Statistical control in hydrologic forecasting

by

H. G. Wilm

RESEARCH NOTES

Issued by the

PACIFIC NORTHWEST FOREST AND RANGE EXPERIMENT STATION

0-61 Portland, Oregon January 1950

STATISTICAL CONTROL IN HYDROLOGIC FORECASTING

By

H. G. Wilm
Pacific Northwest Forest & Range Experiment Station
Division of Flood Control Surveys

With rapidly growing development and uses of water, a correspondingly great demand has developed for advance estimates of the volumes or rates of flow which are supplied by streams. Therefore much attention is being devoted to hydrologic forecasting, and numerous methods have been tested in efforts to make increasingly reliable estimates of future supplies.

This problem can be divided into two main parts: first, finding one or more factors that can be measured in advance of the runoff to be predicted and that are highly correlated with it; and second, working out a reliable method for expressing this correlation and for predicting the expected yields or discharge rates.

An excellent illustration of the first part of the problem is given by Garstka 1/, who describes work that he and other technicians of the Bureau of Reclamation have done in improving the precision of water yield forecasts on the Snake River watershed in Wyoming. Good results were obtained by weighting the water contents of snow measured at each snow course by the relative area of the watershed to which the course data should apply. From this work Garstka obtained an "elevation-weighted" average water content of snow for each year, which could be correlated with the water yield for the subsequent period from April through July. Such investigations are

1/ Garstka, W. U., Interpretation of snow surveys. Trans. American Geophysical Union 30(3): 412-20. 1949.

very important as the precision of forecasts can be progressively improved only by finding and measuring variables that are more highly correlated with runoff.

After some factor or factors have been measured, however, the second phase of the problem enters the picture. It becomes necessary for the forecaster to work out a reliable means of expressing the relation of water yields to the measured factor, and of obtaining forecasts of future yields through the use of this relation and new data on the correlated factor. Because the relation cannot be exact and the forecasts are therefore subject to error, it is also necessary to provide an estimate as to the magnitude of this error.

This statistical aspect of forecasting has been receiving increased attention. Although a number of workers still employ the graphic methods which are commonly used by hydrologists, others are applying the methods of regression and correlation analysis which have been presented by Ezekiel 2/ and others. When these are modified to fit the requirements imposed by relatively short-term records, they are well adapted to use in water-yield forecasting.

It is the purpose of this article to assemble and present available knowledge on statistical methods that can be applied to these problems, and particularly to small samples; to review briefly the assumptions that underlie and limit the methods; to outline an efficient technique for analyzing the required data and for making the forecasts, together with their errors and fiducial limits; and to give a detailed illustration and test of the technique, using actual data from the Snake River watershed. A useful part of this technique is the detection and adjustment of discrepancies in the data and shifts in the population, so as to provide improved accuracy and efficiency with a given set of variables. The discussion will be pointed mainly at those who are familiar with hydrologic forecasting and mathematics, but not with this kind of statistical analysis.

The Method Without Regression

To start examining the statistical techniques of forecasting, it may be desirable to imagine that we are just beginning the study of a watershed with a view to forecasting annual water yields from snow stored on the area. At this early stage in our work we have obtained only a short period of records from a gaging station at the mouth of the watershed. Also, we are installing a series of snow courses at which we can sample

---

2/ Ezekiel, Mordecai. Methods of Correlation Analysis. 2d edition, 531 pages. John Wiley and Sons, New York. 1941.

the amount of water stored in snow about April 1 of each year, before the spring runoff begins. This case may seem academic, as forecasts are not ordinarily made under such conditions. But it will serve as a basis for outlining the basic assumptions and principles that underlie hydrologic forecasting.

At this time we have no statistical control through which forecasts may be increased in precision; all we have is a set of data from which can be calculated an average water yield and an estimate of its variation from year to year. Hence any forecasts will necessarily be quite unprecise, though even now we can give the water user some idea as to what he may expect. As we do so, one or two fundamental assumptions need to be made. These should be kept in mind as we proceed with statistical analyses, although experience has shown that moderate deviations from their requirements have only minor effects on the reliability of forecasts.

For one thing, we assume that the available data may be considered a random sample of a larger population of data, and that the next year's runoff (as yet unmeasured) can be considered as drawn independently from this same population. In thinking of the sample set of n years, we say that its average water yield ($\bar{y}$) is a sample estimate of the true average (μ) for the indefinitely longer series of N years. Second, we assume that the deviations of the annual water yields around their average value may be taken as random and normally distributed, at least within reasonable limits. At first glance these assumptions may not seem well-founded, especially because we are dealing with a consecutive series of observations rather than a true random sample. It has been repeatedly shown,[3] however, that deviations of individual yearly records like these can be taken as random, and that ordinarily cyclic trends are not detectable within short series of records.

With the assumptions in mind, we can try the method by estimating the lowest and highest amounts of water that the water user may expect in the first season following our short series of data. We know that the range of past yields is not likely to be as great as for a longer record, but we can estimate the possible deviation of a new year from our sample average. This is derived from the standard deviation, calculated as follows:[4]

$$s_y = \pm\sqrt{Sy^2/n-1} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

3/ See, for example, "The Yield of Streams as a Measure of Climatic Fluctuations," by W. G. Hoyt and W. B. Langbein. Geographical Review, Volume 34, No. 2, Pages 218 to 234, 1944.

4/ For definitions of symbols, see Glossary of Symbols and Terms, at end of the paper.

Expressed in words, the standard deviation of Y equals the square root of the summed squared deviations ("sum of squares") of the sample observations around their mean, divided by the number of degrees of freedom.

In thinking about the possible magnitude of a new annual water yield, we realize that it may deviate in two ways from the sample average: the deviation of the new yield from the true population average, and the deviation of the sample average from this true average of N years. These deviations may be expressed as:

$$Y - \bar{y} = (\bar{y} - \mu) - (Y - \mu); \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

and their combined variance (squared standard deviation) is estimated by:

$$s_E^2 = s_y^2/n + s_y^2, \quad \text{or} \quad s_y^2(1/n + 1) \ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

where s_E^2 is the variance of the yield to be forecast.

The square root of this variance, the standard error of the forecast, provides us with a basis for estimating the fiducial limits between which the next year's water yield should lie, with any desired degree of likelihood. This estimate is given by multiplying the standard error by a tabulated value of Student's "t", obtained from any published table. 5/ A portion of a table of t is reproduced in Table 1. For samples of any size, t gives values corresponding to the conventional multiples of the standard deviation which are used in large samples. For a likelihood of "19:1," for example (probability 0.05 in the table), t is about 2 in large samples.

