Hydrologic Regionalization of Watersheds. II: Applications

Shih-Min Chiang[1]; Ting-Kuei Tsay[2]; and Stephan J. Nix[3]

Abstract: Multiple regression analysis (MRA) and a time series model (TSM) are developed and validated for using watershed characteristics to synthesize streamflow hydrographs. Relationships between the streamflow parameters and watershed variables are evaluated by canonical correlation analysis at 94 candidate watersheds. These relationships are constructed using MRA to predict streamflow parameters at six validation stations in two main hydrologic regions. The predicted streamflow parameters are applied to synthesizing specific monthly streamflows by using the developed TSM. The synthetic hydrographs are found to be mostly improved over those found from traditional simple regression equations. Statistical properties and reliability curves of the synthetic Qs are compared with those of the historical records. The statistical properties seem to be well preserved, and the reliability curves are reasonable in one hydrologic region but somewhat biased in the other. The proposed regionalization scheme is validated and therefore considered feasible and reliable for estimating monthly streamflows at ungauged sites.

DOI: 10.1061/(ASCE)0733-9496(2002)128:1(12)

CE Database keywords: Watersheds; Hydrologic properties; Regression models; Streamflow.

Introduction

Previous Studies in Streamflow Estimation

Streamflow information is required to reduce uncertainty and permit decisions on water resource planning and design to be made with increased confidence. Problems of water resources planning and design are usually classified into two categories: (1) water use (e.g., water supply, hydropower generation); and (2) water control (e.g., flood control). The design for water use concentrates on the complete hydrograph over a period of years. The design for water control focuses on extreme events of short duration (e.g., peak flows).

Generally, monthly streamflows satisfy the basic data requirements for many types of water resources projects and are usually employed for purposes of water use in time series analysis. Monthly streamflows are applied to estimating reservoir storage capacity for different purposes (e.g., water supply analysis and potential hydropower) (McMahon 1993). Streamflows required for water resource projects located at ungauged sites must be estimated. Streamflow variables of interest may include mean annual or monthly flows, seasonal patterns, flood and low-flow quantiles, maximum and minimum flows, and percentiles of the flow duration curve or flood frequency curve (Mosley and Mckerchar 1993). Alternatively, it may be necessary to estimate a hydrograph that would result from a particular sequence of rainfall events or to estimate streamflow records over several years.

In order to estimate streamflows at ungauged sites, multiple regression analysis (MRA) equations such as

$$Y = k + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_p X_p \tag{1}$$

are often used to develop the relationships for gauged watersheds. Here, dependent variable Y is the streamflow parameter of interest; independent variables $X_1, X_2, \ldots, X_p$ = watershed and climatic characteristics; k = regression intercept; and $\alpha_1, \alpha_2, \ldots, \alpha_p$ = regression coefficients. Since discharge is strongly controlled by watershed area, Aw, Eq. (1) is often reduced to a simple regression analysis (SRA) equation, such as

$$Y = k + \alpha X \tag{2}$$

or

$$Y' = k' X^{\alpha} \tag{3}$$

where Y' = streamflow parameter and k' and α = regression coefficient. With these equations, the streamflow information at an ungauged site can be extrapolated from a gauge located within the same watershed (Gulliver 1991). The value of α depends on watershed characteristics and is approximately equal to one (Black 1988). If the value of α is unavailable, it is typically set to one (Murdock and Gulliver 1993). Note that SRA only uses one independent variable to predict a single mean streamflow value.

If hydrologic regions are defined first at watersheds in Eq. (1) based on similarities of streamflow properties, these properties at an ungauged site could be more precisely estimated in the region. In the first paper (Chiang et al. 2001), a hydrologic regionalization scheme was proposed, with hydrologic regions defined by the classification of 15 streamflow parameters (i.e., 12 monthly streamflow parameters, two moving average parameters, and one variance parameter) from a time series model (TSM):

$$\ln(Q_t) = \sum_{j=1}^{12} w_j D_{ji} + (1 - \theta_1 B - \theta_2 B^2) a_t \tag{4}$$

[1]Senior Environmental Specialist, Office of Secretary General, Environmental Protection Administration, 41, Sec. 1, Chung-Hwa Rd., Taipei, Taiwan. E-mail: smchiang@sun.epa.gov.tw

[2]Professor, Dept. of Civil Engineering, National Taiwan Univ. 1, Sec. 4, Roosevelt Rd., Taipei, Taiwan.

[3]Professor, Dept. of Civil and Environmental Engineering, Northern Arizona Univ., Flagstaff, AZ 86011.

Note. Discussion open until June 1, 2002. Separate discussions must be submitted for individual papers. To extend the closing date by one month, a written request must be filed with the ASCE Managing Editor. The manuscript for this paper was submitted for review and possible publication on April 3, 2000; approved on February 28, 2001. This paper is part of the ***Journal of Water Resources Planning and Management***, Vol. 128, No. 1, January 1, 2002. ©ASCE, ISSN 0733-9496/2002/1-12–20/$8.00+$.50 per page.

where Q_t = current observation (i.e., specific monthly streamflow); and ln = natural logarithmic function. The first term on the right side of Eq. (4) represents 12 monthly seasonal components; w_j = coefficient for the jth period (e.g., w_i: specific monthly streamflow for January); t = time index; D_{jt} is an indicator variable that equals 1 if t corresponds to seasonal period j (e.g., D_{11}, D_{22}) and 0 otherwise (e.g., D_{12}, D_{21}). The second term is the irregular component and is expressed as a moving average pattern MA(2), representing a weighted aggregate of errors resulting from the current month (i.e., a_t) and the prior two months (i.e., Ba_t, B^2a_t). The two weighted values, θ_1 and θ_2, are moving average (MA) parameters; B is a backshift operator indicating the prior two months (i.e., B, B^2), and a_t are random errors assumed independently and identically distributed with normal distribution, $N(0,\sigma^2)$. The parameter σ^2 is the variance of the errors and provides a random variation of the streamflow residuals after the seasonal pattern is removed. Therefore, Eq. (4) consists of a total of 15 streamflow parameters.

The 12 monthly streamflow parameters (MSPs) are separated into "high" MSPs ($w_1, w_2, w_3, w_4, w_5, w_{12}$) and "low" MSPs ($w_6, w_7, w_8, w_9, w_{10}, w_{11}$). Mean values of the "high" (positive) and "low" (negative) MSPs characterize monthly streamflows in the wet and dry season, respectively. The 15 streamflow parameters from 94 studied watersheds were classified into six regions by cluster analysis (Chiang et al. 2001). The classified hydrologic regions [see Fig. 1 in Chiang et al. (2001)] seem to be separated by physiographical boundaries, especially the two main clusters (Cluster-1 and Cluster-3). Discriminant analysis (DA) and principal component analysis (PCA) were used to test and interpret the regional differences and similarities. Regional membership is mainly identified by watershed variables such as elevation, forest area, channel slope, and precipitation based on calculation of scores of canonical discriminant variates. If relationships between streamflow parameters and watershed variables (including regional membership) can be constructed, the streamflow parameters at ungauged sites can be estimated more accutrately. Therefore, the developed TSM [Eq. (4)] can be applied to synthesize a large number of streamflow sequences extended to a desired period.

