WATER RESOURCES RESEARCH, VOL. 34, NO. 1, PAGES 107–119, JANUARY 1998

Disaggregation procedures for stochastic hydrology based on nonparametric density estimation

David G. Tarboton, Ashish Sharma,[1] and Upmanu Lall
Utah Water Research Laboratory, Utah State University, Logan

Abstract. Synthetic simulation of streamflow sequences is important for the analysis of water supply reliability. Disaggregation models are an important component of the stochastic streamflow generation methodology. They provide the ability to simulate multiseason and multisite streamflow sequences that preserve statistical properties at multiple timescales or space scales. In recent papers we have suggested the use of nonparametric methods for streamflow simulation. These methods provide the capability to model time series dependence without a priori assumptions as to the probability distribution of streamflow. They remain faithful to the data and can approximate linear or nonlinear dependence. In this paper we extend the use of nonparametric methods to disaggregation models. We show how a kernel density estimate of the joint distribution of disaggregate flow variables can form the basis for conditional simulation based on an input aggregate flow variable. This methodology preserves summability of the disaggregate flows to the input aggregate flow. We show through applications to synthetic data and streamflow from the San Juan River in New Mexico how this conditional simulation procedure preserves a variety of statistical attributes.

1. Introduction

A goal of stochastic hydrology is to generate synthetic streamflow sequences that are statistically similar to observed streamflow records. Such synthetic streamflow sequences are useful for analyzing reservoir operation and stream management policies. Often, multiple reservoir sites and stream sections need to be considered as part of a system operation plan, and the operating horizon may extend from a few days to several years. For proper system operation it is important that the streamflow sequences generated for the different sites and/or time periods be "compatible." Practically, this suggests that (1) the flow recorded at a downstream gage be represented as the sum of the tributary flows and channel losses/gains, (2) the annual flow represent a sum of the monthly flows, (3) the monthly fractions of flows in wet/dry years be representative of wet/dry years respectively, and (4) the relative delay between the rise and fall of streams in the basin be reproduced. Statistically, this implies that the joint probability distribution of the flow sequences at the different sites and time periods needs to be preserved. As the number of sites/time periods increases this entails the estimation/specification of a high dimensional density function from a relatively small number of data points. Recognizing this problem, a significant body of hydrologic literature evolved on disaggregation models [*Harms and Campbell*, 1967; *Valencia and Schaake*, 1972; *Mejia and Rousselle*, 1976; *Curry and Bras*, 1978; *Lane*, 1979; *Salas et al*., 1980; *Svanidze*, 1980; *Stedinger and Vogel*, 1984; *Bras and Rodriguez-Iturbe*, 1985; *Stedinger et al*., 1985; *Grygier and Stedinger*, 1988; *Santos and Salas*, 1992]. The essence of these models is to develop a staging framework [e.g., *Santos and Salas*, 1992], where flow sequences are generated at a given level of aggregation and then disaggregated into component flows (e.g., seasonal from annual, or monthly from seasonal). At each stage a low dimensional estimation problem is solved. Summability of disaggregated flows and their mutual correlation structure (after some transformation) is preserved. Historically, a parametric structure has been used for this process.

In this paper we present a nonparametric approach to the disaggregation of streamflow. Disaggregation is the simulation of the components of a vector of disaggregated variables $\mathbf{X}$ given (i.e., conditional on) an aggregate variable Z. The problem is posed in terms of sampling from the conditional probability density function.

$$f(\mathbf{X}|Z) = f(\mathbf{X}, Z) \Big/ \int f(\mathbf{X}, Z)\, d\mathbf{X} \tag{1}$$

In this equation $f(\mathbf{X}, Z)$ is the joint probability density function of the vector $\mathbf{X}$ of disaggregate variables (monthly or tributary streamflows) and Z the aggregate variable (annual or main stem streamflow) obtained from an aggregate model at each aggregate time step. The denominator in (1) above is the marginal probability density function of the aggregate variable Z derived by integrating the joint distribution over all the components of $\mathbf{X}$. Specifically, we consider a d-dimensional vector $\mathbf{X} = (X_1, X_2, \cdots X_d)^T$ with aggregate variable $Z = X_1 + X_2 + \cdots + X_d$. The superscript T denotes transpose. Vectors are taken to be column vectors. The model is estimated from n observations of $\mathbf{X}$ and Z, denoted $\mathbf{x}_i$ and z_i. The components of $\mathbf{x}_i$ are the historical disaggregate components, such as monthly, seasonal, or tributary flows that comprise the historical aggregate z_i. We use kernel density estimation techniques to estimate the joint and conditional densities in (1). These methods are data adaptive; that is, they use the historical data (the historical aggregate and component time series) to define the probability densities. Assumptions as to the form

[1]Now at Department of Water Engineering, School of Civil Engineering, University of New South Wales, Sydney, Australia.

Paper number 97WR02429.
0043-1397/98/97WR-02429$09.00

of dependence (e.g., linear or nonlinear) or to the probability density function (e.g., Gaussian) are avoided.

Historically, disaggregation approaches to streamflow synthesis have involved some variant of a linear model of the form

$$\mathbf{X}_t = \mathbf{A}Z_t + \mathbf{B}\mathbf{V}_t \tag{2}$$

Here $\mathbf{X}_t$ is the vector of disaggregate variables at time t, Z_t is the aggregate variable, and $\mathbf{V}_t$ is a vector of independent random innovations, usually drawn from a Gaussian distribution. **A** and **B** are parameter matrices. **A** is chosen or estimated to reproduce the correlation between aggregate and disaggregate flows. **B** is estimated to reproduce the correlation between individual disaggregate components. The many model variants in the literature make different assumptions as to the structure and sparsity of these matrices and which correlations the model should be made to directly reproduce. They also consider a variety of normalizing transformations applied to the data, prior to use of (2), to account for the fact that monthly streamflow data is seldom normally distributed. In these models summability, the fact that disaggregate variables should add up to the aggregate quantity, has also been an issue. It can be shown [see *Bras and Rodriguez-Iturbe*, 1985, p. 148] with a model of the form of (1) that summability of the disaggregate variables to the aggregate variables is guaranteed. However, when a normalizing transformation is used, or when various elements of the matrices are taken as zero during simplification, summability is lost. In these cases investigators [e.g., *Grygier and Stedinger*, 1988] have suggested empirical adjustment procedures to restore summability.

The key idea to recognize from these models, exemplified by (2), is that they provide a mathematical framework where a joint distribution of disaggregate and aggregate variables is specified. However, the specified model structure is parametric. It is imposed by the form of (2) and the normalizing transformations applied to the data to represent marginal distributions.

Some of the drawbacks of the parametric approach are the following.

1. Since (2) involves linear combinations of random variables, it is mainly compatible with Gaussian distributions. Where the marginal distribution of the streamflow variables involved is not Gaussian (e.g., perhaps there is significant skewness), normalizing transformations are required for each streamflow component. Equation (2) would then be applied to the normalized flow variables. It is difficult to find a general normalizing transformation and retain statistical properties of the streamflow process in the untransformed multivariable space.

2. The linear nature of (2) limits it from representing any nonlinearity in the dependence structure between variables, except through the normalizing transformation used. Given the current recognition of the importance of nonlinearity in many physical processes [e.g., *Tong*, 1990; *Schertzer and Lovejoy*, 1991], we prefer at the outset not to preclude or limit the representation of nonlinearity.

Following in the spirit of our recent work [*Lall and Sharma*, 1996; *Sharma et al*., 1997], the purpose of this paper is to develop a nonparametric disaggregation methodology. The necessary joint probability density functions are estimated directly from the historic data using kernel density estimates. These methods circumvent the drawbacks of the parametric methods that were listed. The methods are data driven and relatively automatic, so nonlinear dependence will be incorporated to the extent suggested by the data. Difficult subjective choices as to appropriate marginal distributions and normalizing transformations are avoided.

This paper is organized as follows. First the multivariate kernel density estimator used in the disaggregation model is presented. This is followed by a description of our nonparametric disaggregation approach. The performance of the nonparametric disaggregation procedure is then evaluated by applications to synthetic data from a known nonlinear model and to streamflow from the San Juan River near Archuleta, New Mexico, United States. Results from our approach are compared to those from SPIGOT [*Grygier and Stedinger*, 1990], a popular disaggregation software package based on linearizing transformations of the historical streamflow time series.

2. Kernel Density Estimation

Kernel density estimation entails a weighted moving average of the empirical frequency distribution of the data. Most nonparametric density estimators can be expressed as kernel density estimators [*Scott*, 1992, p. 125]. In this paper we use multivariate kernel density estimators with Gaussian kernels and bandwidth selected using least squares cross validation [e.g., *Scott*, 1992, p. 160]. This bandwidth selection method is one of many available methods. Its performance was compared with various cross-validation estimators for samples of sizes typically encountered in hydrology using a simulation study [*Sharma*, 1996]. Our methodology is intended to be generic and should work with any bandwidth and kernel density estimation method. Procedures for bandwidth and kernel selection are an area of active research in the nonparametric statistics community and as better methods become available they can be easily incorporated into our model. For a review of hydrologic applications of kernel density and distribution function estimators, readers are referred to *Lall* [1995]. *Silverman* [1986] and *Scott* [1992] provide good introductory texts.