For the forecast we are about to make, suppose we select t to give a probability of 0.10. Then we can say that, unless next year's yield is so extreme that it is likely to occur less than once in 10 years, its magnitude should lie between the following limits:

5/ Snedecor, George W. "Statistical Methods," Table 3.8, page 65, 4th Edition, 485 pages. Iowa State College Press. 1946.

Table 1

Values of Student's t

Probability of a Larger Value of t, Sign Ignored

D/F	0.5	0.2	0.1	0.05	0.02	0.01	D/F
1	1.00	3.08	6.31	12.71	31.82	63.66	1
2	.82	1.89	2.92	4.30	6.96	9.92	2
3	.76	1.64	2.35	3.18	4.54	5.84	3
4	.74	1.53	2.13	2.78	3.75	4.60	4
5	.73	1.48	2.02	2.57	3.36	4.03	5
6	.72	1.44	1.94	2.45	3.14	3.71	6
7	.71	1.42	1.90	2.36	3.00	3.50	7
8	.71	1.40	1.86	2.31	2.90	3.36	8
9	.70	1.38	1.83	2.26	2.82	3.25	9
10	.70	1.37	1.81	2.23	2.76	3.17	10
11	.70	1.36	1.80	2.20	2.72	3.11	11
12	.70	1.36	1.78	2.18	2.68	3.06	12
13	.69	1.35	1.77	2.16	2.65	3.01	13
14	.69	1.34	1.76	2.14	2.62	2.98	14
15	.69	1.34	1.75	2.13	2.60	2.95	15
16	.69	1.34	1.75	2.12	2.58	2.92	16
18	.69	1.33	1.73	2.10	2.55	2.88	18
20	.69	1.32	1.72	2.09	2.53	2.84	20
21	.69	1.32	1.72	2.08	2.52	2.83	21
23	.68	1.32	1.71	2.07	2.50	2.81	23
25	.68	1.32	1.71	2.06	2.48	2.79	25
26	.68	1.32	1.71	2.06	2.48	2.78	26
28	.68	1.31	1.70	2.05	2.47	2.76	28
30	.68	1.31	1.70	2.04	2.46	2.75	30
∞	.675	1.282	1.645	1.960	2.326	2.576	∞

This is a shortened edition of Table 3.8 in "Statistical Methods," by Snedecor (see footnote 5).

$$\text{Fiducial limits} = \bar{y} \pm t_{.10} s_y \sqrt{1/n + 1} \quad \ldots\ldots\ldots\ldots (4)$$

To show how this method is used, let us test it on some of the data from Table 2 [6], looking only at Column 3: Water Yield, Inches. We may imagine also that at present only the years 1919-26 are available, and we want to estimate the fiducial limits between which the 1927 yield may lie. Based on the variation of these 8 years' data, we estimate:

$$s_y = \pm \sqrt{128.96/7} = \pm 4.29 \text{ inches}$$

Then the expected fiducial limits, with odds of 0.10 and on 7 degrees of freedom, are

$$\text{Limits} = \bar{y} \pm (1.90)(4.29)\sqrt{1.125}$$

$$= 15.6 \pm 8.6 = \text{from 7.0 to 24.2 inches of water.}$$

If desired, similar calculations may be made with a likelihood other than 0.10. If we want to be more conservative, we might decide to use an 0.05 probability: since $t_{0.05}$ on 7 degrees of freedom is 2.36, the expected yield should fall between 4.8 and 26.4 inches (that is, 15.6 ± 10.8). And at the 0.50 level of t, the corresponding range of the expected yield should be from 12.4 to 18.8 inches.

As it happens, the yield in 1927 was unusually high, exceeding the upper fiducial limit at the 0.10 level of t — an event that may happen about once per ten trials, in the long run. This occurrence may make you wonder how other years behaved; so, although it is not considered a safe procedure to forecast so far into the future from a small sample, let us look at the rest of the yield data in Column 3 of both Tables 2 and 3, and see how many fell outside the various fiducial boundaries as estimated from our 8-year sample. In the 19 years after 1926, 11 yields (58 percent) remained inside the limits specified on 1:1 probability; only 2 fell outside the 0.10 limits; and none exceeded or fell below the 0.05 limits. Thus, it seems that these data conform reasonably well to the mathematical model associated with our fundamental assumptions.

6/ Data in these tables obtained from "Interpretation of Snow Surveys," by W. U. Garstka (see footnote 1).

Table 2

Annual Water Yields and Water Content of Snow
Snake River Above Jackson Lake, Wyoming
(1919-1930)

(1) Year	(2) Water content of snow a/	(3) Water Yield b/
1919	23.1	10.5
1920	32.8	16.7
1921	31.8	18.2
1922	32.0	17.0
1923	30.4	16.3
1924	24.0	10.5
1925	39.5	23.1
1926	24.2	12.4
1927	52.6	24.9
1928	37.9	22.8
1929	30.6	14.1
1930	25.1	12.9
Total	383.8	199.4
Average	31.98	16.62
1931	12.4	

a/ Elevation-weighted average water content of snow on snow courses, about April 1 each year.

b/ Total yield of water from watershed above Jackson Lake, April through July each year.

Forecasting With the Aid of Regression

Linear Regression

Results like these give the forecaster confidence in statements based on statistical methods; although, as time goes on, he is sure to run into enough of the "unusual" years to keep him conservative. But even so, it is obvious that forecasts with such wide fiducial limits are not very useful, and that it is essential to increase their precision by any practical means. This is the reason for the perennial search of forecasters for variables that are highly correlated with water yield and can be measured ahead of time; then the probable magnitude of the new yield can be estimated with a smaller error.

In the Snake River data of Tables 2 and 3, such a variable is supplied by the elevation-weighted average water content of snow on the watershed, as sampled about April 1 of each year. Although for our preceding example we assumed that snow surveys were not being made until after 1926, actually the Bureau of Reclamation started these measurements in 1919 and has obtained them each year since then.

With these snow survey data available we can use a better forecasting tool: the regression of water yields on snow water content. In Garstka's article this relation is expressed graphically, and the fiducial limits shown as a "band of error." This method of analysis is widely used and gives satisfactory results where moderate precision is required. If the data fail seriously to meet the necessary statistical assumptions, it may be the soundest method. But in most cases mathematical control, using least-squares regression and associated statistics, is likely to give more reliable results.

