IMPROVED TECHNIQUES IN REGRESSION-BASED STREAMFLOW VOLUME FORECASTING

By David C. Garen[1]

ABSTRACT: Although multiple linear regression has been used for many years to predict seasonal streamflow volumes, typical practice has not realized the maximum accuracy obtainable from regression. Several techniques can help provide superior forecast accuracy using regression models: (1) Using only data known at forecast time; (2) principal components regression; (3) cross validation; and (4) systematic searching for optimal or near-optimal combinations of variables. Using no future data requires that a separate equation be used each month that forecasts are made rather than using a single equation throughout the forecast season. Consistency of month-to-month forecasts can be obtained by judicious selection of variables to maintain a high degree of similarity in the monthly equations. Results for the South Fork Boise River at Anderson Ranch Dam and other basins in the West indicate that these new regression procedures can give substantial improvements in forecast accuracy over existing procedures without sacrificing month-to-month forecast consistency.

INTRODUCTION

The Soil Conservation Service (SCS) of the U.S. Department of Agriculture and the National Weather Service of the U.S. Department of Commerce are responsible for producing and publishing seasonal streamflow volume forecasts once per month from January through May (in some areas through June) for approximately 600 points throughout the western United States. These appear in "Basin Outlook Reports" (previously called "Water Supply Outlooks"), issued by each Western state office of the SCS and in "Water Supply Outlook for the Western United States," issued jointly by the SCS and the National Weather Service. The period ("season") that is forecast typically covers the months when most of the snowmelt and summer runoff occurs (e.g., March–July, April–September). Virtually all of these forecasts are generated using multiple linear regression models. Input variables to these models are generally snow water equivalent at one or more snowcourses or SCS SNOTEL (snow telemetry) sites (Barton and Burke 1977; Rallison 1981; Crook 1984), monthly precipitation at one or more sites, and streamflow for previous months at the forecast point. Typical practice has relied on standard multiple regression either using data for individual sites and individual months or using indexes as independent variables. These indexes combine data for sites and/or months into weighted sums. Construction of indexes and selection of sites to use have been based on data quality, correlation analyses, conceptual appropriateness, professional judgment, and trial and error. Several references describe the conventional use of regression techniques in water supply forecasting, including Schermerhorn and Barton (1968), Soil Conservation Service ("Snow" 1972), Zuzel and Cox (1978), McCuen et al. (1979), and Stedinger et al. (1988).

Although multiple linear regression has been used for many years to

[1]Hydro., USDA, Soil Conservation Service, Water Supply Forecasting Staff, 511 NW Broadway, Room 248, Portland, OR 97209-3489.

Note. Discussion open until April 1, 1993. To extend the closing date one month, a written request must be filed with the ASCE Manager of Journals. The manuscript for this paper was submitted for review and possible publication on September 6, 1991. This paper is part of the ***Journal of Water Resources Planning and Management***, Vol. 118, No. 6, November/December, 1992. ©ASCE, ISSN 0733-9496/92/0006-0654/ $1.00 + $.15 per page. Paper No. 2454.

predict seasonal streamflow volumes, the results of the present study indicate that typical practice has not realized the maximum accuracy obtainable from regression. Several techniques can help provide superior forecast accuracy using regression models: (1) Basing the regression model only on data known at forecast time (no future data); (2) principal components regression; (3) cross validation; and (4) systematic searching for optimal or near-optimal combinations of variables.

Seeking maximum accuracy in regression-based forecasts is useful for several reasons. First, it is (and will remain for some time to come) the mainstay of seasonal streamflow volume forecasting. Conceptual watershed models, which can produce forecasts of the daily hydrograph as well as the seasonal or monthly runoff volumes, hold promise as an improved forecasting technique (Pearson 1974; Twedt et al. 1977; Kuehl 1979; Speers and Versteeg 1982; Druce 1984; Day 1985), but they are used now only on a limited basis, and their widespread use is still many years away. Second, in some cases, regression may remain the forecast method of choice if adequate water management decisions can be made without more detailed hydrologic information or if more complex methods (such as conceptual models) are not sufficiently accurate. Finally, regression forecasts provide a baseline level of accuracy against which to test conceptual models. Of course, it would be an unfair test to compare conceptual model forecasts with non-optimal regression forecasts.

The aforementioned regression techniques are discussed in turn, then some example results are given.

Future Variables

Usage of Future Variables and Forecast Accuracy

Typical practice in water supply forecasting has often included variables in multiple regression equations that describe future snow accumulation or precipitation and hence are unknown at the time the forecast is made (Schermerhorn and Barton 1968; Snow 1972; Stedinger et al. 1988). The practice has been to calibrate a single equation for a given forecast period using all data through the end of the forecast period. For example, an April–July equation would be calibrated using data through July, which often included precipitation through June or July, and snow water equivalent for the maximum accumulation of the year (typically March or April). This single equation was used in all months that forecasts were made. If a precipitation or streamflow variable was in the future at forecast time, long-term averages were used. If a snow water equivalent variable was in the future, the observed value at forecast time was extrapolated to the target month by adding to it the average accumulation in the intervening months.

As shown by Stedinger et al. (1988), some work by Koch (1990), and the results of the present study, the use of future variables and the substitution of averages can degrade forecast accuracy. An equation calibrated with all input data known is optimal only when all of those data are known; it is no longer an optimal forecaster when some of the input data are unknown. Improvements in forecast accuracy by not including future variables can be substantial, especially early in the forecasting season. The results from one example basin will be given in a subsequent section.

Monthly Equations and Forecast Consistency

If we reject the use of a single equation to avoid using future variables, then we must use a different equation, containing only variables known at

forecast time, whenever a forecast must be made. For routine monthly water supply forecasting, this means each forecast point and period must have a separate equation for each month that forecasting is done.