A multivariate Gaussian kernel density estimate for a d-dimensional vector $\mathbf{x}$ can be written as

$$\hat{f}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(2\pi)^{d/2}\det(\mathbf{H})^{1/2}} \cdot \exp\left(-\frac{(\mathbf{x}-\mathbf{x}_i)^T\mathbf{H}^{-1}(\mathbf{x}-\mathbf{x}_i)}{2}\right) \tag{3}$$

where det( ) denotes determinant, n is the number of observed vectors $\mathbf{x}_i$, and $\mathbf{H}$ is a symmetric positive definite $d \times d$ bandwidth matrix [*Wand and Jones*, 1994]. This density estimate is formed by adding multivariate Gaussian kernels with a covariance matrix $\mathbf{H}$ centered at each observation $\mathbf{x}_i$.

A useful specification of the bandwidth matrix $\mathbf{H}$ is

$$\mathbf{H} = \lambda^2\mathbf{S} \tag{4}$$

Here $\mathbf{S}$ is the sample covariance matrix of the data, and λ^2 prescribes the bandwidth relative to this estimate of scale. These are parameters of the model that are estimated from the data. The procedure of scaling the bandwidth matrix proportional to the covariance matrix (equation (4)) is called "sphering" [*Fukunaga*, 1972] and ensures that all kernels are oriented along the estimated principal components of the covariance matrix.

The choice of the bandwidth, λ, is an important issue in

kernel density estimation. A small value of λ can result in a density estimate that appears "rough" and has a high variance. On the other hand, too high a bandwidth results in an "oversmoothed" density estimate with modes and asymmetries smoothed out. Such an estimate has low variance but is more biased with respect to the underlying density. This bias-variance trade-off [*Silverman*, 1986, section 3.3.1] plays an important role in choice of λ.

Several methods have been proposed to estimate the "optimal" bandwidth for a given data set. Least squares cross validation (LSCV) [*Silverman*, 1986, pp. 48–52] is one such method that is based on minimizing an estimate of the integrated square error of the kernel density estimate.

Sain et al. [1994] provide an expression for the LSCV score in any dimension with multivariate Gaussian kernel functions and **H**, a diagonal matrix. *Adamowski and Feluch* [1991] provide a similar expression for the bivariate case with Gaussian kernels. Here we generalize these results for use with the multivariate density estimator (3) which allows off diagonal terms in **H**:

$$\mathrm{LSCV}(\mathbf{H}) = \left\{ 1 + \frac{1}{n} \sum_{i=1}^{n} \sum_{j \neq i} \left[\exp\left(-L_{ij}/4\right) - 2^{d/2+1} \exp\left(-L_{ij}/2\right) \right] \right\} \Big/ \left\{ (2\sqrt{\pi})^d n \det(\mathbf{H})^{1/2} \right\} \quad (5)$$

where

$$L_{ij} = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{H}^{-1} (\mathbf{x}_i - \mathbf{x}_j) \quad (6)$$

We use numerical minimization of (5) over the single parameter λ with bandwidth matrix from (4) to estimate all the necessary probability density functions. We recognize that LSCV bandwidth estimation is occasionally degenerate, and so on the basis of suggestions by *Silverman* [1986, p. 52] and the upper bound given by *Scott* [1992, p. 181], we restrict our search to the range between 0.25 and 1.1 times the mean square error Gaussian reference bandwidth. This is the bandwidth that would be optimal if the data were from a Gaussian distribution.

3. Nonparametric Disaggregation Model, NPD

In this section a d-dimensional disaggregation model (denoted NPD) is developed. The model can be used to simulate d-dimensional disaggregate vectors $\mathbf{X}_t$ based on an input aggregate series Z_t. Z_t can be obtained from any suitable model for the aggregate streamflow series; however, we recommend a nonparametric model such as those described by *Sharma et al*. [1997] or *Lall and Sharma* [1996]. Since the same procedure is applied for each time step, from here on the subscript t on $\mathbf{X}_t$ is dropped to save notation.

Disaggregation is posed in terms of resampling from the conditional density function of (1). We need a model that given Z, provides realizations of $\mathbf{X}$. To use (1), an estimate of the $d + 1$ dimensional joint density function $f(X_1, X_2, \cdots X_d, Z)$ is required. However, because of summability, this has all its mass on the d-dimensional hyperplane defined by

$$X_1 + X_2 + \cdots + X_d = Z \quad (7)$$

This probability density can then be represented as

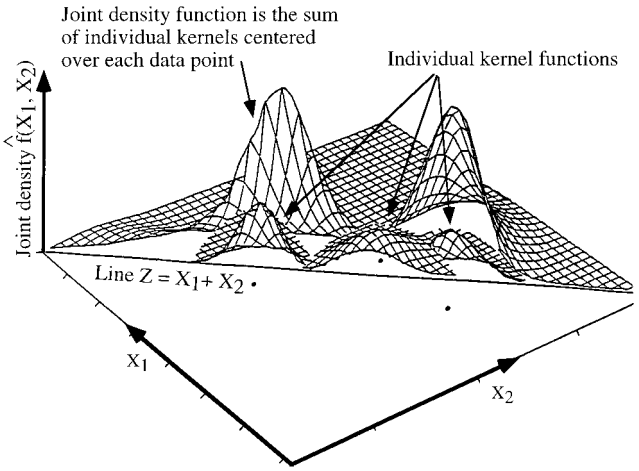

Figure 1. Illustration of a conditional density estimate $\hat{f}(X_1, X_2|Z)$ with $Z = X_1 + X_2$ as a slice through the joint density function. This illustration for clarity uses only three data points, shown as dots in the X_1, X_2 plane. Since the joint density estimate is formed by adding bivariate kernels, the conditional density is estimated as a sum of kernel slices.

$$f(X_1, X_2, \cdots X_d, Z) = f(X_1, X_2, \cdots X_d) \cdot \delta(Z - X_1 - X_2 \cdots - X_d) \quad (8)$$

where $\delta(\ )$ is the dirac delta function. The dirac delta function is a density function that integrates to one with all its mass concentrated at the origin. Kernel density estimation is used to estimate $f(X_1, X_2, \cdots X_d)$ based on the data. The conditional density function is then

$$f(X_1, X_2, \cdots, X_d|Z) = \frac{\delta(Z - X_1 - X_2 - \cdots - X_d)\, f(X_1, X_2, \cdots, X_d)}{\displaystyle\int_{\text{over plane } X_1+X_2+\cdots+X_d=Z} f(X_1, X_2, \cdots, X_d)\, dA} \quad (9)$$

For a particular Z this conditional density function can be visualized geometrically as the probability density on a $d - 1$ dimensional hyperplane slice through the d-dimensional density $f(X_1, X_2, \cdots, X_d)$, the hyperplane being defined by $X_1 + X_2 + \cdots + X_d = Z$. This is illustrated in Figure 1 for $d = 2$. There are really only $d - 1$ degrees of freedom in the conditional simulation. The conditional probability density function (pdf) in (9) can then be specified through a coordinate rotation of the vector $\mathbf{X} = (X_1, X_2, \cdots X_d)^T$ into a new vector $\mathbf{Y} = (Y_1, Y_2, \cdots Y_d)^T$ whose last coordinate is aligned perpendicular to the hyperplane defined by (7). Gram Schmidt orthonormalization [e.g., *Lang*, 1970, p. 138] is used to determine this rotation.

The appendix gives the derivation of the rotation matrix **R** such that

$$\mathbf{Y} = \mathbf{R}\mathbf{X} \quad (10)$$

R has the property that $\mathbf{R}^T = \mathbf{R}^{-1}$ (see appendix). With this rotation the last coordinate of **Y**, Y_d, is in fact a rescaling of Z, denoted Z'.

$$Y_d = Z/\sqrt{d} = Z' \quad (11)$$

We also denote the first $d - 1$ components of **Y** as $\mathbf{U}^T = (Y_1,$

$Y_2, \cdots Y_{d-1})^T$. These reflect the true $d-1$ degrees of freedom in the conditional simulation. With this $\mathbf{Y} = (\mathbf{U}^T, Z')^T$. Now we actually resample from $f(\mathbf{U}|Z') = f(Y_1, Y_2, \cdots Y_{d-1}|Z')$ and recover the disaggregate components of $\mathbf{X}$ by back rotation. The kernel density estimate $f(\mathbf{U}|Z')$ is obtained by applying (3) in rotated coordinates. Substituting $\mathbf{X} = \mathbf{R}^T\mathbf{Y}$ into (3) with bandwidth matrix $\mathbf{H}$ from (4), one obtains

$$\hat{f}(\mathbf{Y}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(2\pi)^{d/2}\lambda^d \det(\mathbf{S})^{1/2}} \cdot \exp\left(-\frac{(\mathbf{Y}-\mathbf{y}_i)^T\mathbf{R}\mathbf{S}^{-1}\mathbf{R}^T(\mathbf{Y}-\mathbf{y}_i)}{2\lambda^2}\right) \quad (12)$$

Now recognize that $\mathbf{R}\mathbf{S}^{-1}\mathbf{R}^T = (\mathbf{R}\mathbf{S}\mathbf{R}^T)^{-1} = \mathbf{S}_y^{-1}$ represents a rotation of the covariance matrix $\mathbf{S}$ into $\mathbf{S}_y$. Also $\det(\mathbf{S}_y) = \det(\mathbf{S})$. The resulting density estimate is therefore the same no matter whether the original or rotated coordinates are used. The conditional density function we resample from is

$$\hat{f}(\mathbf{U}|Z') = \hat{f}(Y_1, Y_2, \cdots Y_{d-1}|Z') = \frac{\hat{f}(\mathbf{U}, Z')}{\int \hat{f}(\mathbf{U}, Z')\, d\mathbf{U}} \quad (13)$$

where $\hat{f}(\mathbf{U}, Z') = \hat{f}(\mathbf{Y})$ is obtained from (12). Recalling that $\mathbf{U}$ denotes $(Y_1, Y_2, \cdots Y_{d-1})^T$, the vector $\mathbf{Y}$ without the last component, the covariance matrix $\mathbf{S}_y$ is partitioned as follows:

$$\mathbf{S}_y = \begin{bmatrix} \mathbf{S}_u & \mathbf{S}_{uz} \\ \mathbf{S}_{uz}^T & S_z \end{bmatrix} \quad (14)$$

