Chapter 9
Statistical Analyses

9-1. Overview

Statistical methods are most frequently applied in snow hydrology for water supply forecasting, where measurements of snow and other variables are used to predict spring-summer snowmelt runoff. Because of this widespread practice, the topic will be covered in this chapter as a somewhat special application that is unique to snow hydrology. Aside from this application, statistical methods are also used in other aspects of snow hydrology in much the same way they are used in general hydrology. Frequency methods are used for determining extreme values of SWE or other parameters for design flood or low-flow analyses; multiple regression is employed for regional analyses of various kinds; and stochastic methodologies are used in long-term forecasting. Because they are not especially unique to snow hydrology, however, they will not be described in this manual except in passing. The general principles for applying statistical techniques in hydrology practice are covered in EM 1110-2-1415.

9-2. Data Analysis

Analyzing data for statistical as well as conceptual modeling in snow hydrology requires additional considerations owing to the nature of the environment and data involved. Some factors involved are as follows:

a. Snow data sampling is sometimes not consistent over a period of record. Sampling techniques have changed (e.g., manual to snow pillow) and station sites are sometimes moved.

b. Snow data often have relatively short periods of record compared with precipitation data.

c. Precipitation monitoring is more difficult in a mountainous environment involving snow. Unattended stations, which are subject to capping, are often used and are generally less accurate in measuring short-duration incremental changes.

d. Estimating SWE from precipitation stations may be feasible for high-elevation areas, but it is questionable for areas subject to rain during the winter.

e. Orographic effects and sparse gauge density make it difficult to estimate missing data or area-mean quantities.

Given the above, extra care should be taken in preparing data for use. Double-mass analysis is recommended to evaluate the consistency of the record, and visual or computer screening should be employed to check temporal consistency. It is a common practice to correct older snow-course data to be consistent with more recent snow-pillow records using correlation, if a sufficient overlapping record exists. Cumulative precipitation data sometimes can be used to supplement or simulate high-elevation snow data if a station-to-station correlation exists.

9-3. Frequency Analysis

Frequency analysis in a snow environment is likely to be done on precipitation, SWE, and, perhaps, temperature records. The PMF study described in Chapter 10 employed an extensive meteorological analysis that used depth-duration frequency curves for numerous precipitation stations in the basin being analyzed. Normal annual precipitation (NAP) maps were employed to convert station frequencies to areal representations. Precipitation analysis procedures are described in EM 1110-2-1415. Frequency analysis of SWE data should generally employ the same procedures as those used for precipitation data.

9-4. Water-Supply Forecasting

Water-supply forecasting is the long-term prediction of runoff volume of a specified duration. This term comes from the practice, originating in the 1930s in the western United States, of sampling the winter accumulation of snow to provide an index of runoff in the succeeding spring. Over the years, this basic methodology has evolved into an important and widespread practice that is used for crop management, irrigation planning, flood warning, and reservoir operations. An extensive network of automated snow-monitoring stations, called SNOTEL, have been set up

by the Natural Resource Conservation Service (NRCS) for this purpose, and many agencies process data and coordinate forecasts, such as the NWS, NRCS, USACE, and Bureau of Reclamation, as well as numerous State agencies and electrical utilities. In the West, the NWS and NRCS publish forecasts for over 600 points that appear in the Basin Outlook Reports published by the NRCS and in the Water Supply Outlook for the Western United States, issued jointly by the NWS and NRCS. In California, forecasts are prepared by the State Department of Water Resources, and in the Northeast, NWS publishes water-supply forecasts for the public.

a. Forecasts are usually expressed in terms of a volume of runoff during the months that have operational importance, i.e., April through August, March through July, etc. Winter runoff can also be included to produce January through July forecasts that are important in hydroelectric operations in the Northwest. Traditionally, forecasts have been produced once each month, beginning in January, immediately following measurements of snow and precipitation made on or near the first of the month. In recent years, more frequent forecasting has been made possible by automated hydromet systems.

b. Water-supply forecasts have typically used classic multiple linear regression techniques that incorporate two to five independent variables, as described below. An alternative to the use of multiple regression has been instituted by several agencies in recent years and shows promise as a viable technique for long-range volumetric forecasting. Termed Extended Streamflow Prediction (ESP) by the NWS, this methodology employs continuous simulation models to generate alternative streamflow time series, each reflecting the current state of the basin's condition (snowpack, soil moisture, etc.), combined with future weather conditions from a given historical year. The resulting traces of possible future alternative streamflow are processed as a data sample for statistical analysis. Chapter 10 has further discussion of this approach.