Before we demonstrate this tool let us reconsider our preliminary assumptions (page 3) and see how they need to be modified. The first assumption is unaltered by the introduction of an independent variable (X), the water content of snow. The second, however, must be changed somewhat; we now assume that the deviations of water yields (Y) around any single value of X are random and normally distributed. Note that the values of X themselves need not meet this requirement; but any values of Y are assumed to be drawn in a random manner from all possible Y's corresponding to a particular value of X. Ordinarily this assumption and the first one may be considered fairly safe in the analysis of water-yield data over relatively short periods; the second, in fact, is likely to be safer than its analogue when regression is not employed. But there is another assumption that may occasionally be important; if it is seriously violated, the use of least-squares regression in this kind of analysis may give unreliable results. It is assumed that X is measured without error, either errors of measurement or so-called "biological variation." If this assumption is untrue, the slope of the regression is biased toward the horizontal, and the error of the forecast may be

Table 3

Annual Water Yields and Water Contents of Snow
Snake River Above Jackson Lake, Wyoming
(1931-1947)

(1) Year	(2) Water content of snow	(3) Observed water yield	(4) Forecast water yield	(5) Difference (3) - (4)	(6) Standard error of forecast	(7) Degrees of freedom	Student's t [a]
	Inches	Inches	Inches	Inches	Inches		
1931	12.4	8.8	5.9	+2.9	2.3	10	1.26
1932	35.1	17.4	18.3	-0.9	2.0	11	.46
1933	31.5	14.9	16.4	-1.5	1.9	12	.79
1934	21.1	10.5	11.2	-0.7	2.4	13	.29
1935	27.6	16.1	14.4	+1.7	1.8	14	.95
1936	30.7	18.9	16.0	+2.9	1.8	15	1.61*
1937	23.2	13.6	12.4	+1.2	1.9	16	.63
1938	28.6	20.0	15.2	+4.8	1.8	17	2.67**
1939	28.2	14.8	15.2	-0.4	2.1	18	.19
1940	19.2	13.6	11.4	+2.2	2.3	13	.96
1941	17.0	12.2	10.8	+1.4	2.1	13	.67
1942	19.1	14.5	12.0	+2.5	2.2	13	1.13
1943	40.1	25.2	21.0	+4.2	2.5	13	1.68*
1944	17.7	13.0	11.4	+1.6	2.5	13	.64
1945	24.5	15.1	14.8	+0.3	2.4	13	.12

a/ Asterisks indicate the range of probabilities which include these values of t:

No asterisk, probability greater than 0.2
* Probability 0.2 - 0.05
** Probability less than 0.05

increased.[7] Ordinarily, however, variations from even this assumption need not cause any material concern unless the errors in X become large in relation to the total variation in X.

Finally, the assumption is made that the variation in Y is not correlated with the magnitude of its mean, but remains essentially homogeneous throughout the range of the regression. A notable and fairly frequent exception to this assumption occurs when the variation in Y may properly be expressed as a percentage of $\bar{Y}$, and the "scatter" of plotted points around the regression line grows wider with increasing magnitudes of X and Y. In such cases it is advisable to transform all the data into logarithms and analyze them in this form.

As an accessory reservation, when working with small samples (say, 10 to 15 years), the forecaster should use care and conservatism in interpreting the results of his analyses and in depending on the calculated accuracy of the forecasts. This caution applies especially if the range of available data is relatively small, so that further sampling might disclose a wider range and perhaps a greater degree of variation. Much assistance can be provided, of course, by past experience and the characteristics of other sets of data from similar sources.

After thinking over all of these precautions, you may feel that it is hardly safe to make forecasts by statistical methods. Like any other sharp-edged tool, to be sure, these techniques may cause unfortunate consequences if they are misused; but when carefully and skillfully handled they will give precise results with a great deal of efficiency. Thus the answer to any doubts is not to leave this tool alone, but to handle it with the necessary skill.

With the fairly safe belief that our data fit the various assumptions reasonably well, we can now determine the relation between water yields and snow water contents, and can set up the technique for forecasting and estimating fiducial limits by the least-squares method. First, as in the first example we must consider that our short series of years (say 1919-1930 as presented in Table 2) provides only a sample estimate of the true average values for both X and Y, and of the true relation between these factors if it were to be calculated from the whole population of N pairs of values. We must also set up some logical hypothesis about the true shape of the regression line: whether it is straight, or what form of curve it should assume. Although we might argue theoretically that the regression of water yields on snow water content should be a curve that is concave upward, this shape is generally undetectable within the range of observed data.

7/ C. Eisenhart, The Interpretation of Certain Regression Methods, Annals of Mathematical Statistics, X, (2): 162-186, 1939.

Because this seems true of the Snake River data, it seems desirable to make use of the simplest hypothesis: that the true regression is satisfactorily fitted by a straight line, a sample estimate of which is provided by the familiar linear equation:

$$E = a + bX = \bar{y} + b(X - \bar{x}) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (5)$$

Associated with this sample regression equation are two kinds of error in its estimate of the true equation: the error of the "origin," a, and that of the regression coefficient, b. The variance of the origin is estimated by:

$$s_{\bar{y}}^2 = s_{y.x}^2/n, \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$$

where $s_{y.x}$ is the "standard error of estimate" of the regression equation. The square root of $s_{\bar{y}}^2$ may be stated as the "standard error of the mean of Y, when X is held at the mean of X." Correspondingly, the variance of the regression coefficient is estimated by:

$$s_b^2 = s_{y.x}^2/Sx^2, \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (7)$$

where S_x^2 is the summed squared deviations of the sample observations of X around their mean.

When these variances are combined, we can obtain an estimate of the variance of the sample regression line in estimating a point on the true line, for any single value of X:

$$s_{\hat{Y}}^2 = s_{y.x}^2(1/n + x^2/Sx^2), \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (8)$$

where $\hat{Y}$ is a sample estimate of the point at which an ordinate erected at any value of X (as X_1) will intersect the true regression line; and x is the deviation of the desired single value of X from the sample mean of X.

If you will glance again at our first forecasting method without the use of regression, you will note that the whole expression in Equation (8) is analogous to the first term (s_y^2/n) on the right side of Equation (3): it expresses the error of a sample average trend in estimating the true trend. Thus $\hat{Y}$ is in effect a value of $\bar{y}$ calculated for a single X. But if, instead of estimating a new mean $\bar{y}$, we wish to obtain a single new forecast of Y corresponding to a new value X, we know that it will vary about the sample mean $\bar{y}$ so that an additional element of error must be included in calculating fiducial limits. In order to complete our statement of the variance of a forecast, we must then add to Equation (8) a term which corresponds to s_y^2, the squared standard deviation. As you will realize, in our linear regression analysis this term is given by $s_{y.x}^2$. Thus the final expression for the variance of a forecast in a linear regression analysis may be stated as:

$$s_E^2 = s_{y.x}^2(1 + 1/n + x^2/Sx^2) \quad \ldots\ldots(9)$$

For those who are not familiar with these methods, very clear and readable discussions are given by Snedecor [8].