$\mathbf{S}_u$ is the $d-1 \times d-1$ covariance matrix of $\mathbf{U}$. S_z is the 1×1 variance of Z', and $\mathbf{S}_{uz}$ is a vector of cross covariance between each component of $\mathbf{U}$ and Z'. Substituting (12) in (13) we obtain

$$\hat{f}(\mathbf{U}|Z') = \frac{1}{(2\pi\lambda^2)^{(d-1)/2}\det(\mathbf{S}')^{-1/2}}\sum_{i=1}^{n} w_i \cdot \exp\left(\frac{(\mathbf{U}-\mathbf{b}_i)^T\mathbf{S}'^{-1}(\mathbf{U}-\mathbf{b}_i)}{2\lambda^2}\right) \quad (15)$$

where

$$w_i = \exp\left(-\frac{(Z'-z_i')^2}{2\lambda^2 S_z}\right) \Big/ \sum_{j=1}^{n} \exp\left(-\frac{(Z'-z_j')^2}{2\lambda^2 S_z}\right) \quad (16)$$

$$\mathbf{S}' = \mathbf{S}_u - \mathbf{S}_{uz}S_z^{-1}\mathbf{S}_{uz}^T \quad (17)$$

$$\mathbf{b}_i = \mathbf{u}_i + \mathbf{S}_{uz}S_z^{-1}(Z'-z_i') \quad (18)$$

The conditional density function $\hat{f}(\mathbf{U}|Z')$ can therefore be seen as a weighted sum of n Gaussian density functions each with mean $\mathbf{b}_i$ and covariance $\lambda^2\mathbf{S}'$. Equation (16) shows that the weight w_i which controls the contribution of point i to the conditional density estimate depends on the distance of z_i' from the conditioning value Z'. Observations that lie closer to the conditioning value (i.e., where $(Z'-z_i')$ is small) receive greater weight. The weights are normalized to add to unity.

Resampling from (15) proceeds as follows.

Preprocessing:

1. Compute the sample covariance matrix $\mathbf{S}$ from the data $\mathbf{x}_i$.
2. Solve for λ by numerically minimizing (5) with $\mathbf{H}$ from (4), using 0.25 and 1.1 times λ_{ref}:

$$\lambda_{\text{ref}} = \left(\frac{4}{d+2}\right)^{1/(d+4)} n^{-1/(d+4)} \quad (19)$$

which is the mean square error Gaussian reference bandwidth, to bracket the search.

3. Compute $\mathbf{R}$, $\mathbf{S}_z$, and $\mathbf{S}'$ from (A2), (14), and (17).
4. Use singular value decomposition to obtain $\mathbf{B}$ such that $\mathbf{B}\mathbf{B}^T = \mathbf{S}'$.

At each time step:

5. Given Z from the aggregate model at each time step, first calculate the weight w_i associated with each observation, using (16).
6. Pick a point i with probability w_i.
7. Generate a $d-1$ dimensional unit Gaussian vector $\mathbf{V}$. Each component in $\mathbf{V}$ is independent $N(0, 1)$.
8. The simulated $\mathbf{U}$ is obtained from $\mathbf{U} = \mathbf{b}_i + \lambda\mathbf{B}\mathbf{V}$.
9. Augment this to obtain $\mathbf{Y}$, $\mathbf{Y} = (\mathbf{U}^T, Z')^T$.
10. Rotate back to the original coordinate space. $\mathbf{X} = \mathbf{R}^T\mathbf{Y}$.

Steps 5–10 are repeated for each aggregate time step. A complication can arise because the Gaussian kernels used in the kernel density estimate have infinite support. Thus they assign some (hopefully small) probability to regions of the domain where streamflow is negative (i.e., invalid or out of bounds). This leakage of probability across boundaries is a problem associated with kernel density estimates based on kernels with infinite support. Kernel density estimates also suffer from problems of bias near the boundaries. Here we address the leakage by checking the flows for validity (positiveness) and if they are invalid repeat steps 7–10 for a given time step. That is, we regenerate a new vector $\mathbf{V}$ and try again. This amounts to cutting the portion of each kernel that is out of bounds and renormalizing that kernel to have the appropriate mass over the within-bounds domain. We record how often this is done, as frequent boundary normalization is symptomatic of substantial boundary leakage. Alternative approaches that use special boundary kernels [*Hall and Wehrly*, 1991; *Wand et al*., 1991; *Djojosugito and Speckman*, 1992; *Jones*, 1993] or work with log-transformed data could be used in cases where this method for handling the boundaries is found to be unsatisfactory.

4. Model Evaluation

This section explores the use and effectiveness of the NPD approach. It is first applied to data from a specified bimodal distribution. This tests the model's ability to maintain distributional characteristics such as nonlinearity and bimodality. It is then applied to simulate monthly streamflow in the San Juan River.

To provide a point of reference, we also generate results using SPIGOT [*Grygier and Stedinger*, 1988, 1990]. SPIGOT is a parametric synthetic streamflow generation package that includes an annual streamflow generation module and, for annual to monthly disaggregation, the condensed model described by *Grygier and Stedinger* [1988, 1990]. SPIGOT's autoregressive model of order 1 (AR1) was used to generate the annual streamflow. SPIGOT first transforms the historical annual and monthly (or seasonal) flows to Gaussian using four choices for the marginal probability densities. These are (1) Gaussian, (2) two-parameter lognormal, (3) three-parameter lognormal, and (4) an approximate three-parameter gamma using the Wilson-Hilferty transformation [*Loucks et al*., 1981,

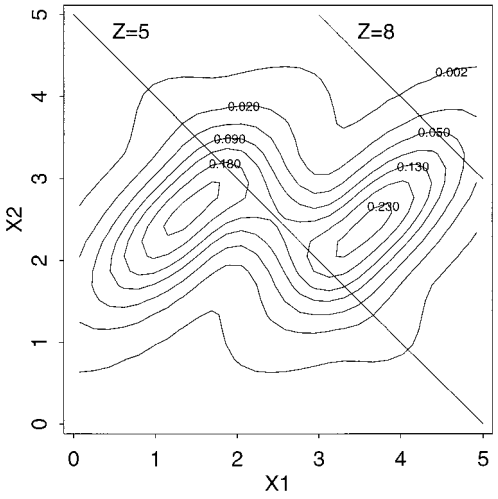

Figure 2. Bivariate distribution used in the synthetic example to test the disaggregation approach. This is a mixture of the three bivariate Gaussian density functions described in Table 1.

p. 286]. The parameters for each distribution are estimated by matching moments and the best-fitting distribution chosen by measuring the correlation of observations to the fitted distribution quantiles (Filliben's correlation statistic [*Grygier and Stedinger*, 1990]).

The next subsection describes the tests for the synthetic data from a specified distribution. This is followed by the San Juan River application.

4.1. Test With Synthetic Data

Here we describe a Monte Carlo investigation to test the ability of the NPD approach to approximate a specified underlying distribution. Our test distribution, illustrated in Figure 2, is based on distribution *J* of *Wand and Jones* [1993]. It consists of a mixture of three bivariate Gaussians having different weights α_i, stated as

$$f = \sum_{i=1}^{3} \alpha_i N(\mu_i, \Sigma_i) \tag{20}$$

where $N(\mu_i, \Sigma_i)$ denotes a Gaussian distribution with mean μ_i and a covariance matrix Σ_i. Individual weights, means, and covariances are shown in Table 1. Simulation from this mixed distribution is achieved by picking one of the three Gaussian distributions with probability α_i, then simulating a value from that distribution.

We simulated 101 bivariate samples, each consisting of 80 data pairs from this distribution. One sample is designated as

Table 1. Parameters of the Test Distribution

Gaussian Density	α_i	μ_i	Σ_i
1	0.4	(1.3, 2.5)	$\begin{pmatrix} 0.36 & 0.252 \\ 0.252 & 0.36 \end{pmatrix}$
2	0.4	(3.7, 2.5)	$\begin{pmatrix} 0.36 & 0.252 \\ 0.252 & 0.36 \end{pmatrix}$
3	0.2	(2.5, 2.5)	$\begin{pmatrix} 0.36 & -0.252 \\ -0.252 & 0.36 \end{pmatrix}$

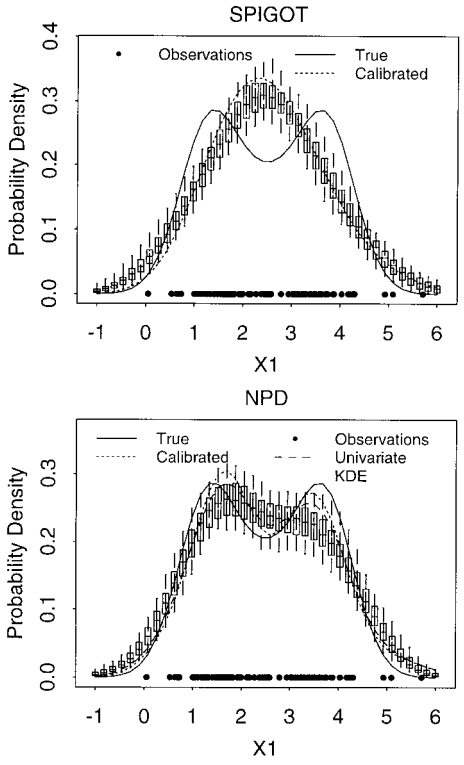

Figure 3. Marginal distribution of synthetic distribution variable X_1 for the calibration and disaggregation samples from SPIGOT and NPD. The true marginal density is obtained by integrating the pdf in Figure 2. The calibration pdf is estimated by integrating the sample joint density of variables X_1 and X_2 (a parametric distribution in case of SPIGOT and a kernel density estimate in case of NPD). The boxes show the ranges of the univariate kernel density estimates applied to the 100 disaggregation samples with a common bandwidth chosen as the median amongst set of optimal LSCV bandwidths for each sample. The univariate KDE in the NPD case is a univariate density estimate based on the calibration data with the same bandwidth as for the box plots. The dots above the *x* axis represent the calibration sample data points.

