STATISTICAL FORECAST MODEL FOR LIBBY BASIN, MONTANA

by

Randal T. Wortman[1]

INTRODUCTION

Water supply forecasts in the western United States provide information critical to both local and regional interests. Accurate and timely prediction of the spring and early summer streamflow allows reservoir operators to plan effective strategies for the storage and release of the anticipated runoff so as to provide the maximum benefits to the water users.

The two principle methods used to provide water supply forecasts are the conceptual hydrologic model and the statistical model (Soil Conservation Service, 1970; Zuzel and Cox, 1978). This paper will focus on the statistical model, and a particular variation know as principal components regression. A new model validation statistic will be discussed, and application of principal components regression to the Libby, Montana water supply forecast will be presented.

LINEAR REGRESSION MODELS

A statistical procedure often used in water supply forecasting is the multivariable linear regression model (Cassidy and Lettenmaier, 1985; Palmer, 1988). Historical hydrometeorological variables in the streamflow basin (typically antecedent streamflows, monthly accumulated precipitation depths, and date-specific measurements of snow water equivalent) are regressed against the historical seasonal runoff volume to develop a least-squares best fit linear model. In practice, the independent variables fitted in the regression model are often surrogate variables, each representing a generic hydromet parameter, such as winter precipitation. The surrogate may be computed as a weighted combination of several similar stations, and often combines station values across several months in a weighting scheme that attempts to reflect the relative importance of certain months in contributing to the seasonal runoff. Both station and monthly weighting schemes have often been highly subjective. The practice of weighting and summing snow water equivalent (SWE) values can sometimes mask important intrabasin variation, failing to improve the prediction accuracy over simpler models (McCuen, et al. 1979).

In an effort to increase the lead time and produce a more timely forecast, the regression model is frequently calibrated for a particular runoff season, but used for making forecasts in advance of the availability of some of the required input data. Station mean values ("normal subsequent" values) are routinely substituted into the regression equations in lieu of the yet-to-be-observed data. It has been shown that 1) the forecast accuracy, in terms of the regression mean square error, is overly optimistic when a mean value is substituted for a "future regressor" value, 2) inclusion of a future regressor can bias the estimation of the regression coefficients, especially if the future regressor is significantly correlated to the other regressors, and 3) the future regressor model may keep a truly better forecasting model from receiving just consideration (McCuen, et al. 1979; Stedinger, et al. 1988).

Intercorrelation (dependence) of the functional independent variables in a least squares fit regression model can produce irrational regression coefficients and use excess degrees of freedom (McCuen and Snyder, 1986). In water supply forecasting it is not uncommon for the winter precipitation to be marginally correlated with SWE, high cross correlations to exist between proximal SWE stations, and for there to be high autocorrelations between SWE variables representing successive monthly readings. It would be preferable in least

---

[1] Hydraulic Engineer, U.S. Army Corps of Engineers, North Pacific Division, P.O. Box 2870, Portland, OR 97208-2870

Presented at the Western Snow Conference 1989, Fort Collins, CO

squares regression modeling to not have to deal with the complexities of highly correlated regressors.

Determination of the appropriate number of regressors to include in a regression model is often as difficult as deciding which variables or stations to use. Although snowmelt is a vital component to the seasonal runoff, the antecedent precipitation and streamflow variables contain certain important information that can also be used by the regression model. Since a single SWE station is often inadequate to provide accurate information on the distribution of snow across a mountainous basin, it is usually desirable, and beneficial, to use data from several SWE stations. Typically, this brings the modeler to the dilemma of too many regressors, or else to the surrogate variable approach (composite, weighted SWEs) discussed previously. Model parsimony is further constrained by the short period of record (often < 50 years) common in the western United States. Model design theory suggests that a minimum of 2^p observations be available to define a p-dimensional orthonormal model space; 64 years for 6 regressors, 128 years for 7 regressors. Stepwise regression procedures can easily suggest models that exceed a reasonable parsimony, although any and all variables meet the specified criteria for significance (Hocking 1976; Breiman and Freedman, 1983).

