WATER RESOURCES RESEARCH, VOL. 33, NO. 2, PAGES 291–308, FEBRUARY 1997

Streamflow simulation: A nonparametric approach

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

Abstract. In this paper kernel estimates of the joint and conditional probability density functions are used to generate synthetic streamflow sequences. Streamflow is assumed to be a Markov process with time dependence characterized by a multivariate probability density function. Kernel methods are used to estimate this multivariate density function. Simulation proceeds by sequentially resampling from the conditional density function derived from the kernel estimate of the underlying multivariate probability density function. This is a nonparametric method for the synthesis of streamflow that is data-driven and avoids prior assumptions as to the form of dependence (e.g., linear or nonlinear) and the form of the probability density functions (e.g., Gaussian). We show, using synthetic examples with known underlying models, that the nonparametric method presented is more flexible than the conventional models used in stochastic hydrology and is capable of reproducing both linear and nonlinear dependence. The effectiveness of this model is illustrated through its application to simulation of monthly streamflow from the Beaver River in Utah.

1. Introduction

A goal of stochastic hydrology is to generate synthetic streamflow sequences that are statistically similar to observed streamflow sequences. Statistical similarity implies sequences that have statistics and dependence properties similar to those of the historical record. These sequences represent plausible future streamflow scenarios under the assumption that the future will be similar to the past. In this paper we present a nonparametric approach for the generation of synthetic streamflow sequences. This approach is appropriate for the simulation of stationary unregulated streamflow inputs that are needed in simulation studies to analyze alternative designs, operation policies, and rules for water resources systems. The utility of this approach relative to conventional parametric methods is demonstrated through applications to monthly streamflow from the Beaver River, near Beaver, Utah, and to samples generated from linear and nonlinear models with known statistical attributes.

Consider a time series $\{X_1, X_2, \cdots, X_t, \cdots\}$ where X_t represents streamflow quantities at time t. In practice, the dependence structure of streamflow sequences is often assumed to be Markovian, that is, dependent on only a finite set of prior values. With this assumption, *Bras and Rodroguez-Iturbe* [1985] note that stochastic streamflow models are an exercise in conditional probability. An order p model simulates X_t on the basis of the previous values, that is, $X_{t-1}, X_{t-2}, \cdots, X_{t-p}$. This requires that a $d = p + 1$ dimensional joint probability distribution be specified. Simulation can proceed from the conditional density function, defined as

$$f(X_t|X_{t-1}, X_{t-2}, \cdots, X_{t-p}) = \frac{f(X_t, X_{t-1}, X_{t-2}, \cdots, X_{t-p})}{\int f(X_t, X_{t-1}, X_{t-2}, \cdots, X_{t-p})\, dX_t} \quad (1)$$

Traditional parametric models specify (1) through assumed distributions. Here, it is suggested that streamflow may instead be directly modeled from empirical, data-driven estimates of the joint and conditional density functions given in (1). Nonparametric estimates of these density functions are developed directly from the historical data. A method is considered nonparametric if it can reproduce a broad class of possible underlying density functions [*Scott*, 1992, p. 44]. Nonparametric methods for density estimation strive to approximate the underlying density locally using data from a small neighborhood of the point of estimate [*Lall*, 1995]. They impose only weak assumptions, such as continuity of the target function, rather than a priori specification or choice of a particular parametric probability distribution (Gaussian, lognormal, etc.). A perusal of the statistical literature shows that nonparametric statistical estimation, using splines, kernel functions, nearest neighbor methods and orthogonal series methods, is an active area, with major developments still unfolding. *Silverman* [1986] and *Scott* [1992] provide good introductory texts. Applications of nonparametric methods in hydrology are reviewed by *Lall* [1995].

