Section 8-II

STATISTICAL AND PROBABILITY ANALYSIS OF HYDROLOGIC DATA

PART II. REGRESSION AND CORRELATION ANALYSIS

VUJICA M. YEVDJEVICH, *Professor of Civil Engineering, Colorado State University.*

I. INTRODUCTION

The regression and correlation analysis is one of the oldest statistical tools used in hydrology. It was first used for filling missing data and extending short records at one hydrologic station by relating the available data at this station with those at adjacent stations. Now its application has been broadened to cover the study of the relationship between two or more hydrologic variables and also the investigation of dependence between the successive values of a series of hydrologic data.

This section deals with the basic definitions and presents discussion necessary for the understanding of the concepts of regression and correlation and the methods of their applications.

The simple linear regression and correlation are presented according to classical treatments, with special emphasis placed on the statistical inference for computing the parameters. For both grouped data in bivariate distribution and ungrouped data, the methods of computing the parameters are discussed. The simple curvilinear regression and correlation, mostly employing the polynomials as the regression function, is only briefly presented.

The multiple linear and curvilinear regression and correlation are discussed in some detail, especially with respect to the various ways of measuring the multiple and partial correlation and to the statistical inference of regression coefficients. The multiple regression and correlation analysis is used a great deal nowadays because such complicated analyses can be made practicable through numerical computations and thus can be economically executed with the aid of electronic digital computers.

II. BASIC DEFINITIONS AND TOOLS

A. Hydrologic Variables and Series

A *variable* in hydrology can be represented by either a *continuous series* (such as a recorded hydrograph) or a *discontinuous series* (such as annual flow values, flood peaks, etc.). The measuring interval may be selected in terms of time (hour, day, month, year), length (foot, mile), or surface area (acre, square mile). By use of the mean or total value of the variable in such intervals, a continuous series can be transformed into a discontinuous series. Generally, only the discontinuous series of data is statistically analyzed in hydrology.

Inasmuch as the hydrologic data are obtained by observations and by further appraisal of observed values, the hydrologic series are subject to human errors (random and systematic) and are often *nonhomogeneous*. *Random errors* are always

present because of the inaccuracy in measurements and observations. *Systematic errors*, or *errors of inconsistency*, refer to errors occurring in one direction, such as trends or jumps in the series. Nonhomogeneity of the data results from changes due either to natural catastrophes, such as fires, landslides, etc., or to man-made developments. It is advisable to appraise the data in terms of their probable errors and nonhomogeneity before using them for statistical analysis and to consider the validity of the data in drawing conclusions concerning the reliability of the statistical parameters and relationships determined from them. This is especially important when the available data series represents a small sample, that is, where the sample size is smaller than about fifty items.

The relationships of variables in hydrology may either show the cause and effect at one hydrologic site (such as the precipitation-runoff relationship at one river basin) or correlate the effects only at neighboring sites (correlating precipitations at two stations or runoffs of two adjacent river basins). The most common case attempted is that of showing the relationship of an effect to many causes, of which a small number of the causes exert greater influence than do all others. In such a case, when neglected variables and inherent errors and nonhomogeneity of data have relatively small effects, the relationship between the remaining limited number of variables would indicate a narrow spread around a basic function. This is the form of relationship generally required and used, since pure, functional relationships in hydrology are rare.

B. Definitions of Regression and Correlation

If two variables, given as a series with concurrent values (x_i, y_i), show a concentration around an imaginary curve when plotted on a graph (Fig. 8-II-1), then for a large

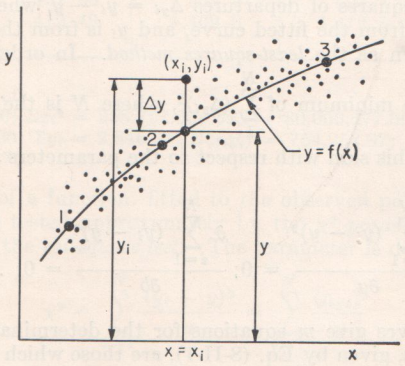

FIG. 8-II-1. Schematic representation of a simple regression and correlation analysis.

series there will always be a distribution of y values for a given value of x_i, or more precisely, a distribution of y values for a given interval Δx around x_i. The mean value y_0 of all y values for this given interval Δx around x_i is the expected value of y for the given $x = x_i$. A curve fitted to all mean values, y_0, is called the *regression line of y versus x*. On the other hand, the curve fitted to all expected (mean) values, x_0, for the given $y = y_i$, defines the *regression line of x versus y*. These two lines do not coincide, but have different parameters, showing the regressional relationships between the variables.

A pure functional relationship between variables assumes that all points would follow a curve, without spread. Inasmuch as the spread of points around the regression lines may actually be great or small, the degree of association of the variables involved is generally called *correlation* and is defined by the parameters of correlation. The correlation is greater when the points are closer to the lines.

In short, a regression problem considers the frequency distribution of one variable when another is held fixed at each of several levels. A correlation problem considers the joint variation of two measurements, neither of which is restricted by the experimenter or observer.

C. Curve Fitting

The methods of curve fitting may be graphical or analytical. If any curve of the type

$$y = f(x, a, b, c, \ldots) \tag{8-II-1}$$

is fitted to the plotted points, the graphical procedure is simply that of tracing a curve by eyes through the mean of the spread (Fig. 8-II-1). This procedure is often subject to large errors, but is practical and expedient.

The simplest analytical procedure is to fit a function employing as many parameters $a, b, c, \ldots$, in Eq. (8-II-1), as the number of selected points. These points can be selected from the sample on the basis of their positions within the sample. Thus, for three parameters, the use of points 1, 2, and 3 is shown in Fig. 8-II-1. The values (x_i, y_i) for each point must satisfy Eq. (8-II-1), so that there are m equations for the determination of m parameters. Better yet, they may be points which represent a weighted average of the groups of points in the sample. Thus, for three parameters, the points would be divided into three adjacent groups, the weighted mean of each group would be determined, and the equation fitted to pass through these three means. These points may be fitted graphically to produce curves. The parameters are then determined from the curves. This procedure is called the *three-point method.*

The currently approved analytical method of fitting curves to scattered points is to minimize the sum of squares of departures $\Delta_{yi} = y_i - y$, where, for a given x_i, the value y is determined from the fitted curve, and y_i is from the observed point (Fig. 8-II-1). This is known as the *least-squares method.* In order that the line of Eq. (8-II-1) may have the minimum of $\sum_1^N (\Delta_{yi})^2$, where N is the number of points, all partial derivatives of this sum with respect to the parameters $a, b, c, \ldots$ should be zero, so that

$$\frac{\partial \sum_{i=1}^{N} (y_i - y)^2}{\partial a} = 0; \quad \frac{\partial \sum_{i=1}^{N} (y_i - y)^2}{\partial b} = 0; \ldots \tag{8-II-2}$$

These partial derivatives give m equations for the determination of m parameters. The simplest functions, given by Eq. (8-II-1), are those which are linear with respect to the parameters. Among such functions the polynomial function for x is a general case, since other functions can be approximated by polynomials if they are developed in power-series form. There must be more points than parameters, or $N > m$. Generally, N should be much greater than m, if the derived line is to be used for prediction purposes.

For example, the fitting of a quadratic parabola of the form

$$y = a + bx + cx^2 \tag{8-II-3}$$

gives the following three equations by using Eq. (8-II-2):

$$\begin{aligned} aN + b\Sigma x_i + c\Sigma x_i^2 &= \Sigma y_i \\ a\Sigma x_i + b\Sigma x_i^2 + c\Sigma x_i^3 &= \Sigma x_i y_i \\ a\Sigma x_i^2 + b\Sigma x_i^3 + c\Sigma x_i^4 &= \Sigma x_i^2 y_i \end{aligned} \tag{8-II-4}$$

with the summations taken from $i = 1$ to N. The solution of these three equations gives a, b, and c.

Table 8-II-1 gives the results of fitting a parabolic rating curve to measured discharges Q and stages H. The sums in Eq. (8-II-4) are given in the table. The computation must be carried out to the same decimal point for all sums, because some of the sums are of the same order of magnitude as the differences involved in the use of much larger numbers. The rating curve by the least-squares method is found to be $Q_1 = 29.9 + 0.567H + 0.00720H^2$. The parabola through three selected points (6, 11, and 13 in Table 8-II-1) is $Q_2 = 38.6 + 0.396H + 0.007597H^2$.

Table 8-II-1. Fitting of Parabola to Rating Curve

	x_i (stage H), cm*	y_i (discharge Q), m^3/sec†	Estimates of Q		χ_1^2	χ_2^2
			Least-squares Q_1	Three-point Q_2		
1	−23	15.55	20.7	35.5	1.25	9.55
2	−22	15.46	20.9	33.6	1.39	9.78
3	−16	20.07	23.6	34.2	0.26	5.00
4	−16	21.99	23.6	34.2	0.11	4.33
5	14	36.11	39.3	45.6	0.26	1.98
6	33	59.82	56.5	59.9	0.19	0.00
7	46	86.58	70.9	72.9	3.50	2.57
8	69	110.96	103.3	102.1	0.57	0.77
9	88	136.52	135.6	132.3	0.01	0.13
10	120	204.40	200.7	190.5	0.07	1.00
11	136	232.87	240.2	233.0	0.22	0.00
12	220	492.50	503.2	493.4	0.23	0.00
13	400	1,412.48	1,409.0	1,412.5	0.01	0.00
					8.07	35.12

* 1 cm = 0.3937 in.
† 1 m^3/sec = 35.314 cfs.
$N = 13$; $\Sigma x_i = 1{,}049.00$; $\Sigma x_i^2 = 258{,}727.00$; $\Sigma x_i^3 = 80{,}006{,}477.00$; $\Sigma x_i^4 = 28{,}580{,}940{,}099.00$; $\Sigma y_i = 2{,}846.31$; $\Sigma(x_i y_i) = 754{,}258.87$; $\Sigma(x_i^2 y_i) = 258{,}951{,}971.73$.