PRINCIPAL COMPONENTS REGRESSION

Principal components analysis is a technique that uses a weighted combination of the observed functional independent variables to form a set of artificial variables. The artificial variables are called principal components. Although it is possible to manually derive the principal components for a set of independent variables, computer programs are readily available that speedily perform the numerical manipulations (SAS, 1985). The theory and techniques of principal components regression, with hydrologic applications, are aptly presented by McCuen and Snyder (1986), and additional examples of its application to hydrologic modeling have been in the literature for quite some time (Marsden and Davis, 1968; Kisiel, 1972; McCuen, et al. 1979). Table 1 contains an example listing from a principle components analysis.

TABLE 1

PRINCIPAL COMPONENTS ANALYSIS FOR LIBBY BASIN
APR-AUG FORECAST ON 1 APR USING 9 VARIABLES AND 9 COMPONENTS

VARIABLES:	FSTNOV	WASBP11	BRIBP12	WASBP1	WASBP2	GRPBP2	APR4	APR6	APR9
		TOTAL RSQ	ADJST RSQ		STD ERROR				
		0.9292	0.9086		524 kdam^3 (425 KAF)				
COMPONENTS :	1	2	3	4	5	6	7	8	9
EIGENVALUES	3.2365	1.3956	1.0789	0.9835	0.7574	0.6003	0.4644	0.2744	0.2090
COMPONENT R2	0.6138	0.1179	0.0010	0.0107	0.0144	0.0003	0.0679	0.0010	0.1022
EIGENVECTORS	-0.0935	0.6895	-0.1930	-0.3199	0.1230	0.1788	0.5005	0.2450	-0.1365
	0.0298	0.4345	-0.3381	0.7321	0.2238	-0.0576	-0.2845	0.0696	0.1443
	0.2639	-0.2264	-0.4024	-0.4096	0.6490	-0.0827	-0.2229	0.1843	0.1942
	0.3663	-0.3158	0.1272	0.3328	0.0712	0.5495	0.3587	0.4426	0.0874
	0.0137	0.2711	0.8091	-0.0227	0.4807	-0.0130	-0.1605	0.0121	0.1185
	0.3561	0.2982	-0.0023	-0.2794	-0.4087	0.4600	-0.5109	-0.0134	0.2575
	0.4210	0.1088	0.1034	-0.0193	-0.2988	-0.6476	0.2056	0.3245	0.3790
	0.5027	0.0690	0.0466	0.0294	0.0053	-0.1499	-0.1816	0.0924	-0.8219
	0.4792	0.0763	-0.0533	0.0577	0.1452	0.0487	0.3580	-0.7688	0.1250
PRINCIPAL COMPONENTS REGRESSION COEFFICIENTS FOR STANDARDIZED VARIABLES									
	0.2523	0.2312	0.1436	0.3534	0.1949	0.1506	0.5237	-0.3979	0.4342

REGRESSION COEFFICIENTS FOR ORIGINAL VARIABLES

INTERCEPT	FSTNOV	WASBP11	BRIBP12	WASBP1	WASBP2	GRPBP2	APR4	APR6	APR9
-2649.21	11.15	433.55	186.62	644.02	479.74	70.66	93.24	-260.25	158.21

Principal components have a variety of useful properties, the two most interesting to us is that the principal components are orthogonal (jointly uncorrelated), and that the iterative method used to derive the principal components orders and presents the components sequenced by the magnitude of their variances. That is, the variances of the principal components are the eigenvalues of the functional independent variables. The eigenvalues are additive and scaled so that their sum is equal to the number of original variates, p. There are initially as many principal components as original independent variates, but if any degree of intercorrelation exists between the independents, there will usually be one or more (say m) principal components with a small and often negligible variance (eigenvalue). If the trailing eigenvalues are nearly zero, we will lose little information by ignoring their associated principal components. By ignoring these m components we can reduce our original set of p independent variables to p-m orthogonal variates. Principal components regression simply regresses our dependent variable on these orthogonal variates.

