WATER RESOURCES RESEARCH, VOL. 36, NO. 11, PAGES 3241–3249, NOVEMBER 2000

Alternative principal components regression procedures for dendrohydrologic reconstructions

Hugo G. Hidalgo
Civil and Environmental Engineering Department, University of California, Los Angeles

Thomas C. Piechota
Department of Civil and Environmental Engineering, University of Nevada, Las Vegas

John A. Dracup
Civil and Environmental Engineering Department, University of California, Los Angeles

Abstract. Streamflow reconstruction using tree ring information (dendrohydrology) has traditionally used principal components analysis (PCA) and stepwise regression to form a transfer function. However, PCA has several procedural choices that may result in very different reconstructions. This study assesses the different procedures in PCA-based regression and suggests alternative procedures for selection of variables and principal components. Cross-validation statistics are presented as an alternative for independently testing and identifying the optimal model. The objective is to use these statistics as a measure of the model's performance to find a conceptually acceptable model with a low prediction error and the fewest number of variables. The results show that a parsimonious model with a low mean square error can be obtained by using strict rules for principal component selection and cross-validation statistics. Additionally, the procedure suggested in this study results in a model that is physically consistent with the relationship between the predictand and the predictor. The alternative PCA-based regression models presented here are applied to the reconstruction of the Upper Colorado River Basin streamflow and compared with results of a previous reconstruction using traditional procedures. The streamflow reconstruction proposed in this study shows more intense drought periods, which may influence the future allocation of water supply in the Colorado River Basin.

1. Introduction

Dendroclimatic analysis has long been used to extract hydroclimate signals from tree ring chronologies. The climatic information stored in trees in the form of ring width and wood density allows researchers to reconstruct hydroclimatic time series such as precipitation, streamflow, and the Palmer Drought Severity Index with annual resolution. Expressed as a mathematical transfer function, this relationship allows us to use the information from trees (predictor) to reconstruct past unrecorded hydroclimatic conditions (predictand). In dendroclimatology it is common to use principal components analysis (PCA) in the formulation of the transfer function that relates the variation between the predictor and the predictand. Applications of PCA in dendroclimatology include *Stockton and Jacoby* [1976], *Fritts* [1991], *Meko et al.* [1993], *Brockway and Bradley* [1995], and *Meko* [1997]. Comparison between orthogonal spatial regression and canonical regression is given by *Cook et al.* [1994].

The focus of this paper is to evaluate different PCA regression model procedures used in dendrohydroclimatic reconstructions and to use the best ones to compare the results to traditional PCA-based reconstructions. It will be shown that PCA results and subsequent regression results can vary significantly depending on several PCA procedural choices. The main procedural choices include the number of principal components to retain, whether or not to rotate the principal components, and the measure of skill used to assess the models. Cross validation is also presented as a method for independently testing the model and evaluating the best subset of predictors from a data set. These procedures have been previously used in the field of hydrology to form better hydrologic forecasting models (e.g., *Garen* [1992]), but they have not been used in dendroclimatology for the reconstruction of hydrologic variables such as streamflow.

The procedures selected as the best ones in this paper are evaluated in a streamflow reconstruction case study using standardized tree ring growth indices. The case study is the Upper Colorado River Basin (UCRB), which is the most important river basin in the southwestern United States in terms of water resource usage. This paper presents a comparison of a previous reconstruction by *Stockton and Jacoby* [1976] (hereinafter referred to as SJ) with the streamflow reconstruction performed using the selected procedures for PCA regression and predictor subset evaluation.

Paper number 2000WR900097.
0043-1397/00/2000WR900097$09.00

2. Data Sources

The tree ring index chronologies for the UCRB were obtained from the National Atmospheric and Oceanic Administration (NOAA) International Tree Ring Data Bank (available on the World Wide Web at http://www.ngdc.noaa.gov/paleo/

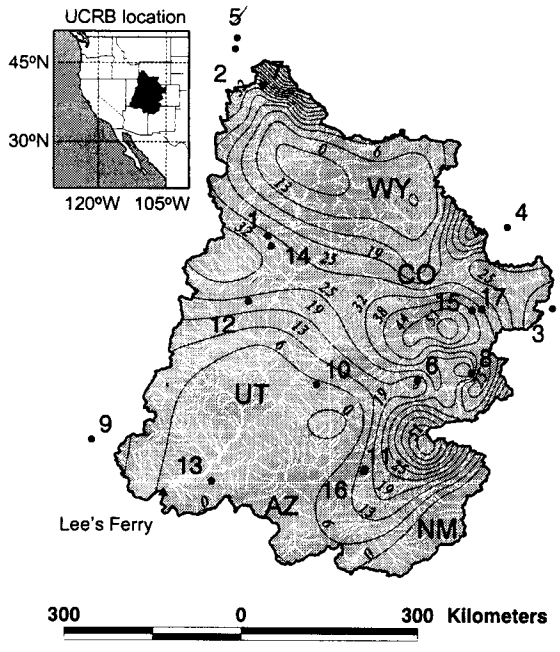

Figure 1. The locations of the 17 tree ring site chronologies in the Upper Colorado River Basin used in this study. Annual water yield contours shown (mm/yr) were computed using data from U.S. Geological Survey (available on the World Wide Web at http://water.usgs.gov).

treering.html). A tree ring index chronology is a standardized record of tree growth. Standardization removes the inherent growth trend in the raw tree ring data due to the normal physiological aging processes. From the 17 chronologies selected to represent the UCRB, 13 of them are the same ones used by SJ. Four of the SJ original sites were not available in the NOAA International Tree Ring Data Bank; these chronologies were replaced by sites located near the original SJ sites and have similar statistical characteristics. Location of the chronologies can be found in Figure 1 and the sites characteristics are listed in Table 1.

The common streamflow data set used for streamflow model calibrations in the Upper Colorado River Basin is the Lee's Ferry record. Lee's Ferry is located at the legal dividing point between the Upper and the Lower Colorado River Basins (Figure 1). An annual unimpaired streamflow record for Lee's Ferry from 1896 to 1995 was obtained from the United States Bureau of Reclamation (USBR) (available on the World Wide Web at http://www.usbr.gov/main/index.html). However, only data from 1914 to 1963 were used owing to the following reasons. First, the majority of the chronologies in the SJ study ended in 1963, which also corresponds to the construction of Glen Canyon Dam and Lake Powell. For consistency, this study only uses streamflow data up to 1963 for calibration to allow comparison of our results to the original 1976 study by SJ. Second, it should be noted that the streamflow data from 1896 to 1913 were extrapolated from distant stations and are not as reliable as the data after 1913 (SJ). The data from 1914 to 1922 were compiled from the three main tributaries of the Upper Colorado River Basin and are judged to be reliable for hydrologic studies (SJ). In 1923, a stream gauge was installed at Lee's Ferry. Only the data deemed more reliable (1914–1963) were used in this study and compared to the SJ model having the same calibration period ("50 year calibration period" model by SJ).

