Hydrologic Regionalization of Watersheds. I: Methodology Development

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

Abstract: A hydrologic regionalization scheme is proposed for classification of watersheds at gauged sites. This scheme used 16 streamflow parameters estimated by a time series model to classify 94 watersheds into 6 regions by cluster analysis. The classified regions seem to be separated by physiographical boundaries, especially the two main clusters. Discriminant analysis tests the significance of the cluster difference; thus, each cluster represents one hydrologic region. Principal component analysis interprets the regional differences and similarities. The regional membership is mainly identified by the watershed variables of elevation, forest area, channel slope, and precipitation based on the calculation of the scores of canonical discriminant variates. This emphasizes the importance of the hydrologic regionalization and the identification of the specific characteristics in each region.

DOI: 10.1061/(ASCE)0733-9496(2002)128:1(3)

CE Database keywords: Watersheds; Methodology; Classification; Regional analysis.

Introduction

Hydrological behaviors of watersheds play an important role in water resource planning and management. It is costly to obtain hydrological information by setting gauge stations for every watershed. Hydrological regionalization, the classification of gauged watersheds into regions according to preset criteria, provides a way to extend information from gauged watersheds to ungauged ones. The preset criteria are generally based on streamflow or watershed and climatic variables. Regionalization techniques provide a mechanism to determine the hydrologic behaviors of gauged watersheds. Streamflow and watershed variables describe streamflow properties such as monthly flows or streamflow parameters and watershed characteristics. A mathematical model (e.g., a time series model) estimates the streamflow parameters. Watershed variables describe the watershed characteristics. If a regionalization scheme is successful, strong relationships between streamflow properties and watershed variables can be realized. These relationships can be utilized to develop useful streamflow information at ungauged watersheds featuring characteristics similar to one of the groups. The purpose of this paper is to develop a regionalization scheme to classify watershed into regions and identify the regional membership. Applications to develop strong relationships and to estimate streamflow information at ungauged sites will be presented in the second paper.

In order to estimate streamflows at ungauged sites, a regression equation such as

$$Q = k W_1^{a_1} W_2^{a_2} \cdots W_n^{a_n} \tag{1}$$

has been developed from information of gauged watersheds (e.g., Fennessey and Vogel 1990; Mosley and Mckerchar 1993). Here, Q = streamflow variable of interest; $W_1, W_2, \ldots, W_n$ = watershed and climatic characteristics; and $k, a_1, a_2, \ldots, a_n$ = empirical coefficients. If hydrologic regions of watersheds can be defined first, the streamflow at an ungauged site within this group could be estimated by employing the regression equation, Eq. (1).

Streamflow parameters of flow duration curves (FDC) and flood frequency curves (FFC) have been used as criteria for hydrologic regionalization (Singh 1971; Quimpo et al. 1983; Cheng 1988; Fennessey and Vogel 1990). The use of FDC and FFC should be limited to problems in which the sequential nature of streamflow is unimportant. However, the sequential nature must be accounted for in many water use and control problems. Therefore, the FDC and FFC methods may not be appropriate for hydrologic regionalization (Fennessey and Vogel 1990). In addition, watersheds with different seasonal patterns may have the same FDC or FFC, and the tails of FDC or FFC are actually fixed by the limited extreme events. Difficulty exists because only limited extreme samples are available to parameterize the FDC or FFC. Some other flow properties have also been used for hydrologic regionalization, such as monthly mean flows or peak flows along with their coefficients of variation (Gottschalk 1985; Tasker 1982; Mosley 1981; Bhaskar and O'Connor 1989; Burn 1989). These flow properties capture some statistical information, but the ''sequential'' or ''stochastic'' nature of streamflows may be lost. Therefore, no generally accepted method of hydrologic regionalization has been developed.

Based on the previous studies, five questions will be answered when developing a regionalization technique: (1) How are watersheds classified into regions; (2) What are the criteria used for regionalization; (3) How are the relationships between streamflow and watershed variables constructed; (4) How are the regionaliza-

[1]Senior Environmental Specialist, Office of Deputy Administrator, 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 November 30, 1999; approved on May 1, 2000. This paper is part of the ***Journal of Water Resources Planning and Management***, Vol. 128, No. 1, January 1, 2002. ©ASCE, ISSN 0733-9496/2002/1-3–11/$8.00+$.50 per page.

tion results applied to ungauged sites; and (5) How is the hydrologic regionalization technique validated. This paper attempts to answer the first two questions. Many researchers have defined regions ''objectively'' using cluster analysis (Mosley 1981; Hawley and McCuen 1982; Tasker 1982; Gottschalk 1985; Cheng 1988; Burn 1989; Bhaskar and O'Connor 1989). They used either flow properties or watershed and climatic characteristics as criteria for classification, and then extract streamflow information. However, none of the previous methods can be used to generate synthetic streamflows. The sequential order and the stochastic nature are lost after the streamflow data are manipulated into FDCs, peak flows, or means flows.

Generally, monthly streamflow data have no trend and have a seasonal component of 12 months with additional biannual, quarterly, or irregular components (Bras and Rodriguez-Iturbe 1993). Many researchers (Roesner and Yevjevich 1966; Rodriguez-Iturbe 1968; Rodriguez-Iturbe and Nordin 1968) have used the component concept to handle the seasonal monthly streamflows. Hipel et al. (1979) removed the seasonal monthly streamflows and then developed an ARMA (autoregressive moving average) model for the irregular component. Salas (1993) introduced the way seasonal series are partitioned into components and the seasonality is removed. It is possible to use seasonal ARMA model instead of removing the seasonal monthly streamflows, or use dummy variables or harmonic analysis to model and simplify the seasonality. However, difficulties may arise in hydrologic interpretation of the parameters from the seasonal ARMA model or harmonic analysis. A lesser number of parameters, limited observations, a general model fit for most watersheds, and comprehensible hydrologic interpretation should all be considered. That is why the seasonal monthly streamflows with an ARMA model for the irregular component was selected in this study.