the "calibration" sample and is used to calibrate the NPD and SPIGOT models. In the case of NPD this involves estimating the sample covariance and bandwidth parameter λ (based on minimizing the LSCV score as described in previous section). Calibration of SPIGOT involves selection of the best marginal density transformation based on Filliben's correlation statistic and estimation of the coefficients in the condensed disaggregation model. The remaining 100 samples are used to form 100 aggregate test realizations by adding the components $Z = X_1 + X_2$. These 100 aggregate test realizations are input to both NPD and SPIGOT to generate 100 disaggregate realizations from both models. These disaggregate series are designated "test" samples and serve as a basis to test how closely the model reproduces statistics of the specified true distribution and of the calibration sample.

SPIGOT was modified to accept the same aggregate flows as the NPD model. Boundary corrections (discussed in the previous section for the NPD approach and specified as an option in the SPIGOT software) were not imposed on either model.

To evaluate the reproduction of marginal distributions by each model, we applied a univariate kernel density estimate to each of the 100 disaggregated samples. Figure 3 illustrates

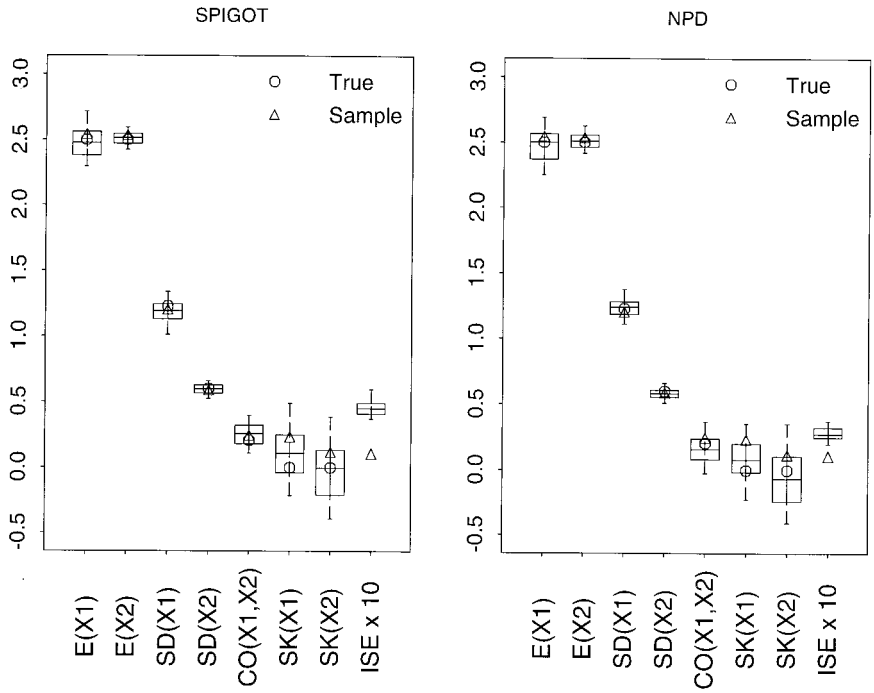

Figure 4. Comparison of statistics of SPIGOT and NPD for the synthetic example. The boxes show the ranges from the regenerated samples. Also shown are the true statistic (based on Figure 2) and the calibration sample statistic. The notation is $E(\ )$, expected value or mean of each variable, X_1 and X_2, respectively; SD( ), standard deviation; CO(X_1, X_2), correlation between variables X_1 and X_2; SK( ), skewness coefficient; and ISE, integrated square error difference from the true distribution, $\int (f(\mathbf{x}) - f(\mathbf{x}))^2 \, d\mathbf{x}$; $\hat{f}(\mathbf{x})$ is obtained for each regenerated sample by applying (3) with bandwidth by minimization of (5).

marginal densities of the calibration and disaggregated samples for variable X_1. Disaggregated sample pdf's are represented using box plots, which consist of a box that extends over the interquartile range of the quantity (in this case the pdf) being plotted. The line in the center of this box is the median, and "whiskers" extend to the 5% and 95% quantiles of the compared statistic. The span of the boxes and whiskers reflect sampling variability. Hence a measure of performance at the 90% level is whether the true statistic falls within the range of the whiskers. In Figure 3 the marginal densities of both the calibration sample and the SPIGOT or NPD disaggregations were estimated using a single common bandwidth taken as the median among the set of optimal bandwidths for the historical sample and the NPD and SPIGOT realizations. This was done to make the plots comparable and free of differences due to different bandwidths. The univariate kernel density estimate (KDE) curve (see NPD results in Figure 3) is also estimated using this median bandwidth. The curve marked "calibrated" represents the marginal density that is theoretically reproduced in simulations from either approach. This is estimated from the joint density of the calibration sample and is a univariate parametric pdf (depending on the transformation used) in the case of SPIGOT results and a numerically evaluated integral of the joint density of X_1 and X_2 (with respect to X_2 for marginal density of X_1) in the case of NPD results. One must note that while disaggregation model marginal densities will always be similar to the calibrated marginals, they are supposed to resemble the true curves instead. For the NPD results in Figure 3, the true, calibrated, and univariate KDE curves all show the same structure as the disaggregations. This is in contrast to SPIGOT disaggregations, where imposition of a three-parameter lognormal distribution on variable X_1 results in realizations with marginal density that bears little resemblance to the sample density estimate or underlying true pdf. The marginal distribution of variable X_2, not shown for brevity, was well reproduced by both SPIGOT and NPD because it is not bimodal and closely resembles a normal distribution.

Figure 4 illustrates statistics for realizations from both approaches. Both models reproduce the moment statistics well. The poor performance of SPIGOT on the marginal density of variable X_1 (Figure 3) only shows up in the comparison of integrated square error which is larger for the realizations generated using SPIGOT than NPD.

The above tests showed that the nonparametric approach is able to model properties of the joint distribution of X_1 and X_2 estimated on the basis of a single sample. We also tested the ability of the nonparametric approach to reproduce the underlying distribution in Figure 2. We rotated the samples in the above test such that each of the 101 samples was used once for calibration. Since disaggregation actually involves resampling from the conditional distribution (1), we used these samples to evaluate how well the kernel density estimation procedure worked for estimating the conditional distribution. This was done for two conditioning values (slices through the joint density function), namely, $Z = 5$, near the center of the distribution, and $Z = 8$, towards the upper tail. These conditioning lines are also shown on Figure 2. Figure 5 shows the conditional density estimates of $f(U|Z)$. Here in two dimensions the rotation (10) results in the independent variable $U = X_2 - X_1$. The boxes in Figure 5 depict the variability with sample size 80 from the 101 NPD samples. Note that these cover the true conditional distribution. The dashed lines give the average

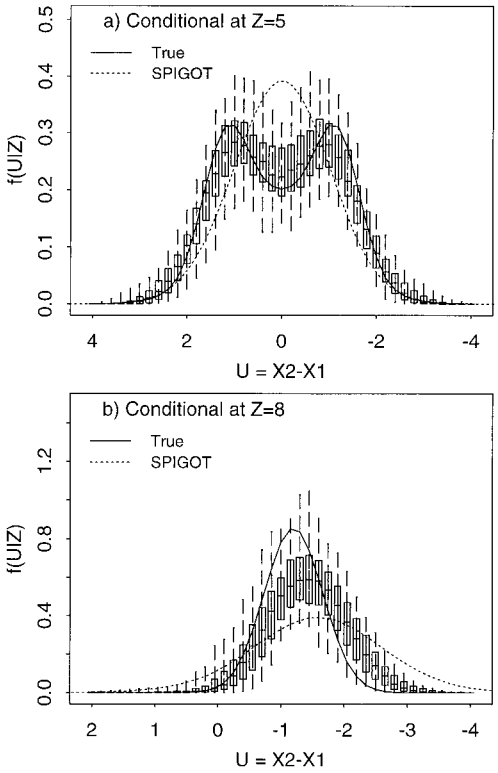

Figure 5. Conditional density estimates for the synthetic example.

of the conditional density resulting from fitting SPIGOT to each sample. The spread of the individual SPIGOT conditional density estimates (not shown) is similar to that in Figure 3. The SPIGOT densities are unable to reproduce the bimodality near the mode present in Figure 5a. In the tail (conditional at $Z = 8$; Figure 5b) the SPIGOT conditional density is more smoothed than the NPD conditional densities which are closer to the true conditional density.