Our model is based on a nonparametric kernel density estimate of the $p + 1$ dimensional density function $f(X_t, X_{t-1}, \cdots, X_{t-p})$, which is then used in (1) to estimate the conditional density function that forms the basis for generation of synthetic streamflow series. This is called a nonparametric order p, or NP_p, model. It has the following advantages:

1. Statistical attributes of the data are automatically honored since one works with a smoothed empirical frequency distribution based directly on the historical data. Such attributes include nonlinear dependence and inhomogeneity (i.e., statistical properties that vary by streamflow state).
2. The somewhat tenuous issue of choosing between different models for the probability distribution is sidestepped.
3. Considerations related to the above two points lead to a procedure that is easy to use and is able to automatically model

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

Paper number 96WR02839.
0043-1397/97/96WR-02839$09.00

the distributional and dependence characteristics of the historical time series. Use of such a procedure should result in improved decisions for reservoir operation and design.

We shall first review some of the traditional approaches, noting their shortcomings and motivating the need for the nonparametric approach. Kernel density estimation is reviewed next. We then describe the NP_p model and illustrate its use with synthetic data from a linear autoregressive (AR1) model and a self-exciting threshold autoregressive (SETAR) model [*Tong*, 1990, section 3.3.1.1]. These tests demonstrate the effectiveness of the NP_p approach in representing both linear and nonlinear systems, without prior specification of the model equations. An application of our model to simulate monthly streamflow from the Beaver River, near Beaver, Utah, is then presented and results are compared to those from an AR1 model with marginal densities chosen from the best fitting of four commonly used probability density functions.

2. Background

Annual and monthly streamflow has been modeled extensively using autoregressive moving average (ARMA) type models [*Bras and Rodriguez-Iturbe*, 1985; *Salas et al*., 1980; *Pegram et al*., 1980; *Loucks et al*., 1981; *Stedinger et al*., 1985b; *Stedinger and Vogel*, 1984; *McLeod et al*., 1977; *Hipel et al*., 1977; *Yevjevich*, 1972]. The early Thomas-Fiering model [*Thomas and Fiering*, 1962; *Fiering*, 1967; *Beard*, 1967], an autoregressive lag 1 model with seasonally varying coefficients, is a good example of this approach.

$$(X_{t,j} - m_j) = \rho_j \frac{\sigma_j}{\sigma_{j-1}} (X_{t,j-1} - m_{j-1}) + \sigma_j (1 - \rho_j^2)^{1/2} W_{t,j} \qquad (2)$$

where $X_{t,j}$ is the seasonal streamflow at year t and season (month) j, ρ_j is the lag 1 correlation coefficient between seasons j and $j - 1$, m_j is the mean streamflow in season j, σ_j is the standard deviation of flow in season j, and $W_{t,j}$ is an independent random variable with mean 0 and variance 1. By allowing the noise term $W_{t,j}$ to be from a skewed distribution [*Lettenmaier and Burges*, 1977; *Todini*, 1980], streamflow from a skewed distribution can be approximated. Thus this model reproduces the mean, variance, and correlations between monthly streamflows and approximates the skewness. These are the variables traditionally considered as most important by stochastic hydrologists. As written, this model applies only to a single site; however, it is illustrative of a very general class of ARMA models for single sites and in a multivariate context for multiple sites or seasons that have been developed and applied extensively in hydrology over the years and described at length in texts on the subject [e.g., *Salas et al*., 1980; *Loucks et al*., 1981; *Bras and Rodriguez-Iturbe*, 1985].

Such models can be viewed as special cases of a general multivariate ARMA(p, q) model:

$$X_{t+1} = \sum_{j=0}^{p} A_j X_{t-j} + \sum_{j=0}^{q} B_j W_{t-j} + U \qquad (3)$$

where X_t is a vector of the variables of interest, including annual and seasonal flows at all sites; A_j and B_j are coefficient matrices; U is a vector of coefficients, and W_t is a vector of independent random innovations. The first term represents an autoregressive component, and the second term represents a moving average component. In all but the simplest univariate models it is impractical to assume anything but a Gaussian distribution for the W_t. This is equivalent to the assumption of a multivariate Gaussian distribution for the time series dependence structure. The ARMA model is then defined through the estimation of the parameters A_j, B_j, U and the model order (p, q). To account for the fact that the real streamflows are not Gaussian, the flows are often first transformed to a Gaussian distribution and then the transformed variables are used with (3) [*Stedinger*, 1981; *Stedinger and Taylor*, 1982; *Stedinger et al*., 1985a]. Reproducing moments in the original coordinates may then be difficult.