The *goodness of fit* of a function, fitted to the observed points by any procedure, can be measured and tested approximately by the χ^2 *parameter* (*fitting parameter*). This test is known as the *chi-square test.* The parameter is defined as [1, p. 203]

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - y)^2}{y} = \sum_{i=1}^{N} \frac{(\Delta_{yi})^2}{y} \qquad \text{(8-II-5)}$$

For the above example of fitting the parabolic rating curve, $\chi_1^2 = 8.07$ for the least-squares method and $\chi_2^2 = 35.12$ for the three-point method (Table 8-II-1), indicating that the least-squares method gives a χ^2 value four times smaller than that given by the three-point method.

It is assumed for the chi-square test that values of y_i are normally distributed around y_0 for any given interval Δx for x_i. This assumption is used as a basis either for comparison of the different procedures of fitting the functions to observed data or for statistical inference.

A similar procedure is used when the sum of $(x_i - x)^2$ is minimized for the regression of x versus y.

D. General Model for Regression and Correlation Analyses

A variable that occurs as a sequence of consecutive values may be composed of values that are *internally independent* of each other (*random variable*) or *internally dependent* (*autocorrelated,* or *serially correlated, variable*). These conditions are called

here *internal dependence* and *internal independence of a variable.* A series of monthly streamflow values is usually internally dependent, because of the carryover of stored water and the influence of the annual cycle of climate.

The regressional or correlative relationship of m variables has two extreme cases. When one variable is dependent on all other $m - 1$ variables, the latter may be either mutually independent, known as *externally independent,* or interdependent, known as *externally dependent.* Runoff data depend on such variables as precipitation, evapotranspiration, topography, geology, and water retention. These are, in turn, generally interdependent i.e., externally dependent.

The frequency distribution of each individual variable (called the *marginal distribution* in multivariate distributions) is usually skewed in hydrology. The normal distribution is generally used as a standard for comparison of such distributions.

The basic relationship for the external or internal dependence of variables can be expressed arbitrarily by any desired function.

When a statistical model is being set up for regression and correlation analyses, the statistical relationships should comply with the true hydrologic behavior. Several of the most important hydrologic variables probably are related among themselves. They show a residual scatter of points around the curve of a selected functional relationship because of the effect of other neglected variables and because of inherent errors. Most hydrologic variables are mutually dependent externally. Nearly all hydrologic variables are skew-distributed in their marginal distributions. And many hydrologic variables are internally dependent (autocorrelated).

Difficulties in statistical treatment of the complex model are usually encountered in hydrology, especially in appraising the reliability (or errors) of the computed statistical parameters defining the regression or correlation. Such difficulties may be overcome by many simplifications in the use of a general hypothetical model.

When only two variables are related, the analysis is one of a *simple regression or correlation.* When three or more variables are involved, the analysis is one of a *multiple regression or correlation.* The approach to simplification is to limit the number of variables to the most significant ones and thereby reduce the labor of computation. Current hydrologic practice generally restricts the analysis to that of simple regression or to multiple regression involving variables limited to a maximum of 5 to 10. Generally, one variable is considered dependent, and all others are considered to be independent. It is also assumed that the marginal distributions of all variables can be approximated by a normal distribution or that, by transformation, they can approach to a normal distribution.

Internal dependence is important only in making statistical inference.

The complexity of the mathematical functions involved can be avoided either by transformation to linearize the relation or by replacing the complex function with a polynomial, provided the function can be developed to a power series with a limited number of terms. Currently, the linear relationship is most used in regression and correlation analyses in hydrology.

E. Transformation of Variables

The transformation of variables has three basic advantages, which may or may not all occur in a given transformation: (1) A simple linear relationship is obtained among the transformed variables. (2) The marginal distributions of the transformed variables approach more closely to the normal distribution than do those of the untransformed variables. (3) The variation of the points along the regression line is more homogeneous. In other words, the variance, or standard deviation, is stabilized.

Table 8-II-2 gives the transformations for linearizing a two-variable relationship. A logarithmic transformation is generally used in hydrology more frequently than others.

F. Use of Ungrouped and Grouped Data

In the regression and correlation analysis of a small number of variables with a large sample size, the use of ungrouped data can be cumbersome and time-consuming if

Table 8-II-2. Transformations for Linearization of Different Functions

	Type of function	Straight-line coordinate		Equation in linear form
		Abscissa	Ordinate	
1	$y = a + bx$	x	y	$[y] = a + b[x]$
2	$y = be^{ax}$	x	$\log y$	$[\log y] = \log b + (a \log e)[x]$
3	$y = ax^b$	$\log x$	$\log y$	$[\log y] = \log a + b[\log x]$
4	$y = a_0 + a_1x + a_2x^2$	$x - x_0$	$\dfrac{y - y_0}{x - x_0}$	$\left[\dfrac{y - y_0}{x - x_0}\right] = a_1 + 2a_1x_0 + a_2[(x - x_0)]$
5	$y = a + b/x$	$1/x$	y	$[y] = a + b[1/x]$
6	$y = x/(a + bx)$	x	x/y	$[x/y] = a + b[x]$
7	$y = a/(b + cx)$	x	$1/y$	$[1/y] = (b/a) + (c/a)[x]$
8	$y = c + be^{ax}$	x	$\log \dfrac{\Delta y}{\Delta x}$	$\left[\log \dfrac{dy}{dx}\right] = \log (ab) + (a \log e)[x]$
9	$y = c + ax^b$	$\log x$	$\log \dfrac{\Delta y}{\Delta x}$	$\left[\log \dfrac{dy}{dx}\right] = \log (ab) + (b - 1)[\log x]$
10	$y = c + \dfrac{b}{x - a}$	$x - x_0$	$\dfrac{x - x_0}{y - y_0}$	$\left[\dfrac{x - x_0}{y - y_0}\right] = -\dfrac{a - x_0}{c - y_0} + \dfrac{1}{c - y_0}[x - x_0]$
11	$y = c + \dfrac{x}{a + bx}$	x	$\dfrac{x - x_0}{y - y_0}$	$\left[\dfrac{x - x_0}{y - y_0}\right] = (a + bx_0) + \dfrac{b(a + bx_0)}{a}[x]$
12	$y = d + cx + be^{ax}$	x	$\log \dfrac{\Delta^2 y}{\Delta x^2}$	$\left[\log \dfrac{d^2y}{dx^2}\right] = \log (a^2b) + (a \log e)[x]$
				$[y - be^{ax}] = d + c[x]$
13	$y = dc^xb^m$, where $m = a^x$	x	$\log \dfrac{\Delta^2(\log y)}{\Delta x^2}$	$\left[\dfrac{\log d^2(\log y)}{dx^2}\right] = \log\left[\dfrac{(\log b)(\log a)^2}{(\log e)^2}\right] + (\log a)[x]$
14	$y = de^{cx} + be^{ax}$	$\dfrac{y_{k+1}}{y_k}$	$\dfrac{y_{k+2}}{y_k}$	$[\log y - a^x \log b] = \log d + (\log e)[x]$
				$\left[\dfrac{y_{k+2}}{y_k}\right] = -e^{(a+c)\Delta x} + (e^{a\Delta x} + e^{c\Delta x})\left[\dfrac{y_{k+1}}{y_k}\right]$
				$[ye^{-cx}] = d + b[e^{(a-c)x}]$
15	$y = e^{ax}(d \cos bx + c \sin bx)$	$\dfrac{y_{k+1}}{y_k}$	$\dfrac{y_{k+2}}{y_k}$	$\left[\dfrac{y_{k+2}}{y_k}\right] = -e^{2a\Delta x} = (2e^{a\Delta x} \cos b\,\Delta x)\left[\dfrac{y_{k+1}}{y_k}\right]$
				$\left[\dfrac{yc^{-ax}}{\cos bx}\right] = d + c[\tan bx]$

REMARK: In types 14 and 15, y_k, y_{k+1}, and y_{k+2} are consecutive values for an increment Δx.

ordinary numerical methods of computation are used. The grouping of the data by deriving a *bivariate distribution* (for simple regression and correlation) or a *multivariate distribution* (for three or four variables) may save much of the computational work. However, the use of grouped data gives somewhat greater correlation parameters and decreases the measure of spread of points around the regression line in comparison with the case for ungrouped data. This bias must be considered when statistical inference gives results close to the limits of the prescribed intervals. An example of analysis for grouped data will be given later.

G. Statistical Inference in Regression and Correlation Analyses

As the regression lines are defined by parameters and the correlation is measured by parameters, the reliability of these computed parameters is important.

The sample of the data available for an analysis is considered as a small part of an infinite *universe* of values. The general case, treated in hydrology, is that the sample is a part of a homogeneous universe, assuming thus that the same cause-effect relation applies to all other data from the universe. A change in homogeneity in either time or space, such as the change of climate, the change of river-basin factors, etc., must be corrected for nonhomogeneity of the data. In other words, any trend of sample changes must be corrected.

If a large number of samples of the same size as the analyzed sample were available, then any parameter describing the samples would have its own distribution. The *standard deviation* of this distribution, which generally is unknown in exact form because only a small sample is usually available in hydrology, is considered as the measure of reliability of the determined parameter. The general rule is that the larger the sample size, the smaller the standard deviation of the computed parameter, and thus the more reliable the parameter, all other conditions being the same.