A third unique feature in principal components regression is that the squared partial correlations for the fitted model are independent and additive. Unlike a traditional regression model, the partial correlations do not need to be recomputed whenever a variate is entered or removed. In a similar fashion as in the eigenvalue analysis, the squared partial correlations can be uniquely associated with a component, and the effect of including or removing that component from the component regression model can be directly evaluated.

In all the discussion that follows, it is assumed that each of the original variables have been standardized by subtracting its mean and dividing by its standard deviation to give a unit-norm variate. Among other things, this provides the scaling whereby the variances sum to p. In some situations it is optional whether or not to standardize the dependent variable, but for this analysis it is assumed that standardization has already been performed.

To summarize the principal components regression procedure, a set of correlated variables X_1, X_2, ..., X_P, each X a vector with N observations, can be transformed into a set of uncorrelated variables C_1, C_2, ... C_P by taking a special weighted combination of the original X's as

$$C_k = \sum_{j=1}^{p} L_{kj} X_j \qquad k=1,2,\ldots p \tag{1}$$

The scoring coefficients (L_{kj}) are the eigenvectors of the X_j's and the variance of the C_k's are the eigenvalues of the X_j's. We will consider retaining only those C_k's that have a meaningful contribution to the total variance. The regression model in terms of the original variates

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p \tag{2}$$

can alternatively be reconstructed in terms of the principal components as

$$Y = \alpha_1 C_1 + \alpha_2 C_2 + \ldots + \alpha_k C_k \qquad k=1,2,\ldots,p-m \text{ ; and } 0 \leq m < p \tag{3}$$

where m components have been discarded due to extremely small contributions to the variance of X's. After fitting the principal components regression model, and deciding on which components to retain, it is easy to convert the α's back to be expressed as coefficients to the original X's. If m=0 and all components are used, then the back-transformed coefficients will be identical to the β's fitted with a standard regression model.

MODEL VALIDATION

An important step in hydrologic model development is the validation of the model. How well does it perform the task it was designed to do? It is one thing to have a model that has a great fit to the calibration data; but it is also highly desirable for the model to provide a reasonable prediction when given data that was not used in its calibration.

The split-sample scheme of model validation has frequently been used with both conceptual and statistical hydrologic models. The split- sample scheme withholds a selected subset of the historic data for validation, and proceeds with calibration of the model to the remaining data. An appropriate error metric, typically the root mean square of the errors, is calculated for both subsets and then compared for significant non-equality. It

is assumed that the data are not autocorrelated across observations. The size of the calibration and validation subsets are often arbitrary, as is the assignment of which observations should go into which subset. A serious shortcoming of split-sample validation in hydrologic modeling is imposed by the limited historic sample sizes available for study. Further subdividing the historic sample into split-sample subsets often 1) withholds critical information from the calibration step, and 2) provides unstable error statistics highly dependent on the subset size and information content.

A variation on the Jackknife method is recommended as a reasonable and robust approach to model validation. The Jackknife procedure (Mosteller and Tukey, 1977) is a logical extension of the split-sample scheme, but minimizes its shortcomings. In the Jackknife procedure a split sample is created by withholding one observation and fitting the model to the other N-1 observations. A prediction ($\hat{Y}$) is then generated using the withheld observation in this model. The process is repeated for each observation, each time fitting a model to N-1 observations, and generating a prediction for data not used in the model calibration. A standard error statistic, herein referred to as the iterative forecast standard error (IFCST SE), can be calculated for the N Jackknife predictions:

$$\text{IFCST SE} = \text{SQRT}\left[\sum_{i=1}^{N}(Y_i-\hat{Y}_i)^2/(N-(p+1))\right] \qquad (4)$$

For a p parameter model (plus an intercept) use the traditional N-(p+1) degrees of freedom in the standard error calculation. Table 2 shows an example of the iterative forecast standard error procedure for the principal component regression model introduced in Table 1.

