ENGINEERING MONOGRAPHS No. 2

United States Department of the Interior
BUREAU OF RECLAMATION

MULTIPLE CORRELATION IN FORECASTING SEASONAL RUNOFF

By Perry M. Ford

Denver, Colorado

June 1959 50 cents

United States Department of the Interior

FRED A. SEATON, Secretary

Bureau of Reclamation

FLOYD E. DOMINY, Commissioner
GRANT BLOODGOOD, Assistant Commissioner and Chief Engineer

Engineering Monograph

No. 2
(second revision)

MULTIPLE CORRELATION IN FORECASTING SEASONAL RUNOFF

by Perry M. Ford
Engineer, Hydrology Branch
Division of Project Investigations

Technical Information Branch
Denver Federal Center
Denver, Colorado

ENGINEERING MONOGRAPHS are published in limited editions for the technical staff of the Bureau of Reclamation and interested technical circles in government and private agencies. Their purpose is to record developments, innovations, and progress in the engineering and scientific techniques and practices that are employed in the planning, design, construction, and operation of Reclamation structures and equipment. Copies may be obtained from the Bureau of Reclamation, Denver Federal Center, Denver, Colorado, and Washington, D. C.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Most hydrologic phenomena are products of multiple causation. Flood season discharge is associated with several variables, antecedent in time, among which are accumulated precipitation, ground-water conditions, and temperature. The effect of each of these causal or associated factors (independent variables) upon flood season run-off (dependent variable) may be determined by multiple correlation, and the resulting equation applied to estimating run-off volume in advance of the flood season. Studies initiated by the writer in 1938 have shown that the use of multiple correlation in forecasting seasonal run-off is a practicable and useful tool in the analysis of hydrologic data.

This monograph was prepared for use by Bureau employees who must deal with problems of hydrological forecasting. It assumes at the very least a reasonably thorough grounding in the fundamentals of statistics. Anyone not sure of these fundamentals, or who is using multiple correlation for the first time, should either make his initial study with someone experienced with the procedure, or have his results checked by such a person.

Emphasis has been placed on the application of statistical methods rather than on the development of the mathematical theory upon which the methods depend. Attention is directed to fundamental principles which have an important effect on the reliability of the analysis, such as length of record, limitation on the number of independent variables, selection of major causal factors, and significance of relationship.

In developing forecasting procedures a unique problem is confronted in that an important causal factor, run-off season precipitation, is unknown at the date of forecasting. A means of forecasting run-off and of determining the reliability of the forecast under these conditions is included. A procedure is offered to facilitate the screening of likely causal or associated factors other than the more obviously important variables. The examples selected illustrate how widely basins differ, and how a given basin required analysis in accordance with the physical and climatological characteristics peculiar to that basin.

The work involved in problems like those illustrated here consists of computations which can be carried out rapidly on a keyboard-type calculating machine. About three man-days should be adequate for solution of such problems, exclusive of exploratory work and tabulation of basic data. Accuracy of the work is assured by the numerous checks on computations.

The labor involved may be materially reduced through the use of automatic computing machines, especially for problems involving long periods of record and 4 or 5 variables; or where forecasting procedures with possible alternatives are to be developed for several streams requiring forecasts for several dates. Such solutions may conveniently include computation of the departures and extensions required in setting up the normal equations, the regression coefficients and their standard errors, the multiple correlation coefficient, the standard error of estimate, and the constant term. This revision of the monograph has been expanded to include an adaptation of multiple correlation computations to machine processing. The procedure is arranged for programming on various types of high-speed computers.

NOTATION

Y = Estimated value of the dependent variable (run-off)

X_1 = Observed values of the dependent variable

$X_2, X_3, \ldots, X_n$ = Observed values of independent variables

$x_2, x_3, \ldots, x_n$ = Deviations from mean values of $X_2, X_3, \ldots, X_n$, respectively

y = Deviation from mean value of Y

a = Constant term of the regression equation

$b_2, b_3, \ldots, b_n$ = Regression coefficients

b_i = Any particular regression coefficient

σ_b = Standard deviation of a regression coefficient, b

σ_m = Standard error of the mean

σ_{y-x_1} = Standard error of individual estimate

$(Y - X_1)_{0.90}$ = Error of individual forecast for 0.90 probability

n = Number of events, or n^{th} term in a series

m = Number of constants in the regression equation

S = Standard error of estimate

R = Coefficient of multiple correlation

r = Coefficient of simple correlation

M_1 = Mean of the dependent variable

$M_2, M_3, \ldots, M_n$ = Means of independent variables

M_y = Mean value of Y

M_x = Mean value of X

Z = Residual

A bar above a symbol ($\bar{R}$, $\bar{S}$, etc.) means that the degrees of freedom have been taken into account in the evaluation.

The notation follows that of Ezekiel's Methods of Correlation Analysis. For this reason (and because of its excellence), Ezekiel's text is suggested to anyone who wishes to review the fundamentals of correlation analysis. A complete reference is given on page 3.

The symbol n is used in this text to indicate (1) some definite number of observations, and (2) the n^{th} term in a series. In each case, the position of the letter clearly indicates its function, since in the first case it is never a subscript, while in the latter it is always a subscript.

Subscripts of the form $b_{12.34}$ are occasionally used. The first of the two digits to the left of the dot (not a decimal point) represents the dependent variable, while the second of the two represents the independent variable whose effect is stated. In subscripts having only one digit to the left of the dot, such as $R_{1 \cdot 234}$, that digit represents the dependent variable. In all cases, the digits to the right of the dot represent the independent variable or variables held constant during the process.

Frequently the subscript notations may be abbreviated without danger of ambiguity; thus $b_{12.34}$ may be identified by the use of the subscript 2 only, i.e., b_2; such abbreviations are employed in numerous instances in this discussion.

THE FORECASTING PROBLEM

The method of multiple correlation is applied here to estimating run-off volume in advance of the flood season. This run-off comes principally from melting snow during the spring and early summer months. In the following example, a forecast is made on April 1 of the April-through-July flow at a given point on a particular stream.

The problem is to develop a multiple regression equation which summarizes the relationships of past hydrologic events as evidenced in records of natural run-off and of factors contributing to run-off, and to determine the reliability of this equation as a means of forecasting. Such an equation is also referred to as an estimating equation, particularly when applied to a period of known run-off as a means of providing a comparison between estimated and observed flows; or as a forecasting equation when predicting future run-off. Since the equation is based on natural flow, the forecast value will require correction for storage and diversions which affect the flood season run-off volume.

Major Factors Affecting Run-off

Although precipitation and run-off are logically related as cause and effect, the amounts of precipitation occurring in different seasons contribute to run-off in different ways and in varying degrees. Recorded precipitation during the fall and early winter months provides an index of ground-water conditions. In some basins precipitation during the preceding summer shows a relationship with current flood season run-off. Snow surveys[1] or winter precipitation records provide an index of the moisture accumulated in the snow cover that is available for release upon melting. Temperatures during the accumulation period show an appreciable effect on flood season run-off in some basins. An added factor to be considered in the analysis is precipitation during the run-off season. In regions where this factor is of major importance, its inclusion in the analysis contributes to greater precision in computing the regression coefficients which apply to the other independent variables, and to greater confidence in the forecast and usefulness of the forecasting procedure.

It will be noted that seasonal precipitation is used rather than monthly values. This grouping of months is necessary in order to reduce the number of independent variables to a practicable minimum. The reduction in number of variables is particularly important in view of the relatively short periods of record available for the development of forecasting procedures. Seasonal values are desirable, also, because of the tendency toward normality of distribution exhibited by records for longer time intervals. These and other requirements for insuring reliability of the forecast will become more evident as the basic principles of multiple correlation are reviewed.

[1] The method of measuring water equivalents of snow by means of sampling the snow cover at specified courses is explained in Snow Surveying, Miscellaneous Publication No. 380, U. S. Department of Agriculture, June 1940.

PRINCIPLE OF MULTIPLE LINEAR CORRELATION

This brief explanation of the mathematical principle of multiple correlation is included to show how the method may be applied in practical problems of hydrology. Only the more elementary algebraic statements are used. Derivations and proofs of the basic formulas given may be obtained from textbooks on statistics.[2]

Multiple correlation may be regarded as an extension of the simple, two-variable correlation procedure. A linear regression involving two variables may be expressed by an equation of the form

$$Y = a + bX \quad \ldots\ldots\ldots\ldots \quad (1)$$

in which Y is a dependent variable, being dependent upon values assigned to an independent variable, X. When the above

[2] A general treatment of the principle of multiple correlation is contained in: Ezekiel, Mordecai, Methods of Correlation Analysis, 2nd Edition, John Wiley and Sons, New York, 1941.

equation is plotted on rectangular coordinates, the two constants, a and b, are the Y intercept and the slope, respectively, of the resulting regression line.

For two or more independent variables the relationship may be represented by a multiple linear regression equation as

$$Y = a + b_2X_2 + b_3X_3 + \ldots + b_nX_n \quad . \quad (2)$$

in which Y, as before, is dependent upon values assigned to the several independent variables, and a is a constant term. The independent variables are designated as $X_2, X_3, \ldots, X_n$, and the several regression coefficients as $b_2, b_3, \ldots, b_n$, the subscripts relating to their corresponding independent variables.

Upon determination of the values of these respective multiple regression coefficients, b's, and of the constant, a, a value for Y may be computed from equation (2) for any set of values of the independent variables. In developing an equation of this type for forecasting run-off from a particular drainage area, values of a and b are determined from an analysis of pertinent climatological and run-off records. The analysis is based on the premise that the correct values of the constants are those which yield estimates of Y, the dependent variable, which are in closest agreement with the observed values for a period of record.

With this minimum deviation between estimated and observed values as a criterion, the values of the constants may be determined mathematically. In accordance with the method of least squares, the agreement between computed estimates, Y, and observed values, X_1, is closest when the sum of the squared deviations is a minimum, that is, when

$$\Sigma(X_1 - Y)^2$$

has a minimum value.

It can be shown by the calculus that the condition of least squares is satisfied for a two-variable linear relationship by the following normal equations:

$$\Sigma(X_1) = an + \Sigma(X_2)b \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

$$\Sigma(X_2X_1) = \Sigma(X_2)a + \Sigma(X_2^2)b \quad . \quad . \quad . \quad . \quad (4)$$

in which X_2 and X_1 represent observed values of the independent and dependent variables, respectively, and n is the number of observations.

The relationship between independent and dependent variables expressed by equations (3) and (4) may be expressed also in terms of deviations from the means (small x's) by the equations

$$\Sigma(x_2^2)b = \Sigma(x_2x_1) \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

$$a = M_{x_1} - M_{x_2}b \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (6)$$

The values of a and b may be determined from these two equations. Then these values substituted in eq. (1) will permit an estimate, Y, of the dependent variable (run-off) for any magnitude of the independent variable (precipitation).

In the solution of problems in multiple correlation, equations in the form of (5) and (6) are employed in which the several variables are expressed in terms of deviations from means.

For a three-variable problem (two independent and one dependent variables) the equations are

$$\left.\begin{aligned} \Sigma(x_2^2)b_2 + \Sigma(x_2x_3)b_3 &= \Sigma(x_1x_2) \\ \Sigma(x_2x_3)b_2 + \Sigma(x_3^2)b_3 &= \Sigma(x_1x_3) \end{aligned}\right\} \quad . \quad . \quad . \quad (7)$$

$$a = M_1 - b_2M_2 - b_3M_3 \quad . \quad . \quad . \quad . \quad . \quad . \quad (7a)$$

and for a four-variable problem they are

$$\left.\begin{aligned} \Sigma(x_2^2)b_2 + \Sigma(x_2x_3)b_3 + \Sigma(x_2x_4)b_4 &= \Sigma(x_1x_2) \\ \Sigma(x_2x_3)b_2 + \Sigma(x_3^2)b_3 + \Sigma(x_3x_4)b_4 &= \Sigma(x_1x_3) \\ \Sigma(x_2x_4)b_2 + \Sigma(x_3x_4)b_3 + \Sigma(x_4^2)b_4 &= \Sigma(x_1x_4) \end{aligned}\right\} \quad . \quad . \quad (8)$$

$$a = M_1 - b_2M_2 - b_3M_3 - b_4M_4 \ldots\ldots (8a)$$

Values of a and b determined by the solution of these equations are substituted in eq. (2) to yield a forecasting equation.

In order to evaluate the x terms of these equations, preparatory to solving for a and b, it would be necessary to compute the departures from the mean, X - M, for each variable involved, then to perform the indicated multiplications and additions to obtain the product sums. In practice, however, considerable time may be saved by computing the products and product sums of the observations (rather than of the departures from the means), then converting these product sums to the required departures from the means. The following relationships are used in this conversion, or correction, to departures from the means, for a four-variable problem (three independent and one dependent variables):

$$\left.\begin{aligned}
\Sigma(x_1x_2) &= \Sigma(X_1X_2) - n(M_1M_2) \\
\Sigma(x_2^2) &= \Sigma(X_2^2) - n(M_2^2) \\
\Sigma(x_1x_3) &= \Sigma(X_1X_3) - n(M_1M_3) \\
\Sigma(x_2x_3) &= \Sigma(X_2X_3) - n(M_2M_3) \\
\Sigma(x_3^2) &= \Sigma(X_3^2) - n(M_3^2) \\
\Sigma(x_1x_4) &= \Sigma(X_1X_4) - n(M_1M_4) \\
\Sigma(x_2x_4) &= \Sigma(X_2X_4) - n(M_3M_4) \\
\Sigma(x_3x_4) &= \Sigma(X_3X_4) - n(M_3N_4) \\
\Sigma(x_4^2) &= \Sigma(X_4^2) - n(M_4^2) \\
\Sigma(x_1^2) &= \Sigma(X_1^2) - n(M_1^2)
\end{aligned}\right\} \ldots (9)$$

(In equations (5) through (9), the small **x** should not be confused with the capital X.)

APPLICATION OF MULTIPLE LINEAR CORRELATION

The following example illustrates the development of an equation for forecasting on April 1 the April-through-July run-off, principally from melting snow. Three independent variables are included in the analysis, and are identified in Table 1 in which observations covering a period of years are arranged for initial computation.

Precipitation stations and snow courses represented in Table 1 are as follows:

Stations	Courses
Arrowrock	Graham Ranch
Atlanta	Galena
Hill City	Trinity Mountain
Idaho City	Soldier Summit R. S.
Obsidian	Atlanta Summit

The initial computations involve the multiplications of quantities in each row in the body of Table 1; the products and squares required are indicated by the symbol headings in Table 2. If desired, the extensions in Table 2 may be checked row by row and by column totals. In checking the column totals, the first block, Extensions with X_2, add horizontally to give the total of the Check Sum column. However, in totaling all remaining blocks horizontally it is necessary to pick up certain entries from previous blocks. In the Extensions with X_3, for example, the X_2X_3 product is missing and must be obtained from the preceding block when the check is made.