9-5. Multiple Regression Forecast Models

a. Basic equation. The multiple linear regression approach in water supply forecasting uses an equation of the form:

$$Y = a + b_1BF + b_2FP + b_3WP + b_4S + b_5SP \quad (9\text{-}1)$$

where

Y = seasonal streamflow volume

a = regression intercept

b_i = regression coefficients

BF = base-flow index

FP = fall-precipitation index

WP = winter-precipitation index

S = snow-water-equivalent index

SP = spring-precipitation index

(1) The base-flow index is usually the streamflow volume during the fall or early winter, e.g., October-December or November-January. The fall-precipitation index is a sum or weighted sum of monthly precipitation at one or more sites for the fall, e.g., September-November or October-December. The fall-precipitation index and the base-flow index are surrogates for soil moisture. The winter-precipitation index is the cumulative precipitation recorded for that season, say November-March. The snow index is a sum or weighted sum of SWE at several sites for the month usually having the maximum snow accumulation for the season; this is typically April, although it can be March or May. The spring-precipitation index is the same as the winter-precipitation index except for the spring period, e.g., April-June. Not all procedures necessarily use all the variables described above, but as a minimum, winter-snow and precipitation indexes,

a spring-precipitation index, and a fall-soil-moisture index are usually employed.

(2) In some areas of the northwest and southwest United States, another independent variable, the Southern Oscillation Index (SOI), improves the performance of water supply forecasting equations, especially in the early winter before the majority of snow has accumulated. The SOI, an indicator of the El Nino phenomenon, has been shown to be a moderate but significant predictor of winter precipitation and snowpack, with a lead time of as much as 6 months (Koch and Redmond 1991). Historical records of SOI are available, and the index is reported in a timely enough way to be usable in an operational setting.

b. Regression model development. Equation development traditionally begins with an analysis of the station data, employing judgment as to the whether the station should be included in the equation and what the station weighting should be (Hanneford 1993). Such factors as the station's degree of independent correlation with runoff, its location and elevation, and the consistency and viability of past and future data reporting are considered. The station's data are weighted to establish its relative influence in the equation. Remember, in this process relative weighting already exists by virtue of each station's natural mean and variance. A stepwise regression program is then used to select predictor variables and compute the regression equation. At least 15 years of data are necessary for reasonable forecast accuracy. Figure 9-1 is an example of a forecast procedure, laid out in a form that is used operationally in preparing it.

(1) An alternative to the stepwise method of equation development is employing principal components regression. This technique, described by Garen (1992), is used to eliminate aggregating weighted data observations into indices, to address the technical problem of variable intercorrelation, and to more rigorously establish an optimal solution for a given set of data. With it, the independent data are restructured into a equal number of uncorrelated variables via a linear transformation. Each new variable (principal component) is a different linear combination of all the original variables. The new variables are regressed, and variables that have the greatest influence in explaining the variance are selected. These can then be inverted, so that the coefficients are expressed in terms of the original variables. If there was a high degree of inter-correlation in the original data set, this method will result in fewer variables, thereby reducing the loss in degrees of freedom.

(2) With the principal components method eliminating the subjective selection and grouping of data stations for independent variables, a more systematic way of finding the near-optimal combination of predictor variables becomes feasible. Since the number of possible combinations is immense, a computer optimization procedure is required. Garen (1992) has developed an iterative algorithm that appears to be practical and is effective in identifying the strongest variable and constructing a near-optimal model.