The correlation between streamflow at Lee's Ferry and the 17 tree ring chronologies is presented in Table 1 in the correlation criterion column. Of the 17 chronologies, chronologies 4–17 have a significant correlation [*Panofsky and Brier*, 1968; *Fritts*, 1991] with streamflow at the 95% confidence level. The cross-correlation matrix between the tree ring chronologies is presented in Figure 2. The order of the chronologies in this matrix, as well as in Table 1, is based on increasing lag 0 correlation with streamflow. The results show that chronologies with a high correlation with streamflow also have a high cross correlation (Table 1 and Figure 2).

Tree ring chronologies are known to have a relatively high

Table 1. List of Tree Ring Chronologies Used in This Study

Site Number	Site Name	Location	Year	Identification Number	SPID	ELEV, m	Correl. Criterion	s.d.	rlag1	M.S.
1	Unita Mountains A	Utah	1972	277550	PCEN	3353	0.14	0.14	0.67	0.11
2	Gros Ventre	Wyoming	1972	316597	PIFL	2179	0.17	0.28	0.47	0.26
3	Chicago Creek	Colorado	1965	115549	PSME	2835	0.22	0.39	0.26	0.40
4	New North Park	Colorado	1965	110549	PSME	2469	0.31	0.37	0.54	0.31
5	Uhl Hill	Wyoming	1972	318599	PIFL	2225	0.36	0.29	0.52	0.27
6	Black Canyon	Colorado	1965	117549	PSME	2426	0.41	0.35	0.52	0.31
7	Wind River Mountains D	Wyoming	1972	283590	PIFL	2500	0.47	0.26	0.51	0.21
8	Upper Gunnison	Colorado	1965	116549	PSME	2530	0.54	0.34	0.38	0.38
9	Mammoth Creek	Utah	1990	MAM519	PILO	2590	0.56	0.37	0.17	0.41
10	La Sal Mountains A	Utah	1972	285620	PIED	2323	0.57	0.33	0.42	0.34
11	Bobcat Canyon	Colorado	1972	61099	PSME	2042	0.62	0.43	0.25	0.47
12	Nine Mile Canyon	Utah	1965	123549	PSME	1920	0.64	0.41	0.41	0.39
13	Navajo Mountain	Utah	1972	133099	PIED	2286	0.66	0.44	0.21	0.51
14	Unita Mountains D	Utah	1972	280620	PIED	2289	0.69	0.32	0.46	0.31
15	Eagle	Colorado	1965	112549	PSME	1951	0.69	0.35	0.62	0.28
16	Sch. Old Tree 1	Colorado	1964	640106	PSME	2103	0.69	0.45	0.30	0.51
17	Eagle East	Colorado	1965	113629	PIED	2164	0.77	0.29	0.34	0.31

The year column corresponds to the year when the chronology was sampled. SPID refers to the following tree species: PCEN, *Picea engelmannii*; PIFL, *Pinus flexilis*; PSME, *Pseudotsuga menziesii*; PILO, *Pinus longaeva*; PIED, *Pinus edilus*. ELEV is the elevation in meters above sea level; Correl. Criterion is the correlation between the tree ring index and streamflow; s.d. is the standard deviation; rlag1 is the lag 1 autocorrelation coefficient; M.S. is the mean sensitivity [*Fritts*, 1976]. All the statistics are computed for the time period from 1493 to 1963, except the correlation criterion, which is computed over the 1914–1963 time period.

Site	#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Unita Mountains A, UT	1																
Gros Ventre, WY	2	0.07															
Chicago Creek, CO	3	-0.33	0.28								-0.33						
New North Park, CO	4	-0.02	0.31	0.22							0.39						
Uhl Hill, WY	5	0.16	0.37	-0.07	0.46						0.59						
Black Canyon, CO	6	-0.06	0.39	0.40	0.20	0.26					0.77						
Wind River Mtns. D, WY	7	0.04	0.34	0.28	0.24	0.59	0.30										
Upper Gunnison, CO	8	0.29	0.20	0.23	0.30	0.13	0.54	0.25									
Mammoth Creek, UT	9	0.04	-0.04	0.20	0.21	0.23	0.23	0.49	0.32								
La Sal Mountains A, UT	10	0.19	0.23	0.05	0.16	0.20	0.47	0.34	0.49	0.35							
Bobcat Canyon, CO	11	0.05	0.22	0.20	0.17	0.27	0.52	0.53	0.44	0.67	0.52						
Nine Mile Canyon, UT	12	0.21	0.18	0.28	0.30	0.18	0.31	0.33	0.45	0.58	0.58	0.60					
Navajo Mountain, UT	13	0.09	0.11	0.10	0.09	0.17	0.38	0.50	0.37	0.70	0.49	0.80	0.50				
Unita Mountains D, UT	14	0.09	0.24	0.35	0.34	0.51	0.50	0.58	0.41	0.54	0.48	0.50	0.55	0.49			
Eagle, CO	15	0.30	0.28	0.26	0.27	0.28	0.48	0.36	0.67	0.38	0.36	0.54	0.45	0.57	0.59		
Sch. Old Tree #1, CO	16	0.15	0.26	0.22	0.25	0.34	0.50	0.59	0.46	0.65	0.50	0.85	0.61	0.77	0.62	0.57	
Eagle East, CO	17	0.08	0.10	0.11	0.39	0.35	0.18	0.32	0.56	0.36	0.45	0.38	0.49	0.38	0.51	0.53	0.43

-0.33 : r between -0.30 to -0.29
0.39 : r between 0.30 to 0.49
0.59 : r between 0.50 to 0.69
0.77 : r =0.70 or more

Figure 2. The cross-correlation matrix for the 17 chronologies used in this study.

degree of autocorrelation, even after detrending, caused by the biological carryover effects from year to year. To account for this characteristic of tree ring data, it is common practice to include lagged versions of the chronologies of standardized tree ring widths in the reconstruction model [*Fritts*, 1976, 1991; *Cook and Kairiukstis*, 1990]. The lagged chronologies were included in the model used in section 5.3 to compare it with a previous streamflow reconstruction in the basin by SJ.