The column totals in Table 2 are entered in Table 3 in rows designated as Extensions with X_2, Extensions with X_3, etc. The "Corrections" are computed as indicated by equation (9). Thus the first correction, 1152.62 (row 2), is the product $n(M_2^2)$. Or as a desirable alternate, since $nM_2 = T_2$ (total of the X_2 column in Table 1), the product of T_2M_2 or 127.03 x 9.0736 = 1152.62, the same as before. The corrected value, 154.22, (1306.84 - 1152.62), or Extension with x_2, is the same value that would have been obtained had the departures from the mean in the x_2 column of Table 1 been computed, these departures squared and the squares totaled. To obtain the check total after the first three rows, it is necessary to bring down certain values from corresponding lines in rows above. In checking the Extensions with x_3, for

TABLE 1

RECORDS OF PRECIPITATION,
SNOW-WATER EQUIVALENTS, AND RUN-OFF, SOUTH
FORK BOISE RIVER ABOVE ANDERSON RANCH DAM, IDAHO

Water Year Ending September 30	October-January Precipitation (inches)	April 1 Snow Water (inches)	April-July Precipitation (inches)	April-July Natural Run-off (100,000 ac.ft.)	
	X_2	X_3	X_4	X_1	ΣX
1936	8.75	26.96	4.60	5.84	46.15
1937	4.10	17.26	3.53	2.91	27.80
1938	10.09	33.64	5.82	7.88	57.43
1939	8.51	14.40	2.24	3.14	28.29
1940	6.36	19.20	2.98	3.86	32.40
1941	8.18	15.58	7.76	3.52	35.04
1942	9.40	19.42	4.29	4.28	37.39
1943	17.71	37.66	4.94	10.36	70.67
1944	4.78	13.08	6.70	3.18	27.74
1945	6.77	18.88	7.00	4.05	36.70
1946	12.45	29.08	3.41	7.13	52.07
1947	12.08	21.74	4.14	4.90	42.86
1948	8.23	19.36	5.84	4.68	38.11
1949	9.62	26.66	2.67	5.17	44.12
Total	127.03	312.92	65.92	70.90	576.77
Mean	9.0736	22.3514	4.7086	5.0643	41.1979

example, the value 259.21 would be included in the horizontal addition to obtain 1143.04.

The extensions with x_2 (x_2^2, x_2x_3, x_2x_4, x_2x_1), with x_3 (x_3^2, x_3x_4, x_3x_1), and others in Table 3 are constituents of equation (8). For brevity the b coefficients and equal sign are to be omitted during computations. Also the repetitious products appearing in equation (8) have been omitted in computing the extensions in Table 3.

The extensions with x which were computed in Table 3 are carried forward to Table 4 as eq. (I) through (III) in the same order as in equation (8), but with the b coefficients and equal signs omitted. (The repeated products are here shown in parentheses. In subsequent computations the values in parentheses drop out as certain totals equal zero.) To illustrate, eq. I in its complete form would be:

$$(154.22)b_2 + (259.21)b_3 - (9.75)b_4 = 81.83$$

TABLE 2
EXTENSIONS OF DATA FROM TABLE 1
(Products and Product Sums)

	Extensions with X_2				Check Sum	Extensions with X_3			Check Sum	Extensions with X_4		Check Sum	Extensions with X_1	Check Sum
	X_2^2	X_2X_3	X_2X_4	X_2X_1	$X_2\Sigma X$	X_3^2	X_3X_4	X_3X_1	$X_3\Sigma X$	X_4^2	X_4X_1	$X_4\Sigma X$	X_1^2	$X_1\Sigma X$
1936	76.56	235.90	40.25	51.10	403.81	726.84	124.02	157.45	1244.20	21.16	26.86	212.29	34.11	269.52
1937	16.81	70.77	14.47	11.93	113.98	297.91	60.93	50.23	479.83	12.46	10.27	98.13	8.47	80.90
1938	101.81	339.43	58.72	79.51	579.47	1131.65	195.78	265.08	1931.95	33.87	45.86	334.24	62.09	452.55
1939	72.42	122.54	19.06	26.72	240.75	207.36	32.26	45.22	407.38	5.02	7.03	63.37	9.86	88.83
1940	40.45	122.11	18.95	24.55	206.06	368.64	57.22	74.11	622.08	8.88	11.50	96.55	14.90	125.06
1941	66.91	127.44	63.48	28.79	286.62	242.74	120.90	54.84	545.92	60.22	27.32	271.91	12.39	123.34
1942	88.36	182.55	40.33	40.23	351.47	377.14	83.31	83.12	726.11	18.40	18.36	160.40	18.32	160.03
1943	313.64	666.96	87.49	183.48	1251.57	1418.28	186.04	390.16	2661.43	24.40	51.18	349.11	107.33	732.14
1944	22.85	62.52	32.03	15.20	132.60	171.09	87.64	41.59	362.84	44.89	21.31	185.86	10.11	88.21
1945	45.83	127.82	47.39	27.42	248.46	356.45	132.16	76.46	692.90	49.00	28.35	256.90	16.40	148.64
1946	155.00	362.05	42.45	88.77	648.27	845.65	99.16	207.34	1514.20	11.63	24.31	177.56	50.84	371.26
1947	145.93	262.62	50.01	59.19	517.75	472.63	90.00	106.53	931.78	17.14	20.29	177.44	24.01	210.01
1948	67.73	159.33	48.06	38.52	313.64	374.81	113.06	90.60	737.81	34.11	27.33	222.56	21.90	178.35
1949	92.54	256.47	25.69	49.74	424.43	710.76	71.18	137.83	1176.24	7.13	13.80	117.80	26.73	228.10
Totals	1306.84	3098.51	588.38	725.15	5718.88	7701.95	1453.66	1780.56	14034.68	348.31	333.77	2724.12	417.46	3256.94

The procedure now is to determine the values of the b coefficients by simultaneous solution of the three equations at the top of Table 4. Since three b values are to be determined, three equations enter into the solution. The Doolittle method is illustrated. The prime equations, as I', are obtained by dividing each term in the preceding row by the first term of that row with its sign changed. Operations on equations (II) and (III) require certain multiplications which are indicated in the first column; e.g., (-1.680781)I, means that the multiplier taken from I' (I prime) is to be applied to each term in equation (I). The remaining operations are clear. If additional equations were present, as in a problem involving four unknowns, the multiplicands in the first column would be I, Σ_2, and Σ_3, and for n unknowns I, Σ_2,

TABLE 3
PRODUCT SUMS CORRECTED TO
DEPARTURES FROM THE MEAN, WITH CHECK

Read this table as follows: At the intersection of the Extensions with X_2 row and the X_2 column is the sum of the products X_2^2, 1306.84; at the intersection of the Extensions with X_2 row and the X_4 column is the sum of the products X_2X_4, 588.38, etc. However, at the intersections of, say, the Extensions with x_2 row and the X_4 column is the sum of the products x_2x_4, not x_2X_4. The Sums and Means are from Table 1.

	X_2 or x_2	X_3 or x_3	X_4 or x_4	X_1 or x_1	Σ
Sums	127.03	312.92	65.92	70.90	576.77
Means	9.0736	22.3514	4.7086	5.0643	41.1979
Extensions with X_2	1306.84	3098.51	588.38	725.15	5718.88
Corrections	1152.62	2839.30	598.13	643.32	5233.37
Extensions with x_2	154.22	259.21	-9.75	81.83	485.51
Extensions with X_3		7701.95	1453.66	1780.56	14034.68
Corrections		6994.20	1473.42	1584.72	12891.64
Extensions with x_3		707.75	-19.76	195.84	1143.04
Extensions with X_4			348.31	333.77	2724.12
Corrections			310.39	333.84	2715.77
Extensions with x_4			37.92	-0.07	8.34
Extensions with X_1				417.46	3256.94
Corrections				359.06	2920.93
Extensions with x_1				58.40	336.00

$\Sigma_3, \ldots, \Sigma_{n-1}$.

The back solution is obtained as follows: Quantities in the top row of the back solution are obtained from the X_1 column and prime rows of Table 4. Signs are changed. The quantity at the right in the second row is brought down from the top row; other quantities are obtained by multiplying this right hand member by values in the X_4 column and prime rows. In the remaining rows the right hand members are obtained by addition, and other members are obtained by multiplying the right hand members by values in successive columns progressing to the left. Entries in Table 5 are as follows: Column 1, the coefficients of multiple regression from Table 4 back solution; column 2, the values of eq. III from Table 4; column 4, the x_1 extensions from Table 3; and column 6, the means of the original variables, shown in Table 1. Op-

TABLE 4
EVALUATION OF MULTIPLE REGRESSION COEFFICIENTS
(Doolittle Solution)

	X_2	X_3	X_4	X_1	ΣX
Eq. I	154.22	259.21	-9.75	81.83	485.51
Eq. II	(259.21)	707.75	-19.76	195.84	1143.04
Eq. III	(-9.75)	(-19.76)	37.92	-0.07	8.34
I	154.22	259.21	-9.75	81.83	485.51
I'	-1.000000	-1.680781	0.063221	-0.530606	-3.148166
II	(259.21)	707.75	-19.76	195.84	1143.04
(-1.680781) I	(-259.21)	-435.675	16.388	-137.538	-816.036
Σ_2	0	272.075	-3.372	58.302	327.005
II'		-1.000000	0.012394	-0.214287	-1.201893
III		(-19.76)	37.92	-0.07	8.34
(0.063221) I		16.388	-0.616	5.173	30.694
(0.012394) Σ_2		3.372	-0.042	0.723	4.053
Σ_3		0	37.262	5.826	43.088
III'			-1.000000	-0.156352	-1.156352

Back Solution

0.530606	0.214287	0.156352
0.009885	0.001938	0.156352
-0.363427	0.216225	
0.177064		

erations are as indicated. The agreement in the figures at the foot of columns 2 and 3 provides a check on computations in Table 4.

The computations in Table 5 for determining $\bar{R}$ and $\bar{S}$ are based on the following formulas:

$$\bar{S}^2_{1\cdot 234\ldots n} = \frac{\Sigma(x_1^2) - \left[b_{12\cdot 34\ldots n}(\Sigma x_1 x_2) + b_{13\cdot 24\ldots n}(\Sigma x_1 x_3) + \ldots + b_{1n\cdot 23\ldots(n-1)}(\Sigma x_1 x_n)\right]}{n - m} \quad \ldots\ldots (10)$$

$$\left.\begin{aligned} R^2_{1\cdot 234\ldots n} &= \frac{\left[b_{12\cdot 34\ldots n}(\Sigma x_1 x_2) + b_{13\cdot 24\ldots n}(\Sigma x_1 x_3) + \ldots + b_{1n\cdot 23\ldots(n-1)}(\Sigma x_1 x_n)\right]}{\Sigma(x_1^2)} \\ \bar{R}^2_{1\cdot 234\ldots n} &= 1 - \left[(1 - R^2_{1\cdot 234\ldots n})\frac{(n-1)}{(n-m)}\right] \end{aligned}\right\} \quad \ldots\ldots\ldots (11)$$

TABLE 5
CHECK ON COMPUTATION OF REGRESSION COEFFICIENTS, AND SOLUTIONS FOR $\bar{R}$, $\bar{S}$, AND a

	1	2	3	4	5	6	7
	(Reg. Coef.)	(Eq. 3 from Table 4)	(1·2)	(X_1 from Table 3)	(1·4)	(Means from Table 1)	(1·6)
X_2	0.1771	-9.75	-1.73	81.83	14.492	9.074	1.607
X_3	0.2162	-19.76	-4.27	195.84	42.341	22.351	4.832
X_4	0.1564	37.92	5.93	-0.07	-0.011	4.709	0.736
		-0.07	-0.07	58.40	56.822	5.064	7.175

$$R^2 = \frac{56.82}{58.40} = 0.973 \qquad R = \sqrt{0.973} = 0.986$$

$$\bar{R}^2 = 1 - (1-R^2)\frac{n-1}{n-m} = 1 - (1-0.973)\frac{13}{10} = 0.965 \qquad \bar{R} = \sqrt{0.965} = 0.982$$

$$\bar{S}^2 = \frac{58.40 - 56.82}{n - m} = \frac{1.578}{10} = 0.1578 \qquad \bar{S} = \sqrt{0.1578} = 0.3972 = 39,720 \text{ acre-feet}$$

$$a = 5.064 - 7.175 = -2.111$$

Multiple regression equation:

$$Y = 0.177X_2 + 0.216X_3 + 0.156X_4 - 2.111$$

$\bar{R}$ may be computed more directly, however, with the equation

$$\bar{R}^2 = 1 - \frac{(n-1)\,\bar{S}^2}{\Sigma x_1^2} \quad \ldots\ldots\ldots (11a)$$

The value of the constant a is determined by the equation

$$a_{1\cdot 234\ldots n} = M_1 - \left[b_{12\cdot 34\ldots n} M_2 + b_{13\cdot 24\ldots n} M_3 + b_{1n\cdot 23\ldots(n-1)} M_n \right] \quad \ldots\ldots (12)$$

Substitution of the a and b constants in eq. (2) results in the following forecasting equation for the South Fork of the Boise River at Anderson Ranch Dam:

$$Y = 0.177X_2 + 0.216X_3 + 0.156X_4 - 2.111 \quad \ldots\ldots\ldots (13)$$

where Y is in 100,000 acre-foot units. This equation is a summarizing expression of the observed data.

To test the correctness of this equation, it may be applied to climatological and run-off data for the period of record, and the agreement noted between estimated run-off, Y, and observed run-off, X_1.

A visual comparison of estimates and observations may be made from a plotting of estimated run-off against values of observed run-off as in figure 1.

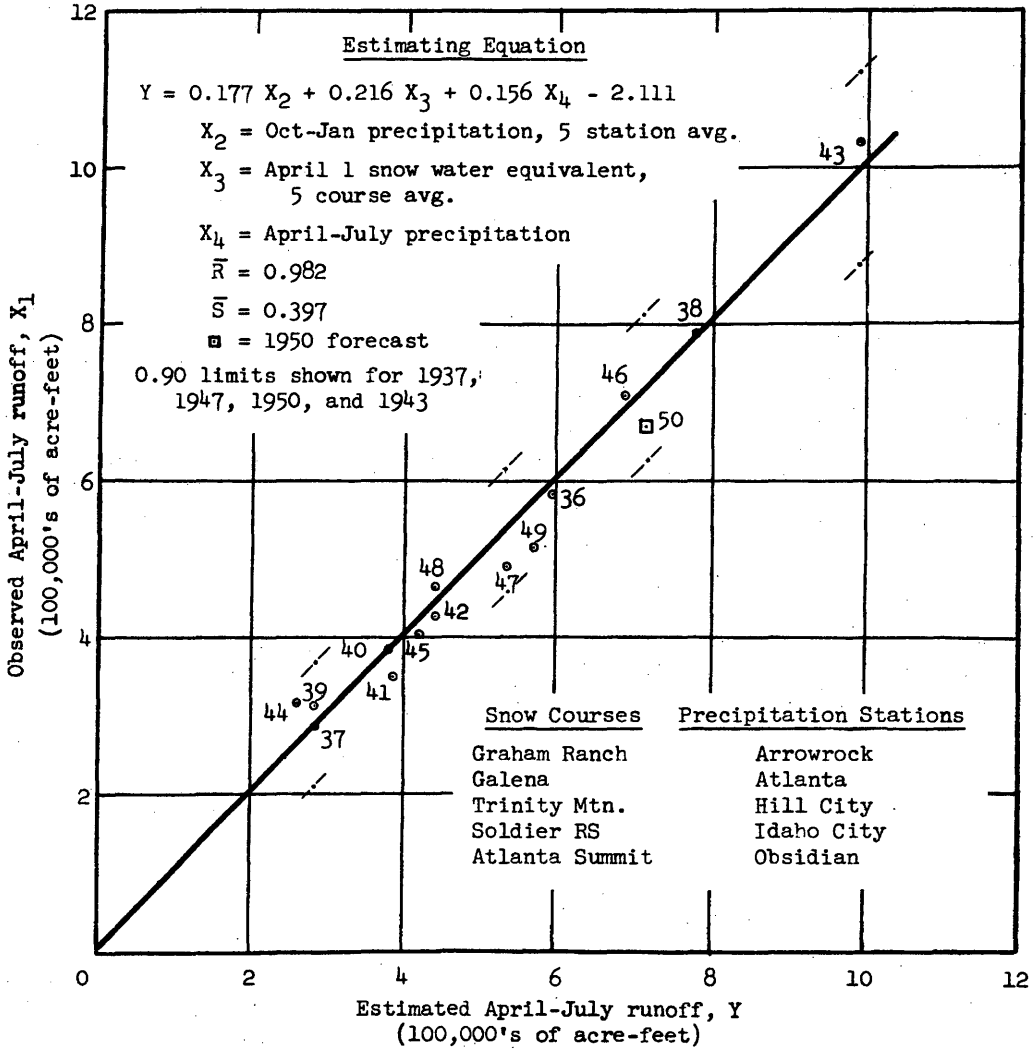


Figure 1 - Correlation of Observed vs. Estimated April-July Run-off, South Fork Boise River.