In order to predict flow properties in a region, several relationships between streamflow properties and watershed variables have been constructed using SRA or MRA. Singh (1971) constructed the relationship between watershed area and streamflow parameters derived from flow duration curves (FDC). MRA has been applied by many researchers (e.g., Benson and Matalas 1967; DeCoursey 1973; Hawley and McCuen 1982; Tasker 1982; Gottschalk 1985; Bhaskar and O'Connor 1989; Fennessey and Vogel 1990) using many independent variables, including watershed and climatic characteristics. Note that predicted streamflow properties using MRA or SRA cannot be applied to synthesize streamflow hydrographs in these previous studies, unless the predicted streamflow properties can be used by a TSM such as Eq. (4) to synthesize streamflow.

After watersheds are classified into regions and regional relationships constructed, the results can be applied to ungauged sites in a classified region. This application involves two procedures: (1) identify the regional membership; and (2) estimate streamflows at ungauged sites. For example, Quimpo et al. (1983) utilized two parameters derived from FDC to regionalize watersheds and then constructed a regression equation to predict the parameters and synthesize the FDCs. Fennessey and Vogel (1990) applied the values of the two parameters derived from FDCs and regressed with watershed areas and elevation differences between basin summit and channel outlet. In addition, DeCoursey (1973) used canonical correlation analysis (CCA) to investigate the relationships between dependent and independent variables. In summary, the hydrologic regions are defined first, then the relationships between streamflows and watershed variables are constructed or investigated by SRA, MRA, or CCA, and streamflows at ungauged sites are predicted based on these relationships. Unfortunately, the validity of the regionalization models is seldom investigated. Only Fennessey and Vogel (1990) selected three gauging stations in order to validate their regionalization model. The major limitation in validation is that only gauged sites can be selected for validation, since there are no streamflow records available at ungauged sites.

Study Objectives

There are two different approaches to predicting streamflows: (1) information based on watershed characteristics using MRA; and (2) synthesis of streamflow using TSM. Either one has its own application limitations on synthesizing streamflows or using watershed characteristics as independent variables. In order to use watershed characteristics to synthesize streamflow hydrographs, a method of integrating MRA and TSM is developed and validated herein. MRA is used to construct relationships between the streamflow parameters in Eq. (4) (dependent variables) and watershed characteristics (independent variables). The predicted streamflow parameters from the MRA equations are used to synthesize hydrographs by the TSM in Eq. (4). The resulting streamflow synthesis produces both stochastic characteristics as well as a complete hydrograph of monthly streamflows over a desired period (e.g., 25 years). The validation procedure is emphasized, and streamflow predictions from MRA over simple regression analysis (SRA) are also compared to demonstrate the improvement of streamflow predictions. The objectives are; (1) to investigate and develop the relationships between streamflows and watershed variables; (2) to demonstrate improvements in streamflow estimations from the proposed MRA over traditional SRA models; and (3) to synthesize complete hydrographs and construct validation procedures for a hydrologic regionalization model. The general goal is to develop an integrated (MRA and TSM) model for generating reliable estimates of monthly streamflows at ungauged sites over a desired period.

Proposed Methodology

Study Area

In the first paper (Chiang et al. 2001), 94 candidate stations including 20 stations ($A1, A2, A3, ..., A20$) in Alabama, 44 stations ($G1, G2, G3, ..., G44$) in Georgia, and 30 stations ($M1, M2, M3, ..., M30$) in Mississippi were used for analysis. Fifteen streamflow parameters, developed by the time series model (TSM) in Eq. (4), characterize specific monthly streamflow (in cfs/mi^2) properties from 1959 to 1983 (25 years) at each station. These parameters, including 12 monthly streamflow parameters (MSPs, w_j), 2 MA (moving average) parameters (θ_1, θ_2), and variance of residuals (σ^2), were used as variables to classify 94 watersheds into six regions [see Fig. 1 in Chiang et al. (2001)]. In this study, regression analysis is applied to construct relationships between the fifteen streamflow parameters (dependent variables), seven watershed variables (independent variables), and six regional membership ($G1, G2, ..., G6$). Watershed variables include

watershed area, *Aw* (mi^2), forest area, *Af* (percentage of contributing drainage area), area of storage, *As* (percentage of contributing drainage area), elevation, *E* (ft, above mean sea level), stream length per unit area, *L* (mi/mi^2), and main channel slope, *S* (ft/mi). In addition, the mean annual precipitation, *P* (in.) is included as one watershed variable, even though it is considered as the input to the system.

Canonical Correlation Analysis and Regression Modeling

Canonical correlation analysis (CCA), multiple regression analysis (MRA), and simple regression analysis (SRA) including SAS procedures CANCORR and REG are used to investigate and construct the variable relationships between streamflow parameters and watershed variables (*SAS* 1990a, b). CCA is a generalization of multiple regression in which several *Y* variables are simultaneously related to several *X* variables (Thompson 1984; Manley 1995). If there are *p* independent variables, $X_1, X_2, \ldots, X_p$ (e.g., watershed variables), and *q* dependent variables, $Y_1, Y_2, \ldots, Y_q$ (e.g., streamflow parameters), then linear relationships of the following forms are established:

$$U_r = a_{r1}X_1 + a_{r2}X_2 + \cdots + a_{rp}X_p \tag{5}$$

and

$$V_r = b_{r1}Y_1 + b_{r2}Y_2 + \cdots + b_{rq}Y_q \tag{6}$$

Here, U_r and V_r = canonical variates and are calculated from linear combination of independent and dependent variables, respectively; r = minimum of p and q; and $a_{r1}, a_{r2}, \ldots, a_{rp}$, $b_{r1}, b_{r2}, \ldots$, and b_{rq} are coefficients. The correlation between U_r and V_r is maximized subject to these variables uncorrelated with U_{r-1} and V_{r-1}. Each pair of the canonical variates (U_1, V_1), $(U_2, V_2), \ldots$, and (U_r, V_r) then represents an independent "dimension" in the relationship between the two sets of variables $(X_1, X_2, \ldots, X_p)$ and $(Y_1, Y_2, \ldots, Y_q)$. The first pair of canonical variates (U_1, V_1) has the highest possible correlation and is the most important; the second pair of canonical variates (U_2, V_2) shows the next highest correlation, and so on.