3. Principal Component Analysis (PCA) Regression Models

A common problem in dendroclimatological reconstructions is the presence of multicollinearity or linear codependancy among the predictors, in this case tree ring chronologies. Because of the high autocorrelation of tree ring chronologies (section 2.1) the inclusion of lagged time series in dendroclimatological reconstruction models increases the possibility of having problems associated to multicollinearity on the results of these models.

Linear regression is based on the assumption that the independent variables are not significantly correlated. When highly intercorrelated predictors are used in a multiple linear regression model, multicollinearity can become the cause of statistically imprecise and unstable estimates of regression coefficients, incorrect rejection of variables, and numerical inaccuracies in computing the estimates of the model's coefficients [*Cureton and D'Agostino*, 1983; *Weisberg*, 1985; *Fritts*, 1991; *Jennrich*, 1995]. In addition, including too many variables may result in an undesirable effect of "over fitting" the model, making it able to predict even the smallest variations from noise in the observed data but with a low predictive skill [*Jackson and Chan*, 1980; *Cureton and D'Agostino*, 1983; *Jennrich*, 1995].

By using PCA the original data set can be transformed into linear combinations of the original variables to create a new set of variables or principal components (PCs) that are independent of one another (i.e., orthogonal). PCs are extracted using an eigenmode analysis from either the correlation or the covariance matrices of the original variables. In this study, the PCs were extracted from the correlation matrix. In PCA the number of PCs is equal to the number of original variables, and the PCs are usually presented in order of greatest to least amount of variance explained from the original data set. If there is a high degree of multicollinearity in the data set, most of the variance can be explained with a fewer number of PCs than original variables. The PCs can also be used as predictors in a regression model, removing multicollinearity problems among the independent variables.

In the case of streamflow reconstructions using tree ring chronologies, the number and selection of which PCs and predictors to be included in the final model and deciding whether or not to rotate (section 3.4) the PCs must be carefully evaluated. The possible models that can be built using these alternatives are shown as models A–H in Figure 3. A more detailed explanation about each of the alternatives shown in Figure 3 will be given in sections 3.1–3.4.

3.1. Truncation or Preselection of Principal Components

The selection of significant PCs, or truncation, is accomplished by prescreening the PCs, using an objective criterion before they are included in the regression part of the model. Truncation of PCs is a topic of conflicting opinions. Some authors [*Haan*, 1977; *Garen*, 1992] (hereinafter referred to as GA) suggest that there is no need to truncate PCs because the t test in a regression model will identify the significant PCs. Other authors [*McCuen*, 1985; *Cook and Kairiukstis*, 1989; *Fritts*, 1991] prefer to truncate PCs based on the assumption that the final PCs represent variations that belong to small-scale features. It is assumed that these PCs do not increase the overall skill of the model.

Several truncation procedures have been developed for identifying the significant modes from a PCA. For PCA-based dendroclimatic reconstructions, a list of the most commonly used procedures is given by *Fritts* [1991]. In the present study, the critical eigenvalue rule [*Kaiser*, 1958] is used for PCs rotation. The critical eigenvalue rule keeps only the PCs that have an eigenvalue ≥ 1 (corresponding to the amount of information contained in a single variable). A PC with an eigenvalue <1 is not considered to be significant.

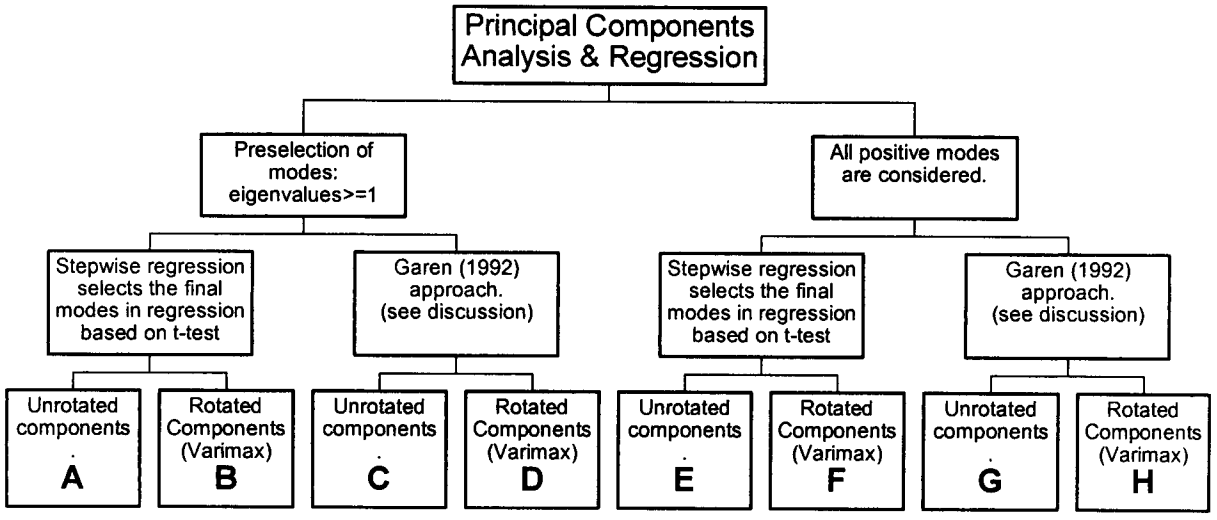

Figure 3. Schematic of the eight different modeling approaches investigated in this study. All models are tested using the cross-validation standard error (CVSE).

3.2. Stepwise Regression and Principal Component Selection

After the truncation of PCs, stepwise regression is used to select which PCs will be part of the final regression model [*Haan*, 1977; *Cureton and D'Agostino*, 1983]. SJ used this type of selection on the original reconstruction of the Upper Colorado River streamflow.

An undesirable effect of stepwise regression is that it allows selection of nonconsecutive PCs (GA). For example, the first, second, fifth, and tenth PCs could be selected for a regression model according to stepwise regression procedures. The skipping of PCs may result in regression coefficients for some of the original predictor variables that have the opposite sign of their initial correlation with the predictand. A model of this type may give results that are neither consistently accurate over time nor conceptually acceptable. Skipping PCs also suggests that there are major modes of variability in the data set that are unrelated to the dependent regression variable. If this is the case, it would be preferable for the variables that represent this variability to be removed from the analysis.