TABLE 6

AREAS UNDER THE NORMAL PROBABILITY CURVE

(From the mean to distances $\frac{x}{\sigma}$ from the mean, expressed as decimal fractions of the total area, 1.0000.)

The proportional part of the curve included between an ordinate erected at the mean and an ordinate erected at any given value on the X axis can be read from the table by determining x (the deviation of the given value from the mean) and computing $\frac{x}{\sigma}$. Thus if $M_x = 25.00$, $\sigma = 4.00$, and it is desired to ascertain the proportion of the area under the curve between ordinates erected at the mean and at 20.00; $x = 5.00$ and $\frac{x}{\sigma} = \frac{5.00}{4.00} = 1.25$. From the table it is found that .3944, or 39.44 percent, of the entire area is included.

$\frac{x}{\sigma}$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.49865	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.49903	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993

From Rugg's Statistical Methods Applied to Education, reprinted by arrangement with the publishers, Houghton Mifflin Company.

Data on run-off season precipitation (X_4 in eq. 13), are not available as of the date of forecasting. For convenience, an equation which will yield a "mean" forecast, subject to a calculable plus or minus error, may be obtained from eq. (13) by assuming normal precipitation during the run-off season. That is, by adding the product b_4M_4 to the constant of eq. (13), a new constant term is obtained, which in this example equals (0.156) (4.709) - 2.111, or -1.376, and the equation becomes:

$$Y = 0.177X_2 + 0.216X_3 - 1.376 \ldots (13a)$$

As was previously mentioned, the a and b constants determined by multiple correlation, when employed in a forecasting equation such as eq. (2), yield estimates Y which are in closest agreement with observed values, X_1, in accordance with the theory of least squares. Owing to inaccuracies in records, inadequacies of data, etc., differences between estimates and observations are to be expected. Consideration must be given to the magnitude of these errors so that the investigator may know how closely the observed values of the dependent variable may be expected to approximate estimated values in future forecasts.

Estimates of errors of forecast which may be expected with given probability, taking into consideration the length of record, number of variables used, and the variation in run-off season precipitation will be treated in the following sections.

Multiple Correlation Coefficient, Standard Deviation, and Standard Error of Estimate

The coefficient of multiple correlation, R, (0.986 in this problem), is a measure of the strength of the relationship in the sample (14 years of record) between the independent and dependent variables. A value of R equal to zero indicates no relationship; an R of one indicates perfect correlation. Whether a relationship between any particular independent variable and the dependent variable is direct or inverse is determined by the sign of the b coefficient of that variable.

The adjusted value of the coefficient of multiple correlation, $\bar{R} = 0.982$, was obtained by correcting for length of record, n, and for the number of constants, m (a and b's), which were determined (same as number of variables in the type of equation used). In computing $\bar{R}$, Table 5, the expression n-m represents the degrees of freedom. The $\bar{R}$ is an estimate of the true correlation which probably exists in a universe of such data, as distinguished from the sample. It will be seen that an increase in the number of variables has the effect of reducing the estimate of the correlation in the universe. $\bar{R}^2$ is the coefficient of determination, and expresses the portion of the variance in the dependent variable which has been explained.

The standard deviation, σ, is a measure of dispersion, or scatter, of a series of items from the arithmetic mean. For a distribution of errors of estimate about a regression line the measure of dispersion is known as the standard error of estimate, S. A bar over the symbol indicates that this parameter has been adjusted for years of record and number of variables. The use of the adjusted value, $\bar{S}$, as well as $\bar{R}$, is particularly important in multiple correlation, since in such analyses there is a tendency for S to be less for the sample than the true value for the universe.

The parameter, $\bar{S}$, is expressed in units of the dependent variable and its magnitude is such that a range of plus or minus one standard error of estimate embraces about 68 percent of the residuals or differences between estimated and observed values. A range of plus or minus two standard errors of estimate embraces about 95 percent of all residuals. About 90 percent of the residuals could be expected to lie within a range of plus or minus $1.645\bar{S}$. This range is limited by the 5 and 95 percentiles, the lower value of which could be expected to be equalled or exceeded 95 percent of the time; the upper value, 5 percent of the time. The probabilities mentioned above are obtained from tables of areas under the normal probability curve (see Table 6). The ratio, $\frac{x}{\sigma}$, shown above as 1.645 varies for small samples (under 20), and appropriate ratios for given sample sizes may be conveniently obtained from Ezekiel's Fig. A, Appendix 3, or from Students t Table.

The standard error of the individual forecast, $\sigma_{y \cdot x_1}$, will be introduced in a later section.

The $\bar{R}$ of 0.982 indicates that a highly significant relationship exists, and that

about 96.5 percent ($\overline{R}^2$) of the variability in run-off has been accounted for; in this particular drainage basin the factors contributing to variations in run-off have been quite well accounted for. The $\overline{S}$ as determined above will be used in determining the precision of the b coefficients (their standard deviations), and in estimating the error of the individual forecast.

Standard Deviation of Regression Coefficients

The regression coefficients as determined from Tables 1 through 5 were based on a 14-year record, 1936-1949, inclusive. Had longer records been available it is possible that the values of these coefficients would have differed somewhat from those obtained. Although the true values of the regression coefficients which would be obtained from a compilation of data for all time cannot be estimated, an approximation of the limits within which, for given odds, the true values might fall, can be based on the standard deviation of the computed regression coefficients. That is, the odds are about 68 in 100 that the true value of b for a compilation of data for all time would lie within plus or minus one standard deviation of that obtained from the available data. These odds increase for multiples of the standard deviation of the regression coefficients as was previously discussed in connection with the standard error of estimate, S.

Computations for determining the standard deviations of the regression coefficients are shown in Table 7. This involves the simultaneous solution of equations of the following form, known as the covariance matrix, to determine the c values for a 4-variable problem:

$$\left.\begin{aligned} \Sigma(x_2^2)\, c_{22} + \Sigma(x_2x_3)\, c_{23} + \Sigma(x_2x_4)\, c_{24} &= 1 \\ \Sigma(x_2x_3)\, c_{22} + \Sigma(x_3^2)\, c_{23} + \Sigma(x_3x_4)\, c_{24} &= 0 \\ \Sigma(x_2x_4)\, c_{22} + \Sigma(x_3x_4)\, c_{23} + \Sigma(x_4^2)\, c_{24} &= 0 \end{aligned}\right\} \quad . . (14)$$

[3] Ibid.

These are solved to obtain values for c_{22}, c_{23}, and c_{24}. Similar equations are set up with c_{32}, c_{33}, and c_{34} as unknowns, with 0, 1, 0 on the right of the equal sign; and finally with c_{42}, c_{43}, and c_{44} as unknowns with 0, 0, 1 on the right of the equal sign. Having determined the values of c_{22}, c_{33}, and c_{44}, the standard deviations may be computed by substitution in equations of the form

$$\sigma_{b_2} = \overline{S}\sqrt{c_{22}} \quad . . . . . \quad (15)$$

as shown in Table 8, and in which σ_{b_2} is the standard error of the regression coefficient, b_2.

Inasmuch as the solution of eq. (14) is parallel to the solution shown in Table 4 (both involve the Doolittle method), the two are usually combined and solved in one operation. Table 7 illustrates such a combined solution.

The multiple regression coefficients with their standard deviations are as follows (from Table 8):

Independent Variable	Coefficient and Standard Deviation
X_2	$b_2 = 0.177 \pm 0.052$
X_3	$b_3 = 0.216 \pm 0.024$
X_4	$b_4 = 0.156 \pm 0.065$

where X_2 is October-January precipitation, X_3 is the April 1 snow survey, and X_4 is April-July precipitation. Considered in relation to the values of the coefficients, these standard deviations indicate significance of all factors. A coefficient may be considered significant if its ratio to the standard error is 2 or more. This ratio, 2, is of course an arbitrary selection.

It will be noted that b_3 can be determined with the greatest precision of the three coefficients. Its value will lie between 0.240 (0.216 + 0.024) and 0.192 (0.216 - 0.024) about 68 percent of the time. Precipitation during the flood run-off season, April-July, contributes relatively less to the variation in run-off in this drainage than in most other basins studied.

TABLE 7
COMBINED SOLUTION FOR EVALUATING THE b COEFFICIENTS AND c-MATRIX

	X_2	X_3	X_4	X_1	c_2	c_3	c_4	Σ (X+c)
Eq. 1	154.22	259.21	-9.75	81.83	1	0	0	486.51
Eq. 2	(259.21)	707.75	-19.76	195.84	0	1	0	1144.04
Eq. 3	(-9.75)	(-19.76)	37.92	-0.07	0	0	1	9.34
I	154.22	259.21	-9.75	81.83	1			486.51
I'	-1.000000	-1.680781	0.063221	-0.530606	-0.006484			-3.154650
II	(259.21)	707.75	-19.76	195.84		1		1144.04
(-1.680781) I	(-259.21)	-435.675	16.388	-137.538	-1.680781			-817.717
Σ_2	0	272.075	-3.372	58.302	-1.680781	1		326.324
II'		-1.000000	0.012394	-0.214287	0.006178	-0.003675		-1.199390
III		(-19.76)	37.92	-0.07			1	9.34
(0.063221) I		16.388	-0.616	5.173	0.063221			30.758
(0.012394) Σ_2		3.372	-0.042	0.723	-0.020836	0.012394		4.044
Σ_3		0	37.262	5.826	0.042385	0.012394	1	44.142
III'			-1.000000	-0.156352	-0.001137	-0.000333	-0.026837	-1.184638

Back Solution on b's (as in Table 4)

0.530606	0.214287	0.156352
0.009885	0.001938	0.156352
-0.363427	0.216225	
0.177064		

The back solution on c values is shown in Table 8.

TABLE 8
BACK SOLUTION ON c VALUES

Back Solution on c_2

c_{22}	c_{23}	c_{24}	Eq. 3	Check
0.006484	-0.006178	0.001137		
0.000072	0.000014	0.001137	37.92	0.043
0.010360	-0.006164		-19.76	0.122
0.016923			-9.75	-0.165
			0	0

Back Solution on c_3

c_{32}	c_{33}	c_{34}	Eq. 3	Check
	0.003675	0.000333		
0.000021	0.000004	0.000333	37.92	0.012
-0.006184	0.00379		-19.76	-0.072
-0.006163			-9.75	0.060
			0	0

Back Solution on c_4

c_{42}	c_{43}	c_{44}	Eq. 3	Check
		0.026837		
0.001697	0.000333	0.026837	37.92	1.018
-0.000560	0.000333		-19.76	-0.007
0.001137			-9.75	-0.011
			1	1.000

$\sigma_{b_2} = \bar{S}\sqrt{c_{22}}$ $\sigma_{b_3} = \bar{S}\sqrt{c_{33}}$ $\sigma_{b_4} = \bar{S}\sqrt{c_{44}}$

$= 0.397\sqrt{0.0169}$ $= 0.397\sqrt{0.00368}$ $= 0.397\sqrt{0.0268}$

$= 0.052$ $= 0.024$ $= 0.065$

$b_2 = 0.177 \pm 0.052$ $b_3 = 0.216 \pm 0.024$ $b_4 = 0.156 \pm 0.065$

Note that $c_{23} = c_{32}$, $c_{24} = c_{42}$, and in general $c_{ab} = c_{ba}$. This agreement provides a critical check on c value computations.

The coefficients of multiple regression in eq. (13) provide information as to the average contribution of the various factors to the variation in run-off from the South Fork of the Boise River. For a change of 1 inch of precipitation accumulated during the months of October through January, an average change of 17,700 acre-feet of run-off occurs during the subsequent flood period, April through July. Each inch of snow-water equivalent as of April 1 contributes an average of 21,600 acre-feet to the flood season run-off; each inch of precipitation during the flood season contributes 15,600 acre-feet. The constant term, a, is not to be interpreted as having any hydrologic significance.

The run-off volume indicated by the estimating equation is subject to error of forecast; a statement regarding the error to be expected with given probability is an integral part of the forecast. Methods of computing these limits are discussed in the following section.

Limits of Error Related to Run-off Season Precipitation

In regions where the run-off season precipitation is an important factor contributing to variations in seasonal run-off, as is often the case, the effect of this precipitation is related to the error of forecast to be expected with any specified probability.

Since the April-July precipitation is unknown on April 1, the date of run-off forecasting, certain values of this unknown will be assumed as normal or as departures from normal to be expected with given probability. Substitution of the normal, or average, precipitation would provide a mean forecast. Other values for this unknown independent variable may be assumed and the corresponding limits of error computed. For example, it may be desired to estimate the magnitude of the error of forecast which will be equalled or exceeded 5 percent of the time in a positive direction (or 5 percent of the time in a negative direction, as interests may dictate). The probability that the error of forecast will lie within these limits (5 to 95 percentile range) is 0.90.

The desired limits of error of forecast should include the residual inherent in the correlation along with that which would result from a departure from mean run-off season precipitation. The inclusion of run-off season precipitation as an independent variable, although important in the analysis, presents a unique problem in determining the limits of error of the forecast. A method of estimating the combined error of forecast will be treated in the following section.

Reliability of the Individual Forecast

The accuracy of a forecast can be expected to be greatest for normal values of the several variables. Errors of forecast increase as the values of the variables depart in either direction from their respective means. This is a result of combined errors in the mean and in the slope of the regression line, or plane. Errors in the mean displace the regression plane vertically, while errors in slope, with elevation determined by the mean, tend to widen the range above and below the mean.

The standard error of an individual estimate is expressed by the basic equation:

$$\sigma_{(Y - X_1)} = \left[\overline{S}^2 + \frac{\overline{S}^2}{n} + \sigma^2_{b_2}(x_2^2) + \sigma^2_{b_3}(x_3^2) + \ldots + \sigma^2_{b_n}(x_n^2)\right]^{1/2} \quad \text{(Approx.)} \quad \ldots (16)$$

Eq. (16) is applicable where all independent variables are known (or predicted). For forecasting computations the variance expressed by eq. (16) may be combined with that which would result from an assumed departure of one standard deviation in run-off season precipitation; and the departure to be expected with 0.90 probability expressed by the following general equation:

$$(Y-X_1)_{0.90} = 1.645\left[\sigma^2_{x_u}(b_u^2) + \sigma^2_{b_u}(\sigma^2_{x_u}) + \overline{S}^2 + \frac{\overline{S}^2}{n} + \sigma^2_{b_2}(x_2^2) + \sigma^2_{b_3}(x_3^2) + \ldots + \sigma^2_{b_n}(x_n^2)\right]^{1/2} \quad \text{(Approx.)} \quad \ldots (17)$$

where

$(Y - X_1)_{0.90}$ = error of individual forecast which would be exceeded on the average one time in ten in either a plus or minus sense,

$\overline{S}$ = standard error of estimate for an equation such as (13) in which run-off season precipitation was included in its derivation,

$\sigma_{b_2}, \sigma_{b_3}, \ldots, \sigma_{b_n}$ = standard errors of multiple regression coefficient of known variables for an equation including run-off season precipitation,

$x_2, x_3, \ldots, x_n$ = departures from mean for known variables,

σ_{x_u} = standard deviation of run-off season precipitation (u for unknown),

σ_{b_u} = standard error of regression coefficient for run-off season precipitation, and

b_u = regression coefficient of run-off season precipitation.