The calculation of CCA involves eigenvector and eigenvalue analysis and finding the correlation matrix between variables, $X_1, X_2, \ldots, X_p$ and $Y_1, Y_2, \ldots, Y_q$. The eigenvalues are the squares of the correlations between the canonical variates. Therefore, the eigenvalues provide estimates of the amount of shared variance between the canonical variates, but not the variance extracted from the set of variables. In addition, a redundancy index is used to measure overlapping information between the dependent and independent variables. Redundancy index is based on the amount of variance in one set of variables, as explained by a linear combination of the other set of variables (Hair et al. 1992).

Multiple regression analysis (MRA) in Eq. (1) is applied to construct relationships between the dependent variables (i.e., 15 streamflow parameters) and the independent variables (i.e., seven watershed variables and regional membership). Simple regression analysis (SRA), as in Eq. (2), is also used to construct relationships between monthly streamflows and watershed areas, *Aw*. For simplicity without considering cross-product terms, the MRA in Eq. (1) can be expanded as

$$Y_i = k + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_p X_p + \beta_1 G_1 + \beta_2 G_2 + \cdots + \beta_{n-1} G_{n-1} \tag{7}$$

where Y_i = streamflow parameters of the TSM; $G_1, G_2, \ldots, G_{n-1}$ = regional membership variables for cluster and are equal to 1 corresponding to each classified region, otherwise 0; n = number of regions; and $\beta_1, \beta_2, \ldots, \beta_{n-1}$ = regression coefficients. In addition, the mean square error (MSE) measures the prediction error of the MRA model. Coefficient of determination R^2 is an index to the variation explained by the regression model.

Validation Procedures

In order to validate this regionalization scheme, a certain number of stations from each region are randomly selected. These stations are removed as observations in the MRA equations [Eq. (7)], and the MRA equations are used to predict the streamflow parameters at these validation stations. These predicted streamflow parameters from the MRA equations are compared with those obtained from the TSM [Eq. (4)] to evaluate the performance of the MRA processes. The predicted streamflow parameters are then used to synthesize a number series, *N*, of the specific monthly streamflows, *Qs* (cfs/mi^2) for a period of historical length, n_T, using the TSM [Eq. (4)]. The synthetic specific mean monthly streamflows (*Qsm*) are the average of the monthly *Qs* in n_T years. The mean of the synthetic *Qsm* (i.e., sum of *Qsm* divided by *N*) and the *Qs* obtained from SRA in Eq. (2) (i.e., predicted monthly streamflow divided by watershed area) are compared with the historical average for n_T years to investigate improvement of the streamflow estimations at the validation stations (USBOR 1991; Merzi et al. 1993). The standard deviations of the synthetic mean monthly streamflows, *Qsd*, and autocorrelation function, ACF, of the synthetic specific monthly streamflow, *Qs*, are also compared with those of the historical streamflow records.

The reliability curves based on the historical records are compared with those of a few synthetic streamflow series randomly selected from the *N* synthetic streamflow series at each validation station. The reliability curves evaluate the reliability (*R*) of different outflows for different storage volumes, as calculated by

$$R = \frac{n_T - n_0}{n_T} 100\% \tag{8}$$

where n_T = total length (e.g., 25 years) of S_i series; and n_0 = length of zeros (i.e., no water in a storage volume) in S_i series. Here, S_i is the storage volume required at the beginning of period *i* and is calculated by

$$S_i = \begin{cases} S_{i-1} + I_i - Q_i & \text{if } 0 \leq S_i \leq K \\ K & \text{if } S_i > K \\ 0 & \text{if } S_i < 0 \end{cases} \tag{9}$$

where S_{i-1} = storage volume at the previous period $i-1$; I_i = inflow during period *i*; Q_i = outflow required in period *i*; and K = maximum storage volume. Assuming an initial storage value S_0 ($=K$ in this study); R = plotted versus K and outflow Q, and repeated over the entire procedure for different values of K and Q to develop a family of curves. For this study, the synthetic streamflows are used as inflow (I_i).

Results and Hydrologic Interpretations

Canonical Correlation Analysis

Canonical correlation analysis (CCA) is applied to processing the streamflow parameters and watershed variables to provide suggestions in the multiple regression analysis (MRA). Correlations between watershed elevation (*E*) and the 12 MSPs in the data set are high or moderate (0.49–0.80). In the two main clusters, cor-

Table 1. Correlations Between Streamflow Parameters and Watershed Variables

Watershed variables	Streamflow parameters: West Season (High MSPs) w_1	w_2	w_3	w_4	w_5	w_6	Dry Season (Low MSPs) w_7	w_8	w_9	w_{10}	w_{11}	Wet w_{12}	MA (2) θ_1	θ_2	σ^2
(a) Data set (all observations, $N=94$)															
Aw	−0.097	−0.139	−0.080	−0.014	0.063	0.173	0.168	0.204	0.225	0.244	0.092	−0.044	−0.155	−0.162	−0.348
S	0.384	0.437	0.357	0.346	0.428	0.433	0.338	0.290	0.278	0.303	0.423	0.404	0.199	0.132	−0.212
L	−0.066	−0.103	−0.060	0.033	0.121	0.228	0.227	0.266	0.285	0.306	0.149	−0.001	−0.140	−0.146	−0.418
E	0.514	0.489	0.526	0.639	0.797	0.756	0.600	0.520	0.536	0.682	0.747	0.616	0.232	0.179	−0.669
As	−0.239	−0.272	−0.238	−0.221	−0.075	0.095	0.179	0.199	0.205	0.204	−0.019	−0.188	−0.121	−0.137	−0.125
Af	0.128	0.091	0.071	0.179	0.302	0.301	0.325	0.381	0.320	0.329	0.289	0.175	−0.015	−0.074	−0.282
P	0.629	0.562	0.489	0.594	0.510	0.199	0.179	0.121	0.205	0.184	0.405	0.620	0.175	0.404	−0.068
(b) Cluster-1 ($N=59$)															
Aw	−0.192	−0.169	−0.035	−0.079	0.016	0.116	0.088	0.093	0.128	0.199	0.029	−0.142	−0.126	−0.150	−0.411
S	0.197	0.120	0.087	0.075	0.238	0.229	0.230	0.188	0.142	0.125	0.270	0.260	0.249	0.103	0.068
L	−0.218	−0.204	−0.071	−0.076	0.080	0.195	0.168	0.183	0.212	0.275	0.102	−0.109	−0.121	−0.170	−0.450
E	0.023	−0.158	0.136	0.236	0.660	0.766	0.624	0.536	0.470	0.680	0.650	0.246	0.295	0.155	−0.653
As	−0.155	−0.159	−0.065	−0.124	0.075	0.184	0.163	0.128	0.170	0.264	0.132	−0.008	0.145	0.008	−0.417
Af	0.194	0.066	−0.006	0.124	0.297	0.228	0.321	0.322	0.327	0.229	0.277	0.284	−0.041	0.028	0.067
P	0.518	0.531	0.330	0.536	0.284	−0.058	0.078	0.099	0.171	−0.013	0.121	0.551	−0.112	0.268	0.427
(c) Cluster-3 ($N=22$)															
Aw	0.156	−0.162	0.138	0.285	0.387	0.222	0.375	0.304	0.360	0.392	0.102	0.031	−0.189	−0.263	−0.410
S	0.206	0.346	0.262	−0.033	−0.086	0.184	−0.129	−0.132	−0.054	−0.152	0.139	0.281	0.155	0.460	0.242
L	0.262	0.091	0.209	0.450	0.379	0.176	0.419	0.384	0.418	0.453	0.097	0.072	−0.113	−0.103	−0.497
E	0.841	0.707	0.758	0.726	0.444	0.307	0.326	0.238	0.493	0.481	0.441	0.707	0.111	0.329	−0.263
As	0.022	−0.131	−0.039	0.030	0.087	0.003	0.152	0.183	0.138	0.225	−0.107	−0.155	−0.413	−0.275	−0.254
Af	−0.293	−0.213	−0.198	−0.188	0.050	−0.188	0.010	0.108	−0.171	0.002	−0.006	−0.223	0.036	−0.224	−0.010
P	0.026	0.106	−0.053	−0.002	0.171	0.178	−0.203	−0.074	0.014	−0.170	0.252	0.197	−0.037	0.181	0.063