Different criteria were also adopted for hydrologic regionalization. Hawley and McCuen (1982), Tasker (1982), and Gottschalk (1985) defined regions by 18 watershed variables, 4 watershed variables, and 12 flow variables, respectively. Although researchers produced specific regions, different criteria have not been applied to the same watersheds for comparisons. In this paper, appropriate criteria for classification will be investigated. One critical procedure for hydrologic regionalization is to identify the regional membership, which is based on the spatial contiguous hydrologic regions or watershed variables. Watershed variables can be applied to calculate the scores of the canonical discriminant variables and identify the regional membership, if the regions are significantly different. Discriminant analysis (DA) also tests the regional differences, checks the properness and stability of the classification results, and classifies new observations to an appropriate group (Mosley 1981; DeCoursey 1973; Cheng 1988; Bhaskar and O'Connor 1989). The regional differences and relationships between variables within each group can be examined by principal component analysis (PCA).

In the proposed scheme, a time series model is first developed to determine streamflow parameters, which are used as a set of criteria to classify watersheds into regions by CA. DA and PCA are employed to test and interpret the regional differences and similarity. DA is finally used to identify the regional membership. This scheme is developed to estimate monthly streamflows at ungauged sites. Applications of this scheme in constructing the variable relationships between regions and generating a reliable estimate of monthly streamflows will be presented in a forthcoming paper.

Proposed Methodology and Study Areas

Review of Statistical Techniques

Techniques involved in the proposed methodology are reviewed briefly for completeness. Time series analysis can be applied to build mathematical models and to generate synthetic hydrologic records (Salas 1993). Conventionally, time series have been thought to consist of three components, namely, trend (T_t), seasonal (S_t), and irregular components (N_t) (Wei 1990). If these components are assumed to be independent and additive, the time series Y_t (e.g., monthly streamflows) can be expressed as

$$Y_t = T_t + S_t + N_t \tag{2}$$

A popular approach is to use a polynomial function of time to model the trend component, dummy variables to model the seasonal component, and the ARMA (autoregression moving average) structure to model the irregular component. For example, the trend component can be written as a mth-order polynomial in time as

$$T_t = k_0 + \sum_{i=1}^{m} k_i t^i \tag{3}$$

where $k_i = i$th parameter associated of time, t, of degree i. Generally, monthly streamflows are trend-free with a seasonal change of 12 months and variations (the irregular component) in each month. Thus, the hydrologic processes are assumed to be trend-free in this paper. These processes had been used to handle the streamflows by Tao and Delleur (1976). The seasonal component S_t can be described as a linear combination of seasonal indicator (dummy) variables or harmonic functions of various frequencies. For example, the seasonal component S_t can be written as

$$S_t = \sum_{j=1}^{s} w_j D_{jt} \tag{4}$$

where s = number of seasonal periods; w_j = coefficient of the jth period (e.g., monthly streamflow of January); t = time index; and D_{jt} = indicator variable and equals 1 if t corresponds to the seasonal period j—otherwise it equals 0. When s is chosen to be 12 months, the time series parameters are assumed to change from month to month.

In addition, the irregular component can be expressed as an ARMA model:

$$N_t = \frac{\theta(B)}{\phi(B)} a_t \tag{5}$$

where $\phi(B)$ = autoregressive (AR) operator; $\theta(B)$ = moving average (MA) operator; and a_t = random error. An ARMA model has the form

$$Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2} - \cdots - \phi_p Y_{t-p} = C + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \tag{6}$$

where Y_t = current observation; $Y_{t-1},\ldots,Y_{t-p}$ = past observations; C = constant; and $\phi_1,\phi_2,\ldots,\phi_p$, and $\theta_1,\theta_2,\ldots$, and θ_q = parameters. The sequence of random errors $a_t,\ldots,a_{t-q}$ are independently and identically distributed with a normal distribution, $N(0,\sigma_a^2)$, and σ_a^2 is the variance of the errors. Introducing the backshift operator, B, that is,

$$BY_t = Y_{t-1};$$

$$B^2 Y_t = B(BY_t) = Y_{t-2}$$

and so on, Eq. (6) then can be written as

$$\phi(B)Y_t = C + \theta(B)a_t \quad \text{or} \quad Y_t = C' + \frac{\theta(B)}{\phi(B)}a_t \tag{7}$$

where $\phi(B) = (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)$; $\theta(B) = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)$; and $C' =$ constant. This is known as the Box-Jenkins ARMA(p,q) model (Box and Jenkins 1976). Here, p denotes the order of the AR operator $\phi(B)$ and q denotes the order of MA operator $\theta(B)$. Low-order ARMA models are useful for operational hydrology in general, especially for modeling annual series (Lettenmaier and Burges 1977; Bras and Rodriguez-Iturbe 1993). For a given data set Y_t, and specific values of m and s [Eqs. (3) and (4)], the standard maximum likelihood method can be used to estimate the parameters k_i, w_j, $\theta(B)$, and $\phi(B)$.

Cluster analysis (CA) can be applied to form groups based on the similarity of variables (Manley 1995). Many algorithms have been proposed for CA. Hierarchical methods, a popular group of algorithms, investigate the data structure at several different levels and are used in this paper. The data sets for CA usually consist of the values of p variables $X_1, X_2, \ldots, X_p$ for n objects. The Euclidean distance function, d_{ij}, measures the distance between two objects i and j (Manley 1995). Variables are usually standardized before distances are calculated; thus, all p variables are equally important in determining these distances.

One popular procedure, the average linkage method (Sokal and Michener 1958), was used to extract the clusters. An iterative process and plots of RMSSTD (root-mean-square standard deviation) against number of clusters are used to find an appropriate number to classify the data sets. Other methods such as pseudo F, t^2 statistics, and CCC (cubic clustering criterion) can also be used (Milligan and Cooper 1985; SAS 1990). Although the average linkage method provides a solution, the solution needs to be consistent. Several techniques were discussed to check the consistency of a cluster solution (Aldenderfer and Blashfield 1984; Holgerson 1978; Hartigan 1975). In this paper, the replication technique serves to test the consistency of a cluster solution. The main idea of this technique is to randomly divide a data set into two subsets and then to check the consistency of the clusters. Each of the three data sets was repeatedly divided into two subsets after clusters have been obtained for the whole data set. The consistency of the clusters between the original data set and the two subsets will then be checked.