Note that run-off season precipitation, x_u, is treated in the first two terms of this equation, and is not included in the series. The standard deviation of x_u, may be determined by the general expression

$$\sigma_x = \sqrt{\frac{\Sigma x^2}{n-1}} \quad \ldots\ldots\ldots\ldots (18)$$

An equivalent expression for the standard deviation is

$$\sigma_x = \sqrt{\frac{\Sigma X^2 - nM_x^2}{n-1}} \quad \ldots\ldots\ldots (18a)$$

Eq. (18a) obviates the need for computing the departures from the mean, X - M = x, for each year of record. Substituting in eq. (18) or (18a) the standard deviation of X_4 for the South Fork of the Boise River problem is

$$\sqrt{\frac{37.92}{13}}, \text{ or } +1.708 \text{ inches.}$$

(In most basins fairly long records of precipitation are available for determining the standard deviation of this variable.) About 68 percent of the time the April-July precipitation, X_4, may be expected to lie within a range 1.708 inches above or below the mean value, i.e., within a range of 4.71 ± 1.708. (A distribution of total precipitation for a period of months, such as April through July, usually shows no appreciable departure from the normal. Therefore, asymmetric distribution will not be considered in this discussion.)

The error of forecast for given probability will vary from year to year depending upon the magnitude of departures from normal of the various independent variables. The error must, then, be computed for each forecast. To facilitate the computations the terms of equation (17) may be considered in two groups. Group one, the first four terms within the brackets, contains quantities which remain constant from year to year. Group two, embracing the remaining terms within the bracket, contains terms which have to be evaluated for each year. The terms of group one are evaluated below, for future reference, for the South Fork Boise River problem:

$$\overline{S}^2 = \quad \ldots \quad = 0.1578$$

$$\frac{\overline{S}^2}{n} = \frac{0.1578}{14} = 0.0113$$

$$\sigma_{b_4}^2 (\sigma_{x_4}^2) = (0.065)^2 (1.708)^2 = 0.0123$$

$$b_4^2 (\sigma_{x_4}^2) = (0.156)^2 (1.708)^2 = \underline{0.0710}$$

$$\text{Total} = 0.2524$$

Application of eq. (17) will be illustrated, using data for the 1950 flood season, South Fork Boise River Basin.

Computations for the April 1, 1950 Forecast, with 0.90 Limits of Error

Pertinent records of precipitation and snow-water equivalents for the April 1, 1950 forecast of April-July run-off are:

X_2 (observed October-through-January accumulated precipitation, average of five stations) = 10.44 inches;

X_3 (observed April 1 snow-water equivalent, average of five courses) = 31.00 inches.

A mean forecast, obtained by substituting the above values in eq. (13a) is as follows:

$0.177 \times 10.44 = 1.87$

$0.216 \times 31.00 = 6.70$

Constant term, a = -1.38

Total = 7.19 (100,000 acre-foot units)

The 0.90 probability limits of error are computed as follows:

Values required for substitution in eq. (17) are:

$x_2 = X_2 - M_2 = 10.44 - 9.07$

$= 1.37,$

$x_3 = X_3 - M_3 = 31.00 - 22.35$

$= 8.65,$

$\sigma_{b_2} = 0.052,$

$\sigma_{b_3} = 0.024,$

$\sigma_{b_4} = 0.065,$

$\overline{S}^2 = 0.1578,$

$\sigma_{x_4} = 1.708,$

where M_2 and M_3 are from Table 1; σ_{b_2}, σ_{b_3}, and σ_{b_4} are from Table 8; and $\overline{S}^2$ is from Table 5. The 0.90 limits of error for the 1950 forecast are

$$(Y-X_1)_{0.90} = 1.645 \left[0.2524^* + \sigma_{b_2}^2 (x_2^2) + \sigma_{b_3}^2 (x_3^2)\right]^{1/2}$$

$$= 1.645 \left[0.2524 + (0.052)^2(1.37)^2 + (0.024)^2(8.65)^2\right]^{1/2}$$

$$= \pm 0.900 \text{ (100,000 ac.ft. units)}$$

The anticipated April-July run-off for 1950 would be expected, with 0.90 probability, to fall between 7.19 + 0.90 = 8.09, and 7.19 - 0.90 = 6.29. Or, expressed in another way, for 1950 conditions, the run-off would be expected to equal or exceed 8.09 about 5 percent of the time and to equal or exceed 6.29 about 95 percent of the time, all in 100,000 acre-foot units.

The 1950 forecast is plotted against observed run-off in fig. 1. Also shown on fig. 1 are the 0.90 probability limits for a normal year, 1947; for the high year, 1943; and for a low year, 1937. All are based on eq. (17).

Inspection of eq. (17) will make clear certain characteristics of limits of error of the forecast. For example, the relative departures x of the respective independent variables from their means, will vary from year to year, indicating a different magnitude of error for each year, even for years of equal run-off. Also, it is apparent that as these departures (x's) increase, either in a plus or minus direction, the error of forecast may be expected to increase. Furthermore, an increase in the standard deviation of the regression coefficients (σ_b values), indicates an increase in the error of forecast. This will explain a need for computing the standard error of the regression coefficients, and for eliminating any non-significant factors.

*This term combines the $\overline{S}^2$, $\frac{\overline{S}^2}{n}$, $b_u^2 (\sigma_{x_u}^2)$, $\sigma_{b_u}^2 (\sigma_{x_u}^2)$ terms of equation (17) as discussed in the preceding section.

CHARACTERISTICS OF LIMITS OF ERROR FOR TWO TYPES OF FORECASTING EQUATION

Eq. (13a), described in the preceding pages, was obtained from a basic eq. (13) in which the run-off season precipitation was considered.

Instead of eq. (13a), a biased forecasting equation could have been derived by means of a multiple correlation involving only the X_2 and X_3 variables in relation to run-off, thus neglecting the effect of the X_4 variable and its influence on the remaining regression coefficients. Such an equation for the South Fork of the Boise River is as follows (computations not shown):

$$Y = 0.170\,X_2 + 0.214\,X_3 - 1.273$$
(biased equation)

Owing to the high correlation between run-off and related factors in the South Fork of the Boise River Basin and to the relatively small effect and low significance of run-off season precipitation in this area the two equations do not differ significantly; however, a comparison of results will illustrate certain typical characteristics of the two equations. Of particular interest is a comparison of errors of forecast. (The omission of an important independent variable would introduce bias in the results, and for this reason it is preferable to avoid this practice. However, the practice of neglecting the effect of run-off season precipitation is a possible approach, and the results are worth considering).

Following are the multiple regression coefficients and the standard errors of these coefficients, for eq. (13a) and for the biased equation:

	Equation (13a)	Biased Equation
b_2	0.177 ± 0.052	0.170 ± 0.062
b_3	0.216 ± 0.024	0.214 ± 0.029

For the biased equation the standard error of estimate, $\bar{S}$, equals 0.475. The variance in the estimate for this equation, in which all independent variables are known, would be expressed by eq. (16). Limits of error for 0.90 probability for the 1950 season are:

$$1.645\left[0.2261 + \frac{0.2261}{14} + (0.062)^2(1.37)^2 + (0.029)^2(8.65)^2\right]^{1/2}$$
$$= 0.919$$

The following is a comparison of 0.90 limits of error for the two equations computed for years of different run-off potentialities (in 100,000 acre-foot units):

	Equation (13a)	Biased Equation
1943 (high year)	1.27	1.39
1950 (forecast year)	0.90	0.92
1947 (average year)	0.87	0.87
1937 (low year)	0.95	0.97

It is important to note that although the $\bar{S}$ for eq. (13a) is about the same in this problem as that for the biased equation (0.477 as compared with 0.475), the error of forecast is less for eq. (13a). This is the result of the smaller errors in the regression coefficients of eq. (13a). For the same reason the errors of forecast increase less rapidly with departures from normal conditions for eq. (13a).

For the above comparison the standard error of estimate for eq. (13a) was computed from the expression

$$\bar{S} = \sqrt{\frac{n\sigma_Z^2}{n-m}} \quad \ldots\ldots\ldots\ldots\ldots (19)$$

where Z is the residual, or difference between estimated and observed run-off. The σ_Z is expressed by

$$\sigma_Z^2 = \frac{\Sigma Z^2}{n}$$

Substituting $\frac{\Sigma Z^2}{n}$ in eq. (19) gives

$$\bar{S} = \sqrt{\frac{\Sigma Z^2}{n - m}}$$

In this comparison, equations (13a) and the biased equation do not differ significantly. In regions where the run-off season precipitation is a more important factor, the differences in errors of forecast for the two equations becomes more pronounced, and in abnormal years the errors are greater for the biased equation. An additional comparison of two such equations involving Colorado River data is contained in a later section.

If the values of the X_4 variable were known (or predicted) then estimates of run-off based on eq. (13) would be subject to errors of forecast determined by eq. (16). These errors, for 0.90 probability, would be as follows for a high, an average, and a low year (in 100,000 acre-foot units):

Error for high year (1943) = $\pm$ 1.17

Error for average year (1947) = $\pm$ 0.73

Error for low year (1937) = $\pm$ 0.83

PRACTICAL CONSIDERATIONS PERTAINING TO CORRELATION ANALYSES

Precision of Multiple Regression Coefficients

The precision with which a particular multiple regression coefficient can be determined in any analysis is a function of parameters among which are the magnitude of the standard error of estimate, and of the correlation between that variable and the remaining independent variables. This relationship for b_2 is expressed by the following formula from page 322 of Methods of Correlation Analysis, by Ezekiel:

$$\sigma_{b_2} = \sqrt{\frac{\bar{S}^2_{1\cdot 234\ldots n}}{n\sigma_2^2 (1-R^2_{2\cdot 34\ldots n})}} \quad \ldots . (20)$$

Significance of the b_2 coefficient increases (its standard deviation decreases) as $\bar{S}$ becomes smaller and as the indicated correlation between independent variables decreases. This requirement for a low value of $\bar{S}$ would indicate the need for considering all major factors influencing stream flow, and will explain the presence of precipitation during the flood season in the analysis for the derivation of forecasting equations in regions where this factor is an important contributor to variations in flood season run-off. Short periods of record accentuate the importance of close relationships between estimated and observed events, as expressed by the adjusted coefficient of multiple correlation, $\bar{R}$, and adjusted standard error of estimate, $\bar{S}$.

Significance of Difference Between Correlation Coefficients

In the selection of the better of two or more forecasting procedures, a mechanical procedure would consist of trying several combinations of independent variables affecting run-off and of selecting the best combination as determined by the highest coefficient of multiple correlation. As a basic analytical procedure such an empirical, trial-and-error method is to be discouraged. (A short-cut procedure for screening variables is described later.)

Certain reservations are in order in following out this empirical procedure. That is, the difference in the correlation coefficients, R, must be significant if the conclusion as to the best combination is to have a real meaning. For if the difference in R's is subject to chance occurrence, then the selection of variables based on

differences in R's is likewise a matter of chance. Some examples of the significance of the difference between two R's when both values are in a high range (R = 0.90 or over) as compared with the significance of the difference between two R's in a lower range (R = 0.80 or lower) will provide a basis for judgment in interpreting results of analyses.

Consider first the range below R = 0.80. For example, coefficients of 0.80, or even 0.55, are significant for n = 16 and m = 3 (13 degrees of freedom). However, the difference between these two values of R is significant only at the 22-percent level, which means that so large a difference from chance causes may be expected about 22 times in 100. This difference would not be considered significant. By way of comparison it is important to note that for the higher values of R the difference between an R of 0.91 and an R of 0.98 is more significant than the far greater difference between an R of 0.80 and an R of 0.55; it is in fact significant at the 5-percent level. This further illustrates the importance of high correlations to assure realistic conclusions. For lengths of hydrologic records usually available, say 20 years, it is only when values of R are in the 0.90's that significance may be attributed to differences between two significant correlations. Inclusion of flood season precipitation often contributes appreciably to this higher correlation, and, in fact, leads to the difference between significance and nonsignificance in the differences between two R's. Insignificant differences could be a matter of chance, and the apparent order of importance in such cases could change with added years of record. The requirement for significance of the difference should be kept in mind in selecting independent variables in simple or multiple correlations.[4]

Concerning the empirical selection of variables, it should be remembered that if the several variables are selected by considering a large number of possible independent variables, and by retaining only those which show the highest correlation with the dependent variable, there is a large possibility of the correlation in the sample exceeding the true correlation in the universe. If error calculations are to be used in judging the significance of the correlations, the variables must be selected on a logical basis.

Number of Variables

Although the inclusion of major factors contributing to variation in the dependent variable is important, it is equally important that only pertinent variables be included in the analysis, keeping the number of variables as low as practicable consistent with efficient use of data. The effect of increased numbers of variables on the apparent correlation, and the need for adjusting the coefficient of multiple correlation and the standard error of estimate as required to correct for numbers of variables and length of record, may be impressively demonstrated by correlating random, unrelated variables and noting the increase in the value of R and the decrease in S with increased numbers of these unrelated variables. Results of such an example are tabulated below.

For this demonstration, a 10-year period of stream flow record was used as a dependent variable. The independent variables were taken from Table 1.2, Ten Thousand Randomly Assorted Digits, in Snedecor's Statistical Methods, 4th Edition.[5] Four trials were made using 3, 4, 5, and 9 sets of random numbers as independent variables. Values obtained for R (unadjusted) for different numbers of variables are as follows:

Total number of variables	Coefficient of correlation, R (unadjusted)
4	0.15
5	0.55
6	0.84
10	1.00

Computations of significance of difference in the above examples are based on a Z transformation described in Statistical Methods for Research Workers (Revised) by R. A. Fisher, Hafner, New York, 1950; and in applied General Statistics by Croxton and Cowden, Prentice-Hall, New York, 1939. Charts on pages 506 to 509 of Ezekiel's Methods of Correlation Analysis, second edition, facilitate the judging of the significance of correlation coefficients.

[5] Snedecor, George W., Statistical Methods, Collegiate Press, Ames, Iowa, 1946.

It would appear that R increases automatically with m, and that for m = n (zero degrees of freedom) the coefficient of multiple correlation, R, unadjusted for degrees of freedom, has a value of 1.00. This indicates perfect correlation, for a population in which no real relationship exists. However, the adjusted coefficient, $\bar{R}$, for these trials, gives no evidence that the sample was not drawn from totally unrelated series of data. (For values of $\bar{R}^2$ less than zero, the $\bar{R}$ is considered to be zero.)

The foregoing example illustrates the need for considering the degrees of freedom. The meaning of the expression, degrees of freedom, as applied in correlation analysis, may be clarified by reasoning as follows. Suppose, for a two-variable correlation, coordinate points are plotted representing two years of record. (For two years of record and two variables the degrees of freedom, n - m, would be zero.) The two coordinate points would determine a line. However, with these two observations there are no cases of departure from the line, and thus no basis for computing a measure of probable departures from the line. Similarly, for a three-variable relationship, describing a plane, three observations fix the position of the plane. Only when observations exceed three are there any departures from the plane upon which to base estimates of dispersion or strength of relationship. In working with short periods of record it is especially important that the number of variables be restricted if stable results are to be obtained.

As demonstrated in the previous section on reliability of the individual estimate, the inclusion of a variable having a large value of σ_b increases the expected error of forecast for abnormal values of that variable. The requirement for significance of the various independent variables automatically limits the number of variables which may be included.