Figure 6 gives the NPD joint density estimate relative average bias (RAB) and root mean square error (RMSE), normalized by the true density,

$$\text{RMSE} = \frac{\left(\frac{1}{n}\sum\left(f(\mathbf{x}) - \hat{f}(\mathbf{x})\right)^2\right)^{1/2}}{f(\mathbf{x})} \tag{21}$$

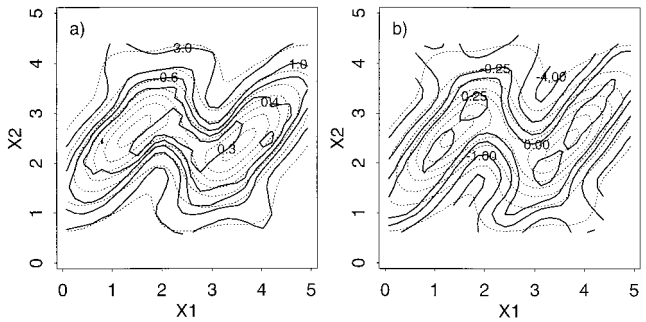

Figure 6. Nonparametric model calibration errors in the evaluation of $f(\mathbf{x})$, $\mathbf{x} = (x_1, x_2)$ for the synthetic example. (a) Relative root mean square error (RMSE). (b) Relative average bias (RAB). The dashed lines show contours of the true density $f(\mathbf{x})$ as in Figure 2.

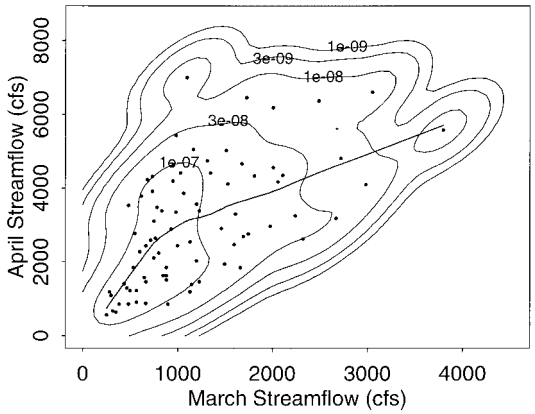

Figure 7. A bivariate kernel density estimate of March–April flows from the San Juan River. The thick line denotes the conditional mean of August flows conditional to July flows. Loess, a locally weighted regression smoother [*Cleveland and Devlin*, 1988], was used in our computations. Default parameter choices (number of iterations = 3; fraction of data used for smoothing at each point = 2/3) were used in the "loess" code in the statistical package S-plus.

$$\text{RAB} = \frac{\frac{1}{n}\sum\left(f(\mathbf{x}) - \hat{f}(\mathbf{x})\right)}{f(\mathbf{x})} \tag{22}$$

This figure indicates a bias towards underestimating the modes by up to 25% because of kernel smoothing. There is also large relative bias in the tails, because the true density used in normalization is so small. The relative root mean square error is around 30% near the modes and increases towards the tails as the normalizing density gets very small. By comparison the relative bias when fitting the SPIGOT distributions (not shown) to this data is 50% at the modes and −50% at the antimodes. The relative root mean square errors from fitting the SPIGOT distributions (not shown) are also larger (50% at the mode and 100% close to the saddle) because of the SPIGOT distributions inability to represent the bimodality. The overall mean integrated absolute error in fitting the NPD model is 0.42 as compared to 0.62 for SPIGOT. These values indicate the performance to be expected from the nonparametric estimate of $f(X_1, X_2)$ when calibrated against a sample of size 80.

4.2. Test With Monthly Flow Data

The application of the nonparametric disaggregation (NPD) model to 80 years (1906–1985) of monthly streamflow in the San Juan River near Archuleta, New Mexico, located at 36°48′05″ N and 107°41′51″ W at an elevation of 5655 feet (1724 m) is described in this section. This is station number AF3555 from the U.S. Bureau of Reclamation Colorado River Simulation System (CRSS) natural flow hydrologic database. This data set was one among many streamflow data sets we tested our model on, all with satisfactory results. This site was chosen because it illustrates well some of the features we wish to emphasize in this paper. Figure 7 shows a contour plot of the bivariate kernel density estimate for the March–April month pair flows. Note that the conditional expectation $E(X_t|X_{t-1})$, estimated using Loess [*Cleveland and Devlin*, 1988], appears to be nonlinear, with slopes different for flows less than and greater than approximately 900 feet^3/s (24,485 L/s). Such a

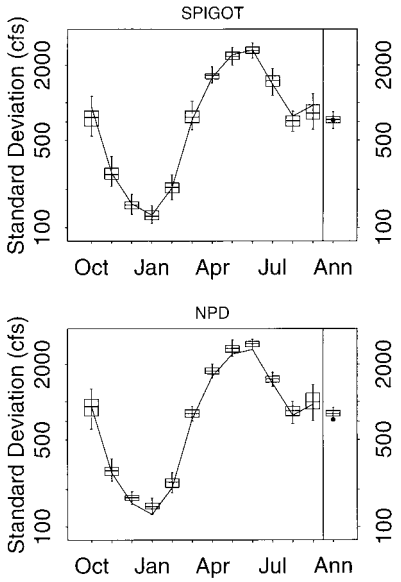

Figure 8. Simulated and observed streamflow standard deviations using SPIGOT and NPD. The line denotes the observed monthly standard deviations. The dot above "Ann" represents the standard deviation for the observed annual flows.

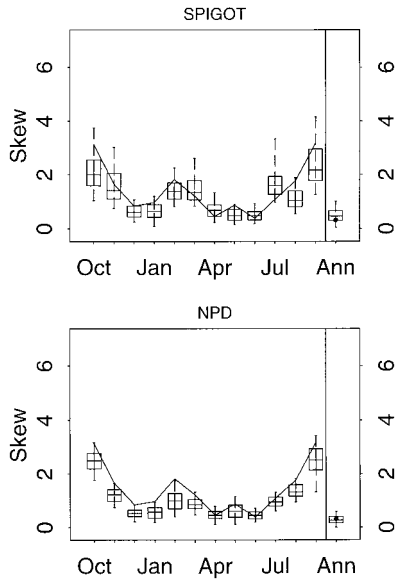

Figure 9. Simulated and observed streamflow skewness coefficients using SPIGOT and NPD. The line denotes the observed monthly skews. The dot above "Ann" represents the skew in the observed annual flows.

nonlinear conditional expectation is difficult to reproduce in simulations from a parametric model.

We compare results from application of the NPD model with those from SPIGOT. Aggregate flows for the NPD application were simulated using the NP1 model [*Sharma et al*., 1997]. The NP1 model is a nonparametric model constructed to preserve first-order Markov dependence in a time series. Flow values are obtained by sequentially resampling from a conditional density estimate $\hat{f}(Z_t|Z_{t-1})$ obtained using (1) and (3). Aggregate flows for the SPIGOT application were simulated using an autoregressive lag 1 (AR1) model. A marginal density transform based on the best Filliben's correlation statistic [*Grygier and Stedinger*, 1990] was used to transform historical annual flows to Gaussian. One hundred realizations, each of length 80 years, were generated from both approaches. Negative flows from NPD (amounting to about 0.1% of the total number of flows simulated) were resimulated using the boundary correction procedure described in the previous section. Both models were tested for their ability to reproduce the following statistics of the historic data: (1) mean; (2) standard deviation; (3) coefficient of skewness; (4) cross correlation between seasonal streamflows and between seasonal and annual streamflow; (5) kernel estimates of marginal distributions; and (6) "state dependent correlations," correlations between different month pairs as a function flow magnitude.

Results are shown in Figures 8–12. The simulation length of 80 years was chosen to be the same as the length of available historic data so that the variability of sample statistics across these realizations is representative of the sampling variability of the historic data. In these box plots the span of the boxes (interquartile range) and whiskers (5% to 95% quantiles) is a measure of the sampling variability associated with each statistic. If the historical statistic falls within the range of the boxes, then differences between model and data can be ascribed to sampling variability. If the historical statistic is outside the range of the boxes, then this indicates a quantity that the model does not reproduce.

The historical mean monthly streamflows for each month (not shown) were well reproduced by both SPIGOT and NPD models. Figures 8 and 9 use box plots to compare standard deviation and skewness of simulated and historical streamflows, respectively. Again both models reproduce these statistics well, though there is some small inflation in the standard deviations and deflation in the skewness of disaggregate flows from the nonparametric model due to the smoothing introduced by the kernel density estimate. On the other hand, SPIGOT tends to inflate the skewness in the months (for example, July) where the marginal density transform is inadequate.

Figure 10 compares the cross correlations of the monthly and aggregate flows from both models. The nonparametric model reproduces this statistic without bias while SPIGOT is unable to model the dependence between certain month pairs (for example, the simulated correlations between flows of month pairs 1–2, 1–3, and 6–11 are lower than the observed). This could be due either to some bias because of the marginal density transform or to the use of a condensed disaggregation model instead of a comprehensive model such as in (2).