Discriminant analysis (DA) has been used to test the clusters to see if they are significantly different and to aid in interpreting the regional differences. DA also classified observations into two or more known groups to estimate the error rate from the observations in groups classified by CA. DA determines the canonical discriminant functions $(\mathbf{Z}_{d1}, \mathbf{Z}_{d2}, \ldots, \mathbf{Z}_{di})$ of the variables $X_1, X_2, \ldots, X_p$ that separate the m groups as much as possible. The simplest approach is based on Mahalanobis distance and takes a linear combination of the original X variables as canonical discriminant functions (Klecka 1980; Manley 1995). The canonical discriminant functions are defined as

$$\mathbf{Z}_{di} = a_{i1}\mathbf{X}_1 + a_{i2}\mathbf{X}_2 + \cdots + a_{ip}\mathbf{X}_p \tag{8}$$

where $\mathbf{Z}_{di} =$ vector of scores for n samples on the ith canonical discriminant function. Variables X_1, X_2, ..., and X_p are vectors for all n samples in the entire data set; $a_{i1}, a_{i2}, \ldots, a_{ip}$ are the canonical discriminant function coefficients for the variables in the ith canonical discriminant function.

The scores for the n samples on the canonical discriminant function variables have possible multiple correlations with the groups. The highest multiple correlation is called the first canonical correlation (SAS 1990). The second-highest canonical correlation is obtained by finding the linear combination uncorrelated with the first canonical discriminant variable (Z_{d1}). The third variable (Z_{d3}) is uncorrelated with Z_{d1} and Z_{d2}. Therefore, Z_{d1} has the highest possible multiple correlation with the groups, Z_{d2} has the second-highest possible multiple correlation with the groups, and so on. In these discriminant functions, the first few functions may be sufficient to account for most of the important group differences. If so, a simple graphical representation of the relationship between the various groups is possible by plotting the values of these functions for the sample individuals.

Principal component analysis (PCA) has been used to examine relationships among the original variables and to aid interpreting each group. The object of PCA is to take p variables, $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p$, and to find linear combinations of these variables to produce uncorrelated principal components $\mathbf{Z}_{C1}, \mathbf{Z}_{C2}, \ldots, \mathbf{Z}_{Ci}, \ldots, \mathbf{Z}_{Cp}$ (Bennett and Bowers 1976; Dunteman 1989; Manley 1995). These principal components can be expressed as

$$\mathbf{Z}_{ci} = b_{i1}\mathbf{X}_1 + b_{i2}\mathbf{X}_2 + \cdots + ib_{ip}\mathbf{X}_p \tag{9}$$

subject to the condition that

$$b_{i1}^2 + b_{i2}^2 + b_{i3}^2 \cdots + b_{ip}^2 = 1 \tag{10}$$

where $\mathbf{Z}_{Ci} =$ vector of scores for n samples on the ith principal component function. Variables $\mathbf{X}_1, \mathbf{X}_2, \ldots$ and $\mathbf{X}_p$, are vectors for all n samples in entire data set; and b_{i1}, $b_{i2}, \ldots$, and b_{ip} are the principal component function coefficients for variables in the ith principal component function. These components are ordered so that Z_{C1} displays the largest amount of variation, Z_{C2} displays the second-largest amount of variation, and so on. Most of the principal components may be negligible. Therefore, the first few Z_C variables account for most of the variation in the data set.

PCA finds the eigenvalues, corresponding eigenvectors, and coefficients in Eqs. (9) and (10). Eigenvalues show the percentage of variance accounted for by each principal component. Eigenvectors are independent, uncorrelated, and orthogonal. The eigenvectors or principal components equal the eigenvectors of the correlation or covariance matrix of the original variables. The lack of correlation implies that the principal components are measured in different and independent ''dimensions'' in the data.

Methodology and Study Area

In this paper, time series analysis is first applied to estimate streamflow parameters at each gauged watershed. Streamflow parameters, watershed characteristics, and the combination of them are used as three sets of criteria to classify watersheds into three sets of groups by CA. The consistence of the classified groups is checked by replication technique to determine an appropriate set of criteria for classification. DA is then employed to test the significance of the classified groups, to investigate the regional differences, and to identify the regional membership by independent variables. PCA aids to interpret the regional differences and similarities.

In order to minimize the snow effect and to have relative climatological homogeneity, watershed areas in Alabama, Georgia, and Mississippi are selected and studied. The data set of 94 candidate stations used in this study—including 20 stations (A1,A2, A3, . . . ,A20) in Alabama, 44 stations (G1,G2,G3, . . . ,G40) in Georgia, and 30 stations (M1,M2,M3, . . . ,M30) in Mississippi—are obtained from the United States Geological Survey (USGS) and are retrieved from the National Water Data Storage and Re-

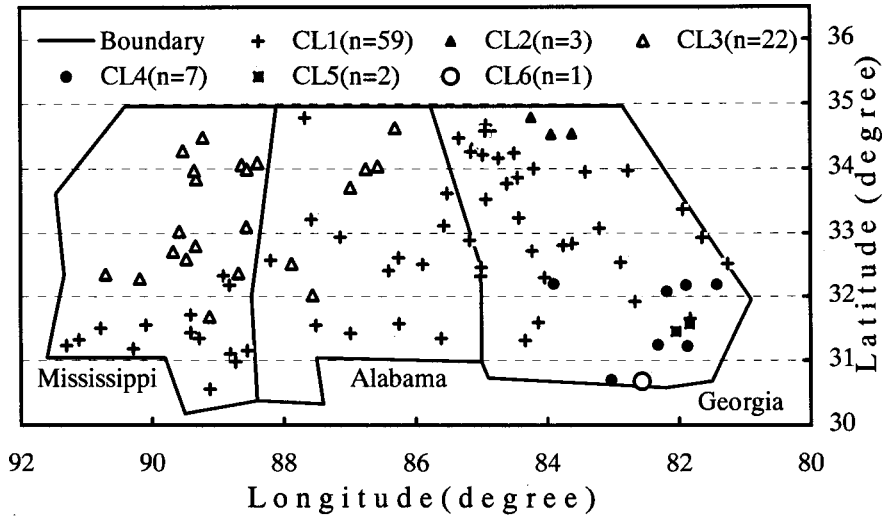

Fig. 1. Hydrologic regions of study area

trieval System. The locations of the 94 candidate stations are shown in Fig. 1 after classification. In order to have a better estimation of streamflow patterns, 25 years (1959–1983) of daily streamflow records (mean value for each day, in cfs) and watershed areas (in mi^2) are used to develop the specific monthly streamflows (in cfs/mi^2).