A concern often raised about the use of separate monthly equations is whether the forecasts will show unexplainable jumps up and down from month to month due solely to the different variables and coefficients in the regression models. One expects forecasts to change for hydrologic reasons, but it is undesirable to have changes due to statistical noise. Such instability in forecasts causes consternation among water managers. Forecasters can use judgment to adjust the equation predictions to smooth out undesirable month-to-month fluctuations, but it is preferable to have equations that give forecasts requiring little adjustment.

Although forecast consistency is difficult to define precisely, it is possible to make some, perhaps crude, measures of it. One possible measure of forecast consistency is the average absolute value of monthly forecast changes. This is computed by taking the average of the absolute values of the differences between the January and February, February and March, March and April, and April and May forecasts of the same seasonal volume for all calibration years. If the forecast period changes during the forecasting season, observed flows will need to be added to the forecast so that all forecasts are for the same period. For example, if April–July volume is forecast in January through April, and May–July volume is forecast in May, then observed April flow will need to be added to the May–July forecast to obtain an April–July volume to compare with previous forecasts. It is desirable that this measure be minimized, because water managers prefer to have forecasts that change as little as possible from month to month as long as hydrologic validity is maintained.

Another possible measure of forecast consistency is the average number of forecast direction changes. A forecast direction change occurs when

$$F_{m-2} > F_{m-1} \text{ and } F_{m-1} < F_m \quad \ldots\ldots (1)$$

or when

$$F_{m-2} < F_{m-1} \text{ and } F_{m-1} > F_m \quad \ldots\ldots (2)$$

where F_m = forecast in month m. In the first case, the forecasts for months $m - 2$ and $m - 1$ established a downward trend, but the forecast for month m is up from month $m - 1$. In the second case, the forecasts for months $m - 2$ and $m - 1$ established an upward trend, but the forecast for month m is down from month $m - 1$. After determining the total number of direction changes for each year, an average is computed. Water managers prefer that forecast seesawing up and down be minimized, again as long as hydrologic validity is maintained.

Forecast consistency can be achieved with monthly equations by selecting variables to maintain a high degree of similarity from month to month without undue loss of forecast accuracy. There is a trade-off between these two goals because the best variables may differ among months. It is a matter of judgment on the part of the forecast developer to strike a compromise between them. Forecast accuracy is the primary goal, but if a significant amount of month-to-month consistency in variable usage can be obtained with only a small loss in accuracy, then this would be a desirable compromise. A comparison of the two measures of forecast consistency for two single-equation procedures and for two sets of monthly equations is presented in a subsequent section.

Usage of Future Variables and Scenario Forecasts

Many forecast users ask the question, "What would the forecast be if we received *xxx*% of average precipitation for the rest of the season?" Users want to obtain an idea of what may happen assuming some future weather scenario, such as current trends persisting or a major change in precipitation patterns. This question presumably gives users an idea of the forecast uncertainty that might be expected. It has hydrologic validity and is understandable to the layperson.

Such a scenario forecast can be easily calculated if the regression equation contains future precipitation variables. Instead of using average precipitation, one uses the desired percent of average precipitation in the calculation. This, however, cannot be done with equations containing no future variables. Some perceive this as a limitation of using monthly equations with no future variables. In fact, it is no limitation, because the generation of scenario forecasts is not the best way to quantify forecast uncertainty. First of all, not all forecast error is due to unknown future weather. Even if the future were known perfectly, there would still be errors in the predictions. A scenario, then, only accounts for part of the forecast uncertainty. Second, the source of forecast error does not matter in decision making as long as the error can be quantified and described probabilistically. Finally, scenario forecasts do not have exceedance probabilities attached to them, so one does not know how likely they are to occur.

The standard process of constructing confidence bands about a prediction provides the means for calculating alternative forecasts with different exceedance probabilities and hence provides the information needed for decision making. For example, the 80% confidence band about a prediction provides forecasts with a 90% and 10% exceedance probability, assuming that the errors are normally distributed. These forecasts are what need to be used in decision making rather than scenario forecasts. To this end the SCS now publishes five forecasts, with the following exceedance probabilities, for each point: 90% ("reasonable minimum"), 70%, 50% ("most probable"), 30%, and 10% ("reasonable maximum"). No assumptions about future weather are made in these forecasts; they simply represent different quantiles of the probability distribution of the seasonal runoff volume conditioned on the current hydrologic state. This information fully expresses the forecast uncertainty and is the proper information necessary for optimal decision making. Krzysztofowicz (1986a, b) gives a detailed analysis of this kind of decision making.

Principal Components Regression

Intercorrelation and Past Practice

The predictor variables used in water supply forecasting are usually correlated with each other, particularly data for different stations of the same data type at the same time (e.g., snow water equivalent on a given date at several snowcourses). If attention is paid to the significance of the regression coefficients, standard multiple regression will keep only a very few such variables in the equation. If all of these variables are retained anyway, the coefficients will not be accurately estimated, and they may not make physical sense (e.g., negative coefficients for variables having a positive correlation with streamflow). Such an equation may not give consistently accurate predictions over time and is not conceptually acceptable (McCuen 1985; Kleinbaum et al. 1988). If only the few significant variables are retained in the equation, too heavy a reliance is placed on a few data sites to represent

spatially variable snowpack and precipitation consistently; again one might expect erratic performance over time. A more robust, accurate, and consistent forecasting equation can be obtained by having several sites for the same data type and time in the equation.