(3) One fundamental consideration in developing a multiple regression forecast equation is how to handle precipitation and snowfall that occur after the date of the forecast. Two alternative methods have been used:

- Develop one equation for the season after all data are known; then, at the time of the forecast, use a median or average value to estimate that part of the input that is yet to occur.

- Develop separate equations for each forecast (usually one per month), using only the data known up to that point.

The former method has the advantage of greater stability from month to month and perhaps an advantage in allowing intuitive judgment of the effects of precipitation being above or below normal. However, it has been shown (Garen 1992) that a loss in accuracy results from this method and that it is less rigorous statistically than the second alternative of using separate equations.

9-6. Assessment of Regression Model Accuracy

Once a multiple regression model has been developed, it is necessary to evaluate its ability to represent the

COMPUTATION FORM

FORECASTING RUNOFF FROM LIBBY LOCAL SUBAREA YEAR _____

Apr - Aug Runoff in Inches = 0.070 (WP) + 0.205 (SP) + 0.047 (SWE) + 0.710 (FRO) - 4.794

1. Forecast Date			1 Jan	1 Feb	1 Mar	1 Apr	1 May	1 Jun		
		Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	
2. FALL RUNOFF (FRO)										
3. Observed Runoff, Inches--(Observed Libby Inflow - Ft. Steele Observed)		____	____							
4. Sum October + November Runoff, Inches				____						
5. Line 4 X 0.710				____	____	____	____	____	____	
6. WINTER PRECIPITATION (WP)	Weight									
7. Elko, B.C.	1.00	____	____	____	____	____	____			
8. Fernie, B.C.	1.00	____	____	____	____	____	____			
9. Fortine 1 N, MT.	1.00	____	____	____	____	____	____			
10. Libby R.S. 1 NE, MT.	1.00	____	____	____	____	____	____			
11. Bonners Ferry 1 SW, ID.	1.00	____	____	____	____	____	____			
12. Polebridge, MT.	1.00	____	____	____	____	____	____			
13. Sum Precipitation by Month (Also Equals Sum of Weighted Precip.)		____	____	____	____	____	____			
14. Sum Precipitation 1 Oct for Forecast Date		____	____							
15. Normal Subsequent Precipitation				43.89	23.58	10.01	0			
16. Sum Winter Precipitation (Oct thru Apr)				____	____	____	____			
17. Line 16 X 0.070				____	____	____	____	____	____	
18. SNOW WATER EQUIVALENT (SWE)	Weight									
19. Sullivan Mine, B.C.	1.00				____	____	____			
20. New Fernie, B.C.	1.00				____	____	____			
21. Red Mtn., MT.	1.00				____	____	____			
22. Kimberly, B.C.	1.00				____	____	____			
23. Weasel Divide, MT.	1.00				____	____	____			
24. Morrissey Ridge, B.C.	0.50				____	____	____			
25. Sum of Weighted SWE by Month					____	____	____			
26. Normal Subsequent SWE					32.58	12.26	0			
27. Sum (Equals 1 Apr SWE) For 1 Jan Only, SWE = 1.191 X Line 16				____	____	____	____			
28. Line 27 X 0.047				____	____	____	____	____	____	
29. SPRING PRECIPITATION (SP)	Weight (Apr & May)									
30. Fortine 1 N, MT.	1.00							____	____	
31. Porthill, ID	1.00							____	____	
32. Kaslo, B.C.	1.00							____	____	
33. Whitefish 5 NW, MT.	1.00							____	____	
34. Sum Spring Precipitation by month								____	____	
35. Accumulated Sum Spring Precipitation								____	____	
36. Normal Subsequent Precipitation (Weighted)				24.87	24.87	24.87	24.87	19.23	11.68	
37. Sum				24.87	24.87	24.87	24.87			
38. Line 37 X 0.205				5.098	5.098	5.098	5.098	____	____	
39. EQUATION CONSTANT				-4.794	-4.794	-4.794	-4.794	-4.794	-4.794	
40. Forecast Apr-Aug Runoff, Inches (Sum of Lines 5, 17, 28, 38, and 39)				____	____	____	____	____	____	
41. Forecast Apr-Aug Runoff KAF = Line 40 X 251.732				____	____	____	____	____	____	