Correlation Between Independent Variables

An added contributor to unstable results is high correlation between independent variables. The effect of such a relationship in increasing the standard error of the regression coefficient is expressed in eq. (20). A high correlation between independent variables may lead to illogical results, possibly to the extent of indicating relationships not in agreement with known physical behavior. If perfect correlation existed between two independent variables there would be no way of separating the effects of each. Examples of data on causal factors which could be expected to be closely correlated are records from two precipitation stations or two snow courses having similar exposure. If units of measure will permit, the closely correlated records may be combined and used as a single causal factor. Otherwise, one of the correlated variables may be dropped.

Some degree of correlation, either real or fortuitous, between independent variables is to be expected. If one of two factors so related is omitted, the factor retained will take on added weight. Whether the relationship is real or fortuitous may have a bearing on the procedure in a given analysis. For example, if flood season precipitation were closely associated with some antecedent event, as winter precipitation, then omission of the former, although significantly related to run-off, might be desirable. However, probably no significant relationship exists between precipitation amounts for different seasons.

Inclusion of Run-off Season Precipitation; Revision of Forecast

In the preceding paragraphs the inclusion of run-off season precipitation as a means of improving reliability was considered. An added practical use of this factor is that of following up the forecast with revised forecasts, as the season advances and precipitation records for the earlier portions of the season become available. These recorded values are substituted in the equation, along with assumptions for the remainder of the season, and a revised forecast prepared. Through a more complete understanding of cause and effect made possible by the inclusion of all significant factors, a better opportunity is afforded for determining the proper course of action in any given situation.

Fortunately, in many basins precipitation during the month of July or during

the June-July period does not contribute significantly to flood season run-off volume. In such cases the precipitation for these periods should be omitted. With this omission the final run-off estimate is known at an earlier date.

SHORT-CUT METHOD FOR SCREENING POSSIBLE CONTRIBUTING FACTORS

In developing forecasting techniques for various basins differing in physiographic characteristics, it may be desired to examine certain factors in addition to such usually dependable indices as snow-water equivalent, fall precipitation, and run-off season precipitation, which might logically be expected to influence run-off to a measurable extent in a particular basin. If a number of possible factors are to be investigated, an approximate, graphical analysis would save considerable labor in conducting exploratory work. For improved accuracy, the graphical analysis would be followed by a mathematical evaluation of the factors which appear, in the graphical explorations, to bear an important relationship with run-off.

In the procedure illustrated below, the factors which are more obviously related to run-off (those which are significantly related in most basins), are summarized in a preliminary estimating equation which is derived by methods of least squares as in the previous example. The residuals, or differences between observed run-off and estimates based on this preliminary equation, are then correlated graphically with an added factor to be investigated. The following preliminary estimating equation is for the May 1 forecast of the flood season run-off, Y, in million acre-foot units, of the Colorado River at Cameo, Colorado:

$$Y = 0.155X_3 + 0.158X_4 + 0.124X_5 - 1.092$$

where

X_3 = October-January precipitation (inches),

X_4 = snow-water equivalent (average of April 1 and May 1 surveys, inches), and

X_5 = May-July precipitation (inches).

Residuals, or differences between observed flows and estimates (fig. 2a) based on the preliminary equation are plotted in fig. 2b as ordinates against the added variable, precipitation during the period July through September of the previous year. The graph indicates a positive relationship between the new variable and the residuals, and hence a positive relationship between this variable and flood season run-off. Having first accounted for the effects of three important factors in the preliminary multiple correlation, the effect of the remaining variables becomes more clearly defined, as in fig. 2b. If in this example the new variable had been plotted directly against run-off, it is doubtful if any relationship would have been in evidence. In such correlations the relationship is obscured by the effects of variables not accounted for.

If additional variables which might logically be expected to influence run-off are to be examined, the work will be facilitated by plotting the points for each new variable on a transparent overlay, rather than directly on fig. 2b. The horizontal, dashed lines on the base figure, positioned by the residuals, are drawn to facilitate such multiple plottings. In this manner a number of possible causal factors may be screened in a much shorter time than by the far more laborious trial multiple regressions involving various combinations of independent variables.

The graphical procedure may be extended by correlating the departures from the trendline in fig. 2b with a second new variable, and so on. In such a series of graphs, the variables would be introduced, as far as practicable, in the order of their relative importance.

Fig. 2b could be used as a correction curve, distances from the zero residual line to be added algebraically to an estimate

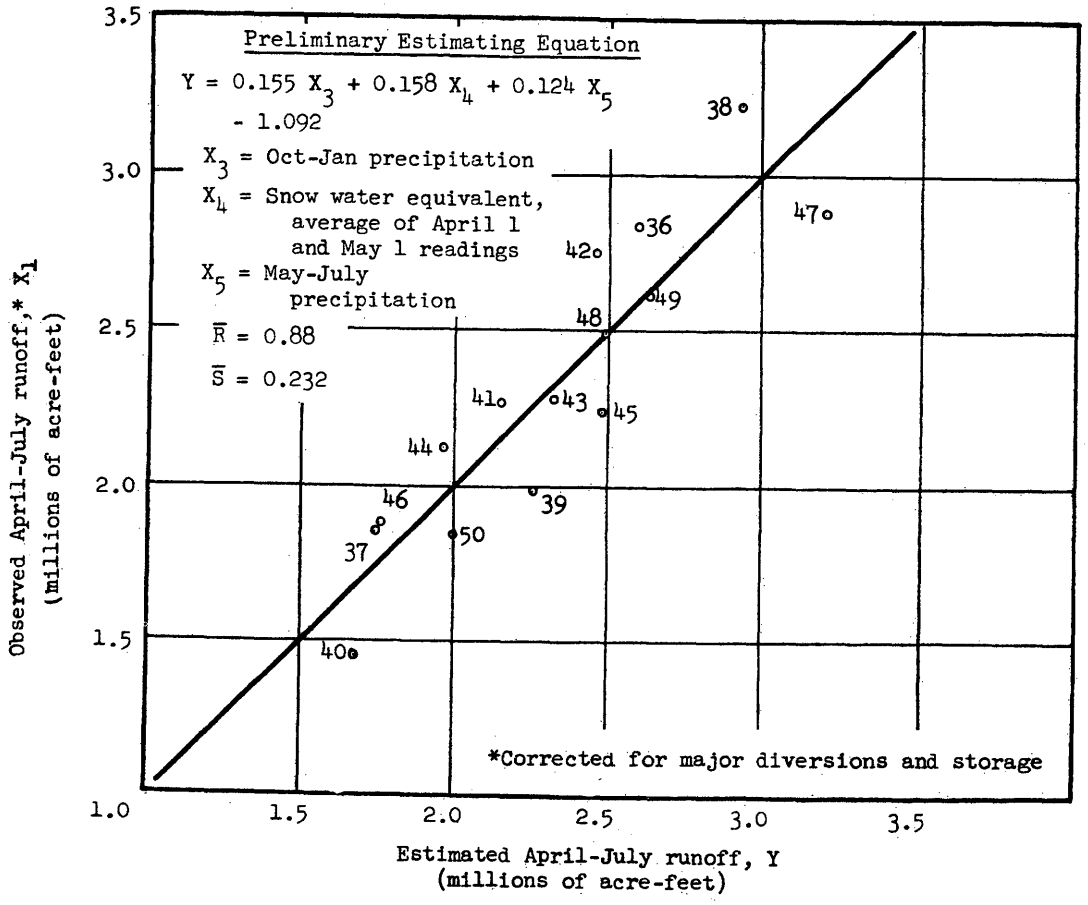


Figure 2a - Relationship between Observed Flows and Preliminary Estimates, Colorado River.

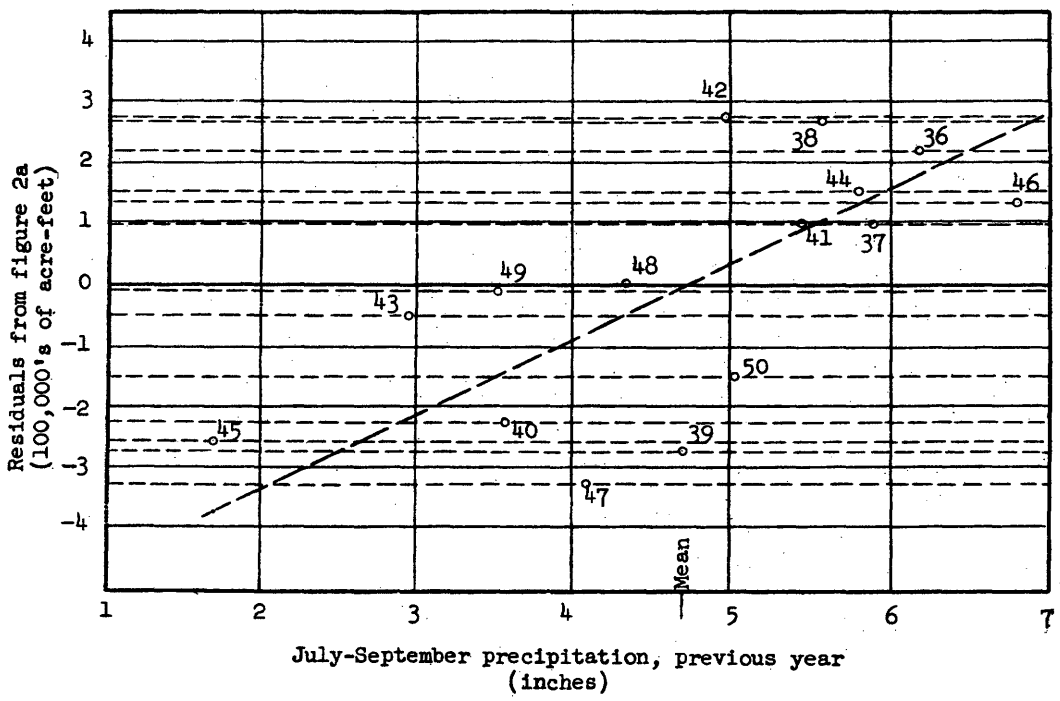


Figure 2b - Relationship between Departures in Figure 2a and Antecedent Precipitation.

of run-off based on the preliminary formula. In this problem, however, more precise information was desired regarding the antecedent July-through-September precipitation as a factor influencing flood-season run-off. Therefore, a new multiple regression equation was derived that included this factor. Other variables are the same as in the first equation. This new relationship is summarized as follows:

$$Y = 0.120X_2 + 0.129X_3 + 0.178X_4 + 0.171X_5 - 1.967$$

$b_2 = 0.120 \pm 0.037$ (July-September precipitation, inches),

$b_3 = 0.129 \pm 0.051$ (October-January precipitation, inches),

$b_4 = 0.178 \pm 0.024$ (snow water, avg. of April 1 and May 1, inches), and

$b_5 = 0.171 \pm 0.038$ (May-July precipitation, inches).

For this equation,

$\bar{R} = 0.94$; $\bar{S} = 0.169$ million acre-feet.

For all independent variables the standard deviation of the regression coefficient is less than 1/2 the value of the coefficient, indicating significance. With the addition of the new variable, X_2, the coefficient of multiple correlation, adjusted, increased from 0.88 to 0.94, and the standard error of estimate, $\bar{S}$, decreased from 0.232 to 0.169 million acre-feet.

Observed vs. estimated run-off volumes are plotted in fig. 3, and the 0.90 probability limits are shown at three levels of recorded run-off: mean (1941), maximum (1938), and minimum (1940). These limits were computed by applying equation (17).

By means of such a graphical analysis as described above and illustrated in fig. 2a and 2b it was found that, in the Swan River Basin in Montana, the temperature during the snow accumulation period had a significant effect on variations in flood-season run-off. This region is frequented by warm winds during the winter months. Temperatures during December and March of Water Year 1934 were far above normal. The resulting snowmelt prior to the nominal flood season was evidenced in part by the occurrence on December 25 of the largest flood of Water Year 1934. The coefficient of multiple regression applying to temperature was, in this case, negative, indicating an inverse relationship between winter temperatures and flood season run-off; i.e., the higher the winter temperature the less the run-off during the following nominal melt season. Such examples further illustrate the individual differences in basin characteristics, and the need for exploratory work in relating factors of cause and effect. The device illustrated in fig. 2a and 2b is offered as a convenient means of selecting added factors in different basins.

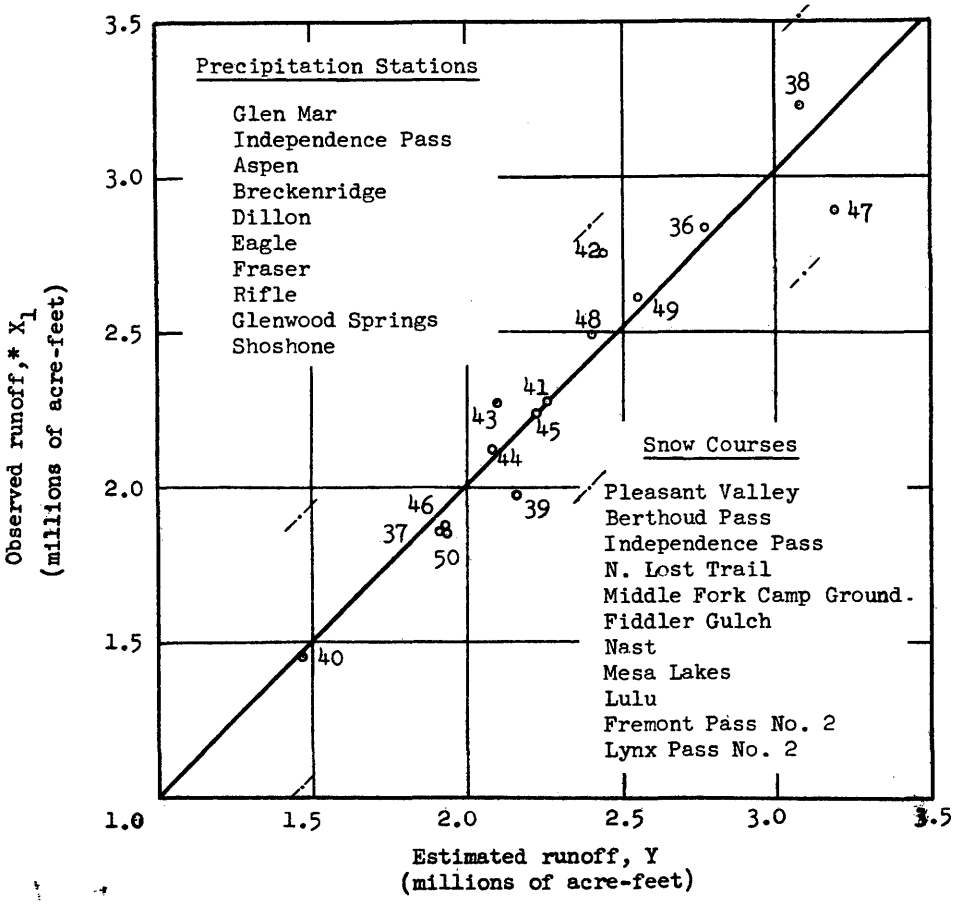
CHANGE OF DEPENDENT VARIABLE

It is sometimes required that the regression of different dependent variables upon the same set of independent variables be determined. For example, estimates of run-off for different periods of months or of flood volume between beginning and end of flood flow (shifting dates) may be required, or estimates of some flood flow characteristic other than total volume, such as peak flows, may be desired, based on the same independent variables. A convenient method of analysis in such cases is described in Statistical Methods for Research Workers by R. A. Fisher. This method will be illustrated, using data of the Colorado River Problem. In this procedure the c values (covariance matrix) are determined, and from these the coefficients of multiple regression are computed from the following relationships:

$$\left.\begin{aligned} b_2 &= c_{22}\Sigma(x_1x_2) + c_{23}\Sigma(x_1x_3) + c_{24}\Sigma(x_1x_4) + c_{25}\Sigma(x_1x_5) \\ b_3 &= c_{23}\Sigma(x_1x_2) + c_{33}\Sigma(x_1x_3) + c_{34}\Sigma(x_1x_4) + c_{35}\Sigma(x_1x_5) \\ b_4 &= c_{24}\Sigma(x_1x_2) + c_{34}\Sigma(x_1x_3) + c_{44}\Sigma(x_1x_4) + c_{45}\Sigma(x_1x_5) \\ b_5 &= c_{25}\Sigma(x_1x_2) + c_{35}\Sigma(x_1x_3) + c_{45}\Sigma(x_1x_4) + c_{55}\Sigma(x_1x_5) \end{aligned}\right\} \quad ..(21)$$

Computations involved in this solution are outlined below. Basic data are shown in Table 9. For this indirect solution, the X_1 values need not be shown in Table 9; they are shown here for future reference.