In water supply forecasting, the practice of constructing composite indexes that were used as independent variables had the effect of at least partially circumventing the intercorrelation problem. These composite indexes were typically weighted sums of data from stations of the same data type to produce snow, fall precipitation, winter precipitation, and spring precipitation terms. By combining data from several highly correlated data sites into a single variable before entering the regression, the major source of intercorrelation was removed. The drawback with this technique, however, is that the weights used in constructing the index were determined outside of the regression (based on correlation analyses, judgment, etc.) and were not, in general, statistically optimal for forecasting. Examples of these indexes are given in a subsequent section.

Principal Components Regression

The most satisfactory and statistically rigorous way to deal with intercorrelation is the use of principal components regression. Previous examples of the use of principal components regression in seasonal streamflow volume forecasting are Marsden and Davis (1968), McCuen et al. (1979), and Wortman (1989). Other examples of principal components regression appear in Haan and Allen (1972), Haan (1977), and McCuen (1985). McCuen and Snyder (1986) give a thorough discussion of the computations involved and some additional examples.

Principal components analysis is a statistical technique that restructures a set of intercorrelated variables into an equal number of uncorrelated variables. Each new variable (principal component) is a different linear combination of all the original variables. The weights used in the linear combinations are from the eigenvectors of the correlation matrix of the original variables. Each component explains a certain percentage of the total variance in the set of variables; the amount is represented by the eigenvalue. Principal components are usually arranged in order of decreasing amount of explained variance. Typically, most of the variance is explained in the first few components.

Principal components analysis is sometimes used to describe modes of variability in a set of data. For example Lins (1985) described major modes of streamflow variability in the United States using the first five of 106 principal components in his data set. This work was purely descriptive, not predictive, in nature.

For prediction, the principal components, calculated from the set of available predictor variables, can themselves be used as independent variables in a regression equation. The number of components retained in the equation depends on how many of them have statistically significant regression coefficients. This is a different selection technique than in descriptive work, where the magnitude of the eigenvalue and the percentage of variance explained are the main criteria. If there was a high degree of intercorrelation in the original data set, the number of components retained in the regression will be much smaller than the number of original variables. This reduces the loss of degrees of freedom, because fewer regression coefficients are being estimated. If all components are retained, it is the same as a standard multiple regression, and there is no advantage in using principal compo-

nents. After the regression coefficients for the components have been computed, the linear transformation can be inverted so that coefficients are expressed in terms of the original predictor variables.

Selection of Principal Components to Retain

For a given combination of variables, one must determine which of the principal components to retain in the regression. This problem is very similar to a standard regression, that is, determining which of the variables (components) are worth keeping. An additional consideration in principal components regression, however, is whether to require the components to be used in sequence (from large to small eigenvalues) or whether any components can be kept regardless of whether any of its predecessors in the sequence are in the equation.

Determining which components ought to be retained in a regression equation is most straightforwardly evaluated by a standard *t*-test (or, equivalently, a partial *F*-test) to determine the significance of the regression coefficient for a variable (component). McCuen (1985) discussed examining the magnitude of the eigenvalues as a preliminary screen for selecting significant components, but this is unnecessary. It only determines which components explain most of the variance in the original variables and says nothing about their ability to explain the dependent variable. The *t*-test is completely adequate to determine which components to keep.

When using the *t*-test, however, it sometimes happens that some components in the sequence will be skipped. For example, the *t*-test may indicate that only components one, two, and five should be retained. Two questions arise: (1) What does it mean that components three and four were skipped? and (2) Should component five be allowed to stay? The answers to these questions first require an explanation of the purpose and philosophy of using principal components.

One of the purposes of principal components regression is to remove the effects of intercorrelation among the original predictor variables. If one has been successful in doing so, one would expect that the regression coefficients would have the same algebraic sign as the correlation coefficient of the predictor variables with the dependent variable (McCuen 1985). In water supply forecasting, then, most regression coefficients should be positive because most predictor variables are positively correlated with streamflow. If a positively correlated variable ends up with a negative regression coefficient, this would suggest that there is intercorrelation among the independent variables; the negative coefficient indicates that this variable is trying to compensate for some of the effect of another independent variable with which it is highly correlated. (If one computes a standard regression with highly correlated independent variables, negative coefficients for some of the variables are frequently obtained, even though they are positively correlated with the dependent variable. The correlated variables are attempting to explain the same thing, and they must compensate for each other's effects.)

In this study, when one or more components were skipped it was often observed that some of the regression coefficients for the original predictor variables were of the opposite sign of their correlation with the dependent variable. This indicates that some of the confounding intercorrelation was reintroduced when components were skipped.

Another purpose of principal components regression is to combine the original variables to create a fewer number of new variables containing

almost the same information as the original ones. It is desired that the original variables be only those most useful for forecasting, so that data for unnecessary variables need not be collected. Since skipping components implies that there are important modes of variability in the original variables that are unrelated to streamflow, one is led to suspect that the combination of variables may contain extraneous information and not be the best for forecasting. It would seem that for a good combination of variables, all major modes of variability ought to be useful for forecasting.

So, one viewpoint might say yes, keep component five even though components three and four were skipped. The components are computed using only data for the original predictor variables and without regard to the dependent variable; it just so happens that a minor mode of variability (component five) is related to streamflow, whereas two larger modes (components three and four) are unrelated to streamflow. Marsden and Davis (1968), Haan (1977), and Wortman (1989) allowed their models to skip components.

The other viewpoint, suggested by McCuen (1985) and the one adopted here, would say no, do not keep component five, and would require components to be included in sequence. Skipping components can reintroduce intercorrelation, and it makes more sense conceptually and from a variable screening point of view not to skip components.

The method used here for selecting principal components to include is as follows.

First, components are added to the model one at a time in sequence, beginning with the one having the largest eigenvalue and progressing in order of decreasing eigenvalue.