REVISED NOV 78

Figure 9-1. Water-supply forecast procedure using multiple linear regression

observed data and to assess its accuracy for use as a forecasting tool. This requires an understanding of how to interpret and use error statistics properly, both as a forecast procedure is being developed, and also as it is being executed in a forecasting situation. An in-depth discussion of this subject is beyond the scope of this manual, but a summary discussion of several analysis methods that are often used in practice will be presented. There is generally no lack of error statistics available for the analyst who is using modern statistical computer programs. The problem in practice generally lies in understanding what the error factor means, and in applying it meaningfully in forecasting or analysis. Further discussion on this topic can be found in EM 1110-2-1415, as well as in numerous textbooks and manuals.

a. Evaluation criteria. There are several criteria that are commonly used to evaluate multiple regression models (McCuen 1985):

(1) Rationality of the coefficients.

(2) Relative importance of the predictor variables.

(3) Characteristics of the residuals.

(4) Coefficient of multiple determination.

(5) Standard error of the estimate.

Coefficient rationality is determined by subjective inspection, by substituting possible values for variables and noting the results in the dependent variable. Basic checks might include the following:

(1) Is the change in forecast logical when a predictor variable is changed in a certain direction?

(2) Is the forecast reasonable when variable extremes are encountered?

A further check of rationality is to examine the relative importance of the predictor variables. This may be subjectively evaluated as above, or analytical procedures can be used. If a certain variable is of little consequence in determining the prediction, it might be deleted from the equation for simplification. On the other hand, if the variable should be more influential than it is showing, then a reexamination of the model is necessary.

b. Analysis of residuals. In correlation analysis, the residual is the unexplained difference between the predicted and observed value of the independent variable, as illustrated in Figure 9-2. By definition, through the application of the least squares objective function, the sum of the residuals must equal zero. However, this does not guarantee that the model is not biased. If, for instance, the residuals tend to be positive for low values of X but negative for high values of X, then bias exists and a nonlinear model may need to be used. Plots of residuals can be made in various ways to check the validity of the model. A plot of residuals as a function of the dependent variables would display the bias as a function of observation magnitude, while a probability plot of the residuals might help verify the assumption that they are normally distributed in the Y direction.

c. Coefficient of multiple determination (R^2). The coefficient of multiple determination is the proportion of the variance of the dependent variable that is explained by the regression equation. A coefficient of determination of 0.25 means that 25 percent of the variance of the Y variable about its mean is accounted for and 75 percent is not explained by the regression equation. The range of R^2 is between 0 and 1.0, with the value of 0 indicating that Y is not related to any of the predictor variables. In general, this statistic provides a relative measure of the accuracy of the equation in making future predictions—assuming, of course, that the data sample is representative of the total population.

d. Standard error of estimate. The standard error of estimate is the standard deviation of the residuals, computed as the square root of the sum of the squares of the errors divided by the degrees of freedom (df). By definition, assuming a normal distribution of the residuals, two-thirds of the estimates will fall within plus or minus one standard error; 16 percent will be above the mean plus one standard error, and 16 percent will be fall below the mean minus one standard error. This is illustrated in Figure 9-2. Other