The *confidence interval* is defined as an interval around the computed parameter within which a given percentage of parameters of a large number of samples is expected to be found. This given percentage is the *level of confidence*. The confidence interval at the 95 per cent level means that, out of 100 samples of equal size, it is expected that 95 values of a parameter would be inside that interval. Sometimes the confidence interval is chosen as the expected percentage of parameters to fall outside the confidence interval; in this case the level is 5 per cent. The *confidence limits* are numerical values describing the boundaries of the confidence interval.

If the computed parameter falls inside the selected confidence limits, it is considered either as *significant* or as *nonsignificant*, depending on the problem at hand. If it falls outside, it is considered either as *not significant*, or *unreliable* for use, or as *significant* as the case may be. The selection of level is made either by convention or by judgment. In hydrology the limits are generally chosen at between 80 and 95 per cent.

If a regression line with m parameters is fitted to a set of data points and if N is the sample size, the number of degrees of freedom is $N - m$. Thus every parameter takes away one degree of freedom in the analysis of the data. When $m = N$, the line passes through all points, so that there are no degrees of freedom. Since the errors of the parameter are inversely related to the degrees of freedom, this line cannot be used for prediction; otherwise, the errors would become infinite. For small-size samples, the sample size N used in related formulas should be replaced by the degrees of freedom $(N - m)$. This correction is necessary for avoiding a bias in the measure of reliability of the computed statistical parameters. If no such correction is made, the computed standard deviations of the parameters are significantly smaller in very small samples than their true values.

III. SIMPLE LINEAR REGRESSION AND CORRELATION

A. Regression Lines

1. Analytical Method of Determining Parameters. The straight-line regression for variable y versus variable x is defined by a straight line which gives the *best*

estimate of y for a given value of x. Similarly, the best estimate of x for a given y is given by the regression line of x versus y.

The line connecting the mean values y_0 of all y_i for given x_i, or more precisely, for the interval Δx around x_i, is for all practical purposes the regression line. In the simple linear regression analysis (Fig. 8-II-2) a straight line is fitted through these values.

The straight regression line is generally fitted analytically by the method of least squares of the departures from the line. The individual mean values y_0 for given x_i, or x_0 for given y_i, however, are not usually computed. For the regression line y versus x, the departures are Δ_{yi} along the ordinate. For the regression line x versus y the departures are Δ_{xi} along the abscissa. The equations of the two lines are

$$y = A_y + B_y x \tag{8-II-6}$$
$$x = A_x + B_x y \tag{8-II-7}$$

where B_y is the *regression coefficient* of y versus x, B_x is the regression coefficient of x versus y, and A_y and A_x are the respective *intercepts* for the regression lines. The regression line of y versus x is not the same as that of x versus y.

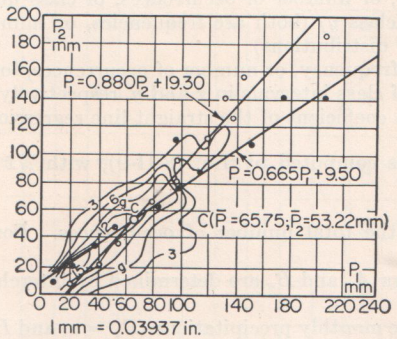

FIG. 8-II-2. Simple linear regression analysis for monthly precipitations (P_1, P_2), using the grouped data in a bivariate distribution. Small circles represent mean values of P_1 in a given class interval ΔP_2; small dots represent mean values of P_2 in ΔP_1.

The sum of $(\Delta_{yi})^2 = (y_i - y)^2$ is minimized, where y_i is the observed value and y is the value from the straight regression line for a given x_i. If $\bar{x}$ and $\bar{y}$ are the mean values of x and y for the sample size N, then

$$B_y = \frac{\Sigma x_i y_i - N\bar{x}\bar{y}}{\Sigma x_i^2 - N\bar{x}^2} = \frac{\Sigma \Delta x_i \, \Delta y_i}{\Sigma (\Delta x_i)^2} \tag{8-II-8}$$

where the summations are taken from 1 to N, with $\Delta x_i = x_i - \bar{x}$ and $\Delta y_i = y_i - \bar{y}$.

The intercept is

$$A_y = \bar{y} - B_y\bar{x} = \bar{y} - \bar{x}\frac{\Sigma \Delta x_i \, \Delta y_i}{\Sigma (\Delta x_i)^2} \tag{8-II-9}$$

The parameters B_x and A_x are obtained in a similar way by interchanging Δx_i with Δy_i, and x with y, in Eqs. (8-II-8) and (8-II-9).

To determine both regression lines the following five summations of the sample should be performed: Σx_i, Σy_i, $\Sigma(\Delta x_i)^2$, $\Sigma(\Delta y_i)^2$, and $\Sigma(\Delta x_i \, \Delta y_i)$.

2. Graphical Method of Determining Parameters. The graphical method is based on the principle that the graphically fitted straight line, minimizing deviations from the line, leaves on both sides the same amount of scattered points and that these points are nearly homogeneously distributed along both sides of the line. Graphical construction of a regression line usually does not give an accurate determination of

the parameters, which are obtained from the line as the intercept at the ordinate A_y and the slope of the line $B_y = \tan \alpha$, α being the slope angle. The computed χ^2 of goodness of fit is generally much greater than that by the analytical least-squares method, because the fit is less accurate.

3. Grouped Data in the Bivariate Distribution. When a large sample necessitates the use of grouped data and the bivariate distribution (x_i, y_i) is used for the determination of parameters (disregarding the small bias obtained by using grouped data), the regression coefficient for the grouped data [2] is computed by the formula

$$B_y = \frac{\sum_{i=1}^{p} \sum_{j=1}^{q} f_{ij} \Delta x_i \, \Delta y_j}{\sum_{i=1}^{p} f_{xi} (\Delta x_i)^2} \tag{8-II-10}$$

where $\Delta x_i = x_i - \bar{x}$, x_i being the middle of each class interval for x
$\Delta y_j = y_j - \bar{y}$, y_j being the middle of each class interval for y
f_{xi} = frequency, or number of occurrences, of class interval x_i, f_{yj} being the same for class y_j (both are frequencies, or numbers of occurrences, of marginal distributions)
f_{ij} = bivariate frequency, or number of occurrences, for (x_i, y_j) class intervals
p, q = number of class intervals in x and y, respectively
B_y = regression coefficient of the straight-line regression y versus x

The parameter A_y is computed by Eq. (8-II-9), with $N\bar{x} = \sum_{i=1}^{p} f_{xi} x_i$ and $N\bar{y} = \sum_{j=1}^{q} f_{yi} y_j$, where N is the total number of occurrences. For the regression line x versus y, the parameters A_x and B_x are determined by interchanging x and y in Eqs. (8-II-9) and (8-II-10).

Table 8-II-3 gives the monthly precipitations ($P_1 = x_i$ and $P_2 = y_j$) for two rainfall stations, with their bivariate distribution f_{ij} and marginal distributions f_{xi} and f_{yj}. It also shows a practical procedure for computing the sums necessary to be used in Eqs. (8-II-9) and (8-II-10). In order to avoid computation with large absolute values of P_1 and P_2, the intervals are coded. For *coding* or *indexing* class intervals, new dimensionless variables u and v are introduced: $u = (113 - P_1)/\Delta P_1$ and $v = (83 - P_2)/\Delta P_2$, where the length of class intervals is $\Delta P_1 = \Delta P_2 = 15$ mm (1 mm = 0.03937 in.). The new variables u and v are selected, with an arbitrary central class interval introduced, such that the central point is equal to zero. The variables u and v have positive and negative integers as marks of successive class intervals. The means of u and v are $\bar{u} = 3.150$ and $\bar{v} = 1.985$, respectively. By decoding class intervals, $\bar{P}_1 = 65.75$ mm and $\bar{P}_2 = 53.22$ mm (1 mm = 0.03937 in.). The regression coefficients are $B_y = 0.665$ and $B_x = 0.880$, and the two regression lines are $P_2 = 9.50 + 0.665P_1$ and $P_1 = 19.30 + 0.880P_2$.

Figure 8-II-2 shows the relation between monthly precipitations P_1 and P_2 for the two stations, with the isolines of equal bivariate frequencies; the points of the mean values y_0 of all y_j for given x_i and the mean values x_0 of all x_i for given y_j of class intervals; and the two regression lines, with means $\bar{P}_1$ and $\bar{P}_2$.