Second, when the first component with a nonsignificant regression coefficient (based on a *t*-test) is found, the components retained are the ones in sequence up to, but not including, the nonsignificant one.

Third, regression coefficients, when expressed in terms of original variables, must have the same algebraic sign as their correlations with the dependent variable. If this condition is not met after a component passes the *t*-test, the components up to, but not including, the last one added are retained temporarily. Keeping the last component added, further components continue to be added as long as they pass the *t*-test. The number of components finally retained is the largest number that passes both the *t*-test and the sign test. For example, if components 1 and 2 pass the *t*-test, but one or more coefficients do not pass the sign test, component 3 will still be tried. If component 3 fails the *t*-test, the final model will contain component 1 only. If component 3 passes the *t*-test and the sign test, and if component 4 fails the *t*-test, the final model will contain components 1, 2, and 3.

Once a combination of predictor variables has been selected for evaluation, this procedure provides properly estimated regression coefficients. Determining how many and which of the (correlated) predictor variables to include requires a search. This topic is discussed in a later section.

Cross Validation

It is well known that a model's performance during the calibration period is often better than its performance during a verification period. In the same way, the standard error for a multiple regression equation can be an overly optimistic measure of the equation's actual forecasting performance. To obtain a more realistic evaluation of an equation's forecasting potential, a cross-validation procedure is used here. Inspired by the jackknife technique

for statistical parameter estimation (Wu 1986) and discussed in some statistics texts [e.g., Kleinbaum et al. (1988)], cross validation is a systematic, iterative variation of split-sample model testing. Beginning with the first year, one year is removed from the calibration data set, and the regression coefficients are calculated. These coefficients are used with the data for the withheld year to predict the streamflow for the withheld year. The withheld year is returned to the calibration set, and the next year is removed. The process is repeated through the entire set of years available so that when finished, a set of "forecasts" generated by equations that have not "seen" the year they were forecasting is available. This procedure somewhat mimics the actual forecasting situation, in which the equation has not been calibrated using the year to be forecasted. In the same manner as the usual regression standard error, a cross-validation standard error (CVSE) can be calculated from these "forecasts." The CVSE, then, is a more realistic measure of the actual forecasting potential of an equation. The CVSE is used as the optimality criterion in the search algorithm for predictor variable combinations, described in the following section.

Systematic Search for Optimal Variable Combinations

Unless there are only a very few variables available, some sort of selection technique to determine which variables to include in a regression equation is necessary to ensure optimality or near-optimality. One way to ensure that the equation with the smallest standard error has been found is to compute regressions for all possible combinations of variables. For most water-supply forecasting applications, the number of variables available makes this computationally infeasible; the number of combinations for n variables is $2^n - 1$. If standard multiple regression is used, one can use the stepwise technique (or one of its several variants) to select variables. For principal components regression, the stepwise technique could conceivably be used to select the principal components to be included for a given combination of predictor variables (although a preferable technique was described earlier), but it does not determine if the given combination of variables is itself optimal.

Haan (1977) and McCuen (1985) discuss a procedure for eliminating predictor variables in principal components regression by examining the magnitudes of the eigenvector values ("loadings" or "scores") for each variable. They suggested that one go through the eigenvectors, picking out the variable(s) in each eigenvector that has (have) the highest loading(s). If there are any variables that do not have high loadings for any component (eigenvector), those variables may be eliminated. McCuen (1985) suggested including the dependent variable in the calculation of the eigenvectors when going through the variable elimination process; it is excluded when the data are prepared for principal components regression.

This approach may give adequate results, but there are several drawbacks. If the eigenvectors are computed from the predictor variables only, it is possible that one or more of those variables may not appear important compared to the others, yet they could still be important predictors for the dependent variable. Eliminating variables in this way never gives those variables a chance to be tested as predictors of the dependent variable. If the dependent variable is included in the calculation of the eigenvectors, it is unclear what extra interpretive value is obtained beyond selecting the variables with the highest correlation coefficients with the dependent variable. One is left with attempting to select variables that have high correlations with the dependent variable but low correlations among themselves.

This is an ill-defined process in that there is no way of knowing the optimal trade-off between correlation with the dependent variable and intercorrelation among the predictor variables. In both of these cases, the elimination of variables is not statistically rigorous, requires subjective judgment, and is unsuitable for automated optimization.