Three data sets [A (streamflow variables), B (watershed variables and precipitation information), and C (the combination of data sets A and B)] are used to classify the 94 watersheds into three sets of groups. Based on the results of the time series model, data set A consists of parameters $k_0, k_i, w_i, \theta(B), \phi(B)$ [Eqs. (3)–(5)], R^2, and σ^2 (variance of the residuals). The parameter R^2 explains the systematic pattern and the variation of the series. In data set B, watershed variables include A_w (watershed area, in mi^2), A_f (forest area, in %, percent of contributing drainage area), A_S (area of storage, in %, percent of contributing drainage area), E (elevation, in ft, above mean sea level), L (stream length per unit area, in mi/mi^2), S (main channel slop, in ft/mi), and P (mean annual precipitation, in in.). Note that precipitation (P) is included as one watershed variable even though it is considered as the input to the system. In addition, latitude (Lat, in decimal degree) and longitude (Lon, in decimal degree) are used to identify the location for each watershed at gauge.

The Forecasting and Modeling Package of the SCA Statistical System (Liu et al. 1992) is used to analyze the 94 streamflow time series and parameterize the streamflow patterns. Both of the procedures, TSMODEL for the model specification and ESTIM for the model estimation, have been used. The normality plots for each variable and the linear scatter plots for all pairs of different variables are investigated first to check whether the data follow a multivariate normal distribution. CA classifies data sets A, B, and C into three sets of groups (e.g., A1,A2, . . .). In addition, each data set is randomly divided into two subsets (e.g., Aa,Ab). As stated in CA, if the watersheds in each group of data set A (or B or C) mostly appear in one group of subsets Aa and Ab, it is concluded that the cluster solutions are very consistent. DA and PCA are applied to test and interpret the differences and similarities between clusters. DA also classifies observations into two or more known groups to estimate the correct classification percentage from the observations in groups classified by CA.

Results and Hydrologic Interpretation

Time Series Analysis

Following the procedures of present proposed methodology, the streamflow pattern for the time series model is identified as

$$\ln(Y_t) = \sum_{j=1}^{12} w_j D_{jt} + (1 - \theta_1 B - \theta_2 B^2) a_t \tag{11}$$

where Y_t represents the specific monthly streamflows and is logarithmically transformed. The first term in the right-hand side is the monthly seasonal component [Eq. (4)], and the second term is the irregular component and is expressed as an MA(2) pattern [Eq. (5)]. The MA pattern is a weighted amount of error (i.e., $\theta(B)a_t$), which occurs at the current month and the prior two months (weighted values are θ_1 and θ_2). ARMA and higher orders of MA models were used in the earlier stage of model specification. It is found that only terms up to the second order make up major contributions to the streamflow with higher value of R^2 parameter (similar to coefficient of determination in regression analysis). It shows that streamflows are affected seasonally by hydrological conditions of the previous two months. The estimated streamflow parameters include monthly streamflow parameters (MSPs) $w_1, w_2, \ldots, w_{12}$ (for January though December), MA(2) parameters (θ_1 and θ_2), the residual variance (σ^2), and R^2 for each of the 94 gauged stations. Parameter R^2 represents a percentage of the total variation explained by the time series model.

The mean values of the estimated streamflow parameters in Eq. (11) for each group are shown in Fig. 2. Except for R^2, the estimated parameters at each station could be applied to synthesize monthly streamflow and extend the streamflow information at the same station. The MSPs characterize the level (but not equal) of the mean monthly streamflows and form a seasonal pattern for each watershed. The mean values of the "high" MSPs, w_1, w_2, w_3, w_4, w_5, and w_{12} are positive and characterize the mean monthly streamflows in the wet season. The mean values of the "low" MSPs, w_6, w_7, w_8, w_9, w_{10}, and w_{11} are negative and represent those in the dry season.

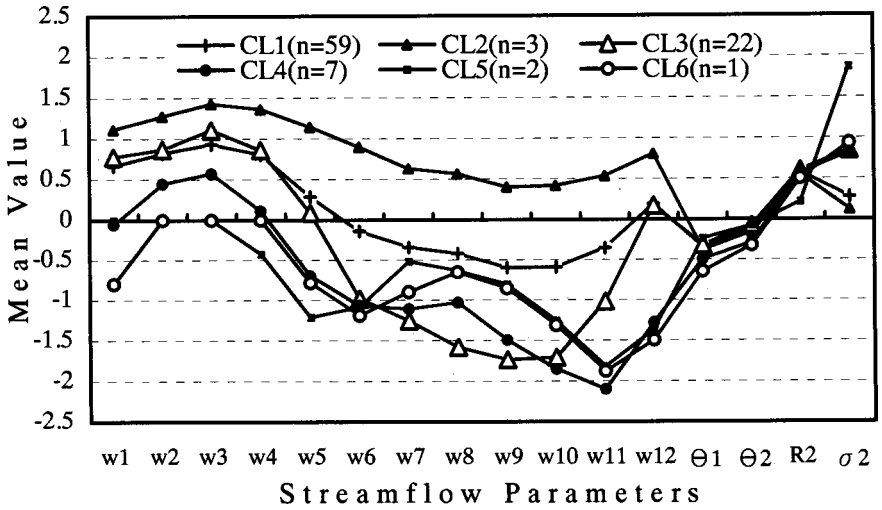

Fig. 2. Regional patterns of streamflow parameters (data set A)

The streamflow random parameters, θ_1 and θ_2, are the weighted number of errors of the past two monthly streamflows to the current monthly streamflow. The absolute value of the mean of θ_1 (.35) is greater than that of the θ_2 (.13). This is reasonable, since the past monthly streamflow is more correlated to the current monthly streamflow. The streamflow residual variance σ^2 provides the stochastic nature in the streamflow synthesis. All streamflow parameters except R^2 are used to synthesize monthly streamflows. The mean of R^2 equals 0.575; thus, the time series model explains 57.5% of the streamflow variations, and the rest of the variations (43.5%) are due to the random component. Small values of R^2 ranging from 0.158 to 0.321 are found at five stations (G15,G17,G23,G24,G40) in Georgia. The main reason for the small R^2 values is that the residual ACFs of these stations are not white noise. Several outliers are found in the five stations. However, the present model [Eq. (11)] fits most stations and is used in this paper.

Testing for Normality

Most of the watershed variables show wide ranges, especially the watershed area, A_w (37 to 30,810 mi^2), and the storage area, A_s (0%–52%), but the range of precipitation, P (41–65 in./year), is relatively narrow. These ranges denote a wide variability of the watershed variables in A_w and A_s in contrast to comparatively homogeneous annual precipitation in the study area. Based on the normality plots and bivariate plots, data sets A, B, and C are considered as multivariate normally distributed after Log transformation of the watershed variables. Most bivariate plots follow linear relationships, especially plots involving the paired variables A_w and L, E and w_5, and most w_i and w_{i+1} or w_i and w_{i+2} ($i=1,2,\ldots,12$; in cyclic) are strongly linear.