It is evident that a minimum amount of information was withheld from each fitted Jackknife model, yet each prediction was generated by a model calibrated to an independent set of data. Each Jackknife prediction is analogous to the prediction of next year's dependent variable, using all the available data through the current year. The IFCST SE statistic can be used to evaluate how well a model performs with data that it hasn't seen.

TABLE 2

ITERATIVE FORECAST STANDARD ERROR COMPUTATION
(Forecasts in KAF)

OBS #	MODEL SE	MODEL R2	Y FCST	Y OBS	OBS #	MODEL SE	MODEL R2	Y FCST	Y OBS
1	430.76	0.9044	8250.03	8480.54	21	428.70	0.9091	5897.85	6239.69
2	428.39	0.9063	5415.97	5058.85	22	428.29	0.9065	7898.84	8247.69
3	418.50	0.9130	8090.04	7396.01	23	431.58	0.9030	4783.99	4653.63
4	408.50	0.9139	7647.96	8529.14	24	431.65	0.9060	8104.88	7982.05
5	431.73	0.9079	6447.73	6336.34	25	431.79	0.9019	8981.95	8869.00
6	430.30	0.9087	6346.33	6590.32	26	426.50	0.9070	5433.59	5027.00
7	430.44	0.9008	9507.69	9142.68	27	426.66	0.9020	8790.34	9215.00
8	431.79	0.9080	6511.95	6612.09	28	411.43	0.9160	6870.49	5980.00
9	429.03	0.9040	9040.47	8728.59	29	388.83	0.9249	8735.29	7411.00
10	428.27	0.9090	6370.16	6026.68	30	425.97	0.8969	3035.70	3493.00
11	428.94	0.9082	6060.06	5731.04	31	432.10	0.9077	6300.04	6288.00
12	431.60	0.9056	7995.82	8124.99	32	425.85	0.9029	4687.23	4210.00
13	428.69	0.9093	6816.03	6463.01	33	414.02	0.9149	5054.63	5979.00
14	427.14	0.9085	7409.21	7820.82	34	402.83	0.9193	6467.39	7457.00
15	428.38	0.9089	6284.59	5965.19	35	429.28	0.9091	6764.79	6484.00
16	429.48	0.9090	6743.89	6440.12	36	430.47	0.9079	6137.28	5925.00
17	407.62	0.9180	6039.28	6937.53	37	422.82	0.9088	5622.94	5072.00
18	428.15	0.9095	6599.13	6964.37	38	429.33	0.9046	5073.59	4775.00
19	398.50	0.9214	6173.30	7183.44	39	422.72	0.9115	5515.23	6098.00
20	432.06	0.9052	8114.90	8161.06	40	430.69	0.9050	5223.90	4991.00

IFCST SE = 728 kdam^3

IFCST SE = 590 KAF

LIBBY STATISTICAL FORECAST MODEL

Background

Libby Dam, located in northwestern Montana on the Kootenai River near the British Columbia border (Figure 1) was completed in 1973. Lake Koocanusa, the reservoir behind Libby dam, impounds 6 million dam^3 (4.9 million acre feet) in active storage, and is an important component in both flood control and hydropower operations in the Pacific Northwest. Seasonal runoff forecasts have been made for the Libby basin since 1972 using a multivariable regression procedure with a database extending back to 1948. The procedure was revised in 1977 (U.S. Army Corps of Engineers, 1977), and updated again in 1986 (Wortman, 1986) to include the additional data through water year 1985 and to eliminate a snow station that had been discontinued. Slight non-stationarity in the Libby local runoff recently detected through double mass curve analysis (U.S. Geological Survey, 1960) may be a result of reservoir bank storage, but variables adjusted for this difference have not shown any additional value over non-adjusted variables in the regression procedures. A rough computation shows the value of the water in the top foot of Lake Koocanusa to hydropower generation to be approximately $32,000 per 1233 dam^3 (1000 AF). Thus a 10% reduction in the 1 April forecast standard error (10% of 872,000 dam^3, SE from the 1986 model) could potentially be worth some $2.2 million.