B. Measures of Linear Correlative Association

1. Correlation Coefficient and Coefficient of Determination. The *correlation coefficient* is the most commonly used statistical parameter for measuring the degree of association of two linearly dependent variables. It is defined [2] as

$$r = \frac{\Sigma(\Delta x_i \, \Delta y_i)}{\sqrt{\Sigma(\Delta x_i)^2 \Sigma(\Delta y_i)^2}} = \frac{\Sigma x_i y_i - N\bar{x}\bar{y}}{s_x s_y (N - 1)} \tag{8-II-11}$$

Table 8-II-3. Computation of Sums for Determination of Parameters in Linear Regression and Correlation Analysis of Precipitations

$v\Sigma f_{ij}u$	$\Sigma f_{ij}u$	$f_{yj}v^2$	$f_{yj}v$	v	f_{yi}	Monthly precipitation, P_1 (mm) = x_i														Monthly precipitation, P_2 (mm) = y_j
						8	23	38	53	68	83	98	113	128	143	158	173	188	203	
30	−6	75	−15	−5	3							1		1					1	158
16	−4	64	−16	−4	4							1	1	1			1			143
15	−5	27	−9	−3	3							1	1						1	128
20	−10	48	−24	−2	12							3	2	5			2			113
−16	16	14	−14	−1	14				2	1	2	3	4	2						98
0	28	0	0	0	27			1	1	3	12	2	2	2	2		1	1		83
90	90	39	39	1	39			2	7	9	9	9	1	2						68
312	156	200	100	2	50		3	7	14	12	6	3	2	2	1					53
534	178	414	138	3	46		6	12	12	7	4	5								38
930	240	800	200	4	50		20	14	7	6	2		1							23
960	192	775	155	5	31	14	11	5		1										8
2921		2456	554		279	14	40	41	43	39	35	28	14	15	3	0	4	1	2	f_{xi}
						7	6	5	4	3	2	1	0	−1	−2	−3	−4	−5	−6	u
					876	98	240	205	172	117	70	28	0	−15	−6	0	−16	−5	−12	$f_{xi}u$
					4546	686	1440	1025	688	351	140	28	0	15	12	0	64	25	72	$f_{xi}u^2$
						70	159	133	97	82	39	9	−6	−15	2	0	−8	0	−8	$\Sigma f_{ij}v$
					2921	490	954	665	388	246	78	9	0	15	−4	0	32	0	48	$u\Sigma f_{ij}v$

NOTATION
$u\Sigma f_{ij}v = v\Sigma f_{ij}u$
1 mm = 0.03937 in.
Entries corresponding to x_i and y_j in the table are f_{ij} values.

where s_x and s_y are the standard deviations of x_i and y_i, respectively. As s_x and s_y are positive, the sign of r depends on the sum of the cross products $\Delta x_i\,\Delta y_i$. Since this sum can vary between $+s_x s_y$ and $-s_x s_y$, the correlation coefficient varies from $+1$ to -1. If the sum of the cross products $\Delta x_i\,\Delta y_i$ is zero, the variables x and y are linearly independent and the correlation coefficient is zero. However, it does not follow that the variables are independent if the sum of the cross products is zero. If the points (x_i, y_i) fall symmetrically around and all along a circle, r is zero, or the linear dependence is zero. However, in this case there is a high correlation for the function which has a circle as the regression line.

The correlation coefficient is unity only if all points fall on a straight line. A positive value of r means that y increases with an increase of x. A negative value of r means that y decreases with an increase of x.

If there is no linear relationship, $r = 0$. If there is a functional linear relationship, $r = \pm 1$. All values of r between these limits describe the various degrees of correlative association. The greater the absolute value of r, the greater is the linear correlation.

By geometrical interpretation, the more the bivariate-frequency isolines (Fig. 8-II-2) are concentrated along a straight line, the greater is the correlation coefficient. The correlation coefficient will be unity for the case where all isolines coincide with the straight line. When the isolines are concentrated approximately in circles, the correlation coefficient is around zero, and there is no significant linear dependence.

In the case of grouped data in a bivariate distribution, the correlation coefficient [2] can be computed from the expression

$$r = \frac{\sum_{i=1}^{p}\sum_{j=1}^{q} f_{ij}\Delta x_i\,\Delta y_j}{\sqrt{\sum_{i=1}^{p} f_{xi}(\Delta x_i)^2 \sum_{j=1}^{q} f_{yj}(\Delta y_j)^2}} \qquad \text{(8-II-12)}$$

Using Eqs. (8-II-8) and (8-II-11), or Eqs. (8-II-10) and (8-II-12), the regression coefficients may be expressed in terms of the correlation coefficient r and the standard deviations s_x and s_y:

$$B_y = r\frac{s_y}{s_x} \qquad \text{or} \qquad B_x = r\frac{s_x}{s_y} \qquad \text{(8-II-13)}$$

Furthermore,

$$D = r^2 = B_x B_y \qquad \text{(8-II-14)}$$

where D is defined as the *coefficient of determination.* It is a measure of the degree to which the variance or square of the standard deviation, s_y^2 and s_x^2, is explained or accounted for by the linear regression. In other words, it is a measure of the difference between the variance of the observed (actual) values y_i and the variance of the values determined for given values of x_i by the use of the linear-regression line. The greater D is, the smaller is this difference.

For the example of grouped data in Table 8-II-3, $D = r^2 = 0.581$ and $r = 0.762$. The coefficient D tells that 58.1 per cent of the variance is accounted for by the linear-regression relationship in this example.

2. Standard Deviation of Residuals. The *residuals* of the straight-line regression y versus x are $\Delta_{yi} = y_i - y$, where y_i is the observed value and y is the value determined from the straight regression line for a given $x = x_i$. The *standard deviation of the residuals* for the straight regression line y versus x is defined [2, p. 72] as

$$S_y = \sqrt{\frac{\sum_{i=1}^{N}(\Delta_{yi})^2}{N}} = s_y\sqrt{1 - r^2} \qquad \text{(8-II-15)}$$

A similar expression for S_x is obtained if Δ_{xi} and s_x replace Δ_{yi} and s_y, respectively.

Taking into account the number of degrees of freedom in the case of small samples, the unbiased standard deviation of residuals, $\hat{S}_y$, is given as

$$\hat{S}_y = \sqrt{\frac{N-1}{N-2}}\, s_y \sqrt{1-r^2} \tag{8-II-16}$$

The greater $\hat{S}_y$ or $\hat{S}_x$, the wider is the spread of the points around the regression line and the less accurate are the values determined from the regression lines. Referring to the example in Table 8-II-3, the standard deviations for the variables u and v are $s_u = 2.52$ and $s_v = 2.20$. By transforming u and v to P_1 and P_2, respectively, the standard deviations for the variables P_1 and P_2 are $s_x = 37.80$ mm and $s_y = 33.00$ mm. Since N is very large, Eq. (8-II-15) gives $S_y = 21.3$ mm and $S_x = 24.4$ mm.

3. Interpretation of Parameters. To interpret properly the practical aspects of linear regression, all three parameters B_y, r, and S_y (or B_x, r, and S_x) are useful and necessary. B_y tells how fast the dependent variable increases or decreases with a change of the other variable. r measures the degree of the association. S_y measures the spread of points around the straight line.

In the example of Table 8-II-3, $B_y = 0.665$, $r = 0.762$, and $S_y = 21.3$ mm. Thus P_2 increases very fast with a corresponding increase of P_1, the association is very high, and despite the high association, the spread of points is also large.

C. Statistical Inference

1. Correlation Coefficient. To make a statistical inference about the computed correlation coefficient r, the type of distribution of r should be known for an infinite number of samples of size N, all drawn from the same universe of data. The general assumption is that in all samples both x_i and y_i are normally distributed and are random in sequence; that is, they are internally independent variables. The correlation coefficient r of the sample available is only an estimate of ρ, the *correlation coefficient for the universe.*

The frequency distribution of r is bounded at both sides ($r = +1$, $r = -1$), but depends only on ρ and N. For $\rho = 0$, it is a bounded symmetrical function, but for large absolute value of ρ, it is a highly skewed distribution. The degree of skewness is a function of ρ.

For ρ close to zero, an approximate method for testing the hypothesis that r is different from zero is based on the assumption that r is normally distributed. It uses the *standard deviation of the correlation coefficient* S_r as a criterion.

$$S_r = \frac{1-r^2}{\sqrt{N}} \tag{8-II-17}$$

where ρ should have been used but is replaced by its sample value r [2, p. 143].

Instead of using the confidence interval (such as 90 per cent, 95 per cent, etc.), the limits for r are sometimes given either as three standard deviations on both sides of r (that is, $r \pm 3S_r$) or as four standard deviations for a stronger test of significance (that is, $r \pm 4S_r$). If both limits ($r + 3S_r$ and $r - 3S_r$, or $r + 4S_r$ and $r - 4S_r$) are of the same sign as r, it is considered by this test that r is *significantly different from zero.* If, however, these limits have different signs, then r is not considered as significantly different from zero.

For the example in Table 8-II-3, $S_r = 0.025$, so that $r + 3S_r = 0.837$ and $r - 3S_r = 0.687$, or $r + 4S_r = 0.862$ and $r - 4S_r = 0.662$. All are positive, so that $r = 0.762$ is significantly different from zero with respect to either of these empirical confidence limits.

If a test of significance of r for ρ close to zero is to be made on the basis of confidence interval and if $N \geq 5$, the distribution of r follows approximately the Pearson type II function, which has the shape of a symmetrical bell-shaped curve. As N becomes

large, or $N \geq 50$, the function approaches the normal distribution and the standardized variable t (with mean equal to zero and standard deviation equal to unity) is

$$t = r\sqrt{N-1} \tag{8-II-18}$$

For any prescribed level of significance, the significant departure of r from zero may be tested by using this function. The table for the Student t distribution should be used for the prescribed level of significance if Eq. (8-II-18) is applied. Such a table is generally available in standard statistical textbooks.