Cluster Analysis

The NCL (number of cluster) is selected to be 6 for data sets A, B, and C because the number of observations in the obtained two main clusters is not changed. In the subsets, the NCL is selected to equal 6, 9, and 7 for subsets Aa, Ba, and Ca and 5, 6, and 5 for subsets Ab, Ba, and Cc, respectively. The NCL selection is based on iterative trial processes and plotting RMSSD (root-mean-square standard deviation) against NCL. The NCL for subsets is to obtain two main clusters. After applying the replication technique, the cluster solutions of subsets seem to be consistent with their mother data sets. The resultant consistent classification percentages of consistency check for CA are 90%, 70% and 78% for data sets A, B, and C, respectively. The consistent classification percentage is the sum of the correctly classified observations in subsets over the total observations for each cluster. Therefore, the cluster solutions are highly consistent, especially in data set A.

The hydrologic regions based on watershed locations and clusters of data sets A (Fig. 1) and C are similar to each other and tend to be spatially contiguous; but the clusters of data set B are not, and there are many overlaps between the areas of the two main clusters. The two main clusters in data sets A and C are located approximately in the northwest and southeast of the study area, respectively. The various cluster solutions in data set B are not as similar as those in data set A. The consistent classification percentage of data set C (78%) is less than that of data set A (90%). Therefore, using data set A (streamflow parameters) as a set of criteria for regionalization is more appropriate than using data sets B and C, and the following results of further analysis focus rather on the cluster solutions in data set A than those in data sets B and C.

In data set A, the two main clusters (Cluster-1 and Cluster-3) tend to be separated by the Appalachian Mountains and the Gulf Coastal Plain. The watersheds of Cluster-2 ($N=3$) are located in the southern end of the Blue Ridge. These watersheds have high elevations and channel slopes. The watersheds of Cluster-4 ($N=7$), Cluster-5 ($N=2$), and Cluster-6 ($N=1$) are located in the Atlantic coastal area. In this area, the watershed elevations are low and the watersheds are covered with many marshes. Also, this area experiences hurricanes and very severe thunderstorms in the summer (USEDS 1959–1983). Since the cluster solutions tend to be separated by the physiographical boundaries, using a "subjective" procedure based on physiographic information would be reasonably appropriate as a first-order guide for regional boundaries. However, an "objective" procedure like CA is more appropriate because it also classifies specific watersheds into different groups within the physiographically defined regions.

In the two main clusters, the average values of the "high" MSPs (monthly streamflow parameters) are similar. But, the pattern of Cluster-3 (Fig. 2) displays smaller values in the "low" MSPs than those of Cluster-1. In addition, the average values of A_w (watershed area) and A_s (area of storage) of Cluster-3 are

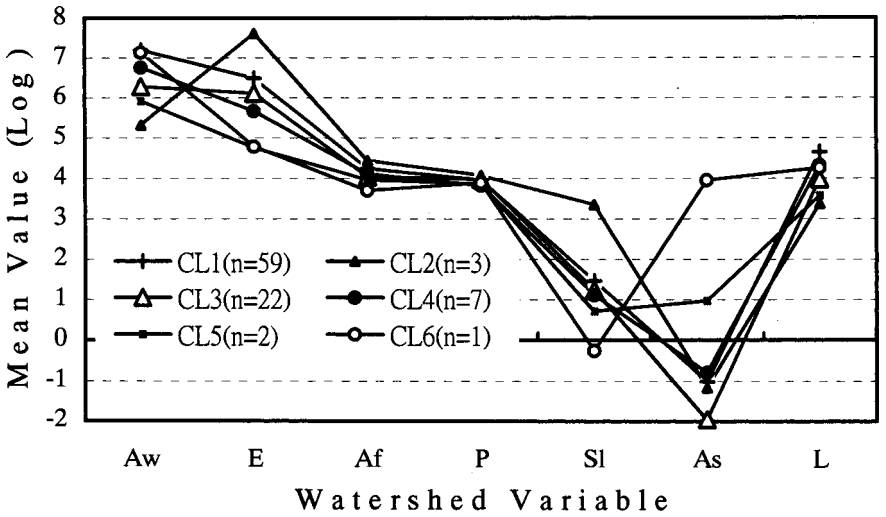

Fig. 3. Regional patterns of watershed variables (data set A)

smaller than those of Cluster-1 (Fig. 3). This may be reasonable, because small watersheds tend to be flashy and smaller surface storage provides less buffering capacity. The three watersheds of Cluster-2 are located in the southern end of the Blue Ridge. As shown in Figs. 2 and 3, the three watersheds have the largest average values of the MSPs, E (basin elevation), A_f (forest area), P (annual precipitation), and S (main channel slope) and the smallest average values of A_w, L (stream length per unit area), and σ^2 (residual variance). In mountain areas, watersheds generally exhibit these characteristics.

In data set A, the streamflow patterns in Clusters-1, -2, and -3 exhibit one peak at w_3, but those in Clusters-4, -5, and -6 located in the Atlantic coastal area exhibit another peak at w_7 or w_8 (Fig. 2). Generally, summer is the dry season in the study area, but the Atlantic coastal area was hit by hurricanes in August and September 1964 and by very severe thunderstorms in July 1964 and August 1969 (USEDS 1959–1983). In addition, the values of the MSPs in Clusters-4, -5, and -6 seem to be lower than the other clusters except for the months experiencing extreme storm events. This is due to low E and S, small A_f, and large A_s, as shown in Fig. 3. In other words, the specific monthly streamflows are low in the low elevation watersheds with large surface storage capacity.

Since the means of the streamflow parameters and watershed variables show different patterns in each cluster of data set A, this emphasizes the importance of classification. Combining all observations without classification would obscure much useful information. If the clusters are tested to be properly classified and significantly different by DA, and the cluster membership can be identified by independent variables (e.g., watershed characteristics), the proposed scheme should be appropriate for hydrologic regionalization and can be applied to synthesize streamflow information at ungauged sites.