3.3. Alternative Procedure for Principal Component Selection

GA, based on *McCuen* [1985], gives an alternative procedure to stepwise regression for PCs selection. This procedure results in a more parsimonious model that better represents the physical system and has better predictive skill than a model created using stepwise regression. This procedure uses the t test and a "sign test" as the criteria for retaining variables. The t test is used to test the significance of the coefficient of the PC in the regression equation. The sign test is passed if the algebraic signs of the regression coefficients of the PCs expressed in terms of the original variables match the algebraic signs of the correlation coefficients (correlation criterion in Table 1) of these original variables with the dependent variable.

The following summarizes the alternative procedure for PC selection. First, test PC1 using a t test. If PC1 passes the t test, compute the regression coefficients in terms of the original variables and perform a sign test. If both the t test and sign test are passed, then accept PC1. Next, test PC2, as skipping PCs is not allowed. If PC2 does not pass the t test, then only retain PC1, and the procedure is finished. If PC2 passes the t test and if the regression coefficients in terms of the original variables using PC1 and PC2 in the model pass the sign test, then continue and test PC3. If PC2 passes the t test but fails the sign test, retain PC2 temporarily and test PC3. Then, if PC3 fails the t test, only retain PC1 in the final model. If PC3 passes the t test and if the sign test passes, then continue and test PC4, and so on. The procedure continues until the next PC does not pass the t test and the addition of this PC to the model causes the sign test to fail.

3.4. Rotation of Principal Components

Rotation is a procedure intended to simplify interpretation of PCs or placing physical significance to the PCs easier. A thorough discussion of reasons for rotation of PCs is given by *Richman* [1986]. In this study, both rotated and unrotated PCA are presented and compared. The method of rotation used here was programmed with Matlab software version 5.0 based on the Varimax criterion for factor rotation [*Kaiser*, 1958] and includes the modifications suggested by *Nevels* [1986] and *ten Berge* [1995].

4. Independent Testing Using Cross Validation

There is a growing body of research that suggests that independent testing techniques can improve the overall accuracy of a regression model [*Jackson and Chan*, 1980; *Michaelsen*, 1987; *Elsner and Schmertmann*, 1994; *Shao and Tu*, 1995; GA]. One of these techniques is minimization of the cross-validation standard error (CVSE) [*Michaelsen*, 1987]. CVSE has been used by GA to select models with better predictive skill and is defined as

$$\mathrm{CVSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2}{n - p}}, \qquad (1)$$

where y_i is the observed streamflow for year i; $\hat{y}_{(i)}$ is the fitted response of the ith year computed from the fit with the ith

observation removed, n is the number of years in the data set, and p is the number of regression coefficients.

The CVSE is used as an objective measure to optimize the different PCA-based models shown in Figure 3. The algorithm used for variable selection for each of the alternative models for Figure 3 is shown in Figure 4. This algorithm determines the model as well as the subset of tree ring variables that has the highest skill (lowest CVSE). First, the algorithm finds the lowest CVSE for each predictor independently. Next, the lowest CVSE for the two variables combination is found. The procedure is continued up to the total number of variables. If the minimum CVSE for combinations with an added variable is larger than the previous minimum CVSE, the program is stopped, and the extra variable is not included. This search procedure is similar to the one used by GA; although it may not necessarily find the global optimum of all combinations of variables, it rewards near-optimal parsimonious models.

For the models that used unrotated PCs, an optimum subset of tree ring variables was found that minimized the CVSE. An independent optimization for the rotated PCs could not be performed because the estimated time to compute the results was prohibitively long when the rotation subroutines were included. Instead, a special logic for selection of the subsets of variables to be tested for rotated components was developed. The rotated PCA-based models start by testing the variables that minimized the corresponding unrotated solution, keeping the other alternatives fixed. That is, the CVSE was calculated for models B, D, F, and H using the variables identified in models A, E, C, and G, respectively. Additional rotated models were tested by adding up the remaining variables one at a time to this basic subset. If the CVSE of the model with the additional variable is larger than with the basic set of variables, then no variable is added. Additionally, starting from the basic set again, rotated models were tested by exploring changing one up to the four of the last variables of the set (constraint dictated by computing time) while keeping the rest of the basic set. For example, for model F, the basic set would be the variables from model E: 17, 16, 14, 13, and 5 (Table 2). Additional models were tested to compute the CVSE by changing (or deleting) the last four chronologies of the set. In other words, we evaluated the models that result from the combination of variable 17 with all remaining combinations from one to four variables.

5. Results

5.1. Cross-Validation Standard Error

The results of the models identified in Figure 3 are presented in Table 2. The CVSE is compared with other verification statistics (explained in section 5.2) commonly used for tree ring reconstruction models [*Fritts*, 1991]. The PCs and the variables that are used to form the different PCs are also shown. There are a total of 17 possible variables in this section, which correspond to the number of tree ring sites. The "complete" model (using all variables) is shown as a comparison with more parsimonious models for each of the alternative procedures. In all cases, the complete model had a higher CVSE than the other models, showing that the inclusion of more variables does not necessarily improve the predictive skill of the model.

All models based on the GA approach were found to retain only the first PC. This suggests that the size of the UCRB is small enough that the climate signal common to all variables belongs to a single climate regime that influences most of the basin. In contrast, the stepwise regression method selected one to four PCs. It should be noted that the correlation coefficients are similar for both approaches.

Truncation of the PCs did not influence the models based on the GA approach because this type of model used only the first PC. For stepwise regression, however, better results are obtained when all the PCs (i.e., no truncation) are considered in the model. The best models using the GA methodology are obtained by using unrotated PCs. In contrast, the stepwise regression approach gives better results using untruncated rotated PCs. This is logical since the rotation of the PCs distributes the variance of the original time series more equally among the PCs. The unrotated solution has a large portion of the variance in the first PC, and the amount of variance in the following PCs drops off much faster than in the rotated solution. The rotation of PCs diminishes the high contribution placed on the first PC, and this affects the GA approach, which favors the first PC. The opposite effect is observed in the stepwise regression selection, which gives importance to some of the latter PCs.