In Figure 11 the marginal probability density estimates of the observed and simulated flows are compared. As in Figure 3 we used a common bandwidth (chosen as the median of a set estimated by minimizing LSCV for historical and simulated samples) to compute these univariate density estimates. The aggregate annual flows from AR1 and NP1 models that drive the SPIGOT and NPD models as well as monthly flows from April and June are compared. The dotted line in the case of SPIGOT flows represents the modeled pdf as suggested by the Filliben's correlation statistic. In comparing the annual flows both models perform reasonably well, although the nonparametric approach is arguably better at representing the flattish top of the distribution around the mode (flows between 1200 and 3500 feet^3/s (33,980 and 99,110 L/s)). The same can be said for the April marginal distributions. The June flows' marginal distribution has a peculiar looking upper tail which the best

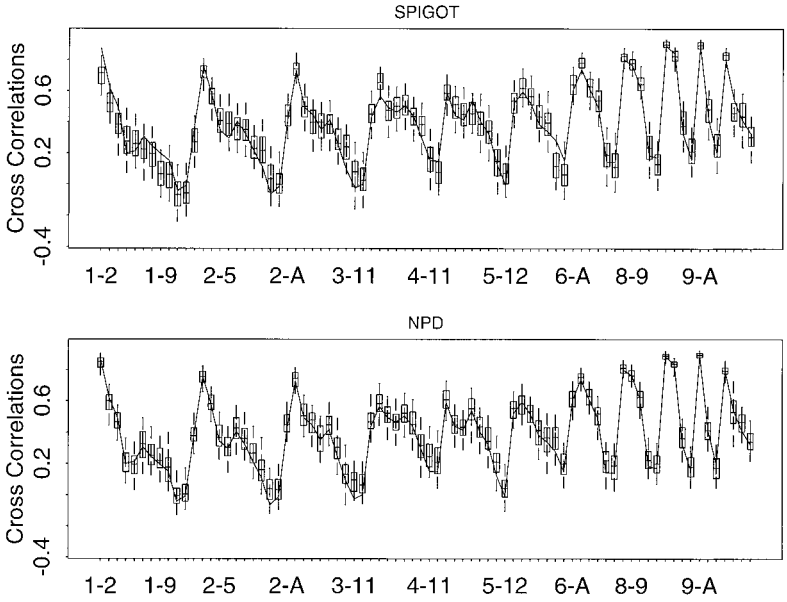

Figure 10. Simulated and observed cross correlation pairs using SPIGOT and NPD. The sequence along the x axis is 1–2, 1–3, $\cdots$, 1–12, 1–A, 2–3, 2–4 $\cdots$, 2–12, 2–A, 3–4, and so on. (1, 2) indicates cross correlation between months 1 and 2, (1, A) indicates cross correlation between month 1 and annual aggregate. Months are numbered according to the water year (1 = October, 2 = November, 4 = January, and so on).

SPIGOT marginal density transformation can only partially represent. The NPD model shows some bias for the lower flows but is able to model the distributional structure for the high flows better than SPIGOT.

It is worth emphasizing here that the NPD monthly flows by construction add up to the simulated aggregate flow used as input. There is therefore no need for adjustments to fix this such as is necessary in SPIGOT [*Grygier and Stedinger*, 1988].

Recall that the March–April month pair flows (Figure 7) suggested dependence of correlation on the flow magnitude

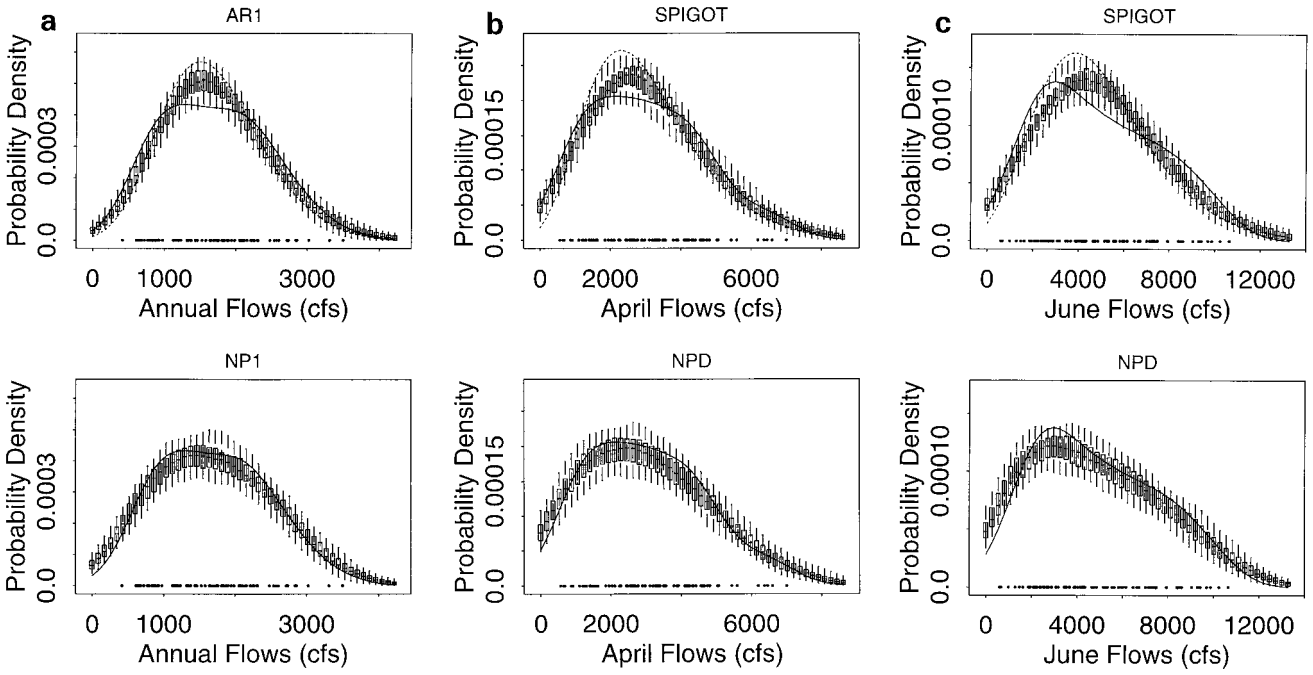

Figure 11. Simulated and observed marginal density estimates using the univariate kernel density estimator. The dotted line in the SPIGOT marginal densities represents the best fitting marginal density chosen using Filliben's correlation statistic. (a) Annual AR1 and NP1 marginal density estimates. A three-parameter lognormal distribution is used in the AR1 model fit. (b) April SPIGOT and NPD marginal density estimates. A three-parameter lognormal distribution is used in the SPIGOT fit for this month. (c) June SPIGOT and NPD marginal density estimates. A three-parameter lognormal distribution is used in the SPIGOT fit for this month.

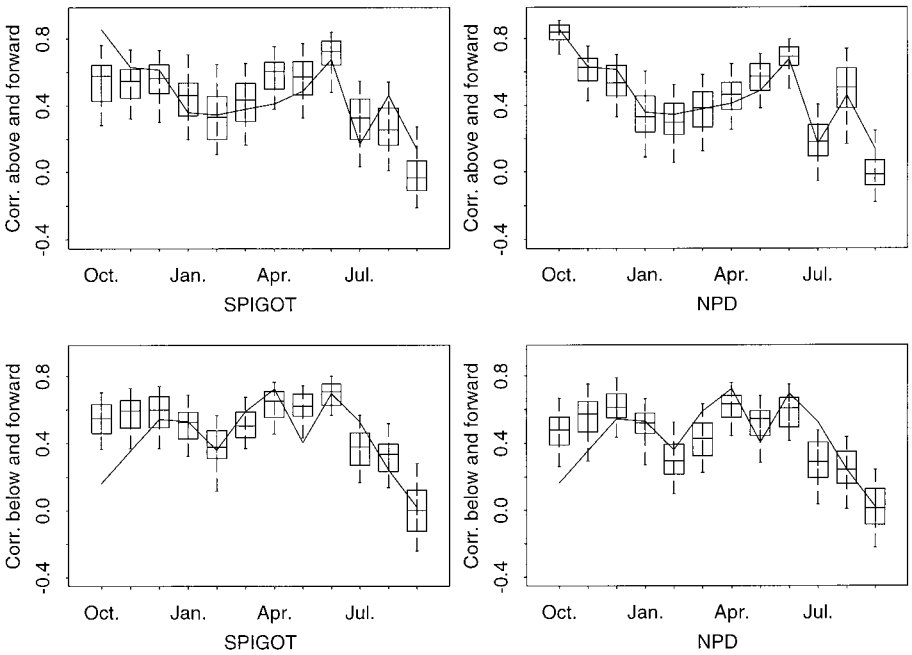

Figure 12. Simulated and observed state dependent correlations for sequential month pairs using SPIGOT and NPD.

(see difference in slopes of conditional mean for flows less than or greater than 900 feet^3/s (25,485 L/s) in Figure 7) that is not easily modeled by parametric models such as SPIGOT. In earlier work [*Sharma et al*., 1997] we used a statistic that quantified the dependence of correlation on the magnitude of flow. This statistic, denoted the state-dependent correlation statistic [*Sharma et al*., 1997, Appendix 1], measures the correlation between flows above or below the median in month i with succeeding flows in month j. For example, the "correlation above and forward" for the March–April month pair would be the correlation between March flows that are above the median March flow and their succeeding April flows. Differences between above median and below median correlations are indicative of nonlinear state dependence in the correlation structure.

Figure 12 shows the "above and forward" and "below and forward" state-dependent correlations for flows in adjacent months from both models. The state-dependent correlations for SPIGOT flows show more bias than those for corresponding NPD flows. For example, the "above and forward" correlations for October–November and the "below and forward" correlations for October–November, November–December, and May–June SPIGOT flows appear biased in contrast to the observed. Some bias is also visible in state-dependent correlations for NPD flows (particularly in the "below and forward" correlations for the October–November month pair). This bias is due to smoothing in the calibration pdf from the NPD model (as illustrated in the synthetic example, Figure 6). On the whole the nonparametric approach shows less bias than SPIGOT.