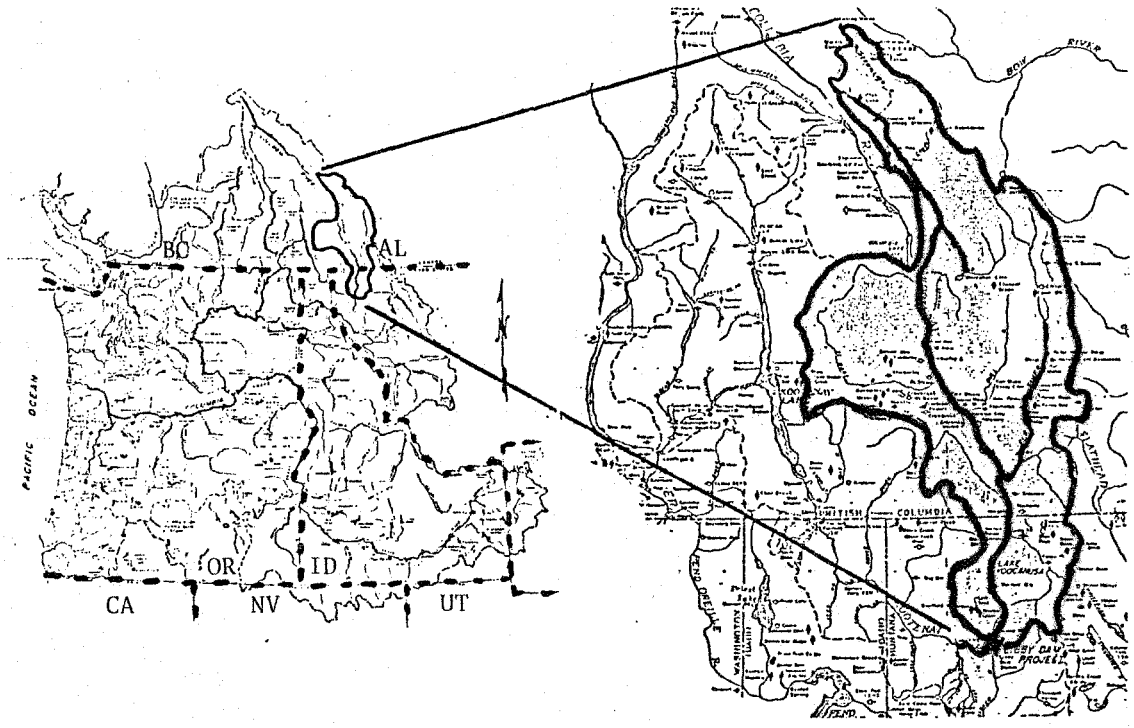

Figure 1. Kootenai River Basin above Libby Dam, Montana

Current Multivariable Model

The current Libby statistical procedure subdivides the Libby basin into two subareas, known as Ft. Steele and Libby local, and fits a multivariable regression model to each subarea. The Ft. Steele subbasin has a streamgage station only dating back to 1961, and there has never been a gage for the Libby local basin, so the runoff variables for both subbasins are calculated values from an assortment of nearby historical stations. The predicted runoff from the two subbasins are combined to give the runoff volume for the total Libby basin. Both regression models calculate the subbasin April through August runoff volume using four structural independent variables:

$$\text{FT STEELE R.O.} = \alpha_0 + \alpha_1 \text{FST\_FRO} + \alpha_2 \text{FST\_WP} + \alpha_3 \text{FST\_SWE} + \alpha_4 \text{FST\_SP} \quad (5)$$

$$\text{LIBBY LOCAL R.O.} = \beta_0 + \beta_1 \text{LIB\_FRO} + \beta_2 \text{LIB\_WP} + \beta_3 \text{LIB\_SWE} + \beta_4 \text{LIB\_SP} \quad (6)$$

where Ft. Steele and Libby local runoff variables are prefixed with FST or LIB, respectively; the FRO suffix denotes the Fall Runoff (October+November streamflow) variable for a subbasin; the WP suffix denotes a Winter Precipitation surrogate variable combining the October through March precipitation values of several stations, with equal monthly weights; the SWE suffix denotes a weighted combination of 1 April snow water equivalent stations; and the SP suffix denotes a Spring Precipitation surrogate variable combining the April though August precipitation values of several stations, with decreasing weights for summer months. The abundance of hydromet stations available for consideration when developing the surrogate independent variables (123 "monthly" stations for the 1 April forecast, without "future" variables) created considerable difficulty in the original model development. The regression equations were fitted to 1 September data (since data through August are required for the SP independent variable).