If ρ is different from zero, Fisher's transformation [2, p. 200] with a new variable

$$z = \tfrac{1}{2} \ln \frac{1+r}{1-r} \tag{8-II-19}$$

should be used, with ln denoting the natural logarithm. For normally distributed x and y, z is normally distributed, with the mean

$$\bar{z} = \tfrac{1}{2} \ln \frac{1+\rho}{1-\rho} \tag{8-II-20}$$

and the standard deviation

$$s_z = \frac{1}{\sqrt{N-3}} \tag{8-II-21}$$

From this the standardized variable, or argument t (with mean equal to zero and standard deviation equal to unity), can be determined by

$$t = \frac{z - \bar{z}}{s_z} \tag{8-II-22}$$

For the example in Table 8-II-3, with $N = 279$, the values of t-distribution function for the 95 per cent confidence level are $t = \pm 1.97$. Replacing ρ by $r = 0.762$ in Eq. (8-II-20) as an estimate of ρ results in $\bar{z} = 1.0$, $s_z = 0.063$ from Eq. (8-II-21), $z_1 = 1.119$, and $z_2 = 0.881$. Finally, from Eq. (8-II-19), $r_1 = 0.807$ and $r_2 = 0.706$. By this test $r = 0.762$ is highly significant, because, on the average, of 95 out of 100 samples of size $N = 279$ and $\rho = 0.762$, r would be between 0.706 and 0.807, a rather narrow confidence interval for the correlation coefficient. This assumes that P_1 and P_2 are internally independent; i.e., they are not serially correlated. This is not the case, however.

It is generally considered in hydrology that a linear-regression approximation is good if r by these tests is significantly different from zero, and if the absolute value of r is greater than 0.6; that is, $r > 0.6$ and $r < -0.6$. However, if the physical relationship can be well justified, the linear regression can be accepted even if the absolute value of r is smaller than 0.6. If there is no physical reason to justify a computed absolute correlation coefficient greater than 0.6, especially for small N, the linear regression must be used discreetly, in spite of the high degree of correlation.

2. Regression Coefficient. Assuming an infinite number of values B_y and an infinite number of samples (x_i, y_i) of size N, the unbiased estimate of the standard deviation s_b of the distribution of B_y is [3, p. 337]

$$\hat{s}_b = \sqrt{\frac{B_y^2(1-r^2)}{(N-2)r^2}} \tag{8-II-23}$$

The standardized variable of the B_y distribution is

$$t = \frac{r(B_y - \beta_y)}{B_y}\sqrt{\frac{N-2}{1-r^2}} \tag{8-II-24}$$

where β_y is the true value of B_y. For $\beta_y = 0$, the standardized variable t is

$$t = r\sqrt{\frac{N-2}{1-r^2}} \tag{8-II-25}$$

which is dependent on B_y only through the relation of r and B_y.

Using Eq. (8-II-25) and the table of the t distribution for the selected level of significance, the test that B_y is not significantly different from zero can be made. If the t value computed from Eq. (8-II-25) is much greater than that which corresponds to the selected level of significance, then B_y is significantly different from zero. For the example from Table 8-II-3, Eq. (8-II-25) gives $t = 19.5$, for $r = 0.762$ and $N = 279$. The t table shows that the level of significance is very close to 100 per cent for $t = 19.5$, so that $B_y = 0.665$ is significantly different from zero at either the 90 or the 95 per cent level of significance. For other values of B_y, the test that B_y does not depart from B_y significantly on the selected level is made by using Eq. (8-II-24). If $\beta_y = 0.50$ in the above example and the level of significance is selected to be 95 per cent, Eq. (8-II-24) for $t = \pm 1.97$ gives the values for the limits of confidence interval: $B_1 = 0.555$ and $B_2 = 0.455$. So, at this level, $B_y = 0.665$ is significantly different from $\beta_y = 0.50$, because it is outside the confidence interval. The variables P_1 and P_2 are assumed internally independent.

3. Intercept. The standardized variable of the distribution of A_y is

$$t = \frac{r(A_y - \alpha_y)}{B_y}\sqrt{\frac{N-2}{(1-r^2)(s_x^2 + \bar{x}^2)}} \tag{8-II-26}$$

where α_y is the true value of A_y.

The standard test is to determine whether or not A_y is significantly different from zero, that is, whether the regression line may be assumed to pass through the coordinate origin. If the regression does pass through the origin, α_y is taken to be zero. Strictly, B_y, s_x, and $\bar{x}$ should be replaced by their true values, or values of the universe, β_y, σ_x, and μ, but the estimates B_y, s_x, and $\bar{x}$ would give results sufficiently accurate for practical purposes, since they are the only values available. Also, the true value of r is unknown; hence the estimate of r is the only value that can be used.

For the above example, with $A_y = 9.50$ mm, $B_y = 0.665$, $r = 0.762$, $N = 279$, $s_x^2 = 1428$, $P_1^2 = \bar{x}^2 = 4{,}330$, and for the 95 per cent level of significance, the limits of the confidence interval are $A_1 = 5.08$ mm and $A_2 = -5.08$ mm. At that level, the value $A_y = 9.50$ mm is significantly different from zero. At the 98 per cent level, $A_{1,2} = \pm 6.04$ mm and $A_y = 9.50$ mm is still significantly different from zero.

4. Inference in the Case of Nonnormal Distributions of Variables Which Are Internally Dependent. If the distributions of x_i and y_i are not highly asymmetrical, the above formulas and procedures can be used as approximations for statistical inference. However, when the distributions are highly asymmetrical, transformations of the variables x and y may give more symmetrical distributions, and thus provide more reliable statistical inference for the regression and correlation parameters.

When x and y are internally dependent (serially correlated), an approximate method of statistical inference can be accomplished through the computation of a sample of reduced size. The standard deviations of the parameters and the confidence intervals are larger in the case of a serially correlated series than in the case of the same series when internally independent.

If the variables x and y follow a cyclic trend, this cyclic trend should be removed. If they do not, but are internally dependent, the sample size N may be reduced to [4]

$$N' = \frac{N}{1 + 2r_1r_1' + 2r_2r_2' + \cdots} \tag{8-II-27}$$

where r_i and r_i' are the serial correlation coefficients of order i for the variables x and y, respectively.

For example, consider the annual flows of the St. Lawrence River at Ogdensburg and the Mississippi River at St. Louis, for which the records for 100 years of concurrent observations are available. The first-order serial correlation coefficients are $r_1 = 0.705$ for St. Lawrence and $r_1' = 0.294$ for Mississippi. All higher-order values for the Mississippi River are not significantly different from zero. Therefore the reduced sample size $N' = N/(1 + 2r_1r_1') = 70.7$, or about 71, a 29 per cent reduction in the size of the sample resulting from the effect of serial correlation of the data.

IV. SIMPLE CURVILINEAR REGRESSION AND CORRELATION

A. Curvilinear Regression

Linear regression is a special case of *curvilinear regression*, where either the variables x and y are originally linearly related or, by a transformation, a curvilinear relationship is linearized. All nonlinear functions which describe the relation between the variables are easier to determine if they can be reduced by transformation to a linear relationship.

Theoretically, any function $y = f(x)$ can be fitted to data by computing its parameters by the method of least squares of the departures $\Delta_{yi} = y_i - y = y_i - f(x_i)$. The other regression line of the same family, given as $x = g(y)$, can be determined by minimizing the squares of departures $\Delta_{xi} = x_i - x = x_i - g(y_i)$.

All functions which cannot be easily transformed to a linear relationship or fitted linearly to the data may be developed in the form of a power series, by neglecting the higher-order terms and reducing the function to a polynomial. In that case, the polynomial can be fitted to the points (x_i, y_i). Linear regression is then a special case, where all terms of the power series of order higher than first are neglected.

The general form of a polynomial regression of y versus x is represented by

$$y = A_0 + A_1x + A_2x^2 + \cdots + A_mx^m \qquad \text{(8-II-28)}$$

or for x versus y by

$$x = B_0 + B_1y + B_2y^2 + \cdots + B_my^m \qquad \text{(8-II-29)}$$

where the coefficients A_i and B_i are the parameters to be determined by the least-squares method. The number m is so chosen as to minimize the sum of squares of departures from the line. It must be stressed, however, that m should be much lower than the number of points N, in order to have enough degrees of freedom $(N - m)$ for making a reliable estimate of the standard deviation. Also, in order to keep the arithmetical work reasonably simple, m must be small. Generally, m is in the order of 2 to 4.

To select the number m, either a graphical plot will suggest the order of magnitude of m or two or three different m values may be selected, and the standard deviation of residuals from the line for each case is taken as a measure to decide the most suitable m for fitting.

When a polynomial is selected for curvilinear regression, two methods of treatment are used. They are (1) to compute the parameters by the least-squares method directly, as was explained earlier by the example in Table 8-II-1, and (2) to consider $x, x^2, x^3, \ldots$ as the variables $u, v, \ldots z$ and apply multiple-linear-regression procedures (described in Subsec. V). The least-squares method in treatment 1 involves the solution of $m + 1$ linear equations involving $m + 1$ moments of x and y and the cross products of powers of x and y, with the highest sum of power $m + 1$.

B. Correlation Index

As was shown earlier in the case of the correlation coefficient for linear association, $r^2 = 1 - S_y^2/s_y^2 = 1 - S_x^2/s_x^2$. In that case $S_y^2/s_y^2 = S_x^2/s_x^2$, where S_x, S_y, s_x, and s_y are standard deviations of residuals and standard deviations of marginal distributions for x and y, respectively. In the case of curvilinear association, these two ratios are not equal, so that the measure of correlation has to depend on the regression curve of either y versus x or x versus y.

For the regression curve of y versus x, the parameter to be used is

$$R_y = \sqrt{1 - \frac{S_y^2}{s_y^2}} \tag{8-II-30}$$

which is called the *correlation index* for y versus x.

For the regression curve of x versus y,

$$R_x = \sqrt{1 - \frac{S_x^2}{s_x^2}} \tag{8-II-31}$$

is the correlation index for x versus y. These two indices indicate the closeness with which the points (x_i, y_i) approximate the regression lines. As s_x and s_y are different from zero, the absolute values of R_y and R_x can be unity only when $S_y = S_x = 0$, or when all points are on the curve. Also, R_x and R_y are zero if $S_y = s_y$ and $S_x = s_x$, respectively.