As in the analysis without regression, the standard error of the forecast ($\pm\sqrt{s_E^2}$) is multiplied by an appropriate value of t in order to provide an estimate of fiducial intervals around the forecast E:

$$\text{Limits} = E \pm ts_E \quad \ldots\ldots(10)$$

For a simple linear regression such as we have been discussing, the fiducial band around the regression line is shown in Figure 1. Because of the likelihood of error associated with the slope of the sample regression line in estimating the true regression, the lines bounding this band are not parallel to the sample regression line. On the contrary, the fiducial limits of a forecast annual water yield grow larger as the average snow water content deviates more and more from the mean; and they may be very large if forecasts are made from values of X that are close to or beyond the range of available sample data.

Now we can illustrate the use of statistical control with an example employing the figures on both water yields and water contents of snow in Table 2. The procedures outlined in Table 4 lead to the calculation of several statistics which are used in obtaining a forecast for 1931, and in estimating the fiducial limits of the forecast.

First of these is the regression coefficient, which is calculated from the corrected squares and products, and incorporated into a linear regression equation such as (5):

$$E = \bar{y} + b\,(X-\bar{x}) = a + bX$$

$$= 16.62 + 0.5477(X-31.98) = +0.5477X - 0.8954$$

By inserting the average result of the 1931 snow surveys into this equation (12.4 inches, see bottom of Table 1) we can obtain a forecast of the most probable water yield for this new year:

8/ See footnote 5. Read especially Chapters 6, 7, 12, 13, and 14, on regression, correlation, and covariance.

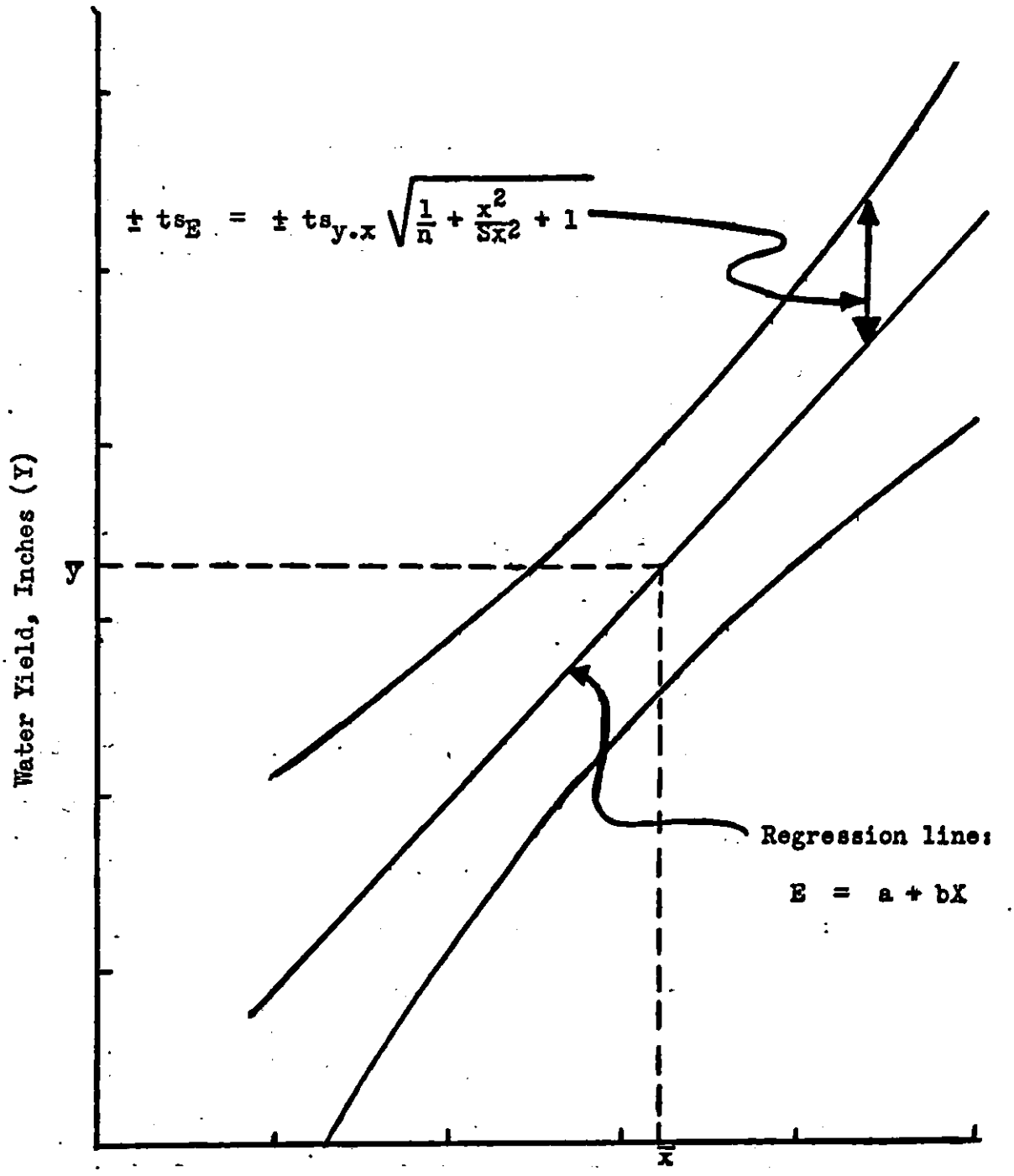

Figure 1. -- Linear regression of water yield on water content of snow, with fiducial limits

Table 4

Regression of Water Yields on Water Contents of Snow
Snake River (1919-1930)

I. Calculation of Sums of Squares			
	Squares and products		
Statistic	X^2	XY	Y^2
(a) Uncorrected sum of squares, etc.	13,043.66	6,798.33	3,577.76
(b) Correction term	12,275.20	6,377.48	3,313.36
(c) Corrected sum of squares, etc.	768.46	420.85	264.40

Procedure:

(a) Uncorrected sum of squares: $X_a^2 + X_b^2 + \ldots\ldots X_n^2 = SX^2$,

$= 23.1^2 + 32.8^2 + \ldots\ldots 25.1^2 = 13,043.66.$

And, for SXY $= (23.1)(10.5) + (32.8)(16.7) \ldots (25.1)(12.9)$

$= 6,798.33;$

with a similar procedure for SY^2, producing the figure 3,577.76

(b) Correction term (for X) $= (SX)^2/n = (383.8)^2/12 = 12,275.20;$

for XY, it is: $(SX)(SY)/n = (383.8)(199.4)/12 = 6,377.48;$

with a similar procedure for SY^2, producing the figure 3,313.36.