During the January to June forecast season a forecast is generated monthly using the data available to date, with historical mean ("normal subsequent") values replacing "future" variables. This procedure produces monthly forecasts that are relatively consistent throughout the forecast season (an important criteria in reservoir operations!), although only moderately accurate. The total basin regression standard errors and iterative forecast standard errors for the monthly forecasts are shown in Table 3.

Table 3

STANDARD ERROR OF APRIL-AUGUST FORECAST
in 1000 dam^3

	JAN	FEB	MAR	APR	MAY	JUN	SEP
Regression	1449	983	905	872	777	603	482
IFCST SE	1462	1009	940	904	814	649	541

The close agreement between the regression and iterative forecast standard errors support the contention that the regression model is not unreasonably overfit.

Stepwise Regression Model

Stepwise regression techniques were employed in an attempt to select the best subsets of variables for each monthly forecast model. The STEPWISE technique (SAS, 1985) performs forward selection with reexamination of the significance of all variables at each step. Variables failing the significance test are removed from the model. Table 4 shows the results of a stepwise regression analysis using a 0.10 entry significance level and a 0.05 significance level to stay.

Table 4

STEPWISE REGRESSION ANALYSIS

	JAN	FEB	MAR	APR	MAY	JUN
No. of variables	4	4	13	11	11	13
R-Square	.5862	.7542	.9595	.9498	.9533	.9808
Adj. R-Square	.5390	.7261	.9393	.9300	.9350	.9712
Std. Error dam^3	1192	919	432	465	432	247
IFCST SE dam^3	1352	1072	670	699	611	385

Examination of the stepwise model statistics shows some interesting results. The March through June models are obviously overfitted. The stepwise technique appears to find more "significant" variables than parsimony would allow us to accept. The large difference between the regression standard error and IFCST SE also suggests overfitting, although all the spring models produce a significantly better forecast than the current regression model. The regression standard errors for the January and February models look better than the current model statistics, but the IFCST SE shows them to have roughly the same prediction capability.

Principal Components Model

The 1 April forecast principal components model was fit to the April-August runoff using the same nine variables as in Tables 1 and 2. Table 5 shows the models suggested by analysis of the principal component eigenvalues and component r-squares.

Table 5

LIBBY APRAUG FORECAST ON 1 APR
PRINCIPAL COMPONENTS REGRESSION WITH 9 VARIABLES

VARIABLES USED: FSTNOV WASBP11 BRIBP12 WASBP1 WASBP2 GRPBP2 APR4 APR6 APR9

COMPONENTS :	1	2	3	4	5	6	7	8	9
EIGENVALUES	3.2365	1.3956	1.0789	0.9835	0.7574	0.6003	0.4644	0.2744	0.2090
COMPONENT R2	0.6138	0.1179	0.0010	0.0107	0.0144	0.0003	0.0679	0.0010	0.1022

COMPONENTS USED	TOTAL RSQ	ADJST RSQ	STD ERROR	IFCST SE
1,2,3,4,5,6,7,8,9	0.9292	0.9086	524	728
1,2,3,4,5,7,9	0.9278	0.9125	513	748
1,2,3,4,7,9	0.9135	0.8982	553	770
1,2,3,7,9	0.9028	0.8890	578	780
1,2,3,9	0.8349	0.8166	743	885
1,2,3,4	0.7434	0.7148	926	1063
1,2,3	0.7327	0.7110	932	1014

