$

ρ	r_1: 0	0.2	0.4	0.6	0.8
0	0.972	0.660	0.432	0.259	0.121
.1	0.977	0.663	0.434	0.260	0.122
.5	1.115	0.756	0.495	0.296	0.139
.9	1.665	0.125	0.734	0.439	0.208
1.0	2.000	1.347	0.878	0.525	0.250

The cross-correlation, if sufficiently large, can provide more information about the mean than is given by the N_1 observations. The variance of the mean for a simple Markov model is given by (5), where N now denotes N_1. If this variance is divided by σ_m^2, given by (14), a measure of information content, I is obtained. The expression for I is sufficiently complex to prohibit a direct evaluation of the influence of N_1 N_2, ρ, and r_1 on the information content. Values of I for various values of ρ and r_1 are given in Table 3 for $N_1 = N_2 = 10$ and in Table 4 for $N_1 = N_2 = 20$.

The values of I given in Tables 3 and 4 show that ρ and r_1 have opposite effects on the information content. As ρ tends to unity, I increases, but as r_1 tends to unity, I decreases.

TABLE 3. Information Content of the Mean Relative to a Nonrandom Series: $N_1 = N_2 = 10$

ρ	r_1: 0	0.2	0.4	0.6	0.8
0	0.939	0.932	0.912	0.859	0.628
.1	0.944	0.937	0.917	0.864	0.631
.5	1.083	1.072	1.048	0.988	0.730
.9	1.646	1.619	1.576	1.486	1.150
1.0	2.000	1.957	1.900	1.793	1.429

TABLE 4. Information Content of the Mean Relative to a Nonrandom Series: $N_1 = N_2 = 20$

ρ	r_1: 0	0.2	0.4	0.6	0.8
0	0.972	0.969	0.960	0.938	0.848
.1	0.977	0.974	0.965	0.943	0.853
.5	1.115	1.111	1.100	1.074	0.974
.9	1.665	1.652	1.632	1.591	1.456
1.0	2.000	1.979	1.951	1.902	1.750

Large values of ρ can offset the loss of information due to r_1, even if r_1 is large.

Information content of a regional mean. It is often desirable to establish a mean, for a particular hydrologic phenomenon, which is representative of a given region. For example, we may wish to determine the mean annual rainfall for a given region. To determine this regional mean, observations must be made at several rainfall stations within the region.

Let us assume that there are n stations and that N concurrent observations have been made at each station. Thus there are n series, each consisting of N observations. Let $\bar{y}_1, \cdots, \bar{y}_n$ denote the estimates of the means for the n series. The estimate of the regional mean is given by

$$\bar{Y} = \sum_{i=1}^{n} \bar{y}_i/n \tag{15}$$

The cross-correlation between the means for series j and $k(j = 1, \cdots, n$ and $k = 1, \cdots, n)$ is denoted by ρ_{jk}. If $j = k$, then $\rho_{jk} = 1$. For n series, there are $n(n - 1)/2$ distinct pairs of correlations. The mean of all these cross-correlations is denoted by $\bar{\rho}$.

If each of the n series is a random series and if the n series have a common variance, σ^2, then the variance of $\bar{y}$ is [*Yule*, 1945]

$$\sigma_m^2 = \frac{\sigma^2}{Nn}[1 + \bar{\rho}(n - 1)] \quad (16)$$

If $\bar{\rho} = 0$, the variance of $\bar{y}$ is σ^2/Nn. The quantity σ^2/Nn is adopted as the reference value for the information content of a regional mean. The information content is given by

$$I = [1 + \bar{\rho}(n - 1)]^{-1} \quad (17)$$

which is always less than unity when $\bar{\rho} > 0$.

The effective number of series, n', is given by (8), where N' and N are now replaced by n' and n, respectively. Thus

$$n' = n[1 + \bar{\rho}(n - 1)]^{-1} \quad (18)$$

From (18), as shown in Figure 2, it can be seen that even small values of $\bar{\rho}$ make n' very much smaller than n. As n tends to infinity, n' tends to $1/\bar{\rho}$. As an example, if there are an infinite number of series, and if $\bar{\rho} = 0.1$, then $n' = 10$. Thus in the derivation of mean precipitation over a drainage basin there are severe limits on substituting density of rain gages for length of record.