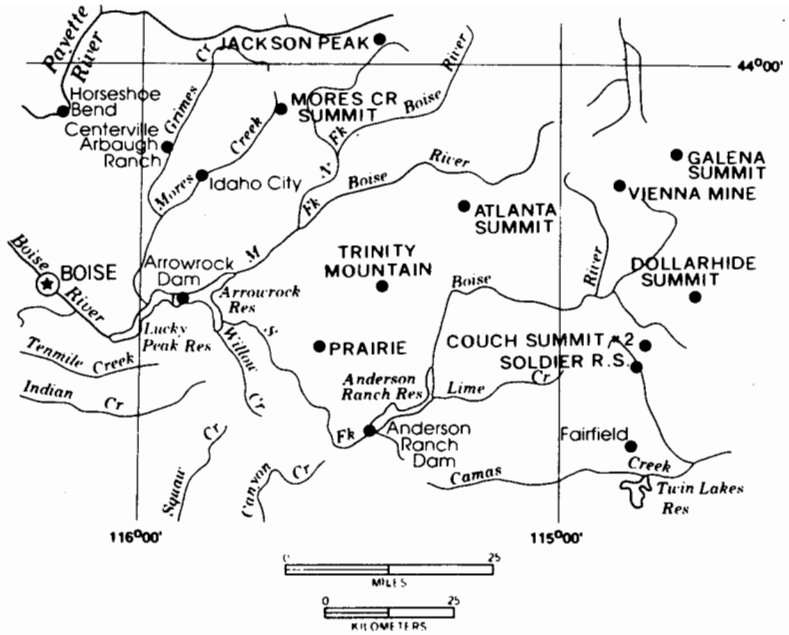
A more straightforward and objective technique for selecting variables is to use an automated search to identify optimal or near-optimal combinations of predictor variables. This eliminates the subjectivity in selecting variables by explicitly testing systematically chosen combinations of variables as predictors of the dependent variable. To this end, a computationally feasible automated search algorithm for principal components regression was developed in this study. This search algorithm is an empirical procedure that evolved from examining how variable combinations were built when all possible combinations were computed beginning with all one-variable equations, then all two-variable equations, and so on. In most cases tested, a pattern developed in which certain variables persistently appeared in combination with others. Based on these observations, an iterative algorithm was developed that begins by computing all one-variable equations and storing the 20 (or all equations, if there are fewer than 20 variables) with the smallest CVSE. Twenty is an arbitrary number that seemed reasonably large enough to allow the algorithm a sufficient number of combinations upon which to build; storing a larger number may be beneficial as the number of variables increases into the thirties or forties, but this has not been tested extensively. Two-variable combinations are then computed by taking each stored one-variable equation and adding one other variable to it. The number of principal components to retain is determined by the sequential t-test and sign test process discussed earlier. Each two-variable equation so constructed is retained in the list of 20 if its CVSE is smaller than any one previously stored. The process continues with three-, four-, five-variable equations and so on until adding one more variable cannot produce a smaller CVSE than the 20 stored ones. The algorithm then stops, and the search is complete.

This search algorithm tends to select for parsimonious models and therefore does not necessarily find the absolute optimum or all combinations of variables with CVSEs between the smallest and the largest in the list of 20. Models that do not perform well until many variables are included are not reached with this algorithm. The algorithm's greatest utility lies in identifying the strongest variables and in constructing near-optimal parsimonious models.

Example of Results

The South Fork Boise River at Anderson Ranch Dam (Idaho), is used here to demonstrate the statistical procedures just discussed and to show the kind of results possible. Results from similar analyses for other basins in the West indicate that the magnitude of the improvements in the standard errors obtained for this basin are typical.

Anderson Ranch Dam (Fig. 1) is one of three reservoirs on the Boise River operated by the U.S. Bureau of Reclamation (USBR). Twice monthly, beginning in January, the USBR forecasts the date-through-July inflow volume to the reservoir for operating purposes. Once per month during January through April, the SCS forecasts the April–July and April–September inflow volumes, and in May it forecasts the May–July and May–September inflow volumes; these forecasts are for the benefit of the general agricultural population. For comparison with the USBR's current forecast-



DATA SITE INFORMATION

Snow		Precipitation	
Name	Number	Name	Number
Atlanta Summit (S)	15F04	Anderson Ranch Dam	0282
Couch Summit #2	14F18	Arrowrock Dam	0448
Dollarhide Summit (S)	14F08	Centerville Arbaugh Ranch	1636
Galena Summit (S)	14F12	Idaho City	4442
Jackson Peak (S)	15E09		
Mores Creek Summit (S)	15F01		
Prairie (S)	15F06		
Soldier Ranger Station	14F11		
Trinity Mountain (S)	15F05		
Vienna Mine (S)	14F04		

(S) denotes SNOTEL sites

FIG. 1. South Fork Boise River at Anderson Ranch Dam Watershed Location and Locations of Data Sites

ing procedure, new date–July forecasting equations were developed using the statistical procedures proposed here and no future variables. The equations developed were for the first-of-month forecasts; mid-month forecasting equations were not developed. In addition, new April–July and May–July forecasting equations were developed to compare with the equations heretofore used by the SCS.

The USBR's current procedure is a single equation used throughout the forecast season, calibrated on October–July volume with all data known, which is as follows:

$$\text{Date–July streamflow } (\times 10^6 \text{ m}^3) = 4.51X_1 + 0.19X_2 + 0.89X_3 + 0.29X_4 - 569.13 - \text{October through date streamflow} \quad (3)$$

where

X_1 = antecedent (October–December) streamflow (10^6 m^3) ..... (4*a*)

X_2 = fall and winter (October–March) precipitation (mm)
= Anderson Ranch Dam + 2 × Arrowrock Dam + Centerville Arbaugh Ranch + Idaho City ..... (4*b*)

X_3 = April 1 snow water equivalent (cm)
= Atlanta Summit + Couch Summit #2 + Jackson Peak + Mores Creek Summit + Soldier Ranger Station + Trinity Mountain + 2 × Vienna Mine ..... (4*c*)

X_4 = spring (April–June) precipitation (mm)
= Anderson Ranch Dam + 2 × Arrowrock Dam + Centerville Arbaugh Ranch + Idaho City ..... (4*d*)

The previous SCS procedure is a single equation for forecasting April–July volume from January through April and a second equation for forecasting May–July volume in May. The equations are as follows:

$$\text{April–July streamflow } (10^6 \text{ m}^3) = 2.90X_1 + 0.18X_2 + 1.53X_3 + 2.18X_4 + 0.36X_5 - 581.56 \quad (5)$$

where

X_1 = antecedent (October–December) streamflow (10^6 m^3) ..... (6*a*)

X_2 = fall and winter (October–March) precipitation (mm)
= Anderson Ranch Dam + Arrowrock Dam + Idaho City .... (6*b*)

X_3 = High elevation seasonal maximum snow water equivalent (cm)
= Atlanta Summit + Dollarhide Summit + Trinity Mountain + Vienna Mine ..... (6*c*)

X_4 = low elevation seasonal maximum snow water equivalent (cm)
= Couch Summit #2 + Prairie ..... (6*d*)