Examination of the eigenvalues and component r-squared statistics show that the ninth component, although of little significance in its contribution to the independent variable total variance, is of certain significance in its contribution to explaining the variance in the component regression model. Most of the r-squared value is contained in components 1, 2, and 9. But in consideration of the eigenvalues, it is wise to attempt to retain as many of the first three or four components as is practical. The model using components 1, 2, 3, 7, and 9 appears to achieve the best balance of the above considerations. Five variates are fitted in the regression model and the standard errors are certainly an improvement over the current regression model. The difference between the regression standard error and IFCST SE is of some concern as it seems to indicate an overly optimistic prediction capability, though substantially better than any previous parsimonious model.

CONCLUSIONS

Historical water supply regression models use several techniques which can often hamper their effectiveness as forecasting tools. Although variable selection by total enumeration is rarely possible in consideration of the large variable pool to choose from, Stepwise regression and principal components regression offer some solutions worthy of further evaluation. A Jackknife statistic derived as an extension of split-sample testing is presented. It is useful in model evaluation by offering additional insights into model overfitting and predictive capabilities.

REFERENCES

Breiman, L. and D. Freedman (1983) How Many Variables Should Be Entered in a Regression Equation? Journal of the American Statistical Association 78 (381) 133-136

Cassidy J.J. and D.P. Lettenmaier, eds., (1985) A Critical Assessment of Forecasting in Western Water Resources Management. American Water Resources Assoc., Bethesda, MD

Hocking, R.R. (1976) The Analysis and Selection of Variables in Linear Regression. Biometrics 32,1-49

Kisiel C.C. (1972) Applications of Principal Components, Canonical Correlation, and Factor Analysis in Hydrology. Institute of Applications of Stochastic Methods in Civil Engineering. Fort Collins, CO

Marsden M.A. and R.T. Davis (1968) Regression on Principal Components as a Tool in Water Supply Forecasting. Proceedings of the Western Snow Conference. Lake Tahoe, CA

McCuen, R.H., W.J. Rawls and B.L. Whaley (1979) Comparative Evaluation of Statistical Methods for Water Supply Forecasting. Water Resources Bulletin 15 (4) 935-947

McCuen, R.H. and W.M. Snyder (1986) Hydrologic Modeling: Statistical Methods and Applications. Prentice-Hall

Mosteller, F. and J.W. Tukey (1977) Data Analysis and Regression. Reading, MA. Addison-Wesley

Palmer P.L. (1988) The SCS Snow Survey Water Supply Forecasting Program: Current Operations and Future Directions. Proceedings of the Western Snow Conference. Kalispell, MT

SAS (1985) Statistical Analysis System Users Guide. Version 5 Edition. SAS Institute, Cary, NC

Soil Conservation Service (1970) Soil Conservation Service National Engineering Handbook, Section 22: Snow Survey and Water Supply Forecasting. Chap. 6

Stedinger, J.R., J. Grygier, and H. Yin (1988) Seasonal Streamflow forecasts Based Upon Regression. Presented at the ASCE 3rd Water Resources Operations and Management Workshop - Computerized Decision Support Systems for Water Managers, Colorado State University, Fort Collins, CO

U.S. Geological Survey (1960) Double Mass Curves. Manual of Hydrology: Part 1. General Surface-Water Techniques. Water Supply Paper 1541-B

U.S. Army Corps of Engineers (1977) Libby Dam and Reservoir Project, Kootenai River, Montana: Draft of Seasonal Volume of Runoff Forecast Procedure. Dept. of the Army, North Pacific Div., Corps of Engineers, Portland, OR

Wortman, R.T. (1986) 1986 Revisions to the Libby Forecasting Procedure (Statistical Model) Memorandum for the Record, Dept. of the Army, North Pacific Div., Corps of Engineers, Portland, OR

Zuzel, J.F. and L.M. Cox (1978) A Review of Operational Water Supply Forecasting Techniques in Areas of Seasonal Snowcover. Proceedings of the Western Snow Conference. Otter Crest, OR