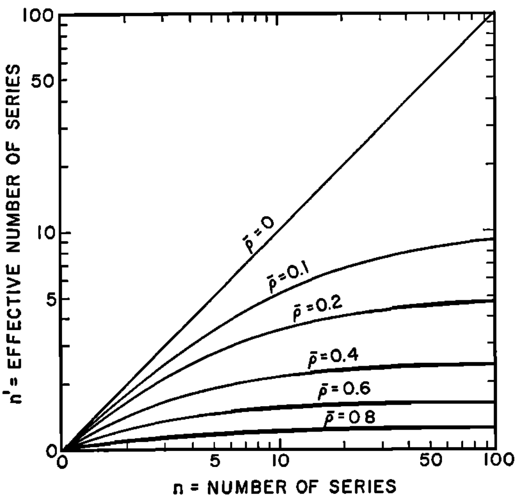

Fig. 2. Effective number of series in computing regional means.

Information Content Inferred from Geochronologic Data

Water-level data. The above discussions have shown that as the autocorrelation increases the information content decreases. However, in the extreme case, as r_1 tends to unity, the autocorrelation can be put to effective use. Let us consider, for example, the fluctuations in water level of a large body of ground water. On an annual basis, the fluctuations may be highly autocorrelated—greater than 0.9. The storage in such a body may integrate the inflows from recharge over long periods of time, and relatively few observations can establish the mean. Moreover, if a relationship can be established between precipitation and fluctuations in water level over even a short period of time, long-term means of precipitation may be inferred.

Considering a body of water whose volume is very large in relation to the average rate of flow through the body, we assume that the recharge to the body of water is from a highly variable source of precipitation or streamflow and that there are no losses from the body of water. Under these assumptions, the mean inflow equals the mean discharge. If the hydraulic system is such that the variable rates of inflow are 'averaged out' in the discharge, a short record of discharge may provide nearly as much information about the mean discharge as a longer record.

If the yearly inflows are randomly distributed in time with variance σ_i^2, the variance of the yearly discharge may be expressed by

$$\sigma_d^2 = \sigma_i^2[(1 - e^{-1/T})/(1 + e^{-1/T})] \quad (19)$$

where T is the hydraulic response time, in years, of the body of water. As T becomes large, say greater than 10 years, σ_d^2 may be approximated by

$$\sigma_d^2 = \sigma_i^2/2T \quad (20)$$

The value of T is considered to be a constant hydrologic characteristic of the body of water and is defined as

$$T = \Delta V/\Delta Q \quad (21)$$

where ΔV is the change in storage and ΔQ is the change in discharge.

If the discharge series consists of N observations, the estimate of the mean, $\bar{Q}$, has a variance equal to σ_d^2/N, provided that the discharge series is random. However, the values of discharge are affected by storage. The average autocorrelation is $T/(1 + T)$. The serial autocorrelation is $T/1 + T$. If this value is substituted for r_1 in (5), the variance of the mean for a nonrandom discharge series can be approximated by the following formula:

$$\sigma_m^2 = \sigma_d^2 \frac{0.5 + T}{0.5N + T} \qquad (22)$$

For values of $T > 10$, σ_m^2 approaches σ_d^2, and there is little to be gained by lengthening the record of discharge. The equivalent length of a short record of discharge is therefore $2T$. This notion is generally applied to the measurements of the discharge of a spring, where a single measurement is often inferred as an average. However, sound practice requires a series of well-spaced observations to assure that the discharge is stable—that T is large.

Ordinarily, data on outflow are not available, so that the above analysis is only of theoretic interest. A more common situation is that records of rainfall, the source of recharge, and records of water level are available, in which case the problem may be to infer the mean precipitation by the following technique.

Whenever annual precipitation is greater than average, recharge is also greater than average, and water levels increase; for the same reason, below-average precipitation results in a decrease in the water level. Thus, if annual precipitation is correlated with net annual changes in water level, the precipitation corresponding to zero net change in water level equals the mean sought. However, this mean is subject to two sources of error. The first error arises because the discharge from the body of water during the period of observation may not correspond to the long-term mean. This error, as explained above, equals σ_d^2. A second source of error arises because the correlation between precipitation and change in water level is not perfect. This error equals $\sigma_i^2 (1 - \rho^2)/N$, where ρ is the coefficient of cross-correlation of the precipitation and the water level series, σ_i^2 is the variance of the precipitation, and N is the number of concurrent observations of the two series.