The general linear model depicted by (3) is a special case of the conditional density function of (1). This multivariate Gaussian structure with transformed marginal distributions (denoted MGTM here) has with few exceptions [*Yakowitz*, 1985; *Smith*, 1991, 1992; *Lall and Sharma*, 1996] underlain practically all stochastic hydrology to date. The Lall and Sharma work is very similar in spirit to this work, though its approach is that of a nearest neighbor bootstrap rather than kernel density estimation. We believe that both are good alternatives that need to be considered for streamflow simulation.

The preceding discussion reveals the basic structure of current time series estimation methods and hints at their restricted view of the possibilities of variation in hydrological time series. The main reasons for the prevalence of linear ARMA models for hydrologic time series analysis may be the following: (1) the framework has been well developed in the statistical literature for stationary processes; (2) the techniques are well understood and taught; and (3) software for multivariate analysis has been developed by a number of people, is readily available, and does not pose a severe computational burden [*Salas*, 1993].

Some drawbacks of the MGTM approach are the following:

1. Only a limited degree of heterogeneity in the statistical dependence structure is admitted through the normalizing transform. The dependence of variance of streamflow on streamflow magnitude is often noted. There is evidence in some streamflow data that correlations are different depending on whether flows are low or high. We give an example of this in section 6 using state-dependent correlation statistics defined in Appendix A. In an MGTM model the correlation structure is fixed regardless of flow magnitude. Among others, *Yevjevich* [1972] has argued for the systematic identification of nonstationarities in the mean of the time series (e.g., jumps, periodicities) and their removal to yield a stationary time series that can be analyzed by standard methods. However, such features may be part of the underlying dynamics and important to model behavior (e.g., to a drought regime) that may be related to threshold dependent processes.

2. The MGTM models impose a time reversible structure. The joint distributions of (X_t, X_{t+1}, $\cdots$, X_{t+m}) and of (X_t, X_{t-1}, $\cdots$, X_{t-m}) are identical. *Tong* [1990, p. 9] shows an example of daily streamflow that is not time reversible and argues that the dynamics of physical processes is time irreversible.

3. The choice of a distribution for W_t or of an appropriate transform can be problematic. For short series, statistical tests are unable to distinguish between candidate distributions [see *Kite*, 1977]. None of the common transformations may be applicable. Figure 1 illustrates this problem with July monthly streamflow from the Beaver River, near Beaver, Utah, located at 38°16′50″N and 112°34′25″W at an elevation of 6200 feet (1890 m) above mean sea level (U.S. Geological Survey station 10234500). *Slack and Landwehr* [1992] report this station as