Stochastic streamflow sequences are frequently used to evaluate storage and water resources issues. Therefore it is necessary to ensure that simulated sequences are representative of the historic data with respect to these variables. Table 2 presents the bias and root mean square error (RMSE) of the reservoir storage capacity required to support yields of 50% and 90% of the mean annual streamflow. These storages were estimated using the Sequent Peak Algorithm [*Loucks et al*., 1981, p. 235] with equal monthly demands (1/12 of the fixed yield fraction) and the bias and RMSE evaluated as fractions of the storage estimated from the historical record

$$\text{bias}/S_h = \left(S_h - \frac{1}{n_r} \sum_{i=1}^{n_r} S_{s_i} \right) \Big/ S_h \tag{23}$$

$$\text{RMSE}/S_h = \frac{\left[\dfrac{1}{n_r} \displaystyle\sum_{i=1}^{n_r} (S_h - S_{s_i})^2 \right]^{1/2}}{S_h} \tag{24}$$

where S_h denotes the historical storage, S_{s_i} is the storage from the ith realization, and n_r is the total number of realizations. The larger yield fraction (90%) represents a higher level of development with the reservoir being required to provide longer term carry over storage. The nonparametric model has a smaller bias and RMSE than SPIGOT for the 90% yield case. The lower yield fraction (50%) represents a lower level of development with the reservoir storage required only for shorter term low flow months. In this case SPIGOT storages are closer to the historical storage. This is because the mass of the marginal density functions in the low flow tails below the

Table 2. Reservoir Capacity Statistics From 100 Realizations Each 80 Years Long of Monthly Streamflow in the San Juan River

Model	Bias/S_h	RMSE/S_h	Bias/S_h	RMSE/S_h
Yield fraction, %	50	50	90	90
SPIGOT	−0.192	0.457	0.412	0.457
NPD	−0.387	0.520	0.284	0.395

lowest observed flow, seen in Figure 11, is more for the nonparametric than SPIGOT distributions. We also tested the ability of the models to preserve long range dependence as quantified by the Hurst coefficient [*Hurst*, 1951] and the minimum average streamflow associated with different averaging durations. We found that both models reproduced these statistics adequately, so for brevity we have not presented the results.

5. Discussion and Conclusions

In this concluding section some conceptual and philosophic issues concerning the use of the method proposed here are discussed. Disaggregation is a method that has been well developed in the hydrology literature using a parametric modeling framework. A number of difficulties with these approaches pertaining to distribution choice and summability were noted earlier. A nonparametric approach to disaggregation of flows was provided as an alternative. Extensive statistical analyses of a synthetic situation and a streamflow record were used to argue that some technical difficulties associated with the traditional approach are avoided without compromising performance of the algorithm and while significantly increasing the generality of application.

Some questions of interest to practitioners are the following.

1. How does the NPD methodology provided here really differ from the existing traditional methods?
2. How does the method perform for extreme values, that is, in the tails of the density function?
3. What is the real advantage of this methodology for disaggregating hydrologic variables? When should it be used?
4. What are the relative data requirements of the proposed method?
5. What is the primary shortcoming of the proposed disaggregation method? When should it not be used?
6. What and how much testing and validation should precede the practical use of a stochastic model?

In the disaggregation context we are primarily interested in identifying reasonable proportions of the aggregate flow to "allocate" to each subset. The parametric methods approach this problem through a "global" prescription of the associated density function and correlation structure in a transformed data domain. The nonparametric methods approach this problem by looking at the relative proportions of the subset variables in a "local" sense; that is, the structure for wet years need not be the same as the structure for dry years. There is growing evidence [*Kahya and Dracup*, 1993; *Dettinger and Cayan*, 1995; *Rajagopalan and Lall*, 1995; *Mann and Park*, 1996] that the seasonality of precipitation, temperature, and streamflow advances/retards systematically over the year in response to low-frequency quasi-oscillatory phenomena. Often these variations are linked to a wet or dry year. The nonparametric approach is better suited to capturing such variations that may lead to heterogeneous density functions. The NPD approach provides for an exploratory analysis of such features through a focus on the visualization of empirical density functions. By contrast a function of best fit and limited flexibility is superposed on the data in the parametric approach and beyond choosing such a function there is little exploration of the data itself. The disaggregated streamflows are used for operating reservoir systems at intra-annual timescales. Exploration of the data to assess whether different regimes and hence distinct possibilities for seasonality phase and amplitude relative to interannual scales exist are more readily accomplished in the nonparametric framework.

Performance of the method with extreme values is an interesting issue. As presented here, the method uses a global bandwidth. This is a search radius that essentially determines how many points are used to determine the nature of the proportions to be ascribed to the subsets during disaggregation. Near the mode of the density this may be a large fraction of points. The resulting disaggregation proportions however need not be at all similar to those from the best fit parametric model. This was illustrated through the synthetic example in this paper, where the NPD method was considerably better. In the tails, given that the number of points available is smaller, the disaggregation proportions will be similar to those for these extreme points. We argue that in the disaggregation context, this may not be any worse than what happens with a parametric method for the following reasons. First, the parametric methods are usually fit such that the data near the modes dominates the fit, and hence the tails can be viewed as an extrapolation of that behavior. Second, if we recognize that the extremes relate to large-scale, low-frequency climate anomalies, such as the occurrence of El Niño events, then the years that correspond to these extremes will also have a distinct seasonal signature that the limited number of points in the tails of the nonparametric density will exploit. Variable or locally adaptive bandwidths as used by the k nearest neighbor method [*Lall and Sharma*, 1996] may further improve the performance of the NPD method in this regard. The tail issue may be more serious for the generation of the annual flow sequences, since there is a rather limited extrapolation of the data beyond the historical extremes.

Summarizing the discussion above, the real advantage of the NPD method is the ability to adaptively model complex relationships between aggregate and disaggregate flows. The state-dependent nature of these relationships and the ability to model them was demonstrated through the use of state-dependent correlations in this paper. A generalization of these measures to local dependence measures [*Jones*, 1996] will likely reinforce this point. Practically, these abilities should lead to the generation of streamflow sequences that better represent seasonality in wet and dry years across the set of sites of interest.

The examples provided in this paper used 80 years of data. This is rather inadequate for accurate estimates of a 13-dimensional density function in terms of statistical efficiency criteria. However, *Scott* [1992] illustrates that even with such small samples, kernel density estimators are able to distinguish modes in the density function and to show its general shape. In the disaggregation context the issue is how the subsets add up to the whole. The mathematics introduced in this paper and the accompanying discussion illustrate that the generated proportions are perturbations of the historical proportions of subperiod flows for a neighborhood of a particular state of annual flow. In this sense the important thing to do is to identify the proper neighborhood to perturb. This is governed by the general shape of the density function in that neighborhood. In this context the sample size requirements may not be as severe as for a precise estimation of the local density. Given the rules we used for the range over which the bandwidth is selected, both severe undersmoothing and oversmoothing of the density function are avoided. The relative degree of smoothing of the density function increases with decreasing sample size to control the variance of the density estimate. Thus for small sample

sizes, the performance of the method may not be too different from that of a parametric model which also smooths the data heavily. We do not expect the NPD approach applied to disaggregate monthly data to be necessarily superior to the parametric models for less than 30–40 years of data. Interestingly, using a different formulation of the kernel density estimator for disaggregation, *Aitchison and Lauder* [1985] (see also *Lall et al*. [1996]) claim superior performance for nonparametric disaggregation relative to parametric models using as few as 20 data points.

The primary shortcoming of the NPD approach is that it is data and computationally intensive. Estimating an optimal bandwidth to use is a computationally demanding task. As with the method of moments and the method of maximum likelihood in the parametric case, different optimality criteria can lead to quite different bandwidths being selected. The choice of the kernel function is not critical, but the parameterization of the bandwidth matrix in the multivariate case may affect the results dramatically. As an example, the use of local rather than global covariance matrices **H** in our scheme will change the scheme and its performance. The sample size required increases as the complexity of the underlying density function increases, thus reducing the advantage of the NPD approach for heterogeneous density functions. However, the approximations of the underlying density for disaggregation purposes may still be good. Another shortcoming of the NPD approach is that no simple equation for the model is available to report. The user needs the historical data set, the kernel representation, and the optimal bandwidth to perform a new analysis. However, given modern computer capabilities, these shortcomings are not critical.

Finally, we feel that it is no longer (and never really was) sufficient to accept a model based on only reproduction of limited moments. Here the disaggregation model was evaluated against a broad range of statistics including the mean, standard deviation, skewness, marginal density function, cross correlations, state dependent correlations, and storage statistics. It was also tested against other statistics, which are not shown for brevity. We feel that this testing against many quantities using graphical devices such as box plots is an important aspect of model validation in stochastic hydrology that we are trying to promote by example in this paper.

In conclusion, we are convinced that nonparametric techniques such as the NPD approach presented here have an important role to play in improving the synthesis of hydrologic time series for water resources planning and management. They can capture the dependence structure present in the historic data without imposing arbitrary linearity or distributional assumptions. They have the capability to reproduce nonlinearity, state dependence, and multimodality while remaining faithful to the historic data and producing synthesized sequences statistically indistinguishable from the historic sequence. Potential applications of the NPD approach extend beyond streamflow to spatial disaggregation or downscaling of climate fields and multiscale representation of hydroclimatic data in a manner that is understood in fundamental probabilistic terms.