In this example, tables of extensions and corrections to departures from the means corresponding to Tables 2 and 3 will be omitted, since the computations involved are the same as in the previous example. In Table 10 is shown a form of the normal equations corresponding to Table 4, except that X_1 values have in this case been omitted to afford a better illustration of the indirect method. The procedure for solving these equations for the c values is the same as in the previous example. Computations for evaluating the c-matrix are shown in Table 10.



*Corrected for major diversions and storage

Estimating Equation

$$Y = 0.120\ X_2 + 0.129\ X_3 + 0.178\ X_4 + 0.171\ X_5 - 1.970$$

X_2 = July-Sept precipitation, antecedent

X_3 = Oct-Jan precipitation

X_4 = Snow water equivalent, average of April 1 and May 1 readings

X_5 = May-July precipitation

$\bar{R} = 0.94$

$\bar{S} = 0.169$

0.90 limits shown for 1938, 1940, and 1948

Figure 3 - Correlation of Observed vs. Estimated April-July Run-off, Colorado River at Cameo, Colorado.

Having determined the c values, the b coefficients for the regression of a new X_1 variable upon the same set of independent variables may be computed from the following relationships (from equation (21)):

$$\left.\begin{aligned} b_2 &= 0.0472\Sigma(x_1x_2)-0.0096\Sigma(x_1x_3) \\ &\quad +0.0753\Sigma(x_1x_4)+0.0183\Sigma(x_1x_5) \\ b_3 &= -0.0096\Sigma(x_1x_2)+0.0918\Sigma(x_1x_3) \\ &\quad -0.1587\Sigma(x_1x_4)-0.0049\Sigma(x_1x_5) \\ b_4 &= 0.0753\Sigma(x_1x_2)-0.1587\Sigma(x_1x_3) \\ &\quad +2.00481\Sigma(x_1x_4)+0.0278\Sigma(x_1x_5) \\ b_5 &= 0.0183\Sigma(x_1x_2)-0.0049\Sigma(x_1x_3) \\ &\quad +0.0278\Sigma(x_1x_4)+0.0508\Sigma(x_1x_5) \end{aligned}\right| \quad (21a)$$

Extensions with X_1 are shown in Table 11 and computations for the extensions x_1^2, x_1x_2, x_1x_3, x_1x_4, and x_1x_5 are shown in Table 12.

The values for the multiple regression

TABLE 9

RECORDS OF PRECIPITATION, SNOW-WATER EQUIVALENTS, AND RUN-OFF, COLORADO RIVER AT CAMEO, COLORADO

Water Year Ending Sept. 30	July-Sept. Precipitation for Previous Calendar Year* (inches)	Oct.-Jan. Precipitation (inches)	Snow Water** (10-inch units)	May-July Precipitation (inches)	April-July Natural Run-off (million ac. ft. units)	
	X_2	X_3	X_4	X_5	X_1 ***	ΣX
1936	6.20	5.74	1.42	4.44	2.83	20.63
1937	5.90	3.83	0.98	5.55	1.85	18.11
1938	5.58	6.31	1.55	4.77	3.22	21.43
1939	4.71	6.60	1.28	2.60	1.97	17.16
1940	3.57	4.39	1.06	3.30	1.45	13.77
1941	5.47	4.55	1.30	3.99	2.27	17.58
1942	4.95	6.19	1.38	3.37	2.75	18.64
1943	2.94	6.11	1.08	6.02	2.27	18.42
1944	5.82	4.33	1.24	3.46	2.12	16.97
1945	1.70	4.29	1.38	5.90	2.23	15.50
1946	6.78	6.16	0.84	4.59	1.87	20.24
1947	4.10	6.63	1.60	5.77	2.88	20.98
1948	4.33	5.80	1.42	3.51	2.49	17.55
1949	3.52	6.04	1.28	6.16	2.61	19.61
1950	5.04	5.28	1.25	2.36	1.84	15.77
Total	70.61	82.25	19.06	65.79	34.65	272.36
Mean	4.7073	5.4833	1.2707	4.3860	2.3100	18.1573

*For example, for Water Year 1936, X_2 is the precipitation in inches during the July to September period of Calendar Year 1935.

**Average of April 1 and May 1 readings.

***This series of X_1 and ΣX values is included for future reference (see Table 11).

coefficients obtained by substituting in eq. (21a) or (21), are shown below with their standard errors (computations for standard errors not shown):

$b_2 = 0.120 \pm 0.037$

$b_3 = 0.129 \pm 0.051$

$*b_4 = 0.178 \pm 0.024$

$b_5 = 0.171 \pm 0.038$

Table 13 provides a check on the solution for the regression coefficients. The final values in columns 2 and 3 are in reasonable agreement; the difference is attributable to the rounding off of values. The constant term and the parameters $\overline{R}$ and $\overline{S}$ are computed as in the previous example.

*This is the decoded value of b_4. The original records were coded by dividing by 10, as indicated in Table 9. Basic data may be coded where desirable by multiplication, division, addition, or subtraction, applying a constant factor throughout a series of values of any variable.

TABLE 10

DOOLITTLE SOLUTION FOR EVALUATION OF c-MATRIX
(Indirect Method)

	X_2	X_3	X_4	X_5	c_2	c_3	c_4	c_5	$\Sigma(X+c)$
Eq. 1	26.081	0.896	-0.786	-8.857	1	0	0	0	18.334
Eq. 2	(0.896)	12.670	0.964	0.383	0	1	0	0	15.913
Eq. 3	(-0.786)	(0.964)	0.604	0.046	0	0	1	0	1.828
Eq. 4	(-8.857)	(0.383)	(0.046)	22.860	0	0	0	1	15.432
I	26.081	0.896	-0.786	-8.857	1				18.334
I'	-1.000000	-0.034355	0.030137	0.339596	-0.038342				-0.702964
II	(0.896)	12.670	0.964	0.383		1			15.913
(-0.034355) I	(-0.89601)	-0.03078	0.02700	0.30428	-0.03436				-0.62986
Σ_2	0	12.63922	0.99100	0.68728	-0.03436	1			15.28314
II'		-1.000000	-0.078407	-0.054377	0.002719	-0.079119			-1.209184
III		(0.964)	0.604	0.046			1		1.828
(0.030137) I		(0.02700)	-0.02369	-0.26692	0.03014				0.55253
(-0.078407) Σ_2		(-0.99100)	-0.07770	-0.05389	0.00269	-0.07841			-1.19831
Σ_3		0	0.50261	-0.27481	0.03283	-0.07841	1		1.18222
III'			-1.000000	0.546766	-0.065319	0.156006	-1.98614		-2.352162
IV			(0.046)	22.860				1	15.432
(0.339596) I			(-0.26692)	-3.00780	0.33960				6.22615
(-0.054377) Σ_2			(-0.05389)	-0.03737	0.00187	-0.05438			-0.83105
(0.546766) Σ_3			(0.27481)	-0.15026	0.01795	-0.04287	0.54677		0.64640
Σ_4			0	19.66457	0.35942	-0.09725	0.54677	1	21.47350
IV'				-1.000000	-0.018277	0.004945	-0.027805	-0.050853	-1.09190

TABLE 10 (Continued)

Back Solution on c_2

c_{22}	c_{23}	c_{24}	c_{25}	Eq. 4	Check
0.03834	-0.00272	0.06532	0.01828		
0.00621	-0.00099	0.00999	0.01828	22.860	0.418
0.00227	-0.00590	0.07531		0.046	0.003
0.00033	-0.00961			0.383	-0.004
0.04715				-8.857	-0.417
				0	0

Back Solution on c_4

c_{42}	c_{43}	c_{44}	c_{45}	Eq. 4	Check
		1.98961	0.02780		
0.00944	-0.00151	0.01520	0.02780	22.860	0.636
0.06042	-0.15719	2.00481		0.046	0.092
0.00545	-0.15870			0.383	-0.061
0.07531				-8.857	-0.667
				0	0

Back Solution on c_3

c_{32}	c_{33}	c_{34}	c_{35}	Eq. 4	Check
	0.07912	-0.15600	-0.00495		
-0.00168	0.00027	-0.00270	-0.00495	22.860	-0.113
-0.00478	0.01244	-0.15870		0.046	-0.007
-0.00315	0.09183			0.383	0.035
-0.00961				-8.857	0.085
				0	0

Back Solution on c_5

c_{52}	c_{53}	c_{54}	c_{55}	Eq. 4	Check
			0.05085		
0.01727	-0.00277	0.02780	0.05085	22.860	1.162
0.00084	-0.00218	0.02780		0.046	0.001
0.00017	-0.00495			0.383	-0.002
0.01828				-8.857	-0.162
				1	0.999

TABLE 11

EXTENSIONS WITH X_1
(Variables as shown in Table 10)

	X_1X_2	X_1X_3	X_1X_4	X_1X_5	X_1^2	$X_1\Sigma X$
1936	17.546	16.244	4.019	12.565	8.009	58.383
1937	10.915	7.086	1.813	10.268	3.422	33.504
1938	17.968	20.318	4.991	15.359	10.368	69.005
1939	9.279	13.002	2.522	5.122	3.881	33.805
1940	5.177	6.366	1.537	4.785	2.102	19.966
1941	12.417	10.328	2.951	9.057	5.153	39.907
1942	13.612	17.022	3.795	9.268	7.562	51.260
1943	6.674	13.870	2.452	13.665	5.153	41.814
1944	12.338	9.180	2.629	7.335	4.494	35.976
1945	3.791	9.567	3.077	13.157	4.973	34.565
1946	12.679	11.519	1.571	8.583	3.497	37.849
1947	11.808	19.094	4.608	16.618	8.294	60.422
1948	10.782	14.442	3.536	8.740	6.200	43.700
1949	9.187	15.764	3.341	16.078	6.812	51.182
1950	9.274	9.715	2.300	4.342	3.386	29.017
Total	163.447	193.517	45.142	154.942	83.306	640.354

TABLE 12

EXTENSIONS CORRECTED
TO DEPARTURES
FROM THE MEAN

$\Sigma x_1x_2 = 163.447 - 15(2.310)(4.7073)$

$= 163.447 - 163.108$

$= 0.339$

$\Sigma x_1x_3 = 193.517 - 15(2.310)(5.4833)$

$= 193.517 - 189.996$

$= 3.521$

$\Sigma x_1x_4 = 45.142 - 15(2.310)(1.2707)$

$= 45.142 - 44.030$

$= 1.112$

$\Sigma x_1x_5 = 154.942 - 15(2.310)(4.3860)$

$= 154.942 - 151.975$

$= 2.967$

$\Sigma x_1^2 = 83.306 - 15(2.310)(2.310)$

$= 83.306 - 80.042$

$= 3.264$

TABLE 13

CHECK ON COMPUTATION OF REGRESSION COEFFICIENTS, AND SOLUTIONS FOR $\bar{R}$, $\bar{S}$, AND a

	1	2	3	4	5	6	7
	(Reg. Coef.)	(Eq. 4 from Tables 10 & 12)	(1·2)	(X_1 from Table 12)	(1·4)	(Means from Table 9)	(1·6)
X_2	0.120	-8.857	-1.063	0.339	0.041	4.707	0.565
X_3	0.129	0.383	0.049	3.521	0.454	5.483	0.707
X_4	1.778	0.046	0.082	1.112	1.977	1.271	2.257
X_5	0.1706	22.860	3.900	2.967	0.506	4.386	0.748
		2.967	2.968	3.264	2.978	2.310	4.277

$$R^2 = \frac{2.978}{3.264} \qquad R = \sqrt{0.9124}$$

$$= 0.9124 \qquad = 0.955$$

$$\bar{R}^2 = 1 - (1-R^2)\frac{n-1}{n-m} \qquad \bar{R} = \sqrt{0.878}$$

$$= 1 - (1-0.912)\frac{14}{10} \qquad = 0.937$$

$$= 0.878$$

$$\bar{S}^2 = \frac{3.264 - 2.978}{n - m} \qquad \bar{S} = \sqrt{0.0286}$$

$$= \frac{0.286}{10} \qquad = 0.169 \text{ (169,000 acre-feet)}$$

$$= 0.0286$$

$$a = 2.310 - 4.277$$

$$= - 1.967$$

Multiple regression equation:

$$Y = 0.120X_2 + 0.129X_3 + 0.178X_4 + 0.171X_5 - 1.967$$

In this indirect procedure the c values serve both to determine the standard deviation of the multiple regression coefficients, and in computing the regression coefficients with equation (21).

ELIMINATION OF AN INDEPENDENT VARIABLE

After a regression equation has been derived, it may happen that one of the independent variables does not appear to bear a significant relationship with run-off, and that the omission of the variable would have been preferable. The regression on the remaining variables could be calculated by repeating a Doolittle solution involving only the equations containing the desired extensions. A more convenient means of eliminating an independent variable is described by R. A. Fisher in Statistical Methods for Research Workers. The method consists of recalculating the new regression coefficients from formulas of the form

$$b'_3 = b_3 - \left(\frac{c_{32}}{c_{22}}\right) b_2 \quad \ldots\ldots\ldots \quad (22)$$

where X_2 is the variable to be eliminated. The values b' are the coefficients which would have been obtained had variable X_2 been omitted. For example, if it were desired to eliminate the X_2 variable in the Colorado River Problem, the b coefficients of the resulting equation would be as follows:

$$b'_3 = b_3 - \left(\frac{c_{32}}{c_{22}}\right) b_2$$

$$= 0.129 - \left(\frac{-0.00961}{0.04715}\right) 0.120$$

$$= 0.154$$

$$b'_4 = b_4 - \left(\frac{c_{42}}{c_{22}}\right) b_2$$

$$= 0.178 - \left(\frac{0.00753}{0.04715}\right) 0.120$$

$$= 0.159$$

$$b'_5 = b_5 - \left(\frac{c_{52}}{c_{22}}\right) b_2$$

$$= 0.171 - \left(\frac{0.01828}{0.04715}\right) 0.120$$

$$= 0.124$$

These regression coefficients are in reasonable agreement with those originally computed and previously listed. Slight differences are due to rounding of figures.

A comprehensive application of Fisher's method, credited to Professor H. Shultz of Chicago, consists of recalculating the c-matrix from formulas of the following form (for 3 remaining independent variables, eliminating X_2):

$$\left.\begin{aligned}
c'_{33} &= c_{33} - \frac{(c_{32})\ (c_{32})}{c_{22}} \\
c'_{34} &= c_{34} - \frac{(c_{32})\ (c_{42})}{c_{22}} \\
c'_{35} &= c_{35} - \frac{(c_{32})\ (c_{52})}{c_{22}} \\
c'_{44} &= c_{44} - \frac{(c_{42})\ (c_{42})}{c_{22}} \\
c'_{45} &= c_{45} - \frac{(c_{42})\ (c_{52})}{c_{22}} \\
c'_{55} &= c_{55} - \frac{(c_{52})\ (c_{52})}{c_{22}}
\end{aligned}\right\} \quad \ldots\ldots \quad (23)$$

In these equations $c_{ab} = c_{ba}$.