Discriminant Analysis

Since there are many variables in the data sets, the SAS *STEPDISC* procedure of discriminant analysis (DA) finds a subset of variables that best reveals differences among classes. After the stepwise entry and deletion of variables (P-value of F statistics <0.05), variables are reduced from 15 to 9 streamflow variables ($w_{12},\sigma^2,w_{10},w_2,w_1,R^2,w_3,\theta_2,\theta_1$) or from 7 to 5 watershed variables (E,P,L,S,A) in data set A. The MANOVA (multivariate analysis of variance) procedure tests the hypothesis that the squared Mahalanobis distances (D_M) of the nine streamflow variables between the class means are significantly different (P-values <0.05). When all watershed variables are used as the entered variables in data set A, the D_M between two pairs of small clusters (Clusters-4 and -5, Clusters-5 and -6) are not significantly different. However, Clusters-4 ($N=7$) and 5 ($N=2$) are significantly different in E (elevation) when the ANOVA (analysis of variance) test is conducted. These tests imply that each cluster represents one ''hydrologic'' region based on streamflow parameters, and the regional differences can be discriminated by watershed variables with the exception of the difference between Cluster-5 ($N=2$) and Cluster-6 ($N=1$).

When the nine streamflow variables are entered as discriminant variables, the correct classification percentage (CCP) is 98%. The CCP is the correctly classified observations over the total observations in the data set. For example, only two observations in data set A ($N=94$) are classified into another cluster. When the entered variables are reduced to 4 ($w_8,w_{10},w_{12},\sigma^2$), the CCP is 94%. The high CCPs indicate that the observations are properly classified and the cluster solutions are very stable. Based on the cluster solutions of data set A, when all and 4 watershed variables are entered as discriminant variables, the CCPs are 86% and 85%, respectively. The high CCPs imply that the data set of the watershed variables contains a lot of overlapped information and that a strong relationship between streamflow and watershed variables exists.

There are six groups for data set A; therefore, five canonical discriminant variables $Z_{d1},Z_{d2},Z_{d3},Z_{d4},Z_{d5}$ are constructed from DA (Table 1). In the data set, Z_{d1}, Z_{d2}, and Z_{d3} account for 89% of the total variance with high canonical correlations ($>.85$) and large eigenvalues (>2.0). Z_{d1} extracts 55% of the variance in the data set and identifies the ''high'' MSPs and R^2 as the best discriminators (coefficient$>.5$). Z_{d2} is associated with 18% of the variance and compares the ''low'' MSPs (coefficients$>.5$) inversely with one of the ARMA parameters σ^2 (coefficient$=-.73$). Z_{d3} explains 16% of the information and mainly involves parameters θ_1 and θ_2 (coefficients$<-.5$). When seven watershed variables are used as discriminators in data set A, Z_{d1} and Z_{d2} extract 88% of the total variations (Table 1). Z_{d1} accounts for

Table 1. Canonical Discriminant Variables of Discriminant Analysis

Canonical discriminant analysis					Total canonical structure				
Correlation	Eigenvalue	Proportion	Cumulation		Z_{d1}	Z_{d2}	Z_{d3}	Z_{d4}	
(a) Data set A (based on streamflow parameters)									
Z_{d1}	0.951	9.503	0.550	0.545	w_1	.729	0.221	−0.344	−0.355
Z_{d2}	0.873	3.207	0.186	0.735	w_2	.706	0.213	−0.009	−0.275
Z_{d3}	0.856	2.734	0.158	0.893	w_3	.766	0.020	−0.048	−0.264
Z_{d4}	0.714	1.037	0.060	0.953	w_4	.799	0.301	−0.064	−0.090
Z_{d5}	0.668	0.808	0.047	1.000	w_5	.673	0.559	−0.000	−0.159
					w_6	.196	0.834	0.143	−0.274
					w_7	−0.035	0.868	0.096	−0.247
					w_8	−0.142	0.876	0.224	−0.211
					w_9	−0.071	0.871	0.119	−0.161
					w_{10}	0.098	0.898	0.063	−0.166
					w_{11}	0.479	0.762	−0.105	−0.144
					w_{12}	0.793	0.434	−0.269	−0.126
					θ_1	0.281	0.082	−0.643	−0.273
					θ_2	0.374	0.008	−0.597	−0.046
					σ^2	−0.427	−0.727	−0.430	0.103
					R^2	0.500	−0.132	0.326	0.081
(b) Data set A (based on watershed variables)									
Z_{d1}	0.803	1.819	0.683	0.683	A_w	0.034	0.420	0.353	0.736
Z_{d2}	0.585	0.522	0.196	0.879	S	0.583	0.056	0.472	−0.601
Z_{d3}	0.430	0.227	0.085	0.964	L	0.079	0.488	−0.419	0.711
Z_{d4}	0.281	0.086	0.032	0.996	E	0.836	0.149	−0.029	−0.117
Z_{d5}	0.100	0.010	0.004	1.000	A_s	−0.146	0.127	0.426	0.795
					A_f	0.435	0.425	0.136	0.038
					P	0.566	−0.509	−0.130	0.177

68% of the information with high canonical correlation (.8) and is associated with E, S, A_f, and P (coefficient$>$.4). Z_{d1} explains 20% of the variance with moderate canonical correlation (.58) and seems to show that A_w, L, and A_f (coefficients$>$.4) are inversely related to P (coefficient$=-.51$).

Based on the canonical discriminant variables, the regional differences depend on the ''high'' MSPs in Z_{d1}, the ''low'' MSPs in Z_{d2}, and ARMA parameters θ_1 and θ_2 in Z_{d3} (Table 1). The major differences between the two main clusters are the ''low'' MSPs in Z_{d2} (Fig. 2). If watershed variables are used, the regional differences depend on E, S, P, and A_f in Z_{d1} and A_w, L, A_f, and P in Z_{d2} (Table 1). Plotting the scores of the first two canonical variables (Z_{d1} vs. Z_{d2}) tends to separate the hydrologic regions, as shown in Fig. 4. Note that the scores of Z_{d1} and Z_{d2} calculated from the watershed variables along with their coefficients of Z_{d1} and Z_{d2} (Table 1) at ungauged sites can be applied to identify their regional membership. The membership identification of DA is a critical procedure in hydrologic regionalization.