The untruncated rotated stepwise regression model (F in Table 2) has the lowest CVSE (2590.34 million m^3/yr) among all the models, although it is not the most parsimonious model (Table 2). The method suggested by GA selected the model with the fewest variables (one less variable than the stepwise regression) and had a CVSE just slightly higher (2659.42 million m^3/yr) than the best stepwise regression model (2590.34 million m^3/yr).

5.2. Other Validation Statistics

Table 2 shows other validation statistics that are commonly used in dendroclimatology studies. Similar to the CVSE, the reduction of error (RE) statistic [*Lorenz*, 1956; *Gordon and LeDuc*, 1981; *Fritts*, 1976, 1991] is a verification tool that is used on independent data to assess the data's reliability. RE varies from negative infinity (infinite error) to 1.0 (perfect estimation). Any positive value of RE indicates some skill of the model compared to a model that uses the calibration mean as the estimate. Negative RE statistics indicate that improvements are needed in the model.

The reduction of error statistic is usually divided into three PCs:

$$\mathrm{RE} = \mathrm{RISK} + \mathrm{BIAS} + \mathrm{COVAR}, \tag{2}$$

where "RISK," "BIAS," and "COVAR" are defined by *Gordon and LeDuc* [1981] and *Fritts* [1991].

RISK is always negative (ideally, RISK = -1) and represents the lower limit of RE, below which the regression reconstructions will exhibit no skill at all in reproducing the variations in physical data. It denotes the risk that the model takes in making independent estimates. Models with small explained variance will characteristically have RISK terms between -0.5 and 0.0, while overrepresented models (too many predictors) will usually have RISK terms smaller than -1 [*Gordon and LeDuc*, 1981].

BIAS can be positive or negative. It is positive if a shift in the mean of the independent sample (in our case the estimates from the deleted-one series) from the calibration sample is reproduced in the estimates. This term of the RE is of particular interest for small sample sizes.

The COVAR term reflects the strength of the correlation

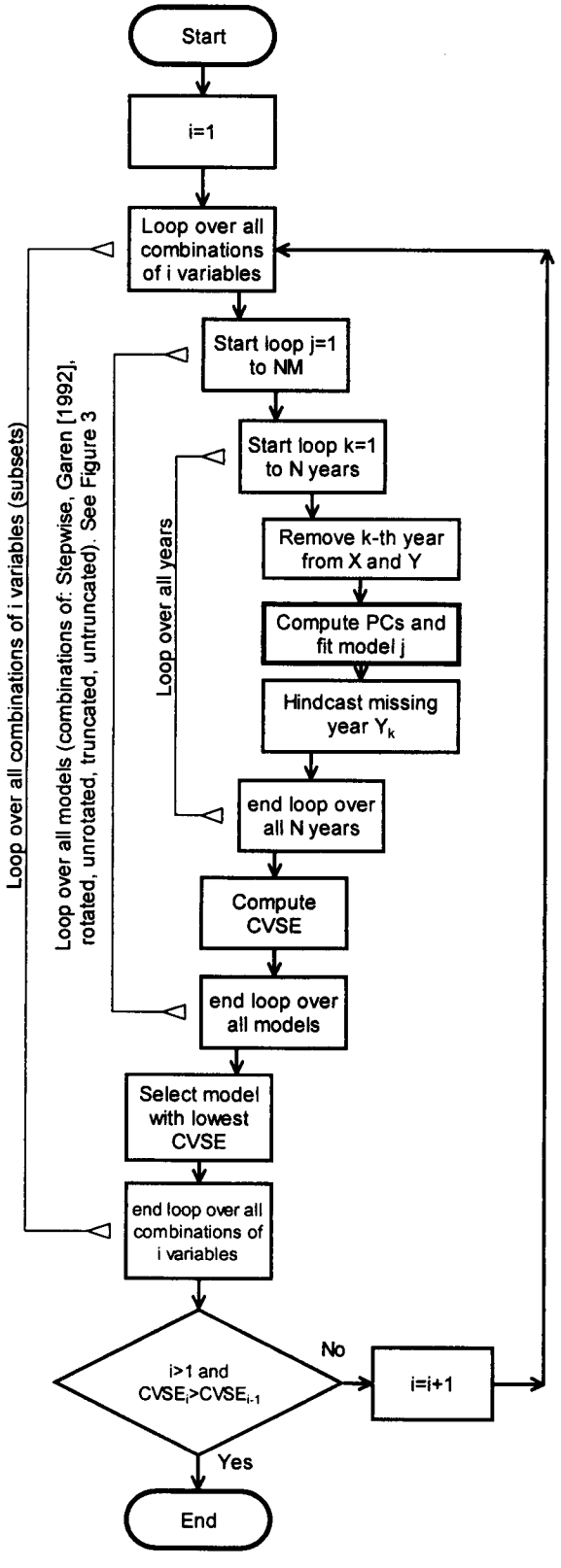

Figure 4. The algorithm used for identification of the optimal model parameters. Given: $X(1 \ldots N, 1 \ldots p)$, predictor variables matrix, where p is the maximum number of predictor variables and N is the number of years. Y: predictand variable matrix, $Y(1 \ldots N, 1)$. NM is number of different models, in this case the eight models shown in Figure 3. CVSE is cross-validation standard error. Note that the best variable subset combination is the one that minimized CVSE_{i-1}.

between the estimated and the observed data and measures the similarity of the temporal patterns in the estimates and observations [*Fritts*, 1991]. It is usually the most important factor of RE.

Dividing RE in this way aids in identifying the limitations of the models, especially the ones with negative RE. For example, models with a good correlation with the independent data but with a small RISK term suggest that the model may duplicate the patterns of variation but contain no appreciable amount of variance [*Gordon and LeDuc*, 1981].

The results in Table 2 show that the RE is not as sensitive for model selection as the CVSE when the verification data is represented by the estimates from the deleted-one (equation (1)) streamflow series. RE gives a coarse estimation of which models perform better, but it may be more valuable for cases where a longer calibration verification period is used (bootstrap) or in cases where the differences between models are more evident.

The C_p statistic [*Mallows*, 1973] for variable subset evaluation is also included in Table 2. The best models have the lowest C_p, and its relative value is dependent on the choice of the estimate for the real error variance. We used the residual mean square of the model using all predictors for the estimate of the error variance. On the basis of the definition of C_p, if n is relatively large, small subsets of variables (p small) may result in valid negative values of C_p. The C_p statistic and CVSE both reward a parsimonious model and a more efficient variable set for prediction. However, the C_p statistic showed more variability than CVSE. It is encouraging that both C_p and CVSE identify the same optimal model using the PCs selection procedure of section 3.3 (GA). For the stepwise approach, C_p and CVSE have different minima, with C_p preferring a stepwise model that does not skip PCs, based on the same principles discussed in section 3.2.