C. Regression and Correlation Inference

When the polynomial regression lines are used, the test of significance for the computed regression coefficients is less simple than it is in the case of multiple linear regression. The general hypothesis to be tested is whether a selected power m of the polynomial represents the relationship or a term of higher power contributes significantly to the relationship. In applying polynomial regression, it is not permitted to omit any intermediate term between the intercept A_0 or B_0 and the regression coefficient of the highest-power term.

Two tests are generally applicable. The first is the test of whether the intercept differs from zero. The second is the test of the highest significant term. Thus the tests are made at the two extremes. The procedure is first to select a polynomial with m power terms. If by test the regression coefficient of the highest-power term is significantly different from zero, then the polynomial with $m + 1$ power terms should be selected and the test repeated. If the regression coefficient of the highest-power term $m + 1$ is not significantly different from zero, then the polynomial with m power terms is finally selected. If the regression coefficient of the highest-power term m is insignificant, the regression coefficients must be computed for the polynomial of $m - 1$ power terms, and the regression coefficient of the highest-power term $m - 1$ tested first, etc. [5, pp. 40–42].

V. MULTIPLE LINEAR REGRESSION AND CORRELATION

A. General

The association of three or more variables can be investigated by the multiple regression and correlation analysis.

The multiple-regression relation may be expressed in the form

$$x_1 = f(x_2, x_3, \ldots, x_m) \tag{8-II-32}$$

where $x_1, x_2, \ldots, x_m$ are m variables. This equation gives the estimate of x_1 for given values of all other variables.

If Eq. (8-II-32) is linear, the regression is referred to as *multiple linear regression* and the association is *multiple linear correlation*.

Because linear equations are easier to treat than nonlinear multiple relations, variables of nonlinear relations in hydrology are often transformed to linear relations for multiple-regression analysis.

B. Linear Regression with Three Variables

The linear-regression relation for three variables is

$$x_1 = B_1 + B_2x_2 + B_3x_3 \tag{8-II-33}$$

where x_1 is the dependent variable and x_2 and x_3 are externally independent variables. In this equation there are three parameters to be determined. The parameter B_1 is the *intercept*. The parameter B_2 is the *multiple regression coefficient* of x_1 on x_2 when x_3 is kept constant. The parameter B_3 is the multiple regression coefficient of x_1 on x_3 when x_2 is kept constant.

The residuals of x_1, or the difference between the observed value x_1' and the estimated value x_1 from Eq. (8-II-33), are $\Delta_1 = x_1' - x_1 = x_1' - B_1 - B_2x_2 - B_3x_3$. By use of the least-squares method to minimize the sum of Δ_1^2, the three partial differential equations in B_1, B_2, and B_3 give three linear equations in B_1, B_2, and B_3 for the fitting of curves:

$$\begin{aligned} B_2\Sigma(\Delta x_2)^2 + B_3\Sigma(\Delta x_2\,\Delta x_3) &= \Sigma(\Delta x_1\,\Delta x_2) \\ B_2\Sigma(\Delta x_2\,\Delta x_3) + B_3\Sigma(\Delta x_3)^2 &= \Sigma(\Delta x_1\,\Delta x_3) \\ B_1 = \bar{x}_1 - B_2\bar{x}_2 - B_3\bar{x}_3 & \end{aligned} \tag{8-II-34}$$

where $\bar{x}_i = \Sigma x_i/N$ and $\Delta x_i = x_i - \bar{x}_i$, with i taken from 1 to 3, and N is the sample size. The sums in the equations may be expressed and computed as $\Sigma(\Delta x_i)^2 = \Sigma x_i^2 - N\bar{x}_i^2$; and $\Sigma(\Delta x_i\,\Delta x_j) = \Sigma x_ix_j - N\bar{x}_i\bar{x}_j$, with i and j having the corresponding values 1, 2, and 3 in the equations. As the sum $\Sigma(\Delta x_1)^2$ is necessary for the computation of other parameters in the multiple-correlation analysis, nine sums of the sample must be computed altogether for the three variables—for the variables themselves, their squares, and their three different cross products. The solution of these linear equations gives the three coefficients B_1, B_2, and B_3.

C. Linear Regression with Several Variables

If there are m variables to correlate, including one dependent and $m - 1$ externally independent, the general equation for multiple linear regression is

$$x_1 = B_1 + B_2x_2 + \cdots + B_ix_i + \cdots + B_mx_m \tag{8-II-35}$$

where B_1 is the intercept and B_i is the multiple regression coefficient of the dependent variable x_1 on the independent variable x_i, with all other variables kept constant.

Applying the least-squares method for the sum of residuals $\Delta_1 = x_1' - B_1 - B_2x_2 - \cdots - B_mx_m$, m partial differential equations in $B_1, B_2, \ldots, B_m$ give m linear equations:

$$\begin{aligned} B_2\Sigma(\Delta x_2)^2 + B_3\Sigma(\Delta x_2\,\Delta x_3) + \cdots + B_m\Sigma(\Delta x_2\,\Delta x_m) &= \Sigma(\Delta x_1\,\Delta x_2) \\ B_2\Sigma(\Delta x_2\,\Delta x_3) + B_3\Sigma(\Delta x_3)^2 + \cdots + B_m\Sigma(\Delta x_3\,\Delta x_m) &= \Sigma(\Delta x_1\,\Delta x_3) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ B_2\Sigma(\Delta x_2\,\Delta x_m) + B_3\Sigma(\Delta x_3\,\Delta x_m) + \cdots + B_m\Sigma(\Delta x_m)^2 &= \Sigma(\Delta x_1\,\Delta x_m) \\ B_1 = \bar{x}_1 - B_2\bar{x}_2 - B_3\bar{x}_3 - \cdots - B_m\bar{x}_m & \end{aligned} \tag{8-II-36}$$

where $\Delta x_i = x_i - \bar{x}_i$, with $i = 1$ to m. These equations enable the determination of m parameters. The computational work increases with the increase of either the number m of the variables or the sample size N.

When there are a small number of variables (say, three or four), either ungrouped or grouped data may be used for a multiple regression and correlation analysis. For more than four variables the use of multivariate distributions in the form of grouped data will become cumbersome, because of the difficulties in obtaining a suitable representation of the multivariate distributions.

The selection of a feasible method of solving the linear equations depends on the number of variables and the available computational devices. When the number of variables and the sample size are relatively small, desk computers can be used for an analytical solution, or a graphical approach can be taken. The practice in the past

has been to restrict the number of variables. Nowadays, card tabulators and digital computers make possible the use of any number of variables which are statistically significant.

The IBM-type tabulators, designed to solve multiple-regression problems [6], can speed up the computational process tremendously, especially if they are equipped with automatic multiplying devices. Electronic computers can perform any operation in solving the multiple-linear-regression problems by following a suitable set of instructions given to the machines.

In a study of floods in New England [7], a multiple linear regression and correlation analysis was made. By this analysis, the annual peak discharge in cubic feet per second for a recurrence interval of T years was found to be

$$Q_T = aA^bS^cS_t^dI^et^fO^g \tag{8-II-37}$$

where A = drainage area, sq mi
S = slope of main channel, ft/mile
S_t = per cent of surface storage area plus 0.5 per cent
I = 24-hr rainfall, in in., of a recurrence interval of T years
t = average temperature in January, °F below freezing
O = orographic factor
b, c, d, e, f, g = multiple linear regression coefficients when a logarithmic transformation is applied to all variables

By a logarithmic transformation of all seven variables, the general relationship expressed by the above equation can be transformed to a multiple-linear relationship of the type represented by Eq. (8-II-35). Table 8-II-4 shows the computation of the coefficients $a, b, c, \ldots$ for three recurrence intervals. The computation work was made by a desk computer. The results given in the table show how the additional variables affect the values of the multiple regression coefficients computed previously.

D. Measures of Multiple Linear Correlative Association

The degree of correlation of a dependent variable x_1 to many externally independent variables in a multiple-linear association is measured by any of the five parameters, the standard deviation of residuals, the multiple correlation coefficient, the coefficient of multiple determination, the partial correlation coefficients, and the beta coefficients.

1. The Standard Deviation of Residuals. This is determined as the standard deviation of the differences between the observed values x_1' and the values of x_1 estimated from Eq. (8-II-35). This parameter is a measure of the closeness with which the observed values approach the estimated values. It will be designated by S_1, to be distinguished from the standard deviation s_1 of the marginal distribution of x_1.

If the total number of points $(x_1, x_2, x_3, \ldots, x_m)$ is small in comparison with the number m of the variables involved, the standard deviation of residuals computed by the usual procedure is biased, since the degrees of freedom $N - m$ are much smaller than the sample size N.

Taking into account m parameters and $N - m$ degrees of freedom,

$$\hat{S}_1 = \sqrt{\frac{NS_1^2}{N-m}} = \sqrt{\frac{\Sigma\Delta_1^2}{N-m}}$$

$$= \sqrt{\frac{1}{N-m}\left[\Sigma(\Delta x_1)^2 - B_2\Sigma\Delta x_1\,\Delta x_2 - \cdots - B_m\Sigma\Delta x_1\,\Delta x_m\right]} \tag{8-II-38}$$

where S_1 is the *biased*, and $\hat{S}_1$ the *unbiased, standard deviation of residuals*. If, however, $\hat{S}_1$ exceeds $\hat{s}_1$ (the unbiased standard deviation of the marginal distribution of x_1), $\hat{s}_1$ would be used for the standard deviation of residuals instead of $\hat{S}_1$. This can occur only for a very small number of degrees of freedom.