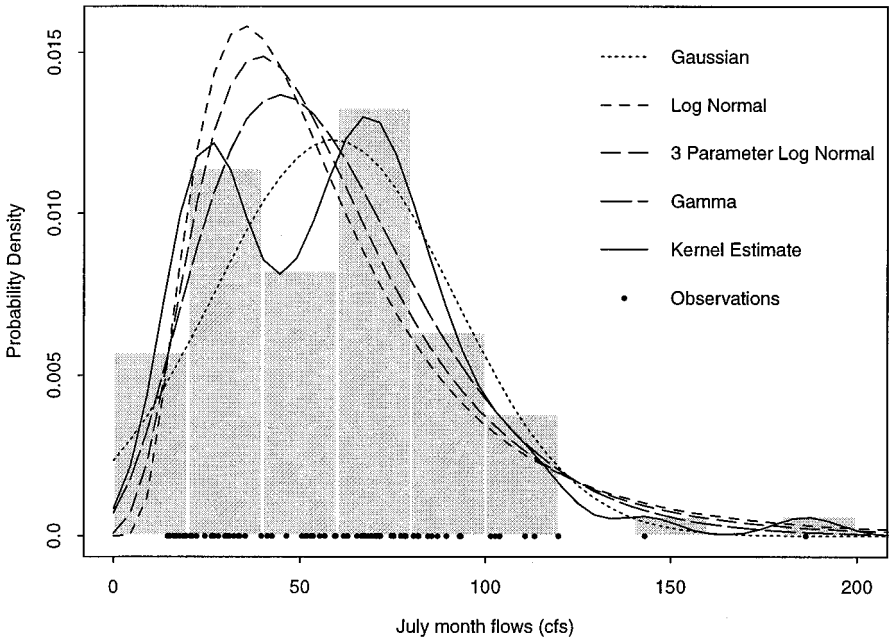


Figure 1. Histogram and probability density estimates of July monthly streamflow (in cubic feet per second; 1 foot^3 s^{-1} is equal to 28.317 L s^{-1}) in the Beaver River near Beaver, Utah. The dots on the x axis denote the individual data points.

unregulated and free from other anthropogenic effects. The figure shows the histogram of monthly flow, with four commonly used distributions fitted to the data. The histogram has bimodality that cannot be reproduced by any of the distributions commonly used. This figure also shows a kernel density estimate. Note that this is effectively a smoothing of the histogram. The following Filliben correlation statistics [*Grygier and Stedinger*, 1990] test the goodness of fit for each distribution in Figure 1: kernel density estimate, 0.998; normal, 0.963; lognormal, 0.979; three-parameter lognormal, 0.982; and gamma, 0.985. The Filliben correlation statistic is the correlation between empirical quantiles from a plotting position and fitted distribution quantiles corresponding to the data values. For a perfect fit the Filliben correlation statistic should be 1, and it is by construction a value that lies close to 1. The relative departure from 1 provides a measure of the relative goodness of fit between the different distributions. By this measure the nonparametric density estimate fits better than any of these commonly used parametric choices. A χ^2 test [*Benjamin and Cornell*, 1970, p. 460] rejected at the 95% level the hypothesis that this histogram was from a normal distribution. However, the χ^2 test would not reject any of the other distributions, including the nonparametric density estimate, which is typical of the inability to distinguish between candidate distributions.

4. The synthetic traces generated by MGTM replicate the first few (2 or 3) moments of the underlying dependence structure. Consequently, the generated series may bear little resemblance to the observed series in terms of persistence and threshold crossings, factors that are of interest to hydrologists. The ARMA models also are incapable of displaying sudden bursts or jumps, a feature that may often be observed during an otherwise prolonged drought.

5. *Salas and Smith* [1981] and *Salas et al.* [1980] discussed physical justifications for ARMA models and showed that a linear control system representation of basin processes can lead to ARMA models of streamflow. However, other factors emerge when one considers the relationship of streamflow to some causative factors. For instance, in snow-fed basins the streamflow response during snowmelt months is a threshold response to temperature. The dynamics of soil moisture is hysteretic and nonlinear. The dynamics of vegetative consumptive use and retention of water is also quite different during wet and dry periods and as a function of temperature. Runoff generation mechanisms during protracted wet or dry periods will consequently be different. While these comments have more direct bearing on streamflow at timescales shorter than a month or a year, they are relevant for the longer timescales in that they influence the variance of the streamflow at these timescales.

6. Despite the fact that nearly 30 years have elapsed since the classical time series (AR) models were introduced to practicing hydrologists, acceptance and application of these models for drought analysis and reservoir operation by practitioners has been limited. They often prefer to base their decisions on the historical record (or a resampled proxy thereof). *Kendall and Dracup* [1991] note that the index sequential method, which is basically a sequential resampling of the historical record, appears to be the procedure of choice in many water management agencies, including the California Department of Water Resources, U.S. Bureau of Reclamation, Los Angeles Department of Water, and the Metropolitan Water District of Southern California. This is practiced in spite of the recognition that history is unlikely to repeat itself and that the record is perhaps woefully short.