Principal Component Analysis

Principal component analysis (PCA) interprets the regional similarities and differences. The lack of correlation between the principal components is a useful property in data analysis because

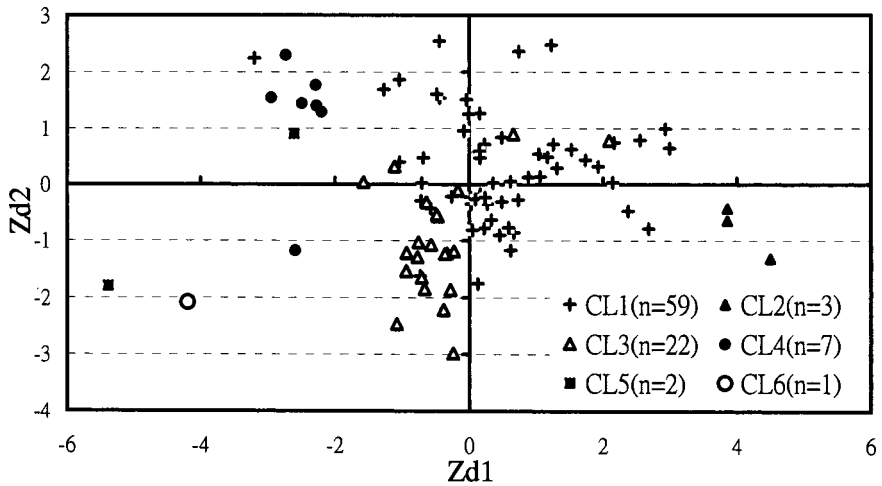

Fig. 4. Z_{d1} versus Z_{d2} of data set A (based on watershed variables)

Table 2. Principal Component Analysis

	Eigenvalues of the correlation matrix				Eigenvectors			
	Eigenvalue	Proportion	Cumulative		$\mathbf{Z}_{C1}$	$\mathbf{Z}_{C2}$	$\mathbf{Z}_{C3}$	$\mathbf{Z}_{C4}$
(a) Cluster-1 (based on watershed variables, 59 observations)								
$\mathbf{Z}_{C1}$	2.988	0.427	0.427	A_w	−0.547	−0.000	0.215	−0.250
$\mathbf{Z}_{C2}$	1.498	0.214	0.641	S	0.444	0.396	0.015	−0.102
$\mathbf{Z}_{C3}$	1.164	0.166	0.807	L	0.521	0.064	−0.145	0.394
$\mathbf{Z}_{C4}$	0.591	0.084	0.892	E	−0.055	0.715	0.192	−0.288
				A_s	−0.395	0.232	0.127	0.823
				A_f	0.188	0.117	0.792	0.102
				P	0.196	−0.510	0.502	−0.036
(b) Cluster-3 (based on watershed variables, 22 observations)								
$\mathbf{Z}_{C1}$	3.286	0.469	0.469	A_w	−0.507	0.100	−0.061	0.374
$\mathbf{Z}_{C2}$	1.398	0.200	0.669	S	0.419	0.165	−0.139	−0.542
$\mathbf{Z}_{C3}$	0.792	0.113	0.782	L	0.465	−0.015	0.366	0.183
$\mathbf{Z}_{C4}$	0.659	0.094	0.876	E	0.128	0.675	0.364	0.417
				A_s	−0.477	0.098	−0.085	−0.222
				A_f	−0.025	−0.674	0.507	0.137
				P	0.325	−0.209	−0.668	0.539

redundant information is omitted. When the 16 streamflow variables are used as entered variables, the eigenvalues (>1.0, a common rule for evaluation; Hamilton 1992) of the first three principal components Z_{C1}, Z_{C2}, and Z_{C3} extract more than 80% of the standardized variance within each of the two main clusters. The variances characterized by Z_{C1}, Z_{C2}, and Z_{C3} show that the differences between the two main clusters are mainly based on the ''low'' MSPs and that the similarities are based on the ''high'' MSPs. Since the regional membership is identified by the watershed variables, using watershed variables as the entered variables in PCA is investigated in detail. The first three principal components extract about 80% of the total variance in the two main clusters (Table 2). Z_{C1} extracts more than 40% of the variance and explains the relationships between S (coefficient >.41) and L (>.46) inversely with A_w (<−.50) and A_S (<−.39) within each of the two main clusters. Z_{C2} accounts for about 20% of the variance within each of the two main clusters, and Z_{C2} is related to E (.72), S (.49), and P (−.51) in Cluster-1 and is associated with E (.68) and A_f (−.67) in Cluster-3. Z_{C3} reflects the positive and inverse correlations between A_f and P in the two main clusters. Therefore, the similarities are based on S, L, A_w, and A_S extracted by Z_{C1}, and the major differences depend on watershed variables E, A_f, P, and S explained by Z_{C2} and Z_{C3}. The similarities imply that (1) surface storage area (A_S, in %) increases with increasing watershed area (A_w), (2) stream length per unit area (L) increases with increasing main channel slope (S), and (3) L and S decrease with increasing A_w or A_S. The differences between the two main clusters are the important watershed characteristics of E, P, A_f, and S, which are used to identify the regional membership in DA (Table 1).

Conclusions and Discussions

A hydrologic regionalization scheme is proposed for the classification of watersheds at gauged sites in this paper. This scheme makes several significant progressions from other studies. In previous studies, a time series model was not applied for hydrologic regionalization, and only streamflow information or watershed variables were used as one set of criteria for CA. In this paper, the component with MA(2) time series model retains the maximum amount of statistical information and the stochastic nature of the monthly streamflows. This model can be applied to synthesize a large number of streamflow sequences at ungauged sites and extended to a desired period. In addition, the comparison of the three sets of cluster solutions provides the best set of criteria (streamflow information) for hydrologic regionalization.

The proposed scheme uses 16 streamflow parameters estimated by a time series model to classify 94 watersheds into six hydrologic regions by CA. The classified regions seem to be separated by physiographical boundaries, especially the two main clusters. Similarities and differences between streamflow parameters and between watershed variables are found in the two main clusters based on the results of DA and PCA. Watershed variables, E, A_f, S, and P are mainly used to identify the regional membership in DA. This emphasizes the importance of the hydrologic regionalization and the identification of the specific characteristics in each region. Combining all observations without classification would obscure much useful information. Since hydrologic regions are significantly different and can be identified by independent variables (i.e., watershed characteristics), the proposed scheme in this paper is appropriate for hydrologic regionalization and can be applied to synthesize streamflow information at ungauged sites. The application of the streamflow synthesis will be presented in a forthcoming paper and will include (1) prediction of streamflow parameters by the watershed variables at ungauged sites via the constructed multiple regression equations, and (2) synthesis of streamflows using the streamflow parameters by the time series model.