The values c' supply the c-matrix which would have been obtained had variable X_2 been omitted. These c' values substituted in eq. (21) yield the desired multiple regression coefficients. The standard error of the regression coefficients is determined by the relationships of the form

$$\sigma_{b_2} = \overline{S}_{1 \cdot 234} \sqrt{c_{22}}$$

as in previous examples. The $\overline{S}$ is the standard error of estimate for a multiple regression equation omitting the X_2 variable.

A convenient procedure for the elimination of an independent variable consists of applying equation (22) for evaluating the

new b coefficients, and equation (23) for evaluating certain c values (c_{22}, c_{33}, ..., c_{nn}) which are required for computing the standard error of the b coefficients. For the purpose of illustrating this procedure, computations are shown below for the elimination of the X_2 variable from the multiple regression based on data in Table 9. First, new values for b_3, b_4, and b_5 would be computed, using equation (22). (These computations are shown above.)

Second, a new value of $\bar{S}$ ($\bar{S}'$) is computed by substituting in equation (10) the b' values as follows:

$$(\bar{S}')^2 = \frac{3.264 - (0.154)(3.521) - (1.590)(1.112) - (0.124)(2.967)}{15 - 4}$$

$$= 0.0533$$

and $\bar{S}' = 0.231$ (231,000 acre-feet)

Third, values of c'_{33}, c'_{44}, and c'_{55} are computed by substituting in equation (23) the indicated c values from Table 10 as follows:

$$c'_{33} = 0.09183 - \frac{(-0.00961)(-0.00961)}{0.04715}$$

$$= 0.08987$$

$$c'_{44} = 2.00481 - \frac{(0.07531)(0.07531)}{0.04715}$$

$$= 1.88452$$

$$c'_{55} = 0.050585 - \frac{(0.01828)(0.01828)}{0.04715}$$

$$= 0.04377$$

From these new c values the standard errors of the b coefficients are computed by the formula, $\sigma_{b_i} = \bar{S}\sqrt{c_{ii}}$,

$$\sigma_{b_3} = 0.231\sqrt{0.08987} = 0.069$$

$$\sigma_{b_4} = 0.231\sqrt{1.8845} = 0.317 \text{ (10-inch units)}$$

$$\sigma_{b_5} = 0.231\sqrt{0.04377} = 0.048$$

and $\bar{R}$ is computed from equation (12a) and Table 12:

$$\bar{R}^2 = 1 - \frac{(15-1)(0.231)^2}{3.264}$$

$$= 0.771$$

$$\bar{R} = 0.88$$

COMPARISON OF ERRORS OF FORECAST

Data for the Colorado River watershed provide an opportunity for additional comparisons of the results of forecasting equations that (1) include the effect of run-off season precipitation on the value and reliability of the multiple regression coefficients, and (2) is biased by neglecting the effect of run-off season precipitation. Corresponding values for the two equations are tabulated on page 35.

The value of $\bar{S}$ for (1) above was computed by substitution in the formula

$$\bar{S} = \sqrt{\frac{\Sigma Z^2}{n - m}}$$

As in the South Fork of the Boise River example, the biased equation, although having an apparently lower $\bar{S}$, is subject to greater error of forecast.

---

(1)	(2)
Including the effect of run-off season precipitation (Equation obtained from a basic relationship in which the effect of run-off season precipitation was considered.)	Neglecting the effect of run-off season precipitation (Derivation of equation involved only independent variables antecedent to date of forecast.)
Equations:	
$Y = 0.120X_2 + 0.129X_3 + 0.178X_4 - 1.217$	$Y = \frac{0.058X_2 + 0.146X_3}{+ 0.168X_4 - 0.908}$
Standard errors of forecast:	
$\overline{S} = 0.298$	$\overline{S} = 0.280$
Corresponding regression coefficients:	
$b_2 = 0.120 \pm 0.037$	$b_2 = 0.058 \pm 0.056$
$b_3 = 0.129 \pm 0.051$	$b_3 = 0.146 \pm 0.085$
$b_4 = 0.178 \pm 0.024$	$b_4 = 0.168 \pm 0.040$
Errors of individual estimate for a high run-off year (1938), average conditions (1948), and for a low year (1940), for 0.90 probability:	
high $= \pm 0.475$	high $= \pm 0.497$
average $= \pm 0.468$	average $= \pm 0.480$
low $= \pm 0.481$	low $= \pm 0.512$
(Errors computed by equation 17)	(Errors computed by equation 16)

CONSISTENCY OF RECORDS

A preliminary investigation of basic records is advisable in the analysis of hydrologic data, particularly if older records are used. In the derivation of a multiple regression (forecast) equation the various constants of the equation are subject to bias from inconsistencies in the records used in evaluating these constants. Before attempting the derivation of a forecast equation for the Swan River in the Flathead River Basin in Montana, stream flow and precipitation records were examined for consistency by means of a graphical procedure based on double-mass diagram[6] as illustrated in figures 4 and 5. Essentially a double-mass diagram consists of a plotting on rectangular coordinates, of accumulated values of paired observations. The slope of a line segment through the plotted points represents the relationship between the variates, and it follows that a change in slope indicates a change in this relationship. (Incidentally, such graph is not to be confused with a correlation chart.)

[6] Application of the double-mass diagram is discussed in Forecasting Water Supply, a report of the Hydraulic Power Committee, Engineering National Section, National Electric Light Association, March 1931; also in "A Comprehensive Study of the Rainfall on the Susquehanna Valley," by C. F. Merriam, Transactions, American Geophysical Union, 1937.

In Figure 4, accumulated values of a 7-station mean precipitation are shown as abscissae and accumulated precipitation recorded at individual stations are shown as ordinates, each station having its own zero point along the vertical axis.

Of the seven precipitation stations, only the gage at Trout Creek was reported in the station histories as having been relocated. According to the history this gage was moved in 1930 and again in 1942. Changes in observed values of precipitation resulting from these relocations are evidenced by abrupt changes in slope of the trend line for the Trout Creek Station, figure 5. To place these records on a basis equivalent to that of current Trout Creek data, adjustment factors of 1.03 and 0.94, respectively, were applied to the Trout Creek records obtained at the two earlier locations. Changes as small as this make no appreciable difference in a seven-station average, nor in the final correlation. However, such an examination of data is desirable, since in many series of records changes appear which are of sufficient magnitude to materially affect the usefulness of the data.

The double-mass diagram in Figure 5 consists of a plotting of accumulated estimates of discharges as abscissae and accumulated observed discharges as ordinates. The estimates of discharge were obtained from a simple regression (not shown) of October-July runoff upon October-March precipitation, the regression line being determined by records of recent years of reasonably consistent data. The change in the slope of the line, about 1932, in Figure 5 indicates a variation in recorded flow as related to estimated flows, or as related to precipitation. Runoff records prior to the indicated date of change in slope should be adjusted by the factor 0.76 if they are to be considered as a part of a consistent record.

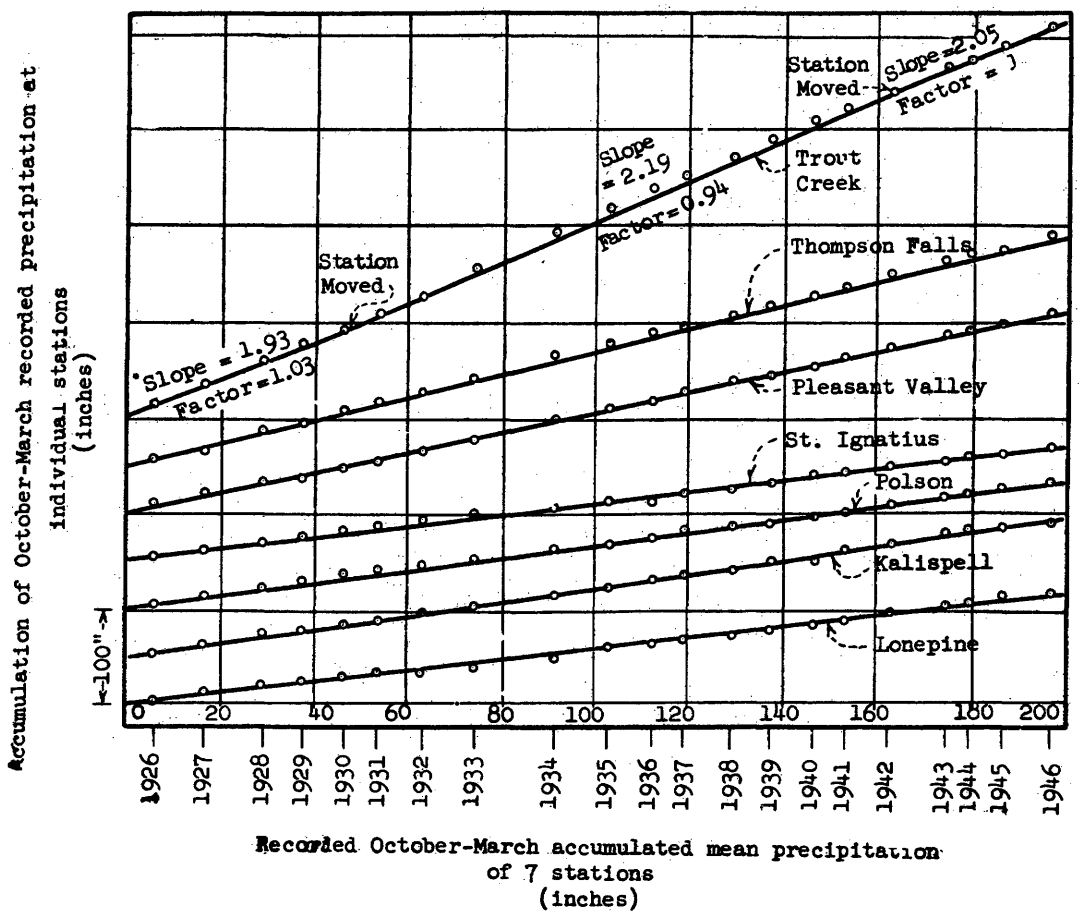


Figure 4 - Double Mass Diagram of Precipitation Data, Individual Stations vs. Average for the Area.

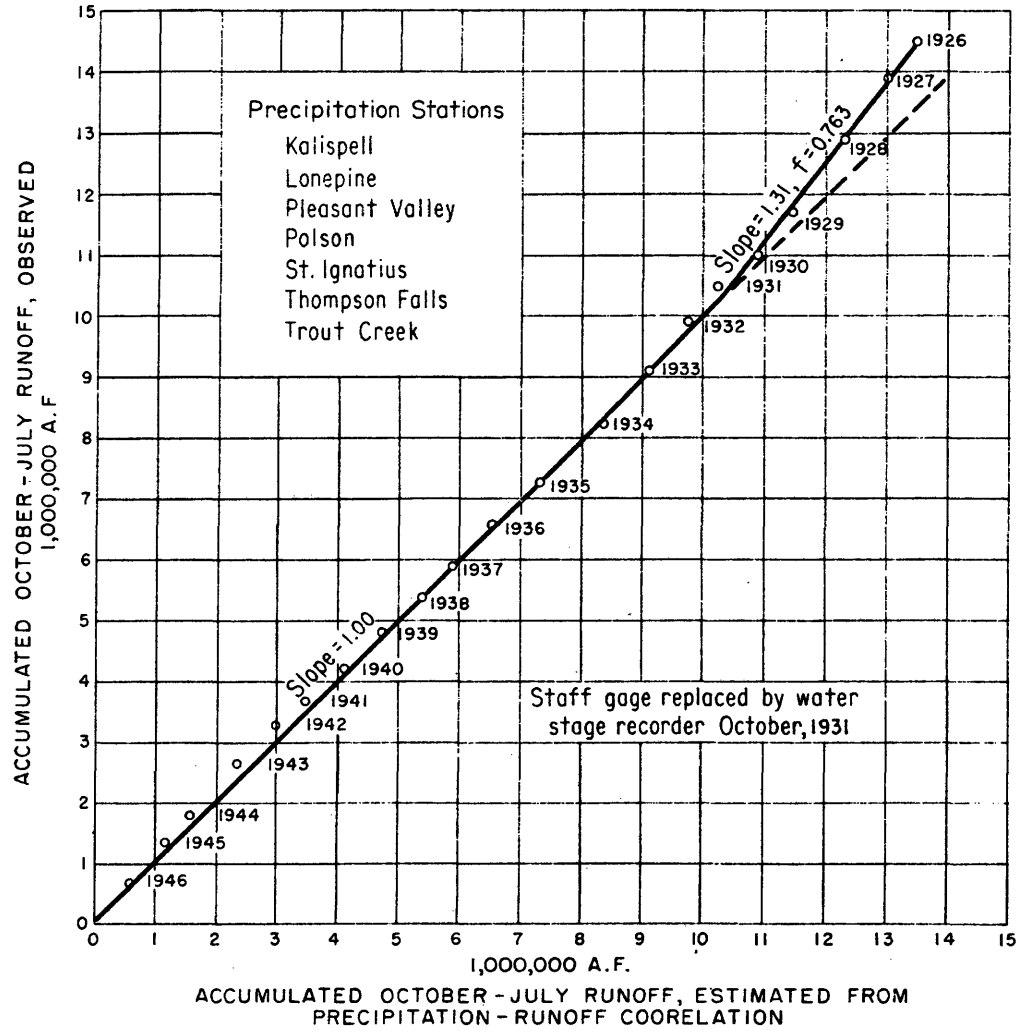


Figure 5 - Double Mass Diagram, Estimated vs. Observed Runoff.

It was mentioned above that the precipitation-runoff regression line used in estimating runoff was positioned by recent years of reasonably consistent data. A decision as to the period to be used requires some exploratory plotting. In this example, a trial plotting would indicate that the period, 1932-1946, would be satisfactory for such correlation. An alternative and somewhat simpler procedure for examining the consistency of runoff records may be used if two essential requirements are met; namely, that the precipitation-runoff regression line pass through or near the origin, and that the precipitation-runoff relationship be essentially linear. If these minimum conditions are satisfied, then the double mass diagram may consist of accumulated precipitation plotted against accumulated runoff.

To satisfy oneself that a change in precipitation-runoff relationship, as shown in Figure 5, is not a result of inconsistencies in the recorded precipitation, it is advisable to use adjusted precipitation data in the initial precipitation-runoff correlation.

The use of long periods of reliable records is, of course, desirable in developing forecasting procedures. However, in some cases the accuracy of earlier records may be in doubt; also, data on certain important causal factors may not be available for early years. In many cases a shorter rec-

ord, such as for 15 years, may contain data on extreme conditions, thus covering a wide range of events. In such cases the use of the shorter period of more accurate and adequate data should be given consideration. With short periods of record the errors of estimate are calculable and allowance can be made for them.

DISCHARGE FROM LARGE RIVER BASINS

In each of the foregoing forecasting problems, consideration was limited to an application of multiple correlation analysis to a single tributary within a large river system. When applying the method illustrated to the forecasting of the probable discharge of the trunk stream of such a river system, a division of the system into subbasins for separate analysis, followed by an integration of these component parts into a basin forecast, would have certain advantages. Benefit of the tributary analyses could derive from the local use of forecasts within individual areas. Furthermore, causal factors within a particular tributary basin would be expected to correlate more closely with discharge of that tributary than with total trunk discharge which reflects events of all portions of the drainage area. This is particularly true in large river systems which drain heterogeneous climatic zones. The closer correlation provides greater reliability in the multiple regression coefficients. Through such a subdivision of the basin for analysis, individual tributary areas in which forecasting relationships are weakest become evident, and an improvement in techniques or need for additional basic data in these particular areas may be indicated.