5.3. Comparison With *Stockton and Jacoby*'s [1976] Previous Reconstruction

An untruncated, unrotated PCA model using the PCs selection procedure described in section 3.3 was used to reconstruct Lee's Ferry streamflow and to compare it with the reconstruction done by SJ with a stepwise regression model that allowed skipping of PCs. The PCs were computed for a calibration period from 1914 to 1963. The calibrated models use lagged $(-1, 0, +1, +2)$ chronologies, so that all 68 variables (17 chronologies times the 4 lag times) were treated as separate variables. The subset of chronologies that resulted in the lowest CVSE was found using the algorithm of Figure 4.

The results of the model developed in this study are presented in Table 3. A comparison between the streamflow reconstructions from the traditional stepwise regression model and the model formed with the procedures from this study is shown in Figure 5.

The use of lag chronologies required a modification to the CVSE criterion as suggested by *Meko* [1997]. When making an independent prediction for year i, the three lag years ($i - 2$, $i - 1$, and $i + 1$) are deleted in addition to the ith year. This procedure is repeated for each ith year, ensuring a truly independent test.

The chronologies selected as the best streamflow predictors from the model using the *Garen* [1992] approach are noted below in the regression equation. In terms of the PCs, the final calibration equation is

$$Q = 3098.91\ \text{PC1} + 16030.13, \qquad (3a)$$

Table 2. Summary of the Results for the Models Identified in Figure 3

PCs	CVSE, $\times 10^6$ m^3	EXP. VAR.	C_p	RMSE, $\times 10^6$ m^3	RISK	BIAS	COVAR	RE	Variables
				Model A: Stepwise and Unrotated					
1	3189.82	0.734	5.59	2692.72	−1.002	1.892	0.085	0.975	17, 14, 13
1	3941.02	0.640	36.53	3135.54	−0.976	1.861	0.088	0.973	1–17
				Model B: Stepwise and Rotated					
1, 2	2640.91	0.790	2.99	2421.35	−0.987	1.859	0.111	0.983	17, 14, 13, 6
2, 4, 8	4013.79	0.744	56.58	2702.59	−1.001	1.876	0.097	0.972	1–17
				Models C and G: Garen *[1992] and Unrotated*					
1*	2659.42*	0.771*	−4.77*	2500.29*	−0.983*	1.857*	0.109*	0.983*	17*, 14,* 13*
1	3770.79	0.680	38.06	2956.69	−0.979	1.861	0.093	0.975	1–17
				Models D and H: Garen *[1992] and Rotated*					
1	3189.82	0.734	5.59	2692.72	−1.002	1.892	0.085	0.975	17, 14, 13
1	3941.02	0.640	36.53	3135.54	−0.976	1.861	0.088	0.973	1–17
				Model E: Stepwise and Unrotated					
1, 3	2591.57	0.798	6.46	2372.01	−0.985	1.857	0.112	0.984	17, 16, 14, 13, 5
1, 5	3863.31	0.722	47.56	2785.23	−0.979	1.854	0.099	0.974	1–17
				Model F: Stepwise and Rotated					
1, 3*	2590.34*	0.795*	13.52*	2390.51*	−0.987*	1.860*	0.111*	0.984*	17,* 14,* 13,* 6*
2, 4, 9, 13	3704.19	0.806	63.82	2375.71	−0.976	1.847	0.105	0.976	1–17

PCs are the principal components included in each model. CVSE is the cross-validation standard error in million cubic meters per year. EXP. VAR. is the explained variance (coefficient of determination). C_p is the C_p statistic for subset evaluation [*Mallows*, 1973]. RMSE is the root-mean-square error in million cubic meters per year. RISK BIAS, and COVAR are the constituents of RE, which is the reduction of error statistic [*Fritts*, 1991]. Variables are the chronologies used in the reconstruction model. Nonsignificant components were truncated from models A to D, while in models E to H, no truncation was performed. Model C gave the same results as model G, so they are shown together; the same applies for models D and H. The first row of each model represents the combination of variables that resulted in the lowest CVSE using that particular type of model. The second row of each model represents the results using all 17 variables.

*These data show the best model for the GA and the stepwise component selection.

where PC1 is the PC for the six chronologies identified as the best predictors. In terms of the original variables,

$$Q = 660.04\ \mathrm{CC} + 880.72\ \mathrm{NN} + 643.39\ \mathrm{UG} + 1191.80\ \mathrm{NM} + 1377.44\ \mathrm{UM} + 2297.51\ \mathrm{EE} + \sout{1377.44\ \mathrm{UM}} + \sout{2297.51\ \mathrm{EE}} + \sout{1377.44\ \mathrm{UM}} + \sout{2297.51\ \mathrm{EE}} + \sout{1377.44\ \mathrm{UM}} + \sout{2297.51\ \mathrm{EE}} + \sout{1377.44\ \mathrm{UM}} + \sout{2297.51\ \mathrm{EE}} + 16030.13, \quad (3b)$$

where Q is the reconstructed annual natural streamflow at Lee's Ferry in million cubic meters and the other abbreviations represent the standardized tree ring growth index for CC, Chicago Creek (site 3) at lag +1; NN, New North Park (site 4) at lag −1; UG, Upper Gunnison (site 8) at lag +1; NM, Nine Mile Canyon (site 12) at lag 0; UM, Unita Mountains site D (site 14) at lag 0; EE, Eagle East (site 17) at lag 0.

As expected, the sites selected are located in the upper part of the Colorado River Basin, where the runoff yield is high. It should be noted that the very high yield sites in the upper part of the Green River, Wyoming, (sites 2, 5, and 7) were not selected by the model over the sites in the upper part of the state of Colorado (sites 3 and 4). One reason may be that the tree species is playing some role in the identification of the best chronologies for streamflow reconstruction in this particular region. In general, the *Pseudosoga mensiesii* and *Pinus edilus* are preferred over *Pinus flexilis* and *Picea engelmannii*.