X_5 = spring (April–June) precipitation (mm)
= Anderson Ranch Dam + Arrowrock Dam + Centerville Arbaugh Ranch + Idaho City ..... (6*e*)

and

$$\text{May–July streamflow } (10^6 \text{ m}^3) = 2.44X_1 + 2.32X_2 + 1.48X_3 + 0.43X_4 - 463.96 \quad (7)$$

where

X_1 = antecedent (October–December) streamflow (10^6 m^3) ..... (8*a*)

X_2 = high elevation seasonal maximum snow water equivalent (cm)
= Atlanta Summit + Dollarhide Summit + Trinity Mountain .. (8*b*)

X_3 = low elevation seasonal maximum snow water equivalent (cm)
= Couch Summit #2 ..... (8*c*)

X_4 = spring (April–June) precipitation (cm)
= Anderson Ranch Dam + Arrowrock Dam + Idaho City .... (8*d*)

The equation coefficients were recomputed for this study, and they are

slightly different from the actual values used by the USBR and SCS. The differences are due to: (1) Calibrating on the period 1961–88 rather than the period actually used by the agencies; and (2) using SNOTEL snow water equivalent instead of manual snowcourse measurements. This involves all snow sites except Couch Summit #2 and Soldier Ranger Station. Observed SNOTEL data are only available for the period 1981 to the present. Before 1981, estimates were used based on linear regression relationships between SNOTEL and the colocated snowcourse.

When forecasting with these procedures before April 1 (or before the seasonal maximum snow accumulation has occurred), snow data are extrapolated by adding the average accumulation between the date and April 1 (or the maximum) to the current value. Averages are used for future monthly precipitation variables. The construction of these equations is fairly typical of many that have been and are still being used by forecasting and water management agencies. Note the use of snow and precipitation indexes as discussed in a previous section.

The new equations for Anderson Ranch Dam are shown in Tables 1 and 2. As with the USBR and previous SCS equations, these equations were calibrated on the period 1961–88, and the SNOTEL data before 1981 were estimates based on linear regression relationships between SNOTEL and the colocated snowcourse. To arrive at these equations, separate analyses using the techniques described earlier were first performed for each forecast month and period. The variables used in the top 20 equations differed somewhat among the forecast months and periods, so the results of each analysis were examined to determine which variables appeared most consistently. By judgment and some trial and error, a set of variables was finally chosen that struck a compromise between optimal CVSE and month-to-month variable consistency.

A comparison of the CVSEs is given in Table 3. The new equations have considerably smaller CVSEs than the USBR and previous SCS equations, particularly for the forecasts before April. Also shown are the CVSEs for the top-ranking equations from the search algorithm for each month. This shows that the increases in CVSE caused by selecting combinations to provide month-to-month consistency are slight.

Using the two measures discussed earlier (the average number of forecast direction changes and the average monthly forecast change), the month-to-month forecast consistency for the USBR equation is compared to that for the new equations in Table 4. The maximum value for forecast direction changes is three in this work; the initial trend is established by the January and February forecasts, and trend changes are set by the March, April, and May forecasts. The volumes used are date–July forecasts plus observed January–date flow to give January–July volume each month. When all three sets of equations are calibrated on the 1961–88 period, the top-ranking equations are the most stable by the direction-change measure. When the new equations are calibrated on the 1951–88 period (but still considering the forecasts for the 1961–88 period), the selected new equations are more stable than the top-ranking equations. In any case, this measure is not greatly different among the three sets of equations. The selected new equations are the most stable according to the average-forecast-change measure for both calibration periods.

The month-to-month forecast consistency for the previous SCS equations is compared to that for the new equations in Table 5. The volumes used are April–July forecasts; in May, observed April flow was added to the

TABLE 1. Regression Coefficients for New Forecasting Equations, South Fork Boise River at Anderson Ranch Dam

Station, data type, and month (1)	FORECAST MONTH AND PERIOD							
	January		February		March		April	May
	January–July (2)	April–July (3)	February–July (4)	April–July (5)	March–July (6)	April–July (7)	April–July (8)	May–July (9)
Anderson Ranch Dam, precipitation								
October (mm)	0.86	0.81	0.99	0.88	—	—	—	—
December (mm)	0.73	0.69	0.78	0.74	0.60	0.57	0.70	0.50
Arrowrock Dam, precipitation								
October (mm)	1.01	0.95	1.02	0.90	0.88	0.84	0.48	0.39
December (mm)	0.95	0.89	1.04	0.98	0.78	0.74	0.89	0.64
March (mm)	—	—	—	—	—	—	0.57	0.55
Centerville Arbaugh Ranch, precipitation								
October (mm)	0.53	0.50	0.47	0.41	0.51	0.48	0.19	0.14
March (mm)	—	—	—	—	—	—	0.39	0.42
April (mm)	—	—	—	—	—	—	—	0.37
Idaho City, precipitation								
October (mm)	0.75	0.70	0.65	0.58	0.70	0.66	0.38	0.31
January (mm)	—	—	0.10	0.13	0.40	0.38	0.45	0.34
April (mm)	—	—	—	—	—	—	—	0.48
Atlanta Summit SNOTEL, snow water equivalent								
January (cm)	2.95	2.78	—	—	—	—	—	—
February (cm)	—	—	2.19	2.11	—	—	—	—
March (cm)	—	—	—	—	2.05	1.95	—	—
April (cm)	—	—	—	—	—	—	1.62	1.24
Dollarhide Summit SNOTEL, snow water equivalent								
January (cm)	3.82	3.59	—	—	—	—	—	—
February (cm)	—	—	2.56	2.46	—	—	—	—
March (cm)	—	—	—	—	2.45	2.33	—	—
April (cm)	—	—	—	—	—	—	2.01	1.54
May (cm)	—	—	—	—	—	—	—	1.35
Galena Summit SNOTEL, snow water equivalent								
January (cm)	5.56	5.23	—	—	—	—	—	—
February (cm)	—	—	3.34	3.22	—	—	—	—
March (cm)	—	—	—	—	3.49	3.31	—	—
April (cm)	—	—	—	—	—	—	3.15	2.42
Prairie snow water equivalent								
February (cm)	—	—	5.32	5.13	—	—	—	—
March (cm)	—	—	—	—	4.42	4.21	—	—
April (cm)	—	—	—	—	—	—	2.91	2.37
Trinity Mtn. SNOTEL, snow water equivalent								
January (cm)	2.32	2.18	—	—	—	—	—	—
February (cm)	—	—	1.51	1.47	—	—	—	—
March (cm)	—	—	—	—	1.48	1.40	—	—
April (cm)	—	—	—	—	—	—	1.42	1.10
May (cm)	—	—	—	—	—	—	—	0.90
Anderson Ranch Dam inflow								
November (10^6 m^3)	4.90	4.60	7.17	6.53	4.36	4.14	3.97	3.34
February (10^6 m^3)	—	—	—	—	2.53	2.40	1.65	1.21
April (10^6 m^3)	—	—	—	—	—	—	—	−1.00
Intercept (10^6 m^3)	−21	−80	−195	−225	−288	−289	−322	−330