MULTIPLE CORRELATION COMPUTATIONS ON ELECTRONIC COMPUTING MACHINES

Increased use of multiple correlation in solving engineering problems, particularly hydrologic problems, and the need for obtaining results in a minimum of time led to investigations into the possible use of electronic computing machines in performing the laborious computations involved.

Multiple correlation provides a means of determining the way in which one variable (dependent variable) changes when two or more other variables (independent variables) change. Since most hydrologic relationships involve multiple causation, the needs for this method of approach are many and varied. The use of multiple correlation in forecasting and the computations involved in the procedure are illustrated in preceding sections of this monograph.

To analyze, adequately and economically, the data involved in hydrologic relationships, a complete solution of multiple correlation problems on an electronic computing machine was considered--the derivation of a forecasting equation and the determination of the coefficient of multiple correlation, the standard errors of estimate, and the standard errors of the partial regression coefficients. The procedure outlined in the following pages provides for the derivation of a number of trial equations involving various factors and combinations of factors which could logically be expected to influence runoff.

The planning for this solution may be divided into two distinct parts:

Part 1. The organizing of the procedures, in accordance with fundamentals of correlation analysis, indicating the operations to be performed and the sequence of operations. (Only these fundamentals will be discussed.)

Part 2. The programing and wiring of the machine to carry out the operations indicated in Part 1. Since this programing will vary for different machines, no detailed discussion is included in the following outline. Part 1 may be followed in any programing.

In organizing the procedures, the writer divided the processing of data into three phases, for which suitable forms were prepared, indicating the operations to be performed in each phase. A fourth form was prepared upon which the results may be summarized. The forms enable the solution on machines of somewhat limited storage capacity. With the more versatile machines, some of the intermediate transcribing of data could be eliminated. Although particular emphasis has been placed on multiple correlation, the procedures described are applicable as well to the simple 2-variable correlation. For analyses involving simple correlation only, a separate form "Simple Correlation Computations," a copy of which is shown as Form E, would simplify operations. However, for mixed simple and multiple correlations, the use of the multiple correlation forms throughout would be preferable to a change in the processing. The machine procedure is applicable to linear, nonlinear, or joint relationships.

Briefly summarized, the three phases are as follows:

1. The summations of the extensions of the basic data, corrected to departures from the means, are computed and tabulated.

2. The normal equations are solved simultaneously to determine the partial regression coefficients and to evaluate the covariance matrix.

3. Values of the coefficient of multiple correlation, standard error of extimate, standard error of the partial regression coefficients, and the constant term of the equation are computed.

To facilitate the adaptation of the procedures to automatic computing machine methods, the writer suggests the following expression for a direct evaluation of $\overline{R}^2$:

$$\overline{R}^2 = 1 - \frac{(n-1)\,\overline{S}^2}{\Sigma x_1^2} \qquad (11a)$$

Details of the operations indicated on the various forms are outlined in the following steps, the underscored paragraph headings indicating the forms used.

Forms and Procedures

Form A

This form is completed in duplicate for machine processing. Information is contained as follows:

A. Problem number, problem name, date of forecast.

B. Basic data--X_1 , dependent variable, X_2 X_3--X_n, independent variables.

C. Description of variables.

D. Totals of lines and of columns.

E. Means of columns. (An added decimal place in the mean provides for a satisfactory check on later computations.)

F. Groups of variables (indicated by check marks) which are to be included in trial equations. Groups are numbered for identification. In numbering problems and groups a simple numbering system is desirable, as Problem 1, 2, 3...., Group 1, 2, 3...

From the data contained on Form A, summations of extensions of departures from the means are computed on the electronic or other high-speed computers (such extensions are illustrated in Tables 2 and 3).

Form B

This form on translucent paper, is intended for duplication when completed. In making the entries, a carbon backup is recommended for better prints.

Procedures are as follows:

A. Enter problem number, problem name, and date of forecast.

B. Enter on Form B summations of extensions of departures from the means for all variables, disregarding groups. These extensions are

taken from the machine tabulations (not shown) resulting from machine processing of data contained on Form A. The entries are located by line and column designations; i.e., the X_2 X_3 value is found in the X_2 line and X_3 column of the machine tabulation, and is entered in the X_2 line and X_3 column of the form, as well as in the X_3 line and X_2 column of the form, making a symmetrical matrix.

C. Obtain prints of the completed form. The number required for each problem is equal to the number of groups (trial equations) indicated on Form A.

D. On the prints, obtained (in "C" above), circle the quantities comprising the normal equations, i.e., corresponding to the groups checked on Form A. Enter the line totals in the summation column.

E. Prints prepared in "D" are ready for simultaneous solutions of the indicated sets of normal equations by machine methods.

Form C

Values to be entered in this form are obtained from three sources as indicated below.

From Form A enter:

A. Problem number, name, forecast date.

B. Group number in Column O.

C. Subscripts of independent variables in Column 3.

D. Means of independent variables in Column 4 opposite the corresponding subscript in Column 3.

E. n-1 in Column 2 (years of record less 1).

F. n-m in Column 2 (years of record less the number of constants in the regression equation; or n-1 less the number of entries in Column 3 of this form).

From machine tabulations of summations of extensions (or from Form B):

G. Values of $\Sigma X_1 X_i$ in Column 6 and of ΣX_1^2 in Column 2.

From the machine tabulations completed in "E" under Form B, above.

H. Values of b_i and c_{ii}.

This form is ready for machine computations to determine values of $\bar{R}$, $\bar{S}$, σb, and a. The column headings indicate the mathematical operations to be performed. Insofar as machine operations are concerned, no entries need be made in Columns 8, 9, 10, 11, 12, 13, 15, 16. These columns facilitate computations by manual methods as desired, as for an initial check on processing, or where a few extra trial equations may be required.

Form D

This form is on a type of paper from which prints may be obtained if extra copies of the final results are desired. The form is completed as follows:

A. Problem number, name, date of forecast, and group numbers.

B. Brief descriptions of variables.

C. Values of the partial regression coefficients, b's; and of the standard errors of the partial regression coefficients, σb; and values of $\bar{R}$, $\bar{S}$, and a.

D. Obtain prints of the completed form D as required.

This is the form which will be of interest to the analyst, since it contains all trial equations, the variates involved, and measures of reliability for each equation and a measure of the significance of each variate. The various trial equations are identified by group (line) numbers in the first column, and these numbers should be made to correspond with those on Form A. Entries on this form complete the cycle.

The above procedures are, as previously described, adaptable to computing machines with somewhat limited capabilities. In recent years, electronic computers which can be programmed to carry out the successive operations outlined above without the manual transcriptions of data shown on Forms B and C, have become readily available. A machine program which offers the most in economy and utility is one which provides

at the end of Phase 1 (extensions of departures from means of data on Form A) for a selection of desired sets of normal equations. This selection eliminates the need for repetitious accumulations of products in Phase 1, which would otherwise be required in investigating alternative combinations of variables.

Advancements in machine processing of data have, through a saving of time and labor, contributed much to the attractiveness of employing correlation analysis.

This monograph was prepared under the general supervision of J. R. Riter, Chief, Division of Project Investigations.

GPO 833303

MULTIPLE CORRELATION DATA SHEET

PROBLEM: MAY 1 FORECAST OF MAY-JULY RUNOFF

Problem Code

FORM A

Date

*	ΣX	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
36	60 15	6 06	2 08	4 77	38	2 13	1 48	7 33	17 26	18 66		
37	48 24	6 19	1 93	4 95	1 76	2 94	1 44	5 04	12 01	11 98		
38	55 90	7 79	2 01	5 91	2 36	2 13	1 13	5 28	14 37	14 92		
39	42 45	3 39	1 70	4 47	27	32	01	7 18	15 59	9 52		
40	41 30	3 22	1 49	5 83	58	83	01	7 04	16 90	5 40		
41	42 63	4 89	1 78	5 65	1 35	1 26	20	6 38	14 35	6 77		
42	51 61	5 85	1 57	4 99	1 93	2 11	69	5 22	13 64	15 61		
43	51 53	5 32	1 26	4 76	2 96	2 70	1 70	4 97	12 98	14 88		
44	44 65	4 57	1 49	4 80	1 80	1 56	66	6 02	14 41	9 34		
45	53 75	8 42	2 25	7 29	2 45	2 62	1 26	5 61	12 72	11 13		
46	42 00	4 44	1 15	3 08	2 15	1 90	1 34	4 29	11 88	11 77		
47	53 40	7 53	2 04	5 10	3 41	3 39	2 16	6 31	12 90	10 56		
48	51 54	5 17	1 80	5 72	71	72	05	6 91	15 92	14 54		
49	64 59	9 72	1 90	7 17	4 15	3 91	2 91	5 06	13 54	16 23		
50	46 08	6 49	2 05	5 74	1 07	1 41	79	4 06	13 10	11 37		
51	44 52	6 24	2 26	5 33	1 94	1 84	70	5 81	12 84	7 56		
52	66 83	9 30	2 68	5 68	1 43	1 68	69	5 74	15 66	23 97		
53	38 92	4 31	1 68	3 82	1 41	1 99	70	4 74	14 45	5 82		
54	37 85	1 65	1 13	3 48	55	69	03	6 96	16 34	7 02		
55	38 13	2 51	1 36	2 65	1 34	1 09	12	6 04	13 90	9 12		
GROUP NO.												
	1	✓	✓	✓	✓					✓		
	2	✓	✓	✓		✓				✓		
	3	✓	✓	✓			✓			✓		
	4	✓	✓	✓	✓				✓	✓		
	5	✓	✓		✓					✓		
	6	✓	✓			✓				✓		
	7	✓	✓				✓			✓		
	8	✓	✓		✓				✓	✓		
	9	✓	✓		✓			✓		✓		
	10	✓	✓			✓			✓	✓		
	11	✓	✓			✓		✓		✓		
	12	✓	✓				✓		✓	✓		
	13	✓	✓				✓	✓		✓		
	14	✓	✓	✓	✓			✓		✓		
	15	✓	✓	✓			✓	✓		✓		
	16	✓	✓				✓		✓	✓		
	17	✓	✓				✓					
T	976.07	113.06	35 61	101.19	34 00	37 22	18 07	115.99	284.76	236.17		
M	48.853	5.653	1.780	5.060	1.700	1.861	0.904	5.799	14.238	11.808		
Variables and Units	(Dependent) MAY-JULY INFLOW 100,000 A.F.	DESCRIPTION OF EACH INDEPENDENT VARIABLE										
b coefficients												

a = , R^2 = , R = , $\bar{R}^2$ = , $\bar{R}$ = , $\bar{S}^2$ = , $\bar{S}$ =

* Year or Item Number

NOTE: Limit variables to 4 digits or less including 2 decimal places. Show 3 decimal places in mean.

MULTIPLE CORRELATION PRODUCT SUMS

FORM B

Problem *MAY 1 FORECAST OF MAY-JULY RUNOFF*

		Summations of Extensions of Departures from Means											c-matrix					Σ
		1	2	3					4			5						
		X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_1	C	C	C	C	C	
1	X_2	3.06	5.77	0.87	2.07	1.38	-0.08	-0.11	16.76			12.89	1					40.35
2	X_3	5.77	26.80	8.15	7.97	6.11	-0.64	-0.99	30.76			36.49		1				108.97
3	X_4	0.87	8.15	19.96	15.79	12.69	-9.98	-20.92	19.39			26.58			1			75.95
	X_5	2.07	7.97	15.79	16.57	13.17	-8.32	-16.23	28.83			27.51						
	X_6	1.38	6.11	12.69	13.17	11.67	-6.28	-11.59	28.85			22.31						
	X_7	-0.08	-0.64	-9.98	-8.32	-6.28	17.97	22.50	-11.16			-14.50						
	X_8	-0.11	-0.99	-20.92	-16.23	-11.59	22.50	47.82	11.09			-20.92						
4	X_9	16.76	30.76	19.39	28.83	28.85	-11.16	11.09	419.51			126.18				1		613.60
	X_{10}																	
	X_{11}																	
	X_1											89.43						

MULTIPLE CORRELATION COMPUTATIONS FOR $\bar{S}, \bar{R}, \sigma_b, a$

FORM C

PROBLEM: *MAY I FORECAST OF MAY-JULY RUNOFF*

(Numerals in headings refer to Columns. 1= unity)

0	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
GROUP	ELEMENTS	i *	M_i	c_{ii}	$\Sigma x_1 x_i$	14 x 6	$\bar{S}^2 = \frac{\Sigma x_1^2 - \Sigma 7}{n-m}$	$\bar{R}^2 = 1-\frac{(n-1)8}{\Sigma x_1^2}$	$\sigma_{b_1}^2 = 8 \times 5$	14 x 4	$\bar{S} = \sqrt{8}$	$\bar{R} = \sqrt{9}$	b_i	$\sigma_{b_i} = \sqrt{10}$	$a = M_1 - \Sigma 11$
1	M_1 5.653	2	1.780	0.6786	12.89								2.5084		
	Σx_1^2 89.43	3	5.060	0.0732	36.49								0.3883		
		4	1.700	0.0611	26.58								0.9387		
	n-m 15	9	11.808	0.0032	126.18								0.1287		
	n-1 19														
						$\Sigma 7$				$\Sigma 11$					
	M_1														
	Σx_1^2														
	n-m														
	n-1														
						$\Sigma 7$				$\Sigma 11$					
	M_1														
	Σx_1^2														
	n-m														
	n-1														
						$\Sigma 7$				$\Sigma 11$					

i = Subscript of independent variables.

PRELIMINARY MULTIPLE CORRELATIONS

Problem *MAY 1 FORECAST OF MAY-JULY RUNOFF*

FORM D

SYMBOLS

X's = independent variables
b = partial regression coefficients
σ_b = standard error of b

$\bar{R}$ = multiple correlation coefficient
$\bar{S}$ = standard error of estimate
a = constant term

General equation: $Y = a + b_2 X_2 + b_3 X_3 + --- b_n X_n$

		Variables										$\bar{R}$	$\bar{S}$	a
		X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}			
1	b	2.5084	0.3883	0.9387					0.1287			0.987	0.340	-3.893
	$\pm\sigma_b$	(2803	0.0921	0.0842					0.0200					
	b													
	$\pm\sigma_b$													
	b													
	$\pm\sigma_b$													
	b													
	$\pm\sigma_b$													
	b													
	$\pm\sigma_b$													
	b													
	$\pm\sigma_b$													
	b													
	$\pm\sigma_b$													
	b													
	$\pm\sigma_b$													

Descriptions of Independent variables

SIMPLE CORRELATION COMPUTATIONS FOR σ_b, $\bar{S}$, $\bar{R}$

(Follows Summation of Extensions of Departure from Means)

FORM E

PROBLEM:

(Figures in parentheses refer to Columns 1 = unity)

2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
(Subscript of Indep. Variable)	Σx_1^2	Σx_i^2	$\Sigma x_1 x_i$	b_i (5÷4)	n-m	$\bar{S}^2$ $\left(\frac{3-6\cdot5}{7}\right)$	n-1	R^2 $1-\left(\frac{9\cdot8}{3}\right)$	$\sigma_{b_i}^2$ (8÷4)	$\bar{R}$ $\sqrt{(10)}$	$\bar{S}$ $\sqrt{(8)}$	σ_{b_i} $\sqrt{(11)}$	M_1	M_i	a (15-6·16)