The SJ model used six PCs that were not consecutive. It is encouraging that our coefficient of determination and the estimate of the root-mean-square error (Table 3) for the calibration over the years 1914–1961 showed that our model has a better fit. Moreover, the six PCs used in the SJ study are composed of 68 variables (representing 17 tree ring chronologies times 4 lags), and there may be some duplicate information that artificially inflates the real predictive skill of the model.

Table 3. Comparison of Statistical Characteristics Between the Model Presented in This Study and *Stockton and Jacoby* [1976] Reconstruction of the Colorado River Streamflow at Lee's Ferry

Model	PCs Used	Var. Used	CVSE, $\times 10^6$ m^3	EXP. VAR.	RMSE, $\times 10^6$ m^3	RISK	BIAS	COVAR	RE
This study	1	6	2344.87	0.824	2158.62	−0.982	1.86	0.109	0.987
50 year calibration model by *Stockton and Jacoby* [1976]	1, 2, 3, 5, 10, 15	68	n.a.	0.740	4711.95	n.a.	n.a.	n.a.	n.a.

PCs are the principal components included in each model. Var. Used are the number of variables used in each model. CVSE is the cross-validation standard error in million cubic meters per year. EXP. VAR. is the explained variance (coefficient of determination). RMSE is the root-mean-square error in million cubic meters per year. RISK, BIAS, and COVAR are the constituents of RE, which is the reduction of error statistic [*Fritts*, 1991]. Unavailable information is denoted by n.a.

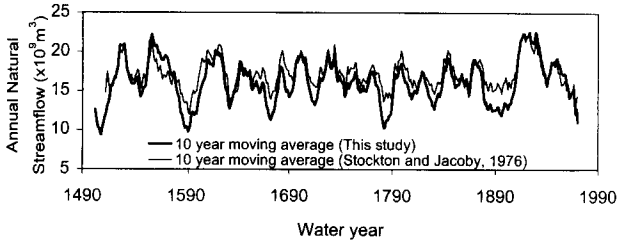

Figure 5. Comparison of the reconstruction results obtained using the *Stockton and Jacoby* [1976] approach and the model from this study. Annual streamflow is expressed in billion cubic meters at Lee's Ferry.

In Figure 5 it is clear that our model responds with more intensity to below average streamflow (droughts) than the SJ model. It is encouraging that both reconstructions show that the lowest streamflow occurred in the 1590s, 1670s, and 1780s. In addition, an extended low-flow period occurred from the 1880s to the 1910s. This suggests a near-centennial return period of extreme drought events in this region. There is also an evidence of a drought in the early 1500s that is similar in magnitude to the drought in the late 1500s, which is considered the most severe drought for water allocation in the basin [*Tarboton*, 1995]. However, this apparent drought is not as reliable as later droughts since the early periods of reconstructions are usually obtained from chronologies composed of small samples of trees to guarantee accurate results [*Fritts*, 1991].

5.4. Characteristics of the Best Predictors

Table 4 shows some statistical characteristics of the chronologies selected by the model as the best streamflow predictors. Except for the mentioned preference of certain tree species and the correlation with streamflow, individual chronologies do not present conclusive characteristics that can be used to infer their potential as good streamflow predictors. Even the correlation with streamflow does not imply a very strong relationship. In fact, some of the predictors have correlation coefficients with the streamflow data series as low as 0.30. Variables with a modest correlation with the dependent variable may contain additional information not contained in other variables with better correlation. Variable selection is determined by a balance between correlation with the dependent variable and intercorrelation among the independent variables.

6. Conclusions

The comparison of PCA-based regression techniques presented in this paper is intended to provide insights to the relative accuracy of these models for streamflow reconstruction using tree ring data. *Garen*'s [1992] methodology for PCs selection resulted in the most parsimonious models, having a low CVSE. This method also produces models that are more physically consistent than those calibrated using stepwise regression. In stepwise regression the undesirable effect of PCs skipping can lead to regression coefficients that are opposite in sign to the physical relationship between the predictor and predictand. It was also found that the minimization of the CVSE is a good tool for determining the most parsimonious model, with a low root-mean-square error (RMSE), while remaining consistent with the underlying physical processes.

A comparison of the optimized model in this study with that of the SJ reconstruction of Lee's Ferry streamflow shows that both models identify the same dry periods; however, the model developed in this study estimates with more intensity the extreme dry periods. It is not clear whether the approach suggested here is superior to the traditional stepwise regression approach; however, the differences in the streamflow reconstruction that each approach gives is worthy of additional study. These differences may be very important for the future allocation of water supply in the Colorado River Basin.

Future work will seek to find more computationally efficient procedures for identifying the best variables to be used in the model. Instead of evaluating all possible variable combinations, the prior information from an analysis of fewer variables may be useful in determining the best predictor variables. Last, it is noteworthy that we have found that the hydrologic data sets in the UCRB show evidence that the climate regime of the post-1976 period is different than the pre-1976 period, a shift that has been observed by researchers in other regions around the Pacific Rim [e.g., *Ebbesmeyer et al.*, 1991; *Graham*, 1994; *Miller et al.*, 1994; *Mantua et al.*, 1997]. It is imperative to update tree ring chronologies so all possible climate scenarios are captured in the tree ring data. This may significantly affect

Table 4. Statistical Characteristics of the Chronologies Used in the Model Selected With the *Garen* [1992] Approach Having the Lowest Cross-Validation Standard Error

Tree Rings	Nine Mile Canyon, Lag 0	Unita Mountains D, Lag 0	Eagle East, Lag 0	New North Park, Lag −1	Chicago Creek, Lag +1	Upper Gunnison, Lag +1
Lag +1 autocorrelation	0.16	0.47	0.11	0.23	0.12	−0.13
Correlation with streamflow	0.64	0.70	0.79	0.32	0.30	0.35
Mean	1.30	1.04	1.03	1.09	1.24	1.09
Standard deviation	0.44	0.34	0.26	0.27	0.38	0.34
Mean sensitivity	0.46	0.33	0.35	0.28	0.38	0.43
Standard deviation/mean	0.34	0.33	0.25	0.25	0.31	0.31

Streamflow	Lee's Ferry
Lag +1 autocorrelation	0.30
Mean, $\times 10^6$ m^3	18497.06
Standard deviation, $\times 10^6$ m^3	5090.59
Mean sensitivity	0.28
Standard deviation/mean	0.28

Lee's Ferry natural streamflow statistics are also shown for comparison purposes.

the identification of severe drought periods as represented in reconstructed streamflow data.