relations between E and the ''low'' MSPs ($w_6, w_7, w_8, w_9, w_{10}, w_{11}$) are mostly moderate (0.47–0.77) in Cluster-1 and are low (0.24–0.49) in Cluster-3. Correlations between E and the ''high'' MSPs ($w_1, w_2, w_3, w_4, w_5, w_{12}$) are low (<.25 mostly) in Cluster-1 and high (0.71–0.84) in Cluster-3. Precipitation (P) is moderately correlated with the ''high'' MSPs (0.41–0.63) in the data set, and correlations between P and the ''high'' MSPs seem to be moderate (0.28–0.55) in Cluster-1 and very low in Cluster-3. These correlations show that high elevation watersheds exhibit high MSP values, and there is more precipitation in the wet season.

Watershed variables of surface storage areas, *As*, and watershed area, *Aw*, seem to be positively correlated with the ''low'' MSPs and negatively correlated with the ''high'' MSPs (Table 1). Generally, watersheds with large areas provide a larger buffer capacity. Streamflows from large watersheds may be affected by groundwater supply in the dry season. For these reasons, watersheds with large *Aw* and *As* display higher MSP values than watersheds with small *Aw* and *As* in the dry season. In the wet season, a large storage capacity will hold surface runoff water and reduce streamflows. Therefore, lower MSP values in the wet season for watersheds with large *Aw* and *As* would be reasonable. In addition, σ^2 (residual variance) is moderately correlated with E (−0.67), L (stream length per unit area, −0.42), and *Aw* (−0.35) in the data set, with E (−0.65), L (−0.45), *As* (−0.42), *Aw* (−0.41), and P (0.43) in Cluster-1, and with L (−0.50) and *Aw* (−0.41) in Cluster-3 (Table 1). The inverse correlations show that monthly streamflow variations are small for watersheds with high E and large L. This may be due to more stable monthly streamflows in the high elevation watersheds than those in the low elevation watersheds.

Based on the above correlation analysis, the moderate and high correlations in the data set and two main clusters (Table 1) suggest that the streamflow parameters can be predicted by the watershed variables, although correlations between the streamflow parameters and watershed variables could be different in the two clusters. When redundancy indices within each of the two main clusters are high, the MRA would be employed to construct relationships within each cluster. Otherwise, dummy variables should be used to indicate the regional membership in MRA.

In CCA, the streamflow parameters are divided into three subsets: (1) ''high'' MSPs; (2) ''low'' MSPs; and (3) MA(2) parameters and σ^2. Relationships between streamflow parameters and watershed variables are similar in each subset. Similar relationships are also obtained from the MRA model. In redundancy analysis, the amount of variance in the dependent variables, indicated by the variance in the independent variables, is calculated. The redundancy indices, the standardized variances of the streamflow parameters explained by the first opposite canonical variate (V_1), are 61, 47, and 15% in subsets I, II, and III, respectively. The respective redundancy indices in subsets I, II, and III of Cluster-1 are 39, 33, and 21%, and those of Cluster-3 are 35, 7 and 20%. The remaining canonical variates are not worthy of consideration because of low index values. The indices of subsets I (61%) and II (47%) in the data set are higher than those in the two main clusters (7–39%). The low redundancy indices in the two main clusters imply that the data set from the 94 stations is more appropriate than the two main clusters in assessing the relationships between the 12 MSPs and the watershed variables. Low indices in subset III (15, 20, and 21%) and low correlations between MA(2) parameters and watershed variables reveal that watershed variables may not be valid independent variables for

Table 2. Canonical Correlation Analysis

(a) Standardized Canonical Coefficients for Streamflow and Watershed Variables

Subset		w_1	w_2	w_3	w_4	w_5	w_{12}		*Aw*	*S*	*L*	*E*	*As*	*Af*	*P*
I	U_1	0.794	0.723	0.723	0.890	0.987	0.881	V_1	0.058	0.453	0.122	0.842	−0.102	0.333	0.597
Subset		w_6	w_7	w_8	w_9	w_{10}	w_{11}		*Aw*	*S*	*L*	*E*	*As*	*Af*	*P*
II	U_1	0.967	−0.481	0.509	−0.731	−0.060	0.736	V_1	0.201	0.194	−0.041	0.806	−0.003	0.042	0.268
Subset		θ_1	θ_2	σ^2					*Aw*	*S*	*L*	*E*	*As*	*Af*	*P*
III	U_1	0.086	0.255	−1.020				V_1	−0.375	0.314	0.857	0.606	0.081	0.030	0.249

(b) Correlations Between Canonical Variates and Streamflow and Watershed Variables

Subset		w_1	w_2	w_3	w_4	w_5	w_{12}		*Aw*	*S*	*L*	*E*	*As*	*Af*	*P*
I	U_1	0.794	0.723	0.723	0.890	0.987	0.881	V_1	0.058	0.453	0.122	0.842	−0.102	0.333	0.597
Subset		w_6	w_7	w_8	w_9	w_{10}	w_{11}		*Aw*	*S*	*L*	*E*	*As*	*Af*	*P*
II	U_1	0.906	0.734	0.639	0.679	0.818	0.945	V_1	0.091	0.569	0.151	0.952	−0.079	0.332	0.351
Subset		θ_1	θ_2	σ^2					*Aw*	*S*	*L*	*E*	*As*	*Af*	*P*
III	U_1	0.220	0.218	−0.929				V_1	0.335	0.361	0.432	0.916	0.122	0.362	0.257

predicting MA(2) parameters. Therefore, dummy variables indicating regional membership are used in further determination of the relationships in MRA because variable correlations are different in the two main clusters.