In summary, while the MGTM ARMA framework is indeed useful in certain contexts, a more flexible time series analysis method capable of reproducing additional features of hydrologic data is needed. The success of linear ARMA models with some hydrologic data sets may be fortuitous and a consequence of short records. The technical issues are nonlinearity, nonstationarity, and inhomogeneity in the underlying dependence structure. Parametric nonlinear models [*Bendat and*

Piersol, 1986; *Tong*, 1990] can be used in place of the linear ARMA models to model nonlinear time series. The use of such models, however, still requires specification of the form of nonlinear dependence, something which may be difficult to do in practice. From a practitioner's perspective the key issues are reproducibility of observed data characteristics, simplicity, and dependability. The nonparametric techniques proposed here avoid the difficult model specification issues associated with parametric linear or nonlinear models. They amount to resampling from the original data, with perturbations, and reproduce directly the characteristics of the original data in a simple and dependable way.

3. 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 estimation methods [*Scott*, 1992, p. 125.]. In this paper we use multivariate kernel density estimators with Gaussian kernels and bandwidth selected using least squares cross-validation (LSCV) [e.g., *Scott*, 1992, p. 160]. This bandwidth selection method is one from among the many available methods. Our methodology is intended to be generic and should work with any bandwidth and kernel density estimation method. This section reviews kernel density estimation first in a univariate and then in a multivariate setting and gives details of the LSCV procedure for estimating bandwidth. For a review of hydrologic applications of kernel density and distribution function estimators, readers are referred to work by *Lall* [1995]; *Silverman* [1986] and *Scott* [1992] provide good introductory texts.

A univariate kernel probability density estimator is written

$$\hat{f}(x) = \sum_{i=1}^{n} \frac{1}{nh} K\left(\frac{x - x_i}{h}\right) \tag{4}$$

where there are n sample data x_i. $K(\ \ )$ is a kernel function that must integrate to 1, and h is a parameter called the bandwidth that defines the locale over which the empirical frequency distribution is averaged. There are many possible kernel functions given in texts such as those by *Silverman* [1986] and *Scott* [1992]. The Gaussian kernel function, a popular and practical choice, is used here:

$$K(x) = \frac{1}{(2\pi)^{1/2}} \exp\ (-x^2/2) \tag{5}$$

The density estimate in (4) is formed by summing kernels with bandwidth h centered at each observation x_i. This is similar to the construction of a histogram where individual observations contribute to the density by placing a rectangular box (analogous to the kernel function) in the prespecified bin the observation lies in. The histogram is discrete and sensitive to the position and size of each bin. By using smooth kernel functions, the kernel density estimate in (4) is smooth and continuous.

A multivariate extension of (4) and (5) for a vector $\mathbf{x}$ in d dimensions can be written as

$$\hat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{(2\pi)^{d/2} \det\ (\mathbf{H})^{1/2}} \cdot \exp\left(-\frac{(\mathbf{x} - \mathbf{x}_i)^T \mathbf{H}^{-1} (\mathbf{x} - \mathbf{x}_i)}{2}\right) \tag{6}$$

where n is the number of observed vectors $\mathbf{x}_i$ and $\mathbf{H}$ is a bandwidth matrix that must be from the class of symmetric positive definite $d \times d$ matrices [*Wand and Jones*, 1994]. The above density estimate is formed by summing 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

$$\mathbf{H} = \lambda^2 \mathbf{S} \tag{7}$$

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 (7)) is called "sphering" [*Fukunaga*, 1972] and ensures that all kernels are oriented along the principal components of the covariance matrix.