TABLE 2. Statistics for New Forecasting Equations, South Fork Boise River at Anderson Ranch Dam

Statistic (1)	FORECAST MONTH AND PERIOD							
	January		February		March		April	May
	January–July (2)	April–July (3)	February–July (4)	April–July (5)	March–July (6)	April–July (7)	April–July (8)	May–July (9)
Correlation coefficient	0.89	0.90	0.93	0.94	0.96	0.95	0.98	0.97
Standard error (10^6 m^3)	167	147	135	117	104	103	72	69
Cross-validation correlation coefficient	0.87	0.89	0.91	0.92	0.94	0.94	0.97	0.96
Cross-validation standard error (10^6 m^3)	176	156	155	134	117	116	88	85
Number of principal components used	1	1	3	3	1	1	3	3

TABLE 3. Cross-Validation Standard Error Comparison

Forecast month and period (1)	Bureau of Reclamation Equation		Previous Soil Conservation Service Equations		Selected New Equations		Rank #1 New Equations	
	Cross-validation standard error (10^6 m^3) (2)	Percent of average (%) (3)	Cross-validation standard error (10^6 m^3) (4)	Percent of average (%) (5)	Cross-validation standard error (10^6 m^3) (6)	Percent of average (%) (7)	Cross-validation standard error (10^6 m^3) (8)	Percent of average (%) (9)
January								
January–July	267.8	33	—	—	176.4	22	170.7	21
April–July	—	—	211.3	30	155.6	22	155.6	22
February								
February–July	236.4	30	—	—	154.7	20	141.0	18
April–July	—	—	162.5	23	134.2	19	121.4	17
March								
March–July	196.6	26	—	—	117.1	15	109.9	14
April–July	—	—	135.0	19	116.2	16	104.6	15
April								
April–July	101.9	14	97.7	14	88.4	12	74.9	11
May								
May–July	104.5	18	110.2	19	85.1	15	74.9	13

Note: Streamflow averages used are for the period 1961–85. All equations were calibrated on the period 1961–88.

May–July forecast. The previous SCS equations are more stable than the new equations according to the direction-change measure but less stable according to the average-forecast-change measure. Neither set of equations, then, has a clear advantage in stability. The direction-change measures for the selected new equations and the top-ranking equations show that some stability improvement of this type was made by careful variable selection. The average forecast changes, however, are nearly the same for the two sets of equations. From the comparisons in Tables 4 and 5, it appears that monthly equations can provide stable forecasts while giving significantly greater accuracy.

TABLE 4. Fore Equations

Equations (1)	Calibration period (2)	Average number of direction changes (3)	Average monthly forecast change (10^6 m^3) (4)
Bureau of Reclamation equation	1961–88	1.46	73.5
Selected new equations	1961–88	1.54	55.4
Selected new equations	1951–88	1.43	53.3
Rank #1 new equations	1961–88	1.36	56.0
Rank #1 new equations	1951–88	1.79	56.1

TABLE 5. Forecast Consistency Comparison—Previous Soil Conservation Service and New Equations

Equations (1)	Calibration period (2)	Average number of direction changes (3)	Average monthly forecast change (10^6 m^3) (4)
Previous Soil Conservation Service equations	1961–88	1.29	70.4
Selected new equations	1961–88	1.61	53.2
Selected new equations	1951–88	1.54	51.6
Rank #1 new equations	1961–88	1.64	53.0
Rank #1 new equations	1951–88	1.86	52.1

COMMENT

The foregoing techniques have been used together as a system to develop new forecasting equations for the example basin and other basins analyzed. One may wonder, however, which of the techniques produces the most forecast improvement. One could imagine an extensive set of experiments to test equations with all the possible combinations of combined indexes versus individual data elements, future data versus no future data, standard regression versus principal-components regression, and so on. A limited number of such experiments was conducted for the South Fork Boise River, and the results suggested that the maximum forecast accuracy gain is obtained by proper selection of variables, followed by the use of principal components regression and using only known data (no future variables). The objective of this work, however, was not to determine the effects of individual past practices, as these have already been determined by either previous research or statistical reasoning to be questionable or nonoptimal. Rather, the objective was to develop a system of techniques that work together to rectify the weaknesses of past practices and to arrive at equations that are based as much as possible on statistically rigorous and appropriate techniques.