The first pair of canonical variates (U_1, V_1) explain 86% (eigenvalue=6.2), 60% (eigenvalue=3.0), and 78% (eigenvalue=1.7) of the variance shared by the dependent and independent variables in the respective subsets I, II, and III. The remaining canonical variates are not investigated because of low eigenvalues (<1.0). In subset I, (U_1, V_1) is associated with w_5 (coefficient=0.96) and w_1 (0.45) and watershed variables *E* (0.84) and *P* (0.54). Variate U_1 is highly correlated with the ''high'' MSPs (correlation >.72), and V_1 is correlated with *E* (0.84) and *P* (0.60). In subset II, (U_1, V_1) shows different weights for the ''low'' MSPs (−0.48–0.96) and *E* (0.81). Variate U_1 seems to be highly correlated with the ''low'' MSPs (0.63–0.94), and V_1 is correlated with *E* (0.81). In subset III, (U_1, V_1) is associated with σ^2 (−1.02) and watershed variables *L* (0.85) and *E* (0.60). Variate U_1 is negatively correlated with σ^2 (−0.93), and V_1 is correlated with *E* (0.92) and *L* (0.42). After summarizing these CCA results, it is clear there exist relationships between the 12 MSPs and watershed variables *E* and *P*, as well as relationships between σ^2 and watershed variables *E* and *L*. Also, the canonical correlations of (U_1, V_1) are highly correlated (0.76–0.97) with the data subsets. This is consistent with the correlations between streamflow parameters and watershed variables as shown in Table 2.

Regression Analysis

To assure the fundamental assumption of multivariate normal distribution, the quantities of watershed variables are logarithmically transformed in the MRA processes. The results of the MRA equations in Eq. (7) contain all significant independent variables except *Af* (Table 3). The regression coefficients of elevation, *E*, and the mean annual precipitation, *P* (α_1 and α_2), are all positive and

Table 3. Regression Equations from MRA Using All Data

Streamflow parameters, Y_i		Intercept, k'	Watershed variables X_p						Regional Membership G_{n-1}					MSE	R^2
			ln(*E*) α_1	ln(*P*) α_2	ln(*S*) α_4	ln(*As*) α_5	ln(*Aw*) α_6	ln(*Af*) α_7	G_1 β_1	G_2 β_2	G_3 β_3	G_4 β_4	G_5 β_5		
Wet season (high MSPs)	w_1	−6.752	0.156	1.504	-	-	-	-	0.426	0.519	0.572	-	-	0.173	0.75
	w_2	−5.389	0.121	1.347	0.059	-	-	-	-	-	0.091	-	−0.434	0.166	0.61
	w_3	−5.317	0.212	1.231	-	-	-	-	-	-	0.234	-	−0.448	0.171	0.66
	w_4	−7.219	0.252	1.614	-	-	-	-	-	-	0.120	−0.262	−0.648	0.151	0.82
	w_5	-	0.462	1.853	-	-	-	-	-	-	-	−0.336	−0.518	0.166	0.87
Dry Seasons (low MSPs)	w_6	−8.299	0.578	1.118	-	-	-	-	-	-	−0.664	−0.316	-	0.279	0.79
	w_7	-	0.504	1.981	-	0.035	-	-	-	-	−0.730	-	0.804	0.302	0.75
	w_8	-	0.448	1.906	-	0.036	-	-	-	-	−1.009	-	0.647	0.336	0.76
	w_9	-	0.514	2.754	-	0.052	-	-	-	-	−0.945	-	0.828	0.376	0.73
	w_{10}	−9.975	0.579	1.442	-	0.052	-	-	-	-	−0.897	−0.623	-	0.333	0.80
	w_{11}	-	0.633	2.002	-	-	-	-	-	-	−0.448	−0.935	-	0.314	0.81
Wet	w_{12}	-	0.333	2.372	-	-	-	-	-	-	-	−0.874	−0.765	0.226	0.84
MA (2)	θ_1	−0.685	0.045	-	-	-	−0.011	-	0.124	-	0.160	-	0.297	0.065	0.46
	θ_2	−0.925	-	0.201	-	-	-	-	-	-	0.051	−0.130	-	0.072	0.36
$\ln(\sigma^2)$		3.093	−0.532	-	-	−0.035	-	-	-	-	−0.026	−0.002	-	0.186	0.80

Note: MRA equation: $Y_i = k + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_p X_p + \beta_1 G_1 + \beta_2 G_2 + \ldots + \beta_{n-1} G_{n-1}$ in Eq. (7).

Table 4. Regression Equations from SRA

Y monthly streamflow	Intercept k	X ln(Aw) a_I	MSE	R^2
$\ln(Q_1)$	1.119	0.955	0.300	0.953
$\ln(Q_2)$	1.250	0.956	0.209	0.977
$\ln(Q_3)$	1.404	0.958	0.238	0.970
$\ln(Q_4)$	1.200	0.972	0.275	0.962
$\ln(Q_5)$	0.488	0.988	0.365	0.937
$\ln(Q_6)$	−0.258	1.016	0.430	0.918
$\ln(Q_7)$	−0.428	1.012	0.428	0.918
$\ln(Q_8)$	−0.747	1.040	0.535	0.884
$\ln(Q_9)$	−0.885	1.042	0.509	0.894
$\ln(Q_{10})$	−0.904	1.049	0.505	0.897
$\ln(Q_{11})$	−0.242	0.989	0.547	0.868
$\ln(Q_{12})$	0.903	0.931	0.422	0.907

Note: SRA equation: $Y=k+\alpha X$ in Eq. (2).

significant in the MRA equations of the MSPs. Precipitation P is the dominant variable in each of the monthly streamflow equations (Table 3), and the regression coefficients of E in the dry season are larger than those in the wet season. Therefore, the specific monthly streamflows, Qs, increase with increasing P and E. Furthermore, elevation E of the watershed affects streamflows in the dry season more than those in the wet season. Surface storage area As seems to be slightly correlated with the "low" MSP, which is probably due to a large As providing buffer capacity that could slightly increase streamflows during the dry season. In the two main clusters, the regression coefficients of β_3 in Cluster-3 ($N=22$) are positive in the wet season and negative in the dry season, and those of β_1 in Cluster-1 ($N=59$) are mostly insignificant. This shows that Qs tend to be higher in the wet season and lower in the dry season of Cluster-3 than those in Cluster-1. This is consistent with the streamflow patterns as shown in the first paper (Chiang et al. 2001).