Under the assumption of a linear relationship between precipitation and net annual change in water levels, the variance of the mean precipitation, σ_m^2, is composed of two terms which are contained in the following expression:

$$\sigma_m^2 = \frac{\sigma_i^2}{N}\left[\frac{N}{2T} + (1 - \rho^2)\right] \qquad (23)$$

where σ_i^2, T, ρ, and N are as previously defined. The reciprocal of the quantity in brackets in (23) is the information content. Hence, by (9), the effective period of record is

$$N' = N\left[(1 - \rho^2) + \frac{N}{2T}\right]^{-1} \qquad (24)$$

If the cross-correlation is perfect ($\rho = 1$), then $N' = 2T$. In this case there is a magnification of $2T/N$ of information about the mean, so that even short response times would produce large extensions. Usually ρ imposes a significant limiting effect on the extension. Unless ρ^2 exceeds $N/2T$, there is no gain of information; when ρ^2 is less than $N/2T$, there is a loss of information. If N exceeds $2T\rho^2$, the mean of the precipitation series is better than that derived by cross-correlation with changes in water levels.

Tree ring data. Studies dealing with the distribution of hydrologic variables with respect to time are handicapped because the lengths of hydrologic records are short and cover approximately the same time span. To increase the information content of hydrologic means, one may study the time distribution of natural phenomena which are related to and which cover a longer span of time than the hydrologic variables. Under certain conditions, the annual growth of trees measured by the widths of annual rings may be used to study the time distribution of hydrologic variables. Trees growing on well-drained sites where rainfall is the limiting climatic factor influencing growth are a possible source of hydrologic information.

For a given tree ring series, the ring widths tend to diminish with the age of the tree, so that the ring widths follow a trend which is monotonic and nonlinear. The variation in ring widths about the trend line also tends to decrease with the age of the tree. Thus the series of annual ring widths are not stationary. Tree ring series are converted to stationary time series by dividing the ring widths by their corresponding trend values. This transformed variable is called a tree ring index.

A tree ring series for a single tree in a given region cannot be considered a representative hydrologic indicator for the entire region. To obtain a representative indicator, the tree ring series for several trees in the region are averaged to form a regional tree ring series. If w_{ij} denotes the index for the ith year for the jth tree, the average index during the ith years is

$$x_i = \sum_{j=1}^{m} w_{ij}/m \tag{25}$$

where m denotes the number of trees. If the indices are randomly distributed in time and if the mean and variance of the indices are the same for all trees in the region, the variance of x is given by

$$\sigma_x^2 = \frac{\sigma_w^2}{m}[1 + (m - 1)\bar{\rho}] \tag{26}$$

where $\bar{\rho}$ is the average of all possible correlations between the m tree ring series of indices. Equation 26 is identical to equation 16.

If $\bar{\rho} = 0$, then σ_x^2 tends to zero as m tends to infinity. In this case, averaging the indices results in a total loss of information regarding the fluctuation of the annual ring indices. On the other hand, if $\bar{\rho} = 1$, then σ_x^2 tends to σ_w^2 as m tends to infinity. In this case, averaging the indices for several trees yields no more information than is contained in a single tree ring series.

In a given region, as m increases, the average cross-correlation between the tree ring series of indices, $\bar{\rho}$, tends to increase. If this does not happen, no system can be operating in the basin, in which case no correlation can exist between the growth of trees and some hydrologic variables such as streamflow.