Appendix: Derivation of Gram Schmidt Rotation Matrix

Gram Schmidt orthonormalization is a procedure for determining an orthonormal set of basis vectors for a vector space from any suitable basis. The standard basis (basis vectors aligned with the coordinate axes) is orthonormal but does not have a basis vector perpendicular to the conditioning plane defined by (7). We therefore drop one of the standard basis vectors and replace it by a vector perpendicular to the conditioning plane. The basis set is now not orthonormal. We then apply the Gram Schmidt procedure to obtain an orthonormal basis vector set that includes a vector perpendicular to the conditioning plane. The result is a rotation matrix **R** such that

$$\mathbf{Y} = \mathbf{R}\mathbf{X} \tag{A1}$$

where **R** has rows that consist of the basis vectors for the rotated coordinate space:

$$\mathbf{R} = \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \vdots \\ \mathbf{e}_d \end{pmatrix} \tag{A2}$$

Denote the standard basis as

$$\begin{aligned} \mathbf{i}_1 &= (1, 0, 0, \cdots 0)^T \\ \mathbf{i}_2 &= (0, 1, 0, \cdots 0)^T \\ &\vdots \\ \mathbf{i}_d &= (0, 0, \cdots 0, 1)^T \end{aligned} \tag{A3}$$

Define

$$\begin{aligned} \mathbf{e}_d &= (1/\sqrt{d}, 1/\sqrt{d}, \cdots 1/\sqrt{d})^T \\ &= 1/\sqrt{d}\,\mathbf{i}_1 + 1/\sqrt{d}\,\mathbf{i}_2 + \cdots + 1/\sqrt{d}\,\mathbf{i}_d \end{aligned} \tag{A4}$$

This is a unit vector perpendicular to the conditioning plane. Now apply Gram Schmidt orthonormalization to obtain an orthonormal basis including $\mathbf{e}_d$.

For j decreasing from $d - 1$ to 1,

$$\begin{aligned} \mathbf{e}'_j &= \mathbf{i}_j - \sum_{k=j+1}^{d} (\mathbf{e}_k \cdot \mathbf{i}_j)\mathbf{e}_k \\ \mathbf{e}_j &= \mathbf{e}'_j/|\mathbf{e}'_j| \end{aligned} \tag{A5}$$

The first step above obtains a vector orthogonal to the basis vectors $\mathbf{e}_k$, $k = j + 1 \cdots d$, obtained thus far and the second step normalizes it to unit length. Since **R** is defined with unit orthogonal basis vectors $\mathbf{R}\mathbf{R}^T = \mathbf{I}$ so $\mathbf{R}^{-1} = \mathbf{R}^T$.

Acknowledgments. This research was supported by the U.S. Geological Survey (USGS), Department of the Interior, under USGS award 1434-92-G-2265. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Government. We would like to thank anonymous *Water Resources Research* reviewers for their input, which resulted in improvements to this paper.

References

Adamowski, K., and W. Feluch, Application of nonparametric regression to groundwater level prediction, *Can. J. Civ. Eng*., *18*, 600–606, 1991.

Aitchison, J., and I. J. Lauder, Kernel density estimation for compositional data, *Appl. Stat*., *34*(2), 129–137, 1985.

Bras, R. L., and I. Rodriguez-Iturbe, *Random Functions and Hydrology*, 559 pp., Addison-Wesley, Reading, Mass., 1985.

Cleveland, W. S., and S. J. Devlin, Locally weighted regression: An approach to regression by local fitting, *J. Am. Stat. Assoc*., *83*(403), 596–610, 1988.

Curry, K., and R. L. Bras, Theory and applications of the multivariate broken line, disaggregation and monthly autoregressive streamflow generators to the Nile River, *Rep. 78-5*, Technol. Adaptation Prog., Mass. Inst. of Technol., Cambridge, 1978.

Dettinger, M. D., and D. R. Cayan, Large scale atmospheric forcing of recent trends towards early snowmelt runoff in California, *J. Clim*., *8*(3), 606–623, 1995.

Djojosugito, R. A., and P. L. Speckman, Boundary bias correction in nonparametric density estimation, *Comm. Stat. Theory Methods*, *21*(1), 69–88, 1992.

Fukunaga, K., *Introduction to Statistical Pattern Recognition*, Academic, San Diego, Calif., 1972.

Grygier, J. C., and J. R. Stedinger, Condensed disaggregation procedures and conservation corrections for stochastic hydrology, *Water Resour. Res*., *24*(10), 1574–1584, 1988.

Grygier, J. C., and J. R. Stedinger, Spigot, a synthetic streamflow generation package, technical description, version 2.5, School of Civ. and Environ. Eng., Cornell Univ., Ithaca, N. Y., 1990.

Hall, P., and T. E. Wehrly, A geometrical method for removing edge effects from kernel-type nonparametric regression estimators, *J. Am. Stat. Assoc*., *86*(415), 665–672, 1991.

Harms, A. A., and T. H. Campbell, An extension to the Thomas-Fiering model for the sequential generation of streamflow, *Water Resour. Res*., *3*(3), 653–661, 1967.

Hurst, H. E., Long-term storage capacity of reservoirs, *Trans. Am. Soc. Civ. Eng*., *116*, 770–799, 1951.

Jones, M. C., Simple boundary correction for kernel density estimation, *Stat. Comput*., *3*, 135–146, 1993.

Jones, M. C., The local dependence function, *Biometrika*, *83*(4), 899–904, 1996.

Kahya, E., and J. A. Dracup, U.S. streamflow patterns in relation to the El Nino/Southern Oscillation, *Water Resour. Res*., *29*(8), 2491–2500, 1993.

Lall, U., Recent advances in nonparametric function estimation, Supplement, *U.S. Natl. Rep. Int. Union Geod. Geophys. 1991–1994*, *Rev. Geophys*., *33*, 1093–1102, 1995.

Lall, U., and A. Sharma, A nearest neighbor bootstrap for time series resampling, *Water Resour. Res*., *32*(3), 679–693, 1996.

Lall, U., B. Rajagopalan, and D. G. Tarboton, A nonparametric wet/dry spell model for resampling daily precipitation, *Water Resour. Res*., *32*(9), 2803–2823, 1996.

Lane, W. L., Applied stochastic techniques, users manual, Eng. and Res. Cent., Bur. of Reclam., Denver, Colo., 1979.

Lang, S., *Linear Algebra*, 2nd ed., 400 pp., Addison-Wesley, Reading, Mass., 1970.

Loucks, D. P., J. R. Stedinger, and D. A. Haith, *Water Resource Systems Planning and Analysis*, 559 pp., Prentice-Hall, Englewood Cliffs, N. J., 1981.

Mann, M. E., and J. Park, Greenhouse warming and changes in the seasonal cycle of temperature: Model versus observations, *Geophys. Res. Lett*., *23*, 1111–1114, 1996.

Mejia, J. M., and J. Rousselle, Disaggregation models in hydrology revisited, *Water Resour. Res*., *12*(2), 185–186, 1976.

Rajagopalan, B., and U. Lall, Seasonality of precipitation along a meridian in the western U.S., *Geophys. Res. Lett*., *22*(9), 1081–1084, 1995.

Sain, S. R., K. A. Baggerly, and D. W. Scott, Cross-validation of multivariate densities, *J. Am. Stat. Assoc*., *89*(427), 807–817, 1994.

Salas, J. D., J. W. Delleur, V. Yevjevich, and W. L. Lane, *Applied Modeling of Hydrologic Time Series*, 484 pp., Water Resour., Littleton, Colo., 1980.

Santos, E. G., and J. D. Salas, Stepwise disaggregation scheme for synthetic hydrology, *J. Hydraul. Eng*., *118*(5), 765–784, 1992.

Schertzer, D., and S. Lovejoy (Eds.), *Non-linear Variability in Geophysics, Scaling and Fractals*, 318 pp., Kluwer, Norwell, Mass., 1991.

Scott, D. W., *Multivariate Density Estimation, Theory, Practice, and Visualization*, 317 pp., John Wiley, New York, 1992.

Sharma, A., Nonparametric approaches for simulation of streamflow sequences, Ph.D. thesis, Civ. and Environ. Eng., Utah State Univ., Logan, 1996.

Sharma, A., D. G. Tarboton, and U. Lall, Streamflow simulation: A nonparametric approach, *Water Resour. Res*., *33*(2), 291–308, 1997.

Silverman, B. W., *Density Estimation for Statistics and Data Analysis*, 175 pp., Chapman and Hall, New York, 1986.

Stedinger, J. R., and R. M. Vogel, Disaggregation procedures for generating serially correlated flow vectors, *Water Resour. Res*., *20*(11), 47–56, 1984.

Stedinger, J. R., D. Pei, and T. A. Cohn, A condensed disaggregation model for incorporating parameter uncertainty into monthly reservoir simulations, *Water Resour. Res*., *21*(5), 665–675, 1985.

Svanidze, G. G., *Mathematical Modeling of Hydrologic Series* (translated from Russian), Water Resour., Fort Collins, Colo., 1980.

Tong, H., *Nonlinear Time Series Analysis: A Dynamical Systems Perspective*, Academic, San Diego, Calif., 1990.

Valencia, D. R., and J. L. Schaake, A disaggregation model for time series analysis and synthesis, *Rep. 149*, Ralph M. Parsons Lab., Mass. Inst. of Technol., Cambridge, 1972.

Wand, M. P., and M. C. Jones, Comparison of smoothing parameterizations in bivariate kernel density estimation, *J. Am. Stat. Assoc*., *88*(422), 520–528, 1993.

Wand, M. P., and M. C. Jones, Multivariate plug-in bandwidth selection, *Comput. Stat*., *9*, 97–116, 1994.

Wand, M. P., J. S. Marron, and D. Ruppert, Transformations in density estimation, *J. Am. Stat. Assoc*., *86*(414), 343–361, 1991.

U. Lall and D. G. Tarboton, Utah Water Research Laboratory, Utah State University, UMC82, Logan, UT 84322-8200. (e-mail: ulall@kernal.uwrl.usu.edu)

A. Sharma, Department of Water Engineering, School of Civil Engineering, University of New South Wales, Sydney, NSW 2052, Australia.

(Received April 21, 1997; revised August 6, 1997;
accepted August 25, 1997.)