The time series model used in this paper fits for most studied watersheds. Other models such as the seasonal ARMA model, the variances of the monthly streamflows, or the irregular component in Eq. (11) exhibit month-to-month variation and may be further considered. Difficulties may arise for the seasonal ARMA model in hydrologic interpretation of streamflow parameters. Different models with too many variables and limited observations may result in difficulties for classification and regression analysis. However, these issues should be further studied.

Acknowledgment

The writers would like to thank the reviewers of this paper for their comments and suggestions.

Notations

The following symbols are used in this paper:

$a1,a2,a3$ = coefficients;
a_{i1},a_{i2},a_{ip} = canonical discriminant function coefficients;
a_t = random error;
b_{i1},b_{i2},b_{ip} = principal component function coefficients;
C,C' = constant;
D_{jt} = indicator variable of jth period with time index t;
k = empirical coefficient;
k_0 = parameter;
k_i = parameter associated with i;
N_t = irregular components;
Q = streamflow variable;
S_t = seasonal component;
T_t = trend component;
t^i = time index of degree i;
W_1,W_2,W_3 = watershed or climatic characteristics;
w_j = coefficient of jth period;
X_1,X_2,X_p = vectors for all n samples in entire data set;
Y_t = streamflow time series;
Z_{ci} = ith principal component function;
Z_{di} = ith canonical discriminant function;
$\theta_1,\theta_2,\theta_q$ = parameters;
$\theta(B)$ = moving average (MA) operator;
σ_a^2 = variance of errors;
ϕ_1,ϕ_2,ϕ_p = parameters; and
$\phi(B)$ = autoregressive (AR) operator.

References

Aldenderfer, M. S., and Blashfield, R. K. (1984). *Cluster analysis*, SAGE, Calif.

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.*, 115(6).

Box, G. E. P., and Jenkins, G. H. (1976). *Time series analysis: Forecasting and control*, 2nd Ed., Holden Day, San Francisco.

Bras, R. L., and Rodriguez-Iturbe, I. (1993). *Random functions and hydrology*, Dover, New York.

Burn, D. H. (1989). "Evaluation of regional flood frequency analysis with a region of influence approach." *Water Resour. Res.*, 26(10).

Cheng, C. C. (1988). "*Hydrologic regionalization based on flow duration curve*." PhD dissertation, Syracuse Univ., Syracuse, N.Y.

DeCoursey, D. G. (1973). "Object regionalization of peak flow rates, floods and droughts." *Proc., 2nd Int. Symposium in Hydrology*, Water Resources, Fort Collins, Colo.

Dunteman, G. H. (1989). *Principal components analysis*, Sage, Calif.

Fennessey, N., and Vogel, R. M. (1990). "Regional flow-duration curves for ungaged sites in Massachusetts." *J. Water Resour. Plan. Manage.*, 116(4).

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

Hamilton, L. C. (1992). *Regression with graphics*, Wadsworth, Inc., Calif.

Hartigan, J. (1975). *Cluster algorithms*, Wiley, New York.

Hawley, M. E., and McCuen, R. H. (1982). "Water yield estimation in Western United States." *J. Img. Drain. Eng.*, 108(1).

Hipel, K. W., McBean, E. A., and McLeod, A. I. (1979). "Hydrologic generating model selection." *J. Water Resour. Plan. Manage.*, 105(2).

Holgerson, M. (1978). "The limited value of cophenetic correlation as a clustering criterion." *Pattern Recogn.*, 10, 187–213.

Klecka, W. R. (1980). "*Discriminant analysis*," Sage, Calif.

Lettenmaier, D. P., and Burges, S. J. (1977). "Operational assessment of hydrologic models of long-term persistence." *Water Resour. Res.*, 12(1).

Liu, L. M., Hudak, G. B., Box, G. E. P., Muller, M. E., and Tiao, G. C. (1992). *Forecasting and time series analysis using the SCA statistical system*, Vol. I, Scientific Computing Associates, Ill.

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

Milligan, G. W., and Cooper, M. C. (1985). "An examination of procedures for determining the number of clusters in a data set." *Psychometrika*, 50, 159–179.

Mosley, M. P. (1981). "Delimitation of New Zealand hydrologic regions." *J. Hydrol.*, 49.

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

Quimpo, R. G., Alejandrino, A. A., and McNally, T. A. (1983). "Regionalized flow duration for Philippines." *J. Water Resour. Plan. Manage.*, 109(4).

Rodriguez-Iturbe, I. (1968). "A modern statistical study of monthly levels in the Orinoco River." *Bulletin of the International Association of Scientific Hydrology*, No. 4, 25–41.

Rodriguez-Iturbe, I., and Nordin, C. F. (1968). "Time series analysis of water and sediment discharges." *Bulletin of the International Association of Scientific Hydrology*, No. 2, 69–84.

Roesner, L. A., and Yevjevich, V. (1966). "Mathematical models for time series of monthly precipitation and monthly runoff." *Hydrology Paper No. 15*, Colorada State Univ., Fort Collins, Colo.

Salas, J. D. (1993). "Analysis and modeling of hydrologic time series." *Handbook of hydrology*, D. R. Maidment, ed. in chief, McGraw-Hill, New York.

SAS/STAT user's guide. (1990). Volume 1, ACECLUS-FREQ, Version 6, 4th Ed., SAS Institute, Cary, N.C.

SAS/STAT user's guide. (1990). Volume 2, GLM VARCOMP, Version 6, 4th Ed., SAS Institute Cary, N.C.

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

Sokal, R. R., and Michener, C. D. (1958). "A statistical method for evaluating systematic relationships." *Univ. Kans. Sci. Bull.*, 38, 1409–1438.

Tao, P. C., and Delleur, J. W. (1976). "Seasonal and nonseasonal ARMA models in hydrology." *J. Hydraul. Div., Am. Soc. Civ. Eng.*, 2(10).

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

USGS (1980). "Data formats for U.S. Geological Survey computer files containing daily values For water parameters." U.S. Geological Survey, Washington, D.C.

USEDS (1959–1983). "*Climatological data*." Dept. of Commerce, National Oceanic and Atmospheric Administration, Environmental Data Service, National Climatic Center, Asheville, N.C.

Wei, William W. S. (1990). *Time Series Analysis—Univariate and Multivariate Methods*, Addison-Wesley, Redwood City, Calif.