The addition of a new independent variable x_{m+1} in a multiple linear regression will decrease the standard deviation of residuals if this variable influences the dependent variable x_1. Whether this variable should be included in the multiple regression depends on how much the standard deviation of residuals is decreased.

For the study of floods in New England (Subsec. V-C), Table 8-II-4 shows how the standard deviation of residuals in percentage of the mean flood discharge changes with a change of the number and selection of independent variables.

2. Multiple Correlation Coefficient. The correlation coefficient is a measure of the linear association of two variables. In this case, the two variables are the estimated and actual values. Thus the correlation coefficient is the ratio of the standard deviation of estimated values to that of actual values, or

$$r = \frac{s_{e1}}{s_1} = \sqrt{1 - \frac{S_1^2}{s_1^2}} \tag{8-II-39}$$

where s_{e1} is the standard deviation of estimated values, s_1 is the standard deviation of actual values of x_1, and S_1 is the standard deviation of residuals.

In a similar manner, the measure of linear association of many variables, including one dependent and the others externally independent of one another, is expressed by the ratio of the standard deviation of estimated values divided by the standard deviation of the actual values of the dependent variable. This is the *multiple correlation coefficient*, or the *coefficient of multiple correlation*. It measures the combined influence of all externally independent variables in terms of the difference between the actual and the estimated values of the dependent variable. The coefficient is expressed as

$$R_1 = \frac{s_{e1}}{s_1} = \sqrt{1 - \frac{S_1^2}{s_1^2}} = \sqrt{\frac{B_2\Sigma\,\Delta x_1\,\Delta x_2 + B_3\Sigma\,\Delta x_1\,\Delta x_3 + \cdots + B_m\Sigma\,\Delta x_1\,\Delta x_m}{\Sigma(\Delta x_1)^2}} \tag{8-II-40}$$

The unbiased value $\hat{R}_1$ is obtained by putting in Eq. (8-II-40) the unbiased values $\hat{S}_1$ and $\hat{s}_1$ for S_1 and s_1, respectively. Table 8-II-4 gives the values of R_1 for the example of floods in New England (Subsec. V-C), showing the change of R_1 with the change of number and selection of variables.

The multiple correlation coefficient is a dimensionless measure of linear association while the standard deviation of residuals is a measure having the dimension of the variable x_1. As s_1 is always the same, regardless of the number of other variables, a decrease of the standard deviation of residuals will cause an increase of the multiple correlation coefficient.

The multiple correlation coefficient can be defined as the correlation coefficient of linear association of two variables, one being the actual values of the dependent variable and the other the estimated values of the dependent variable, based upon all other externally independent variables.

3. Coefficient of Determination. The square of the multiple correlation coefficient, R_1^2, is referred to as the *coefficient of multiple determination* D_1.

The square of the multiple correlation coefficient indicates the part of the variance in the dependent variable s_1^2 which has been mathematically accounted for, whereas $1 - D_1 = 1 - R_1^2$ indicates the part of the variance which has not been accounted for by the multiple linear correlation of the variables $x_1, x_2, \ldots, x_m$.

Table 8-II-4 gives the values of D_1 in percentages for different numbers and selections of variables in the example of floods in New England (Subsec. V-C).

4. Partial Correlation Coefficients. The *partial correlation coefficients* measure the association of the dependent variable x_1 with any given independent variable x_i. Thus they are measures of the variation of x_1, which is explained by, and only by, the given x_i. A simple method of estimating the partial correlation coefficient r_{1-i} involves the determination of (1) the multiple correlation coefficient R_1 between x_1 and all the independent variables, and (2) the multiple correlation coefficient R_{1-i}, between x_1 and all the independent variables except the chosen x_i. The partial correlation coefficient r_{1-i} is then determined from [8, p. 193]

$$r^2_{1-i} = \frac{(1 - R^2_{1-i}) - (1 - R_1^2)}{1 - R^2_{1-i}} = 1 - \frac{1 - R_1^2}{1 - R^2_{1-i}} \tag{8-II-41}$$

Table 8-II-4. Multiple Linear Regression and Correlation Analysis of Floods in New England*

Recurrence interval T	No. stations N	Independent variables included	a	b	c	d	e	f	g	R_1	Standard deviation, % of mean Q	D_1
											$\bar{Q}$ = 164.4 cfs:	
		A	113.06	.7858						.883	63.6	.780
		A, S	5.9808	1.0182	.5720					.942	44.4	.888
10	164	A, S, O	3.4623	1.0804	.6119				.8482	.965	34.5	.932
		A, S, O, S_t	6.1854	1.0508	.5153	.2645			.8543	.972	30.4	.945
		A, S, O, S_t, t	5.1680	.9707	.4306	.2524		.3549	1.0815	.978	27.2	.959
		A, S, O, S_t, t, I	3.6365	.9726	.4255	.2592	.1915	.3963	1.0668	.978	27.2	.959
											$\bar{Q}$ = 145.4 cfs:	
		A	342.32	.6764						.875	59.9	.765
		A, S	17.892	.9197	.5704					.931	44.6	.865
50	116	A, S, O	9.7302	1.0006	.5727				.9894	.962	33.0	.925
		A, S, O, S_t	17.115	.9675	.4978	.2520			.9288	.969	29.9	.940
		A, S, O, S_t, t	13.433	.9354	.4602	.2554		.2165	1.1089	.971	29.0	.940
		A, S, O, S_t, t, I	1.8401	.9458	.4349	.2757	.8773	.4504	1.0866	.975	26.8	.951
											$\bar{Q}$ = 134.6 cfs:	
		A	395.75	.6922						.862	62.8	.743
		A, S	25.504	.9095	.5438					.910	50.6	.827
100	100	A, S, O	13.937	.9922	.5332				1.0627	.950	37.7	.902
		A, S, O, t	10.964	.9480	.4667			.2837	1.2930	.952	36.8	.905
		A, S, O, t, I	1.1194	.9524	.4409		.9507	.4409	1.2730	.958	34.6	.920
		A, S, O, t, I, S_t	1.4203	.9226	.3823	−.2308	1.0685	.5632	1.2129	.965	32.2	.935

*From Benson [7].

Since R^2_{1-i} is smaller than R_1^2, then $r_{1-i} \leq 1.0$. The greater the difference between R_1 and R_{1-i}, the smaller is the ratio on the right side of the equation and the greater is r_{1-i}. It is seen from Eq. (8-II-41) that a considerable amount of computation is necessary in the case of many variables if all partial correlation coefficients are to be determined.

It is important to stress that the partial correlation coefficients should have the sign of their corresponding regression coefficients. If B_i is negative, r_{1-i} is also negative. For a given number of variables and their multiple correlation coefficient R_1, the numerator of the fraction in the right end of Eq. (8-II-41) is a constant, and only the denominator changes from one partial correlation coefficient to another. This agrees with the definition of the coefficient, which measures the reduction of variation of the dependent variable when all other variables except the considered one have been taken into account.

If a new variable x_{m+1} is added, all the partial correlation coefficients computed for the m variables will change.

The partial correlation coefficients are important parameters, since they measure the association of each independent variable with the dependent one after the influence of certain related variables has been accounted for. It is possible that the simple correlation coefficient between the dependent and the new independent variable is very small, and this variable may still have a significant influence on the variation of the dependent variable. It may be reminded that the converse can also occur. The partial correlation coefficient is the measure of the association, after the relation of the dependent variable to other independent variables has been considered and included in the analysis. Therefore the simple correlation coefficient is not as true a measure of association between two variables in multiple linear correlation as is the partial correlation coefficient.

An independent variable may be highly correlated with another independent variable, when simple linear correlation is applied between the two variables, but it may not be highly correlated with the dependent variable in a multiple correlation. The partial correlation coefficient may give a reliable answer as to the real effect of an independent variable on the variation of the dependent variable. In hydrology one may find many examples of this type (e.g., the example given in Table 8-II-4).

5. Beta Coefficients. The effect of individual independent variables on the dependent variable may be measured by a dimensionless form of the regression coefficient. Expressing each variable in ratio to its standard deviation, Eq. (8-II-35) may be written as

$$\frac{x_1}{s_1} = \beta_1 + \beta_2 \frac{x_2}{s_2} + \beta_3 \frac{x_3}{s_3} + \cdots + \beta_m \frac{x_m}{s_m} \tag{8-II-42}$$

where

$$\beta_1 = \frac{B_1}{s_1};\ \beta_2 = B_2 \frac{s_2}{s_1};\ \beta_3 = B_3 \frac{s_3}{s_1};\ \cdots;\ \beta_m = B_m \frac{s_m}{s_1} \tag{8-II-43}$$

These are *beta coefficients*, the dimensionless parameters which measure the effect of the individual independent variables on the variation of the dependent variable.

The comparison of beta coefficients with partial correlation coefficients shows, in general, that the rank or order of magnitude for the effect of the independent variables is the same for both these measures. This does not always hold true, since they are two different sets of measures. It is easier to compute beta coefficients than partial correlation coefficients, since B_i values and s_i values in Eq. (8-II-43) must always be computed for regression analysis. For this reason, there is a general preference for beta coefficients, although the partial correlation coefficients are better measures of the influence of individual independent variables.

E. Statistical Inference of Regression Coefficients

The same general conclusions for the correlation coefficient in simple linear association can be applied to the measures of association in multiple linear regression.