(c) Corrected sums of squares and products are obtained by subtracting Line (b) from Line (a), giving Sx^2, Sxy, and Sy^2; as, $Sx^2 = 768.46$

II. Calculation of Regression Statistics				
(1)	(2)	(3)	(4)	(5)
	Degrees of	Errors of estimate		Regression
$S\hat{y}^2$	freedom	Sum of Squares	Mean square	coefficient
230.48	10	33.92	3.392	+ 0.5477

(1) $S\hat{y}^2 = (Sxy)^2/ Sx^2 = (420.85)^2/768.46.$

(2) Degrees of freedom = n-m, where m is the number of variables; in simple linear regression, n - 2.

(3) Sum of squares for errors of estimate = $Sy^2 - S\hat{y}^2 = 264.40 - 230.48 = 33.92.$

(4) Mean square (errors of estimate) = Sum of squares/DF = 33.92/10 = 3.392

(5) Regression coefficient = $Sxy/Sx^2 = 420.85 / 768.46 = + 0.5477$

$$E = (0.5477)(12.4) - 0.8954 = 5.9 \text{ inches of water.}$$

It is still early in 1931 when this forecast is made, so of course we have no knowledge as to the actual water yield for this year; but we can estimate the probable range within which it should lie, on any desired odds. This procedure is begun by calculating the variance of the forecast, using Equation (9):

$$s_E^2 = s_{y.x}^2(1 + 1/n + x^2/Sx^2)$$

$$= 3.392 \left[1.083 + (12.4 - 32.0)^2/768.46\right]$$

$$= 3.392\ (1.583) = 5.3695$$

You will notice how much the variance of the forecast has been affected by the large deviation of the 1931 water content of snow from the sample mean. Compared to the variance of a forecast when X is held at the mean of X (as estimated by $s_{y.x}^2(1.083)$), this variance is about 46 percent greater.

From the calculated variance we can now obtain fiducial limits of the forecast (Equation 10) at any desired probability level. If we choose the 0.10 level and with 10 degrees of freedom, the limits are:

$$E \pm t_{.10} s_E = 5.9 \pm 1.81\sqrt{5.3695}\text{ , or from 1.7 to 10.1 inches}$$

of water. You might also like to figure the limits at the 0.50 level; these would be:

$$5.9 \pm (0.70)(2.32), \text{ or from 4.3 to 7.5 inches.}$$

Then you will want to convey this information to the water user, perhaps using a statement like this:

> "For 1931, the odds are even that the April-July yield of water will be between 4.3 and 7.5 inches. This might, of course, be an unusual year. Unless it is the kind that comes only about once in a decade, however, you will get more than 1.7 and less than 10.1 inches of water."

Perhaps you can work out a phrasing that is better suited to your needs. But the important thing is to tell the water user the real limits between which his water yield should lie, rather than to give him an "exact" figure which, sooner or later, is sure to weaken his confidence in your predicting ability. As to this particular forecast for 1931, obviously its precision still leaves something to be improved upon; but at least it is far better than the statement we could make without the aid of linear regression.

When the 1931 water yield has actually been measured (8.8 inches, in the top line of Table 3), we can see that it fell safely within the fiducial limits at the 0.10 level; it deviated only 2.9 inches from the forecast value. If you wish, you can calculate the probable frequency with which a deviation of this size is likely to occur, by calculating for this particular forecast:

$$t = (Y - E)/s_E = 2.9 / 2.32 = 1.25 \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (11)$$

By looking up t (on 10 D/F) in Table 1 we can see that this kind of deviation is likely to occur about once in four trials in the long run: the probability lies between 0.2 and 0.3.

To show how this method of forecasting worked out for each of a series of forecasts up to and including 1945, Table 3 gives the forecast value for each year, the difference between observed and forecast yields, the standard error of each forecast, and the value of t obtained by means of Equation (11). If we compare these t's with the tabular values for the associated degrees of freedom, we can see how the forecast deviations for these new years conformed to the theory. As shown by the asterisks, only 1 of the 15 forecasts fell outside the fiducial limits calculated at the 0.05 level, and three (or 20 percent) outside the 0.20 limits.

While examining the successive forecasts in Table 3, you may feel that we have again done something a little risky in making forecasts so far beyond the period included by our first regression. When you look at the "degrees of freedom" column, however, you will guess that up to 1939 each forecast was based on a new regression, containing all of the available observations up to date.

Progressive Development of the Regressions

The process of building up regressions by adding new observations does not involve any new principles; but its mechanics can be greatly simplified by following a special procedure. In the analytic technique outlined in Table 4 you will have noticed that we calculated the "sums of squares and products" of deviations around sample means by a method which may be unfamiliar. For a single short analysis such as this one, the method saves little effort as compared to the more conventional method of calculating and squaring (or multiplying) deviations directly. But it will help streamline the progressive building up of regressions; and this is likely to be important to forecasters who have to repeat these calculations for a number of separate watersheds each year.

The procedure outlined in Table 4 is expanded in Table 5, and may be summarized as follows:

(1) Obtain new sums for both X and Y by adding the 1931 (or $n_1 + 1$) data to the previous sums for 1919-1930.

(2) Calculate new correction terms from these sums.

(3) To the uncorrected sums of squares and products for n years, add the squares and product of X and Y for the one new year.

(4) Subtract the new correction terms from the values calculated in (3), and proceed as before with the computation of regression statistics, remembering that a degree of freedom is added with each new observation.

Especially when an electric calculator is available, this process is extremely rapid and simple, and can be run through in relatively few minutes per watershed once the data are ready for analysis. It may be repeated for each new year, as long as the data are accurate and you have no reason to suspect that the relation between water yields and snow is shifting in some way. With these reservations, each addition of data will tend to improve the regression, and therefore to make the forecasts a little more reliable.

Keeping Control Over the Forecasts

As these techniques for calculating water yield forecasts are carried through, you can exert a form of control over them somewhat like the "quality control" methods employed in manufacturing, and can make any necessary adjustments as a part of the process. Two features may be discussed here: the need to watch for and remove errors in the data which might become persistent, and the use of "moving" regressions to help compensate for population shifts.