CONCLUSIONS

By not using future variables and by using principal components regression, cross validation, and a search technique, substantial improvements in accuracy over current practice in seasonal streamflow volume forecasts can

be obtained. Month-to-month forecast consistency can be maintained with the use of separate equations for each month by judicious selection of variables to maintain some month-to-month similarity in each month's equation. By using these techniques to develop seasonal volume forecasting equations, improved water management is possible under current decision-making procedures. In the future, these improved seasonal volume forecasting techniques can be used as a base from which to judge the success of monthly or full hydrograph forecasts.

Acknowledgments

I wish to thank my colleague, Robert Hartman, now with the National Weather Service, for his many valuable insights that contributed significantly to the development of the algorithms used in this work. I have also benefited greatly from many lengthy discussions with Randal Wortman of the U.S. Army Corps of Engineers on various aspects of the statistical techniques used here. James Doty of the U.S. Bureau of Reclamation provided me with the USBR forecasting equation.

Appendix. References

Barton, M., and Burke, M. (1977). "SNOTEL: An operational data acquisition system using meteor burst technology." *Proc., Western Snow Conference*, 82–87.

Crook, A. G. (1984). "The SNOTEL data acquisition system: A tool in runoff forecasting." *A critical assessment of forecasting in western water resources management; Proc., AWRA Symp.*, J. J. Cassidy and D. P. Lettenmaier, eds., Seattle, Wash., 25–30.

Day, G. N. (1985). "Extended streamflow forecasting using NWSRFS." *J. Water Resour. Planning and Mgmt.*, ASCE, 111(2), 157–170.

Druce, D. J. (1984). "Seasonal inflow forecasts by a conceptual hydrologic model for Mica Dam, British Columbia." *A critical assessment of forecasting in western water resources management; Proc. of AWRA Symp.*, J. J. Cassidy and D. P. Lettenmaier, eds., Seattle, Wash., 85–91.

Haan, C. T. (1977). *Statistical methods in hydrology*. Iowa State University Press, Ames, Iowa.

Haan, C. T., and Allen, D. M. (1972). "Comparison of multiple regression and principal component regression for predicting water yields in Kentucky." *Water Resour. Res.*, 8(6), 1593–1596.

Kleinbaum, D. G., Kupper, L. L., and Muller, K. E. (1988). *Applied regression analysis and other multivariable methods*, 2nd Ed., PWS-KENT Publishing Co., Boston, Mass.

Koch, R. W. (1990). "Influences of climate variability on streamflow variability: Implications in streamflow prediction and forecasting." *Final report for grant award 14-08-0001-G1316*, U.S. Geological Survey, Washington, D.C.

Krzysztofowicz, R. (1986a). "Expected utility, benefit, and loss criteria for seasonal water supply planning." *Water Resour. Res.*, 22(3), 303–312.

Krzysztofowicz, R. (1986b). "Optimal water supply planning based on seasonal runoff forecasts." *Water Resour. Res.*, 22(3), 313–321.

Kuehl, D. W. (1979). "Volume forecasts using the SSARR model in a zone mode." *Proc., Western Snow Conference*, 38–47.

Lins, H. F. (1985). "Interannual streamflow variability in the United States based on principal components." *Water Resour. Res.*, 21(5), 691–701.

Marsden, M. A., and Davis, R. T. (1968). "Regression on principal components as a tool in water supply forecasting." *Proc., Western Snow Conference*, 33–40.

McCuen, R. H. (1985). *Statistical methods for engineers*. Prentice-Hall, Englewood Cliffs, N.J.

McCuen, R. H., Rawls, W. J., and Whaley, B. L. (1979). "Comparative evaluation of statistical methods for water supply forecasting." *Water Resour. Bulletin*, 15(4), 935–947.

McCuen, R. H., and Snyder, W. M. (1986). *Hydrologic modeling: Statistical methods and applications*. Prentice-Hall, Englewood Cliffs, N.J.

Pearson, T. (1974). "Simulating runoff to the Hungry Horse reservoir of western Montana." *Proc., Western Snow Conference*, 96–102.

Rallison, R. E. (1981). "Automated system for collecting snow and related hydrological data in mountains of the western United States." *Hydrological Sci. Bulletin*, 26(1), 83–89.

Schermerhorn, V., and Barton, M. (1968). "A method for integrating snow survey and precipitation data." *Proc., Western Snow Conference*, 27–32.

"Snow survey and water supply forecasting." (1972). *National engineering handbook, section 22*. Soil Conservation Service (SCS), U.S. Department of Agriculture, Washington, D.C.

Speers, D. D., and Versteeg, J. D. (1982). "Runoff forecasting for reservoir operations—the past and the future." *Proc., Western Snow Conference*, 149–156.

Stedinger, J. R., Grygier, J., and Yin, H. (1988). "Seasonal streamflow forecasts based upon regression." *Computerized decision support systems for water managers; Proc. 3rd Water Resour. Operations and Mgmt. Workshop*, ASCE, New York, N.Y., 266–279.

Twedt, T. M., Schaake, J. C. Jr., and Peck, E. L. (1977). "National Weather Service extended streamflow prediction." *Proc., Western Snow Conference*, 52–57.

Wortman, R. T. (1989). "Statistical forecast model for Libby basin, Montana." *Proc., Western Snow Conference*, 100–107.

Wu, C.F.J. (1986). "Jackknife, bootstrap and other resampling methods in regression analysis." *The Annals of Statistics*, 14(4), 1261–1295.

Zuzel, J. F., and Cox, L. M. (1978). "A review of operational water supply forecasting techniques in areas of seasonal snowcover." *Proc., Western Snow Conference*, 69–77.