Let us assume that the relation between $\bar{\rho}$ and m is

$$\bar{\rho} = m/(c + m) \tag{27}$$

where c is a constant characterizing the region. In evaluating an equation of this type, care must be taken not to bias the correlations by a posteriori selection of trees. If (27) is substituted for $\bar{\rho}$ in (26), then

$$\sigma_x^2 = \frac{\sigma_w^2}{m}\left[1 + \frac{m(m - 1)}{(c + m)}\right] \tag{28}$$

For $m = 1$, $\sigma_x^2 = \sigma_w^2$, and as m tends to infinity, σ_x^2 also tends to σ_w^2. Between these two limits, σ_x^2 is a minimum for a finite value of m, since σ_x^2 cannot be greater than σ_w^2. If the derivative of σ_x^2 with respect to m is set equal to zero, σ_x^2 is found to be a minimum for $m = 1 + (c + 1)^{1/2}$. This minimum value is the optimum number of trees to be averaged for a regional sample. As the value of c increases, the minimum value of m increases, which tends to decrease the information content. Thus that system is best in which the minimum value of σ_x^2 is largest. In other words, the smaller c is, the better the system is. Moreover, as c increases, a relation is unlikely to exist between tree ring data and hydrologic data. Unfortunately, the converse does not hold: if $c = 0$, a perfect correlation between tree ring widths and hydrologic phenomena does not necessarily exist. We surmise that low values of c among an unbiased selection of trees may be difficult to achieve.

Summary

From a series of observations for a particular hydrologic phenomenon, an estimate of the mean may be obtained. The reciprocal of the variance of this estimate is a measure of the information given by the observations about the mean. By adopting a random series as the standard of information, the relative information content about the mean for any type of series is defined as $I = (\sigma^2/N)\sigma_m^2$, where σ^2/N is the variance of the mean for a random series and σ_m^2 is the variance of the mean for some given type of series.

Basically, the information content of the mean can be increased by increasing the length of period of observation. Water development cannot be deferred, however, and so advantage may be taken of the fact that, within limits, information may be increased more expediently by utilizing the information contained in related series.

Hydrologic series frequently consist of observations that are dependent upon one another. Such series are referred to as nonrandom series and may be represented by a simple Markov model. The dependence between the observations is measured by the autocorrelation coefficients. In a nonrandom series each observation repeats part of the information contained in past observations. Consequently, a nonrandom series yields less information about the mean than a random series having an equal number of observations.

If a series of a given length is related to a series of longer length, the correlation between

the two series can be utilized to increase the information content about the mean for the shorter series. If the two series are random, the correlation will permit an increase of information providing $\rho^2 > 1/N - 2$, where ρ is the coefficient of cross-correlation between the two series and N is the number of observations in the shorter series. If both series are generated by first-order autoregressive schemes, the cross-correlation serves to increase the information while the autocorrelation within each series serves to decrease the information. However, large values of cross-correlation can offset the loss of information due to autocorrelation.

A regional mean is defined as the mean of the means for several variables in a given region. If these variables are uncorrelated with one another, the regional mean has its maximum information content. The information content of the regional mean decreases rapidly with an increase in the dependence between the variables.

As previously stated, as the autocorrelation increases, the information content decreases. However, in the extreme case, as the autocorrelation approaches unity, its upper limit, the autocorrelation can be put to effective use. Such a situation arises when the outflow from a given system is a function of the integration of the inflows to the system over long periods of time. In such a situation, relatively few observations can establish the mean.

References

Brooks, C. E. P., The variability of means of a series of observations, *Meteorol. Mag., 77,* 283–285, 1948.

Brooks, C. E. P., and N. Carruthers, *Handbook of Statistical Methods in Meteorology, M. O. 538,* Meteorological Office, London, 412 pp., 1953.

Julian, Paul R., *A Study of Statistical Predictability of Stream Runoff on the Upper Colorado River Basin,* High Altitude Observatory, University of Colorado, Boulder, Oct. 1961.

Matalas, N. C., and J. R. Rosenblatt, Use of correlation for augmenting autocorrelated streamflow information, in preparation as *U. S. Geol. Survey Water Supply Paper,* 1962.

Thomas, H. A., Jr., Correlation techniques for augmenting stream runoff information, unpublished report, Harvard University, Cambridge, Mass., 1958.

Yevdjevich, V. M., *Some General Aspects of Fluctuations of Annual Runoff on the Upper Colorado River Basin,* Colorado State University, Civil Engineering Section, Fort Collins, Colo., Oct. 1961.

Yule, G. U., A method of studying time series based on their internal correlations, *J. Roy. Statis. Soc., 108,* 208, 1945.

(Manuscript received January 29, 1962; revised May 11, 1962.)