Silverman [1986, pp. 70–72] cites results indicating that sufficient conditions for convergence of the kernel density estimate to an underlying density function under broad conditions met by any kernel that is a usable probability density function, are that as $n \to \infty$, $h \to 0$ and $nh \to \infty$. This also applies to λ in the multivariate context. However, the rate of convergence depends on how h or λ is chosen. Methods for choosing the bandwidth are based on evaluation of factors such as bias, $E\{f(\mathbf{x}) - \hat{f}(\mathbf{x})\}$; variance, $\mathrm{Var}\{\hat{f}(\mathbf{x})\}$; mean square error (MSE); integrated square error (ISE); and mean integrated square error (MISE) of the estimate:

$$\begin{aligned} \mathrm{MSE} &= E\{[f(\mathbf{x}) - \hat{f}(\mathbf{x})]^2\} \\ &= \{E[f(\mathbf{x}) - \hat{f}(\mathbf{x})]\}^2 + \mathrm{Var}\ \{\hat{f}(\mathbf{x})\} \end{aligned} \tag{8}$$

$$\mathrm{ISE} = \int_{\mathcal{R}^d} (\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2\, d\mathbf{x} \tag{9}$$

$$\mathrm{MISE} = E \int_{\mathcal{R}^d} (\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2\, d\mathbf{x} \tag{10}$$

A small value of the bandwidth (h or λ) can result in a density estimate that appears "rough," and has a high variance. On the other hand, too high an h results in an "over smoothed" 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 h.

Taylor series expansion of the one-dimensional density estimate in (4) can be used to show that the asymptotic mean integrated square error (AMISE) is [*Silverman*, 1986, p. 40; *Sain et al.*, 1994]

$$\mathrm{AMISE}(h) \approx \frac{R(K)}{nh} + \tfrac{1}{4}\sigma_K^4 h^4 R(f'') \tag{11}$$

where $R[g(x)] = \int g(x)^2\, dx$ for any function $g(x)$ (either $K(x)$ or $f''(x)$), f'' is the second derivative, and $\sigma_K^2 = \int u^2 K(u)\, du$. This can be generalized to higher dimensions.

One choice for the bandwidth is one that directly minimizes (11) if the true distribution were known. This value is known as the AMISE optimal bandwidth for that distribution. For a Gaussian distribution with Gaussian kernel functions (estimator defined by (6) and (7)) *Silverman* [1986, pp. 86–87] gives this bandwidth as

$$\lambda = \left(\frac{4}{d+2}\right)^{1/(d+4)} n^{-1/(d+4)} \tag{12}$$

In the univariate case ($d = 1$) this reduces to $h = 1.06\hat{\sigma}n^{-1/5}$

where $\hat{\sigma}$ is an estimate of the standard deviation (Silverman advocates a robust estimate) of the data. An upper bound on bandwidth can be obtained by minimizing $R(f'')$ over a class of probability densities. This leads to the optimal bandwidth for the smoothest possible density function. *Scott* [1992, p. 181] cites results showing that this upper bound ($1 \leq d \leq 10$) is 1.08 to 1.12 times the λ in (12).

Data-driven methods have been developed to estimate the bandwidth when the underlying distribution is not known. They minimize estimates of ISE, MISE, or AMISE formed only from the data. LSCV [*Silverman*, 1986, pp. 48–52] is one such method, based on the fact that the integrated square error (equation (9)) can be expanded as

$$\text{ISE} = R[\hat{f}(\mathbf{x})] - 2\int \hat{f}(\mathbf{x})\, f(\mathbf{x})\, d\mathbf{x} + R[f(\mathbf{x})] \qquad (13)$$

The first term may be directly evaluated. The second term may be recognized as $E[\hat{f}(\mathbf{X})]$ and estimated using leave one out cross validation. The last term, $R(f(\mathbf{x}))$, is independent of the bandwidth and does not need to be considered. The LSCV method in one dimension chooses the bandwidth, h, to minimize the following LSCV score, comprising the first two terms in (13):

$$\text{LSCV}(h) = \frac{1}{n^2 h}\sum_{i=1}^{n}\sum_{j=1}^{n} K^{(2)}\left(\frac{x_i - x_j}{h}\right) - \frac{2}{n}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq 1}}^{n}\frac{1}{nh}K\left(\frac{x_i - x_j}{h}\right) \qquad (14)$$

Here, $K^{(2)}$ denotes the convolution of the kernel function with itself (for example, if K is the standard Gaussian kernel, then $K^{(2)}$ will be the Gaussian density with variance 2).