When only the extreme low and extreme high values of variables in multiple linear regression are available, the multiple correlation coefficient, the partial correlation

coefficients, and the beta coefficients are likely to be higher. Thus, if the sample values of variables are available only for a narrow range, these coefficients are likely to be smaller than the true coefficients for the universe.

The unbiased standard deviation of any multiple regression coefficient B_i is [8, p. 283]

$$\hat{S}_{bi} = \sqrt{\frac{\hat{S}_1^2}{Ns_i^2(1 - R_i^2)}} = \frac{S_1}{s_i}\sqrt{\frac{1}{(N - m)(1 - R_i^2)}} \qquad \text{(8-II-44)}$$

where s_i is the standard deviation of marginal distribution of variable x_i, $\hat{S}_1$ is the unbiased standard deviation of residuals for variable x_1, S_1 is the biased standard deviation of residuals for variable x_1, $N - m$ is the number of degrees of freedom, and R_i is the multiple correlation coefficient of x_i with respect to all variables except the variable x_1, with i varying from 2 to m. If the variables $x_2, x_3, \ldots, x_m$ are externally independent among themselves, $R_i = 0$, and Eq. (8-II-44) becomes

$$\hat{S}_{bi} = \frac{S_1}{s_i\sqrt{N - m}} \qquad \text{(8-II-45)}$$

It may be noted that the more the variables $x_2, x_3, \ldots, x_m$ are related among themselves, the greater is R_i. Consequently, $\hat{S}_{bi}$ is larger, and the regression coefficient is less reliable.

The standard deviation of individual values of B_i can be very high. Some of the values are more accurate and some are less accurate than the others. If x_1 is estimated using all B_i values, the error in one value of B_i may be compensated by an error in another value of B_i, so that the error of estimated x_1 may be smaller than that which may appear from the errors in B_i. If the extreme values of x_1 are estimated by the multiple regression linear equation, the error may be large, because B_i values have much greater error for the extremes.

VI. MULTIPLE CURVILINEAR REGRESSION AND CORRELATION

The use of any function in multiple regression and correlation analysis other than a linear function is referred to as *curvilinear regression and correlation*. Either analytical or graphical methods of curvilinear regression are used in hydrologic analyses.

A. Analytical Method

The analytical method applies the least-squares fitting of a regression function. The computational work increases greatly and the determination of the parameters becomes more difficult as a more complex mathematical function is used or as the number of variables increases. The use of digital computers enables the application of different functions and of many variables in a multiple curvilinear regression analysis. The method is similar to that used for multiple linear regression, except that the derived equations for determining the regression coefficients are more complicated.

B. Graphical Method

The *coaxial graphical method* [8] has been widely used in hydrology [9].

An example, taken from Ref. 9 and represented in Fig. 8-II-3, can be used to explain this method. This example relates the basin recharge as the dependent variable to the antecedent precipitation, date (or week number), the rainfall amount, and the rainfall duration as the independent variables. A three-variable relationship is developed (chart A) by plotting antecedent precipitation vs. recharge, labeling the points with week numbers and fitting a family of curves, with one curve for each week. The second step (chart B) relates the observed recharge to the recharge computed from chart A, and the points are labeled with rainfall duration. A family of curves is fitted to represent the effect of different durations on recharge. The third step (chart C) relates the observed recharge to the recharge computed from charts A and B, labeling the points with rainfall amount. A family of curves is fitted to represent the

effect of rainfall amounts on recharge. Charts *A*, *B*, and *C* constitute the first approximation [9, pp. 173–175]. Chart *D* indicates the accuracy of the derived charts. A process of successive approximations must be used to converge upon the correct graphical solution to the relation. This process becomes more necessary as the interdependence of the independent variables increases. The successive approximation may be accomplished by a reverse process, beginning with the last independent variable used in each approximation and working back to the dependent variable, refining

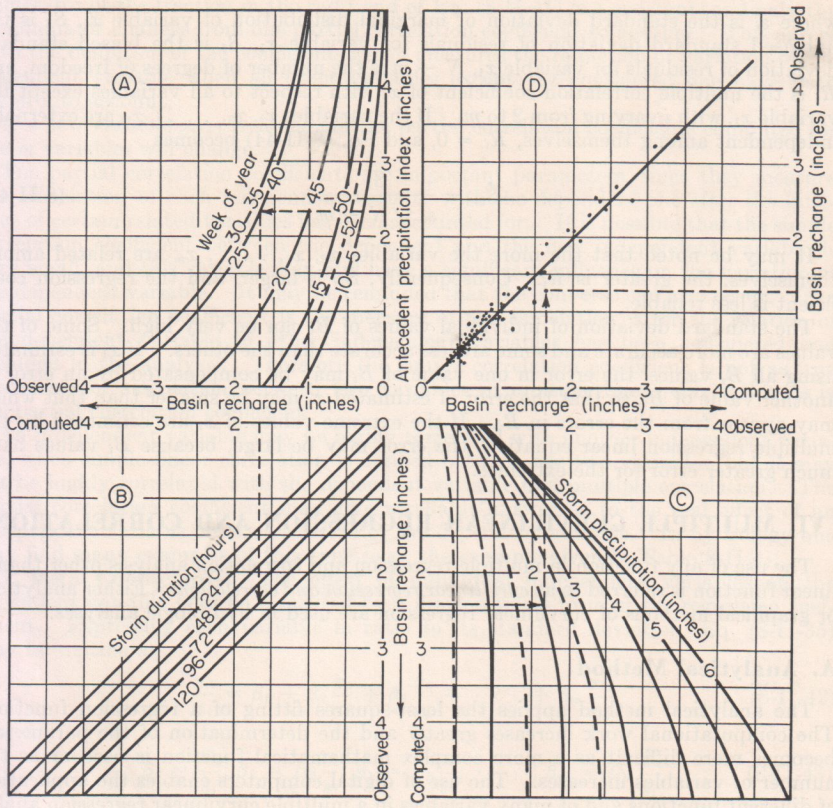

FIG. 8-II-3. Basin-recharge relation for the Monocacy River at Jug Bridge, Md., determined by the coaxial graphical method. (*After U.S. Weather Bureau and Linsley, Kohler, and Paulhus* [9].)

all families of curves along the way. Further approximations then follow the same procedure in both directions alternatively.

The advantage of the graphical coaxial method lies in the fact that the families of curves are independent of any analytical function and its parameters. The disadvantages are the laboriousness of successive approximations and the difficulties in making a reliable statistical inference.

VII. MULTIVARIATE ANALYSIS

Multiple regression is sometimes used to evaluate numerically the assumed model of a hydrologic process with the assumption that independent variables are not correlated. Synder [10] pointed out that in hydrologic analyses the independent variables

may not be uncorrelated. For example, a multiple regression solution does not usually provide a satisfactory estimate of the independent contributions of the rainfall and its duration to the runoff because a positive correlation may exist between the amount and duration of rainfall. In such cases, multivariate analysis appears to offer a workable and mathematically oriented solution in fitting prediction equations to the observed hydrologic data.

Multivariate analysis is the study of arrays of variables. In such analyses, a vector of means and a matrix of covariances of several variables are used instead of the simple mean and the variance of a single variable. This concept allows the association of errors with more than one variable, thus making possible the determination of all the truly independent components of variation in an array of variables. Detailed discussion of the analysis is beyond the present scope and the reader may refer to Refs. 11 to 14.

VIII. REFERENCES

1. Elderton, W. P.: "Frequency Curves and Correlation," 4th ed., Cambridge University Press, Cambridge, and Harren Press, Washington, D.C., 1953.
2. Weatherburn, C. E.: "A First Course in Mathematical Statistics," Cambridge University Press, London, 1952.
3. Kendall, N. G.: "The Advanced Theory of Statistics," 5th ed., Hafner Publishing Company, Inc., New York, 1952, vol. 1.
4. Bartlett, M. S.: Some aspects of the time-correlation problem in regard to tests of significance, *J. Roy. Statist. Soc.*, vol. 98, pp. 536–543, 1935.
5. Williams, E. F.: "Regression Analysis," John Wiley & Sons, Inc., New York, 1959.
6. Allan, D. H. W., and R. F. Attridge: The applications of an IBM calculating punch to solve multiple regression problems, *Proc. Seventh Ann. Conf. Am. Soc. Quality Control*, pp. 521–533, 1953.
7. Benson, M. A.: Factors influencing the occurrence of floods in a humid region of diverse terrain, *U.S. Geol. Surv. Water-Supply Paper* 1580-B, 1962.
8. Ezekiel, Mordecai, and K. A. Fox: "Methods of Correlation and Regression Analysis: Linear and Curvilinear," 3d ed., John Wiley & Sons, Inc., New York, 1959.
9. Linsley, R. K., Jr., M. A. Kohler, and J. L. H. Paulhus: "Hydrology for Engineers," McGraw-Hill Book Company, Inc., New York, 1958.
10. Snyder, W. M.: Some possibilities for multivariate analysis in hydrologic studies, *J. Geophys. Res.* vol. 67, no. 2, pp. 721–729, February, 1962.
11. Kendall, M. G.: "A Course in Multivariate Analysis," Hafner Publishing Co., Inc., New York, 1957.
12. Rao, C. R.: "Advanced Statistical Methods in Biometric Research," John Wiley & Sons, Inc., New York, 1952.
13. Anderson, T. W.: "An Introduction to Multivariate Statistical Analysis," John Wiley & Sons, Inc., New York, 1958.
14. Roy, S. N.: "Some Aspects of Multivariate Analysis," John Wiley & Sons, Inc., New York, 1958.