Of these two controls, the first is relatively simple for the statistical analyst. As each forecast is compared with the subsequent observed water yield, he will naturally notice whether the deviation is larger than might be expected by chance. If a t-test indicates--as it did in the regression forecast computed for 1927--that such a deviation could be expected only rarely, it is wise to look for possible causes and see that they are removed before the next season's data are collected. In a factory the machinery may be stopped while the supervisors and technicians look for otherwise unsuspected changes in the manufacturing process that might have affected the product. Similarly,

Table 5

Building up the Regressions of Water Yields on Water Contents of Snow in Successive Years

I. Tabulation of Data

Period	Observations X	Observations Y	Sums X	Sums Y	Averages x̄	Averages ȳ	No. years n
1919-30			383.8	199.4	31.98	16.62	12
1919-31	12.4	8.8	396.2	208.2	30.47	16.02	13
-32	35.1	17.4	431.3	225.6	30.81	16.11	14
1919-37	23.2	13.6	565.4	299.6	29.76	15.77	19
-38	28.6	20.0	594.0	319.6	29.70	15.98	20
1925-39	28.2	14.8	448.1	245.2	29.87	16.35	15
1926-40	19.2	13.6	427.8	235.7	28.52	15.71	15
1930-44	17.7	13.0	376.6	226.4	25.11	15.09	15
1931-45	24.5	15.1	376.0	228.6	25.11	15.24	15

II. Regression Analysis (1919-1931)

Statistic	Squares and Products X^2	XY	Y^2
(a) Uncorrected SS, etc.	13,197.42	6,907.45	3,655.20
(b) Correction Term, CT	12,074.95	6,345.29	3,334.40
(c) Corrected SS, etc.	1,122.47	562.16	320.80

Procedure:

(a) Uncorrected SS = $SS_{1930} + X_{31}^2$ (or XY, or Y^2) = 13,043.66 + 153.76 = 13,197.42.

(b) Correction term = $(SX_{1930} + X_{31})^2/n_1 + 1$ = $(383.8 + 12.4)^2/13$ = 12,074.95

Correction term for XY = $(383.8 + 12.4)(199.4 + 8.8)/13$ = 6,345.29.

(c) Corrected SS, as usual, is the difference between (a) and (b).

III. Calculation of Regression Statistics (as in Table 3)

s_y^2	Degrees of freedom	Errors of estimate: Sum of squares	Errors of estimate: Mean square	Regression coefficient
281.54	11	39.26	3.569	+ 0.5008

the forecaster immediately examines the snow-course data and streamflow records. He may even take a look at the courses and gaging station themselves, and talk with their operators. If deviations from prescribed techniques are observed, their correction cannot help the past year's forecast; but it may do a great deal to keep future forecasts as reliable as possible. If, on the other hand, no changes in technique or other sources of error can be found, it can only be concluded that this peculiar year was one of the chance events which must come to every forecaster.

A more serious problem may arise if a real and progressive change develops in the relation of water yield to snow water content. The forecaster should keep a constant lookout for this kind of "population shift," so that he can take any necessary measures if it begins to be noticeable. The technique can be nicely illustrated by the Snake River data: apparently such a shift actually occurred in 1936 and continued from that time on. The shift in these records, as it happens, may not have been the result of climate or other uncontrollable changes, but is more likely due to a much simpler cause; one which might have been removed by 1938 or 1939 if the methods of statistical control had been available for its detection. As indicated in Garstka's paper, in 1936 the Snake River courses "were incorporated into the Westwide system of snow surveys coordinated by the Soil Conservation Service, at which time the snow courses were staked out in the standard manner." It seems likely, judging from the population shift which is apparent in the records, that some of the courses may have been altered in this process. Even though no trouble might have been expected to result from these alterations when they were made, the statistical forecaster would have detected the resulting shift before very long. Then any necessary readjustments of the courses could have been made immediately, and the only loss would have been a temporary decrease in the precision of the forecasts.

For our present purpose, however, let us suppose that an examination of the courses and gaging station was made when the shift was first observed, and that no cause was detected. Then we can only assume that there is a real change in the relation of water yields to snow, and must compensate for it as best we can.

But first you will want to know how the analyst can detect the onset of such a shift. The basis for his observations is shown in Table 6, which presents the several statistics that may be affected; a table like this should be kept currently for each watershed.

Except for the "Deviation" column, each line in the table is completed when the forecast is made; and of course that column is filled in after the actual yield has been recorded. As you go down through the table, one line at a time, imagine that the lines below are blank so that the information above is all that you have available. You will

notice several features of the data recorded up to 1935 or even 1936. First, the deviations as observed from forecast yields are reasonable in size and random in direction; none are larger than might easily be expected by chance. Second, the variance of estimate tends to remain roughly the same from year to year. And third, the constants in the regression equation vary only as one might expect when new random observations are added.

In 1936 the deviation of the observed yield from the forecast was a little larger, and both this year and 1937 are on the high side of the regression line. By themselves these facts would not tend to arouse suspicion, as the probability is still low; nor would the small but progressive rise in the regression constant a. If the analyst happened to know of the recent staking-out of the snow courses, these minor shifts might lead him to inquire into any changes in conditions which might have resulted. More likely, though, he would wait to see what another year or two might bring.

By 1938 an unusually large deviation is accompanied by a further rise in the regression constant--as a result of the series of positive deviations--and by 1939 the extra effect of this unusual deviation is expressed in a striking increase in the variance of estimate, as well as still another rise in the magnitude of a.

Let us imagine that by now the analyst has searched for any possible causes of these deviations; has failed to find any; and therefore must conclude that some uncontrollable shift is occurring in the relation of water yields to snow water contents. If it is a true population shift, his forecasts are likely to be affected by it for some time, even though he makes the best effort he can to remove its effects.

He is, however, able to take one useful form of action; he can attempt to keep his forecasting equation as nearly abreast of the population shift as possible. This is done by dropping off some of the earliest years of the snow course and water yield records, basing his regression only on more recent years. And from then on he employs a "moving" regression, dropping off the earliest year of his current regression series each time he adds a new year.

After some preliminary trials, for the 1939 forecast he decides that his best course is to drop off the records for 1919-24, inclusive; and, from 1939 on, to employ regressions based only on the most recent series of 15 years. Thus, the 1940 forecast is calculated from a regression equation based on the years 1925-39, and later years are estimated from similar 15-year series.