On the basis of results by *Sain et al.* [1994] and *Adamowski and Feluch* [1991] the generalization of the LSCV score to higher dimensions with multivariate Gaussian kernel functions and a symmetric positive definite bandwidth matrix $\mathbf{H}$ as specified by, for example, (7) is

$$\text{LSCV}(\mathbf{H}) = \frac{1 + (1/n)\sum_{i=1}^{n}\sum_{j\neq i}[\exp(-L_{ij}/4) - 2^{d/2+1}\exp(-L_{ij}/2)]}{(2\pi^{1/2})^d n \det(\mathbf{H})^{1/2}} \qquad (15)$$

where

$$L_{ij} = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{H}^{-1}(\mathbf{x}_i - \mathbf{x}_j) \qquad (16)$$

We use numerical minimization of (15) over the single parameter λ with bandwidth matrix from (7) to estimate all the necessary probability density functions. We recognize that LSCV bandwidth estimation is occasionally degenerate, 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 $\lambda/4$ to 1.1λ.

4. Nonparametric Order *p* Markov Streamflow Model, NP*p*

To keep the presentation simple, the equations will be presented for a lag 1 (order $p = 1$) model. The formulae presented are readily extended to include higher-order lags. Consideration of higher-order models raises the issue of determination of the correct order p. This is deferred to future work. Here results are presented for the simplest case (NP1) analogous to the simple AR1 model. In the form presented below, the model can be applied to simulate stationary sequences such as annual flows. Section 6 describes how application of the model to pairs of sequential months is used to simulate seasonally nonstationary (e.g., monthly) streamflow sequences.

The joint distribution of X_t and its prior value X_{t-1} is estimated using (6) on the basis of n observed data vectors $\mathbf{x}_i$. For a time series $x_0, x_1, x_2, \cdots, x_n$, the data vector $\mathbf{x}_i$ has elements (x_i, x_{i-1}), where $1 \leq i \leq n$. Hence $\mathbf{x}_1 = (x_1, x_0)$, $\mathbf{x}_2 = (x_2, x_1), \cdots, \mathbf{x}_n = (x_n, x_{n-1})$. These are a series of ordered pairs. There is one less ordered pair than the length of the time series. The conditional density (equation (1)) is written as

$$f(X_t|X_{t-1}) = \frac{f(X_t, X_{t-1})}{\int f(X_t, X_{t-1})\, dX_t} = \frac{f(X_t, X_{t-1})}{f_m(X_{t-1})} \qquad (17)$$

where $f_m(X_{t-1})$ is the marginal density of X_{t-1}. Now applying the estimator in (6), the joint density estimate is obtained as

$$\hat{f}(X_t, X_{t-1}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2\pi\lambda^2 \det(\mathbf{S})^{1/2}} \cdot \exp\left(-\left\{\begin{bmatrix} X_t - x_i \\ X_{t-1} - x_{i-1}\end{bmatrix}^T \mathbf{S}^{-1}\begin{bmatrix} X_t - x_i \\ X_{t-1} - x_{i-1}\end{bmatrix} \Big/ 2\lambda^2\right\}\right) \qquad (18)$$

Note that each observation contributes to this density estimate depending on the distance of the observation (x_i, x_{i-1}) to the point (X_t, X_{t-1}), the bandwidth λ, and the sample covariance matrix $\mathbf{S}$ of (X_t, X_{t-1}). The bandwidth λ is obtained by minimizing the LSCV score function (equation (15)).