The values of the coefficient of multiple determination (R^2) range from 0.61 to 0.89 with the exception of the regression model for θ_1 ($R^2=0.46$) and θ_2 ($R^2=0.36$). The high R^2 values indicate that the MSPs and σ^2 can be reasonably predicted by watershed variables. Note that the dependent variable was already logarithmically transformed. The inverse correlation between σ^2 and E and regional membership G_1 (i.e., Cluster-1) shows that a higher elevation watershed tends to have smaller residual variance. This is consistent with the regional patterns based on streamflow parameters and watershed variables in the first paper (Chiang et al. 2001) and the results of canonical correlation analysis. The values of R^2 and mean standard error (MSE) are low in the regression equations of θ_1 and θ_2. This indicates that the MA parameters are not highly correlated with watershed variables and regional membership; therefore, using the mean values of θ_1 and θ_2 in each cluster to generate streamflows may be acceptable. Since some watershed variables are correlated with each other, the variance inflation factor (VIF) is used to test the multicollinearity. The resulting small VIF values (far less than 10) indicate low intercorrelation among the significant independent variables.

In SRA, as shown in Eq. (2), strong linear relationships can be found between mean monthly streamflows and watershed area, Aw, (coefficient of determination, $R^2>.86$) after data were logarithmically transformed (Table 4). Note that the results of SRA have greater R^2 values but smaller MSE values than those of

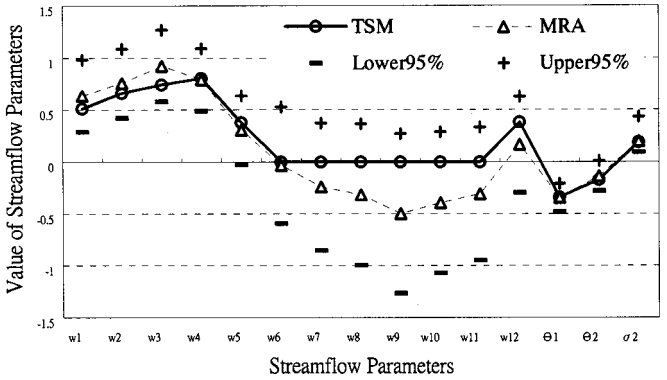

Fig. 1. Comparison of streamflow parameters by TSM and MRA at station A08 of Cluster-1

MRA, and the values of MSE are smaller in the wet season than those in the dry season. The regression coefficients a are slightly less than one in the wet season and slightly greater than one in the dry season, and the values of regression intercept k are positive in the wet season and negative in the dry season. Therefore, large watersheds exhibit lower monthly streamflows in the wet season and higher monthly streamflows than small watersheds in the dry season.

Validations

Three stations in each of the two main clusters, A08, G35, and M05 in Cluster-1 and M11, M18, and M26 in Cluster-3, were randomly selected from the original 94 stations to serve as validation sites. After excluding these six validation stations, the stepwise regression selection procedure were repeated by manipulating the remaining 88 station data. For brevity, results of one or two from the six validation stations are selected for illustration. Streamflow parameters obtained by the MRA equations (Table 3) are compared with those determined by the TSM [Eq. (4)]. These two sets of streamflow parameters, as well as the 95% confidence interval at one of the six validation stations (A08) in Cluster-1, are plotted in Fig. 1. It is observed that most streamflow parameters obtained from MRA and TSM are within the 95% confidence interval at the validation stations. Some values of the streamflow parameters in TSM are equal to zero (e.g., $w_6, w_7, w_8, \ldots, w_{11}$ in Fig. 1) because of its insignificance.

The estimated streamflow parameters from MRA are used to synthesize 50 series of specific monthly streamflows (Qs) over a 25 year period using the developed TSM [Eq. (4)]. Results of the mean of Qsm (specific mean monthly streamflows) by TSM and Qs (specific monthly streamflows) by SRA (i.e., predicted monthly streamflows divided by watershed area) are compared with historical data in Fig. 2. It can be seen that the specific monthly streamflows better agree with the historical data than those of SRA, except in the wet season of Cluster-3. Therefore, the proposed method of integrating MRA and TSM is not only an improvement over the traditional SRA model, but also can be applied to synthesizing streamflow hydrographs. The standard deviations of the synthetic mean specific monthly streamflows, Qsd, seem to be preserved except for those in the wet season of Cluster-3, as shown in Fig. 3. Note that the values of σ^2 in Cluster-3 are higher than those values in Cluster-1 (Chiang et al. 2001). The high values of Qsd and σ^2 in Cluster-3 indicate that large synthetic monthly streamflows are anticipated from the proposed method and the magnitude of monthly streamflow varia-

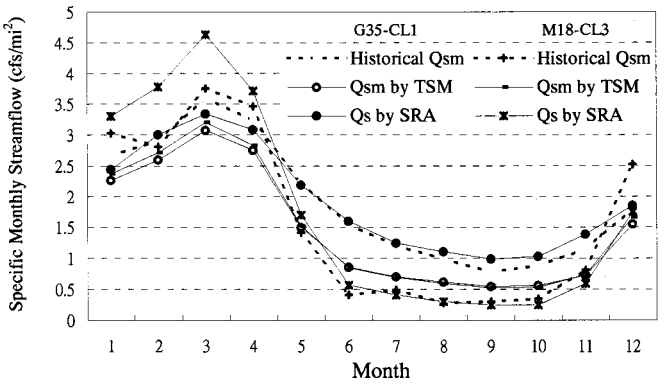

Fig. 2. Comparisons of historical *Qsm*, present *Qsm* by TSM, and *Qs* predicted by SRA at Station G35 in Cluster-1 and M18 in Cluster-3

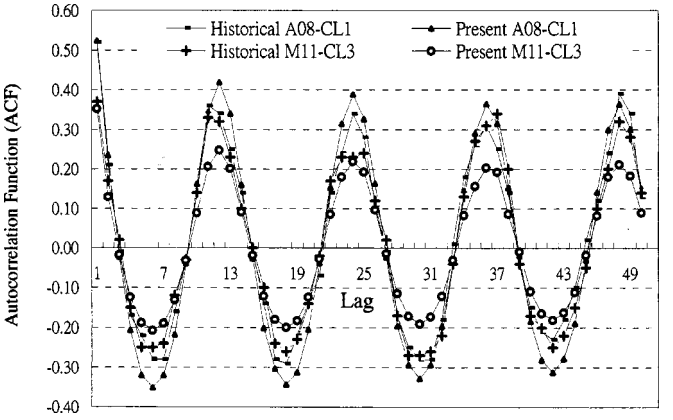

Fig. 4. Historical and present synthetic mean ACF of *Qs* at Station A08 in Cluster-1 and M11 in Cluster-3

tions are significant in this season. This may explain why the proposed TSM cannot satisfactorily predict the streamflows in Cluster-3.

The autocorrelation function (ACF) tends to be well preserved except for a slight departure around lag 6 with a cyclic period of 6. This departure seems to be more significant in Cluster-3 than that in Cluster-1, as shown in Fig. 4. Consequently, the proposed method does better generate the streamflows in Cluster-1 but is slightly biased in Cluster-3.