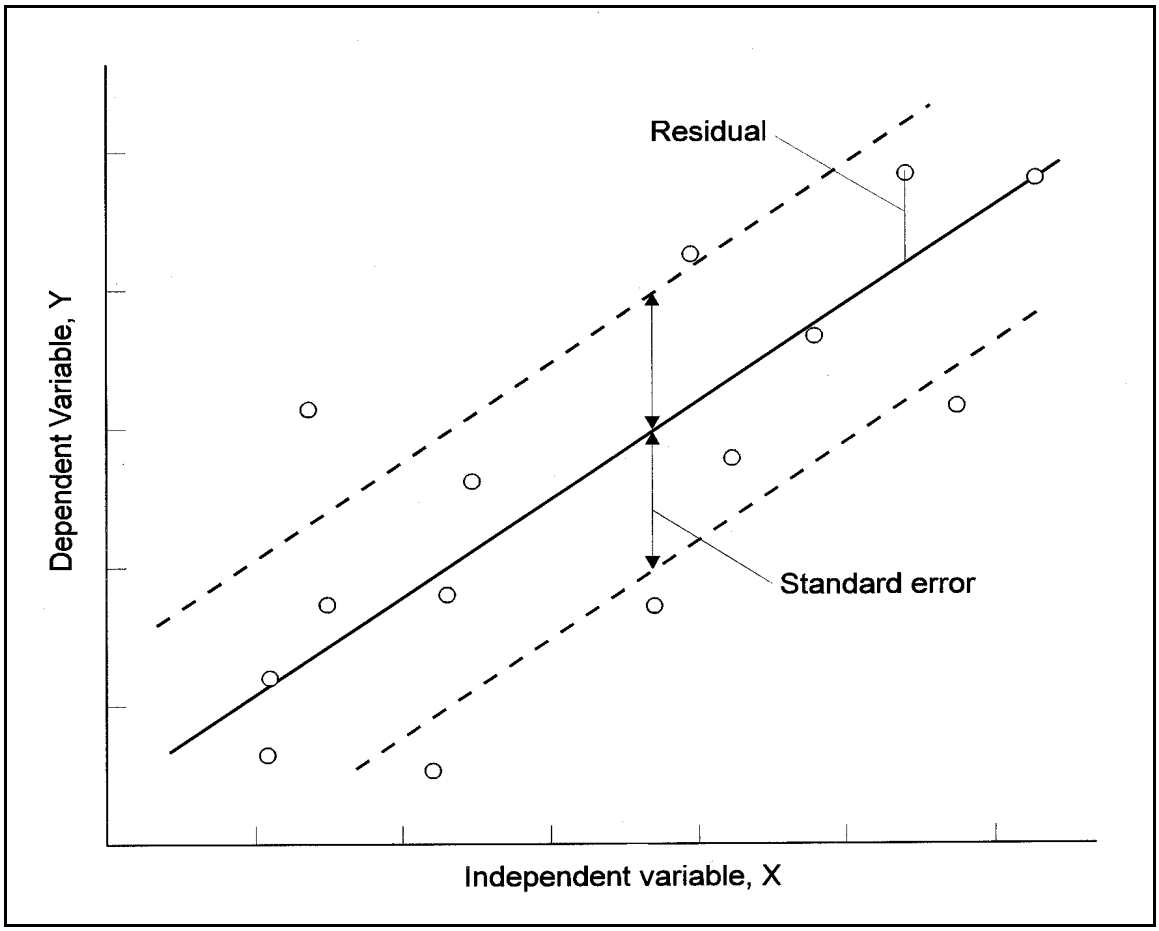

Figure 9-2. Correlation error analysis

cumulative normal curve found in statistical reference books.

(1) A problem frequently encountered in water-supply-forecasting practice is properly accounting for the value of degrees of freedom. The value of the degrees of freedom is obtained by subtracting the number of variables (independent and dependent) from the number of data points (years) defining the relationship. It is common practice to use a df equal to the number of major variables—snow, precipitation, etc. Yet, these variables may in fact have been made up of a number of stations that have been subjectively selected and weighted. In reality, the loss of degrees of freedom may be higher than the number of nominal variables contained in the equation, and a plot such as Figure 9-2 may be optimistically portraying the ability of this forecast to perform in the "real world" of actual future forecasts. This has been borne out in general by comparing with the cross-validation technique described below.

(2) In recent years, a more realistic portrayal of forecast accuracy in an actual forecasting situation has been obtained by using the cross-validation or "jack-knife" procedure. Here, one observation is removed from the data set, and the regression coefficients are calculated. These coefficients are used to predict the dependent variable for the withheld observation. The withheld data are returned and the next observation is removed. This process is repeated until a "forecast" has been made for all of the observations, using coefficients that do not reflect that data. A standard error is then calculated from these "forecasts." Comparison of error estimates using this method with traditionally computed standard errors shows that the traditional errors tend to underestimate the more rigorously computed standard errors.