Now that we know what action the analyst took in 1939, we can look at the rest of Table 6 and observe the results. The effects of the population shift and of the control measures taken to offset it are evident

Table 6

Statistics employed in detecting shifts in the water yield-snow relation, Snake River

Year	Forecast yield inches	Deviation inches	Variance of estimate ($s_{y.x}^2$)	Constants in regression equation a	b
1931	5.9	+ 2.9	3.392[a]	- 0.8939	+ 0.5477[a]
1932	18.3	- 0.9	3.569	+ .7599	+ 0.5008
1933	16.4	- 1.5	3.337	.7874	.4973
1934	11.2	- 0.7	3.255	.7149	.4964
1935	14.4	+ 1.7	3.052	.5224	.5016
1936	16.0	+ 2.9*[b]	3.041	.7302	.4980
1937	12.4	+ 1.2	3.342	.8498	.4993
1938	15.2	+ 4.8**	3.218	1.087	.4934
1939	15.2	- 0.4	4.256	1.448	.4893
1940	11.4	+ 2.2 [d]	4.648	2.486	.4642[c]
1941	10.8	+ 1.4	4.486	3.571	.4256
1942	12.0	+ 2.5	4.399	4.220	.4094
1943	21.0	+ 4.2*	4.756	4.252	.4168
1944	11.4	+ 1.6	5.418	3.368	.4546
1945	14.8	+ 0.3	4.732	3.513	.4611
			3.981	3.312	.4736

a/ Variances and regression constants calculated up to 1939 are based on regressions of data from 1919 to the preceding year.

b/ The asterisks indicate the probable significance of these differences: * = probability less than 0.20; ** = probability less than 0.05 (compare Table 2).

c/ From 1940 through 1946, the variances and regression constants are based on moving regressions, each including 15 years' data up to and including the year preceding the forecast. The 1940 constants, for example, are based on the 1925-39 regression analysis.

d/ If the forecasts had been made from regressions including all data from 1919 to the year preceding each forecast, the deviations in this column would have been as follows, starting with 1940: + 2.8, + 2.1, + 3.2, + 4.4, + 2.1, and + 0.9.

in these statistics. First, of course, the variance of estimate jumps up in a disturbing manner. At least in part, this probably implies a bias resulting from the shift; but it also means that we must attribute less reliability to our forecasts than was previously possible. Next, the deviations remain positive in direction; there is a tendency for the yields to exceed the forecasts. And then, the regression constants exhibit a degree of instability which results from adding more and more values of X and Y which fall above the regression line. Finally, though, we find one good feature; the forecasts adhere more closely to the observed values than they would have if calculated from regressions based on the whole series of data from 1919 on (footnote "d", Table 6). This improvement is, of course, the object of using the moving regressions.

Because the shift in these data was abrupt rather than the slow trend which is more likely to occur in a true population shift, the non-random portions of these effects are probably more pronounced than ordinary. In any case, the data serve to show the symptoms which can be used in detecting this kind of shift and adjusting for it.

The mechanism of calculations in the moving-regression procedure is an extension of the method shown in Table 4, and can be carried through with almost equal facility. It is only necessary to subtract the oldest values of X and Y from the sums of both variables, at the same time that new values are added. Similarly, the uncorrected squared and products of the oldest values are subtracted from the sums of squares and products, in the same process with the addition of the newest values. With practice and the aid of an electric calculator, the whole procedure can be executed with surprising rapidity.

Careful use of these techniques for statistical control of the forecasting procedure, combined with close coordination between field men and statisticians, should provide the maximum degree of precision that can be gained from the combination of water yield data and a single correlated factor employed in a linear regression. Further precision may sometimes be gained by the use of additional factors, or by fitting curvilinear regressions if they are based on a sound hypothesis.

Curvilinear and Multiple Regressions in Forecasting

In an article of this kind it is not necessary to go into detail on the use of these more complex regressions in water yield forecasting, expecially because complete and interesting discussions are given by both Ezekiel and Snedecor (Footnotes 2 and 5). It may be desirable, though, to indicate the roles that may be played by curvilinear and multiple regressions and to show briefly how they may be extended from the simpler linear regression techniques, with a similar streamlining of procedures.

Considering first the use of curved regression lines, it is essential that there should exist a logical hypothesis for curvilinearity, which should also specify the probable shape of the curve in general

terms. As previously suggested, the theoretical relation of water yield to snow water content might be curvilinear, though not pronouncedly so; and this curve might be satisfactorily fitted by a quadratic equation of the form:

$$E = a + bX + cX^2 \quad \text{.....(12)}$$

In other cases the hypothetical curve might be of the logarithmic or exponential form; then the "best fit" might be provided by transforming the data for either X or Y, or both, to logarithms and fitting a straight line to the resulting data:

$$E = a + b(\log X), \text{ or} \quad \text{.....(13a)}$$

$$\log E = \log a + b(\log X), \text{ for example} \quad \text{.....(13b)}$$

Where no particular hypothesis as to the shape of the curve can be set up, it may be desirable to fit a quadratic equation to the data as suggested above. In this case it is easy, in the course of the analysis, to test the validity of the curvilinear hypothesis by statistical methods.

All that is necessary is to add to the analysis a set of values for a second independent variable (X_2), with each item for this variable supplied by squaring the yearly values for snow water content (X_1). Corrected squares and products are obtained for the various combinations of these X's and Y, and then the regression statistics are calculated by means of equations presented in detail by Snedecor (Chapters 13 and 14). As a result, you will obtain a sum of squares for errors of estimate corresponding to that supplied by the linear regression for the same data, but with one less degree of freedom. The difference between these two sums, on one degree of freedom, is a sum of squares attributable to the quadratic regression, and it can be tested for significance by the standard methods of variance analysis.

In interpreting the results, of course you will use the same caution that is required otherwise in dealing with sampling, especially with small samples. In this kind of test, your interpretation should be greatly assisted by a good logical basis for or against the hypothesis of curvilinearity and as to the shape of the curve. If, for example, such an hypothesis is logically untenable, you will probably not bother to make the test. And in any case, remember that about once in every 20 trials chance alone is likely to give you results that are "significant" at or beyond the 0.05 level.

The use of multiple regressions may often be more profitable than the fitting of curves. As our knowledge of variables that are correlated with water yield increases, we are likely to want to associate more and more independent variables with the one we need to forecast: variables such as total winter precipitation as well as the spring water content of snow; solar radiation during the early spring, when evaporation and sublimation losses may be expected to subtract important amounts of water from the stored snow; antecedent soil moisture, as a measure of deficits of water stored in the soil reservoir; stream discharge during the preceding autumn, as an index of antecedent conditions; and others which have been suggested and are being studied.

One variable which might be particularly valuable is the experience of some highly trained and acclimated technician who has worked on a watershed for some time, and has generated a "feel" for watershed conditions. As he collects data on streamflow, snow water contents, and other variables he may develop a current impression that, other measured variables being equal, this year should produce more or less water than usual. Hydrologists are well acquainted with this capacity of trained and experienced technicians, but it is not ordinarily thought possible to incorporate such information into a statistical analysis.

If, however, such a feel can be translated into numerical terms--as an adjustment percent, for example--it can be added to the analysis as one more independent variable, and tested for its significance. This may add materially to the precision of water yield forecasts, expecially because the observer can continually improve the usefulness of his own experience by watching his progressive and repeated successes and failures. The value of this variable will depend, of course, on the likelihood that the technician can remain in a single region long enough to make his estimates contribute materially to the regression. To retain maximum control over the information provided, such estimates should be kept as a separate variable rather than to be combined with the snow water content as some form of adjustment. By this means the statistician can keep a constant watch on the efficiency of the variable, and can help the technician gain experience and thereby add precision to the forecasts.