The reliability curves of different outflows for different storage volumes are calculated by Eqs. (8) and (9) based on the 25 years of historical *Qs* and 50 synthetic *Qs*. Three synthetic *Qs* were randomly selected from the 50 synthetic *Qs* for each validation station. Depending on different maximum storage volumes (K in cfs/mi^2), the reliability curves evaluate the reliability of the required outflow ($Q = 2.2, 2.6, 3.0$ cfs/mi^2) for downstream water use, as shown in Figs. 5 (station G35 in Cluster-1) and 6 (station M11 in Cluster-3). Note that each outflow (e.g., $Q = 2.2$ cfs/mi^2) includes one historical and three synthetic (same symbol) reliability curves; thus, Figs. 5 and 6 shows twelve reliability curves with three different Q values. The required outflows generally depend on downstream requirements for water use. These are obtained through a series of trial and error experiments to approximately satisfy 100% reliability when the maximum storage volume increases. The reliability curve provides an easy way to calculate the storage volume (e.g., a reservoir) and to compare with monthly streamflows. For example, to obtain 100% reliability with $Q = 2.2$ cfs/mi^2 of outflow at station G35, the storage volume K is about 100 cfs/mi^2, or 45 times the outflow (Fig. 5). If the storage volume (in cfs/mi) is multiplied by watershed area (in mi^2) and seconds per year, then the required annual storage volume (m^3) of this watershed is obtained. The reliability curves in each outflow group (e.g., $Q = 2.2$ cfs/mi^2) seem to stay close or even overlap (especially station G35 as shown in Fig. 5) and separate into three different groups in Cluster-1. However, in cluster-3, the reliability curve of the synthetic series tends to be higher than that of the historical records. Thus, the synthetic specific monthly streamflows are reliable in Cluster-1 but a bit biased in Cluster-3.

Conclusions and Discussions

By integrating multiple regression analysis (MRA) and a time series model (TSM), an integrated method of generating synthetic specific monthly streamflows is developed and validated. The streamflow parameters in TSM are predicted by employing MRA equations that construct relationships between streamflow parameters and watershed variables. Canonical correlation analysis (CCA) is employed to provide suggestions for the MRA. These predicted streamflow parameters are applied to synthesize a large number of streamflow sequences at ungauged sites for a desired period by using TSM. Comparisons of results among the proposed method, traditional SRA, and the historical data show that streamflow estimations from the proposed method agree better

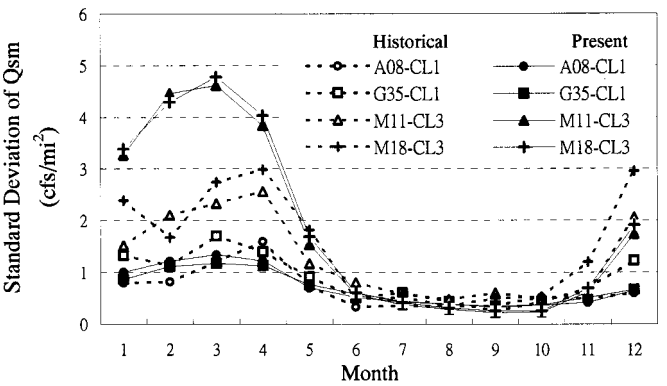

Fig. 3. Comparisons of historical *Qsd* and present *Qsd* by TSM at Station A08 and G35 in Cluster-1 and M11 and M18 in Cluster-3

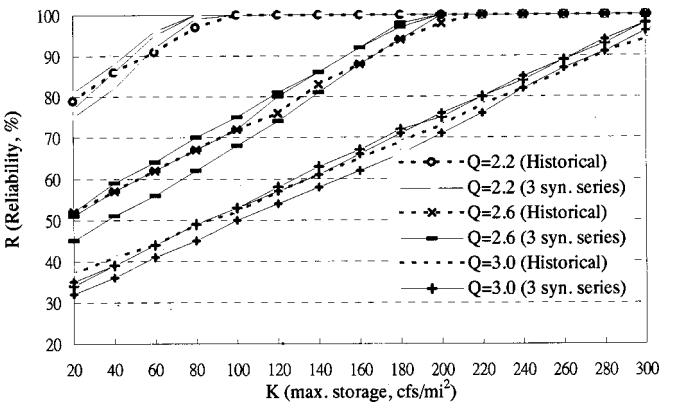

Fig. 5. Reliability curves for validation at Station G35 in Cluster-1

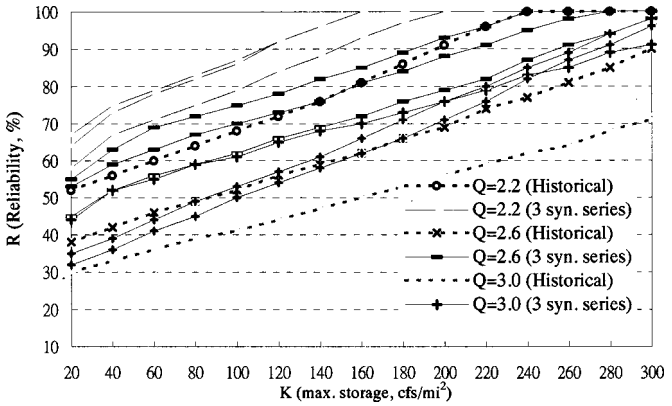

Fig. 6. Reliability curves for validation at Station M11 in Cluster-3

with historical data than the traditional SRA method. This improvement would be crucial to engineering designs in reducing engineering cost, increasing social and economic benefits, and reducing damages.

Streamflow statistical properties (standard deviation, autocorrelation function) and reliability of different outflows for different storage volumes are validated with a proposed regionalization scheme (Chiang et al. 2001) by using historical records and the synthetic streamflows from the TSM. This validation procedure is critically important for a water use project to estimate whether water stored in a storage volume meets the downstream water demands during a design period. To obtain better confidence, it is suggested that reliability curves be calculated for all synthetic series, percentiles of the synthetic series plotted, and comparisons of the percentiles with the historic series made. The TSM in the the proposed method retains the statistical information and the stochastic nature of the monthly streamflows properties in Cluster-1, but is a bit biased in Cluster-3. The biases of the streamflows synthesized by the TSM in Cluster-3 may be caused by the high residual variance (σ^2) or the model itself. The high σ^2 results from abnormal flood events and outliers in the time series analysis. Detecting and removing the outliers and handling the abnormal flood events from the streamflow records are suggested to improve accuracy of the estimations of the streamflow parameters in MRA and the synthesis of the streamflow series. Furthermore, a validation procedure has been demonstrated for practical applications when synthetic streamflows are needed in many water resources problems. It should be noted that accuracy of the synthesized streamflows at ungauged sites by the proposed method ultimately rely on results of regionalization of watersheds (Chiang et al. 2001).