Acknowledgments. This work is supported by the University of California Water Resources Center under award WRC-889 and the National Science Foundation under award EAR 9421030. The authors are obliged to David C. Garen for his helpful comments, to Rod Carson from the United States Bureau of Reclamation for providing the Lee's Ferry unimpaired streamflow record, and to Judith King for editorial assistance. Two anonymous reviewers and the Associate Editor (Peter Rasmussen) are thanked for their comments that have enhanced the overall impact of this paper.

References

Brockway, C. G., and A. A. Bradley, Errors in streamflow drought statistics reconstructed from tree ring data, *Water Resour. Res.*, *31*, 2279–2293, 1995.

Cook, E. R., and L. A. Kairiukstis, *Methods of Dendrochronology: Applications in the Environmental Science*, Kluwer Acad., Norwell, Mass., 1990.

Cook, E. R., K. R. Briffa, and P. D. Jones, Spatial regression methods in dendroclimatology—A review and comparison of 2 techniques, *Int. J. Climatol.*, *14*, 379–402, 1994.

Cureton, E. E., and R. B. D'Agostino, *Factor Analysis, an Applied Approach*, L. Erlbaum, Hillsdale, N. J., 1983.

Ebbesmeyer, C. C., D. R. Cayan, D. R. McClain, F. H. Nichols, D. H. Peterson, and K. T. Redmond, 1976 step in Pacific climate: Forty environmental changes between 1968–1975 and 1977–1984, in *Proceedings of the 7th Annual Pacific Climate (PACLIM) Workshop, April 1990, Interagency Ecol. Study Program Tech. Rep. 26*, pp. 115–126, Calif. Dep. of Water Resour., Sacramento, Calif., 1991.

Elsner, J. B., and C. P. Schmertmann, Assessing forecast skill through cross validation, *Weather Forecasting*, *9*, 619–624, 1994.

Fritts, H. C., *Tree Rings and Climate*, Academic, San Diego, Calif., 1976.

Fritts, H. C., *Reconstructing Large-scale Climatic Patterns From Tree-Ring Data: A Diagnostic Analysis*, Univ. of Ariz. Press, Tucson, 1991.

Garen, D. C., Improved techniques in regression-based streamflow volume forecasting, *J. Water Resour. Plann. Manage.*, *118*, 654–670, 1992.

Gordon, G. A., and S. K. LeDuc, Verification statistics for regression models, paper presented at Seventh Conference on Probability and Statistics in Atmospheric Sciences, Am. Meteorol. Soc., Boston, 1981.

Graham, N. E., Decadal-scale climate variability in the tropical and North Pacific during the 1970s and 1980s: Observations and model results, *Clim. Dyn.*, *10*, 135–162, 1994.

Haan, C. T., *Statistical Methods in Hydrology*, Iowa State Univ. Press, Ames, 1977.

Jackson, D. N., and D. W. Chan, Maximum-likelihood estimation in common factor analysis: A cautionary note, *Psychol. Bull.*, *88*, 502–508, 1980.

Jennrich, R. J., *An Introduction to Computational Statistics, Regression Analysis*, Prentice-Hall, Englewood Cliffs, N. J., 1995.

Kaiser, H. F., The Varimax criterion for analytical rotation in factor analysis, *Psychometric*, *23*, 187–200, 1958.

Lorenz, E. N., Empirical orthogonal functions and statistical weather prediction, *Massachusetts Institute of Technology, Sci. Rep. 1*, contract AF19, Statistical Forecasting Project, Mass. Inst. of Technol., Boston, 1956.

Mallows, C. L., Some comments on Cp, *Technometrics*, *15*, 661–675, 1973.

Mantua, N. J., S. R. Hare, J. M. Wallace, and R. C. Francis, A Pacific interdecadal climate oscillation with impacts on salmon production, *Bull. Am. Meteorol. Soc.*, *78*, 1069–1079, 1997.

McCuen, R. H., *Statistical Methods for Engineers*, Prentice-Hall, Englewood Cliffs, N. J., 1985.

Meko, D., Dendroclimatic reconstruction with time varying predictor subsets of tree indices, *J. Clim.*, *10*, 687–696, 1997.

Meko, D., E. R. Cook, D. W. Stahle, C. W. Stockton, and M. K. Hughes, Spatial patterns of tree-growth anomalies in the United States and Southeastern Canada, *J. Clim.*, *6*, 1773–1786, 1993.

Michaelsen, J., Cross validation in statistical climate forecast models, *J. Clim. Appl. Meteorol.*, *26*, 1589–1600, 1987.

Miller, A. J., D. R. Canyan, T. P. Barnett, N. E. Graham, and J. M. Oberhuber, The 1976–1977 climate shift of the Pacific Ocean, *Oceanography*, *7*, 21–26, 1994.

Nevels, K., A direct solution for pairwise rotations in Kaiser's Varimax method, *Psichometrika*, *51*, 327–329, 1986.

Panofsky, H. A., and G. W. Brier, *Some Applications of Statistics to Meteorology*, Pa. State Univ., University Park, 1968.

Richman, M. B., Rotation of principal components, *J. Climatol.*, *6*, 293–335, 1986.

Shao, J., and D. Tu, *The Jackknife and the Bootstrap*, Springer-Verlag, New York, 1995.

Stockton, C. W., and G. C. Jacoby Jr., Long-term surface-water supply and streamflow trends in the Upper Colorado River Basin based on tree-ring analysis, *Lake Powell Res. Project Bull. 18*, Inst. of Geophys. and Planet. Phys., Univ. of Calif., Los Angeles, 1976.

Tarboton, D. G., Hydrologic scenarios for severe sustained drought in the southwestern United States, *Water Resour. Bull.*, *31*, 803–813, 1995.

ten Berge, J. M. F., Suppressing permutations or rigid planar rotations: A remedy against nonoptimal Varimax rotations, *Psichometrika*, *3*, 437–446, 1995.

Weisberg, S., *Applied Linear Regression*, John Wiley, New York, 1985.

J. A. Dracup and H. G. Hidalgo, Civil and Environmental Engineering Department, University of California, 5732 Boelter Hall, P.O. Box 951593, Los Angeles, CA 90095. (jdracup@ucla.edu)

T. C. Piechota, Department of Civil and Environmental Engineering, University of Nevada, 4505 Maryland Parkway, Box 454015, Las Vegas, NV 89154.

(Received May 17, 1999; revised March 27, 2000;
accepted April 7, 2000.)