Whatever the variables are, they can be added to the regression equation in any desired number and combination of curved and linear relations, as long as enough degrees of freedom are left to provide a reliable estimate of the variance of the forecast. Ordinarily, for water yield data, at least 9 or 10 degrees of freedom are required after all deductions have been made for curvilinearity and independent variables.

As in the test for curvilinearity, the value of each independent variable can be tested by statistical methods. When, as a result of these tests, the forecaster is fairly certain that one or more of the variables being tried out does not contribute significantly to the precision of his forecasts, the variable may be dropped from further consideration; and, if desired, its place can be taken by a more efficient variable.

Another possible elaboration of the linear regression technique is to make preliminary forecasts from the data available early in the season, and then to improve these forecasts as the year goes on by adding later measurements as new independent variables. Where this method has been applied, the precision of the later forecasts is generally seen to be increased.

As more independent variables are added to the analysis, naturally the mechanical work of calculating the statistics becomes more laborious; but these calculations can be streamlined in a manner similar to that outlined for linear regressions. After the corrected squares and products are obtained for all the necessary combinations of X's and Y, the regression coefficients are calculated through the use of simultaneous equations. Their use in this kind of analysis is presented in detail by Snedecor, together with methods for calculating and employing the so-called "Gauss multipliers," required in obtaining the variance of a forecast in multiple regression. When the forecasting equation assumes the form

$$E = a + b_1X_1 + b_2X_2 + b_3X_3 \quad \text{.....(16)}$$

for example, the corresponding variance equation is:

$$s_E^2 = s_{y.x}^2(1 + 1/n + c_{11}x_1^2 + c_{22}x_2^2 + c_{33}x_3^2 + c_{12}x_1x_2 + c_{13}x_1x_3 + c_{23}x_2x_3) \quad \text{.....(17)}$$

where the various c's are the Gauss multipliers, sometimes termed "elements of the inverse matrix" and calculated by the method of determinants from the simultaneous equations; and the x's, as usual, are deviations from the means of X.

With this sketch of the somewhat more complex methods involved in curvilinear and multiple regression, we have completed an outline of statistical procedures that are adapted especially to water yield forecasting. It is sincerely hoped that techniques such as these will help forecasters make their predictions more useful and precise.

Thanks are due to the staff of the Bureau of Reclamation's Division of Hydrology in Denver, Colorado, for supplying the data for this series of analyses and for other aspects of their friendly cooperation in recent years. Special acknowledgment is due also to Dr. R. A. Fisher and his associates and followers, who developed the statistical theory and

methods on which these techniques are based; and to Dr. George W. Snedecor and his associates at Iowa State College, who have done a very great service in making statistical methods widely used and understood by workers in the biological sciences.

GLOSSARY OF SYMBOLS AND TERMS

Symbol	Definition
a	The "y-intercept;" in the regression equation for a simple straight line, $a = \bar{y} - b\bar{x}$.
b	The regression coefficient. In a linear equation, $b = Sxy/Sx^2$.
c	(as c_{11}, c_{12}, etc.) The Gauss multipliers, calculated by the method of determinants from a set of simultaneous equations.
CT	Correction term, used in analysis to correct the sum of squared observations (SX^2, SXY, and SY^2) to the "sum of squares" or "sum of products" of the deviations of observations around their sample average. $CT = (SX)^2/n$, SXY/n, or $(SY)^2/n$, in linear regression.
D/F	Degrees of freedom: n-m, where m is the number of variables (including Y) in an analysis, or the number of constants in the regression equation. Without regression, D/F = n - 1.
E	The forecast value, obtained with or without regression.
Fiducial Limits	The upper and lower limits, within which the actual value is likely to fall, as measured on either side of a forecast value. Limits = $E \pm ts_E$.
μ	The true average value for a population of observations, as water yields.
n	The number of observations in a sample set of data, as compared with ....
N	The total number in a larger set (or population) of data, of which the smaller set may be considered a sample.
r	Correlation coefficient (in simple correlation of two variables): $r = \pm\sqrt{(Sxy)^2/(Sx^2)(Sy^2)}$, or $\pm\sqrt{S\hat{y}^2/Sy^2}$
s_y	Standard deviation calculated from a set of sample observations; a sample estimate of σ, the population standard deviation. The equation for calculating s_y is: $s_y = \pm\sqrt{\frac{SY^2 - (SY)^2/n}{n-1}}$, or $\pm\sqrt{\frac{S(Y-\bar{y})^2}{n-1}}$

Symbol	Definition
$s_{\bar{y}}$	Standard error of a sample average of Y. Without regression, $s_{\bar{y}} = \pm\sqrt{s_y^2/n}$. With regression and for a fixed set of X's, $s_{\bar{y}} = \pm\sqrt{s_{y.x}^2/n}$.
$s_{y.x}$	Standard error of estimate, expressing the variation of observations around the regression line. In a linear equation, $s_{y.x} = \pm\sqrt{Sd_{y.x}^2/n-2}$, or $\pm\sqrt{Sy^2 - S\hat{y}^2/n-2}$.
s_b	Standard error of the regression coefficient. In a linear equation, $s_b = \pm\sqrt{s_{y.x}^2/Sx^2}$
s_E	Standard error of a forecast. Without regression, $s_E = \pm s_y\sqrt{1 + 1/n;}$ with regression $s_E = \pm s_{y.x}\sqrt{1 + 1/n + x^2/Sx^2}$
$s_{\hat{Y}}$	Standard error of a point on the sample regression line in estimating the true population value for any given value of X: $s_{\hat{Y}} = \pm s_{y.x}\sqrt{1/n + x^2/Sx^2}$
S	A symbol meaning "sum of."
SS	A sum of squares or products, as SX^2, Sx^2, SXY, Sxy, etc.
$S\hat{y}^2$	Sum of squares attributable to regression. In a linear regression, $S\hat{y}^2 = (Sxy)^2/Sx^2$
$Sd_{y.x}^2$	Sum of squares of deviations from regression = $Sy^2 - S\hat{y}^2$
t	Any single value of student's distribution of t ; for a complete table, see "Statistical Methods," by George W. Snedecor, 4th Edition, Page 65. Iowa State College Press, 1946.
X, Y	Any sample observation of an independent variable (X) or dependent variable (Y).

Symbol	Definition
$\bar{X}$, $\bar{Y}$	The average of a series of sample observations of X and Y.
x, y	The deviation of any single sample value of X or Y from the sample average.
Variance	Squared standard deviation.