Although this regionalization scheme can be applied to ungauged sites or limited record sites, further research should explore the possibility of improved multiple regression models (e.g., adding cross products of watershed variables and defining regional membership) and TSM. Questions arise as to whether these predicted streamflow parameters can be adjusted to improve streamflow estimation based on a short record.

Acknowledgments

The writers would like to thank the reviewers of this paper and Professor Stephan Rocky Durrans in the Department of Civil and Environmental Engineering, University of Alabama, for their comments and suggestions.

Notation

The following symbols are used in this paper:

Af = forest area;
As = area of storage;
Aw = watershed area;
$a_{r1}, a_{r2}, \ldots, a_{rp}$ = coefficients;
a_I = regression coefficient;
a_t = random error;
B = backshift operator;
$b_{r1}, b_{r2}, \ldots, b_{rq}$ = coefficients;
D_{jt} = indicator variable (equals 1 or 0);
E = elevation;
$G_1, G_2, \ldots, G_{n-1}$ = regional membership variables;
I_i = inflow during period i;
K = maximum storage volume;
k, k', k_i' = regression coefficient or intercept;
L = stream length per unit area;
N = number of synthesized series;
N = number of synthesized series;
n = number of regions;
n_T = total length (e.g., 25 years) of S_i series;
n_0 = length of zeros in S_i series;
P = precipitation;
Q_i = outflow required in period i;
Q_s = specific monthly streamflow;
Q_{sm} = synthetic specific mean monthly streamflow;
Q_t = current observation of specific monthly streamflow;
R = reliability;
R^2 = index to variation explained by regression model;
S = main channel slope;
S_i = storage volume required at beginning of period i;
S_{i-1} = storage volume at previous period $i-1$;
$U_1, U_2, \ldots, U_r$ = canonical variates;
$V_1, V_2, \ldots, V_r$ = canonical variates;
w_j = coefficient of jth period;
$w_1, w_2, \ldots, w_{12}$ = monthly streamflow parameters;
$X_1, X_2, \ldots, X_p$ = independent variables (i.e., watershed variables);
$Y, Y', Y_i, Y_1, Y_2, \ldots, Y_q$ = dependent variables (i.e., streamflow parameters);
$\alpha_1, \alpha_2, \ldots, \alpha_p$ = regression coefficients;
$\beta_1, \beta_2, \ldots, \beta_p$ = regression coefficients;
ε = error;
θ_1, θ_2 = moving average parameters; and
σ^2 = variance of errors.

References

Benson, M. A., and Matalas, N. C. (1967). "Synthetic hydrology based on regional statistical parameters." *Water Resour. Res.*, 3(4).

Bhaskar, N. R., and O'Connor, C. A. (1989). "Comparison of method of residuals and cluster analysis for flood regionalization." *J. Water Resour. Plan. Manage. Div., Am. Soc. Civ. Eng.*, 115(6), 793–808.

Black, P. E. (1988). "Strange attractors in the closet?" *Proc., Annual Fall Meeting of the American Geophysical Union*, American Geophysical Union, Washington, D.C.

Chiang, S. M., Tsay, T. K., and Nix, S. J. (2001). ''Hydrologic regionalization of watersheds. I: Methodology development.'' *J. Water Resour. Plan. Manage. Div., Am. Soc. Civ. Eng.*, in press.

DeCoursey, D. G. (1973). ''Object regionalization of peak flow rates, floods, and droughts.'' *Proc., 2nd Int. Symposium in Hydrology*, E. F. Schulz, V. A. Koelzer, and K. Mohmood, eds., Water Resources, Fort Collins, Colo.

Fennessey, N., and Vogel, R. M. (1990). ''Regional flow-duration curves for ungauged sites in Massachusetts.'' *J. Water Resour. Plan. Manage. Div., Am. Soc. Civ. Eng.*, 116(4), 530–549.

Gottschalk, L. (1985). ''Hydrologic regionalization of Sweden.'' *Hydrol. Sci. J.*, 30(1).

Gulliver, J. S. (1991). ''Preliminary studies: hydrology, hydraulics, and costs.'' *Hydropower engineering handbook*, J. S. Gulliver and R. E. A. Arndt, eds., McGraw-Hill, New York.

Hair, J. F., Anderson, R. E., Tatham, R. L., and Black, W. C. (1992). *Multivariate data analysis with readings*, 3rd ed., Macmillan, New York.

Hawley, M. E., and McCuen, R. H. (1982). ''Water yield estimation in western United States.'' *J. Irrig. Drainage*, 108(1), 25–34.

Manley, B. F. J. (1995). *Multivariate statistical methods, a primer*, 2nd ed., Chapman & Hall, New York.

McMahon, T. A., (1993). ''Hydrologic design for water use.'' *Handbook of hydrology*, D. R. Maidment, ed., McGraw-Hill, New York.

Merzi, N., Usul, N., Ozsaracoglu, Z., and Ozaydin, V. (1993). ''Stochastic modeling of mean monthly runoffs for Coruh Basin, Turkey.'' *Engineering Hydrology: Proc., of the Symposium*, C. Y. Kuo, ed., ASCE, New York.

Mosley, M. P., and McKerchar, A. (1993). ''Streamflow.'' *Handbook of hydrology*, D. R. Maidment, ed., McGraw-Hill, New York.

Murdock, R. U., and Gulliver, J. S. (1993). ''Prediction of river discharge at ungaged sites with analysis of uncertainty.'' *J. Water Resour. Plan. Manage. Div., Am. Soc. Civ. Eng.*, 119(4), 473–487.

Quimpo, R. G., Alejandrino, A. A., and McNally, T. A. (1983). ''Regionalized flow duration for Philippines.'' *J. Water Resour. Plan. Manage. Div., Am. Soc. Civ. Eng.* 109(4), 320–330.

SAS/STAT User's Guide: Volume 1, ACECLUS-FREQ, Version 6, Fourth Edition. (1990a). SAS Institute Inc., Cary, N.C.

SAS/STAT User's Guide: Volume 2, GLM VARCOMP, Version 6, Fourth Edition. (1990b). SAS Institute Inc., Cary, N.C.

Singh, K. P. (1971). ''Model flow duration and streamflow variability.'' *Water Resour. Res.*, 7(4).

Tasker, G. D. (1982). ''Comparing methods of hydrologic regionalization.'' *Water Resour. Bull.*, 18(6).

Thompson, B. (1984). *Canonical correlation analysis*, Sage, Thousand Oaks, Calif.

United States Department of the Interior (USBOR). (1991). ''Lake Nasser simulation model—synthetic inflow trace generators.'' *Irrigation Management System Project Rep., Planning Studies and Model Component (Nile River Projects)*, Bureau of Reclamation, Washington, D.C.