Denote the terms in the covariance matrix:

$$\mathbf{S} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22}\end{bmatrix} \qquad (19)$$

Then for a given X_{t-1}, (18) substituted in (17) reduces to a sum of Gaussian kernels dependent on a single variable X_t:

$$\hat{f}(X_t|X_{t-1}) = \sum_{i=1}^{n}\frac{1}{(2\pi\lambda^2 S')^{1/2}} w_i \exp\left(-\frac{(X_t - b_i)^2}{2\lambda^2 S'}\right) \qquad (20)$$

where

$$w_i = \exp\left(-\frac{(X_{t-1} - x_{i-1})^2}{2\lambda^2 S_{22}}\right) \Big/ \sum_{j=1}^{n}\exp\left(-\frac{(X_{t-1} - x_{j-1})^2}{2\lambda^2 S_{22}}\right) \qquad (21a)$$

$$S' = S_{11} - \frac{S_{12}^2}{S_{22}} \qquad (21b)$$

$$b_i = x_i + (X_{t-1} - x_{i-1})\frac{S_{12}}{S_{22}} \qquad (21c)$$

This is illustrated in Figure 2. The conditional density is a slice through the bivariate density function, composed of a sum

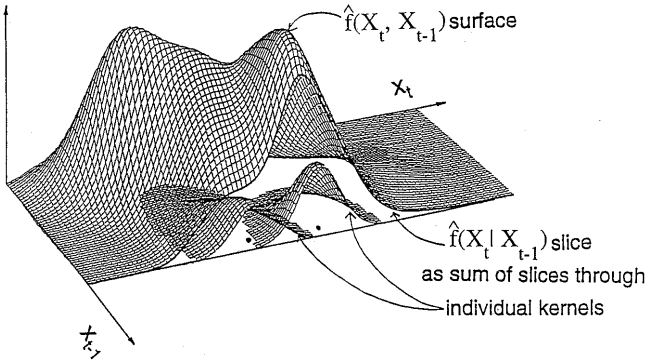


Figure 2. Illustration of conditional probability density function.

of slices through the individual kernels that form the bivariate density estimate. Parameters b_i and $\lambda^2 S'$ give the center and spread of each kernel slice, respectively. The area under each kernel slice is the weight w_i which controls the contribution of x_{i-1} to the conditional density estimate. Observations that lie close to the conditioning plane (i.e., where $(X_{t-1} - x_{i-1})$ is small) receive greater weight. A time series realization is simulated by sampling X_t from (20), given a current value for X_{t-1}. The simulation then proceeds sequentially through time, updating X_{t-1} as the last sampled value. A flowchart describing the steps needed to simulate a sample of size n_r is provided in Figure 3.

Note that in the simulation scheme one does not need to explicitly estimate the conditional density in (20). Since the conditional density function is the sum of n Gaussian kernel slices that each contribute weight w_i (the weights sum to 1, equation (21a)) simulation can be achieved by first picking a slice with probability w_i and then selecting X_t as a random variate from that kernel slice with mean b_i and variance $\lambda^2 S'$ using

$$X_t = b_i + \lambda (S')^{1/2} W_t \tag{22}$$

where W_t is $N(0, 1)$.

A complication can arise because W_t is unbounded and may result in negative X_t. The Gaussian kernels used in the kernel density estimate have infinite support and assign some (small) probability to regions of the domain where the streamflow is negative (i.e., invalid or out of bounds). This leakage of probability across boundaries is a problem when using kernel density estimates based on kernels with infinite support. It is also present in the parametric context where a Gaussian distribution or any parametric distribution with support extending to negative values or beyond a lower or upper bound on the process is used. Here we address the leakage by checking at each step whether the simulated flow values are positive. Whenever a negative X_t is encountered, we generate another sample from the same kernel slice, repeating this process until a positive X_t is obtained. This is achieved by simply generating a new W_t in (22). This is equivalent to cutting the portion of each kernel that is out of bounds and renormalizing that kernel to have the appropriate mass. We record how often this is done, as frequent boundary normalization is symptomatic of a substantial boundary leakage problem. Although the boundary renormalization procedure results in some bias in the simulated density in the neighborhood of the boundary, this was required for less than 1% of total realizations for the streamflow data sets the model was evaluated on.

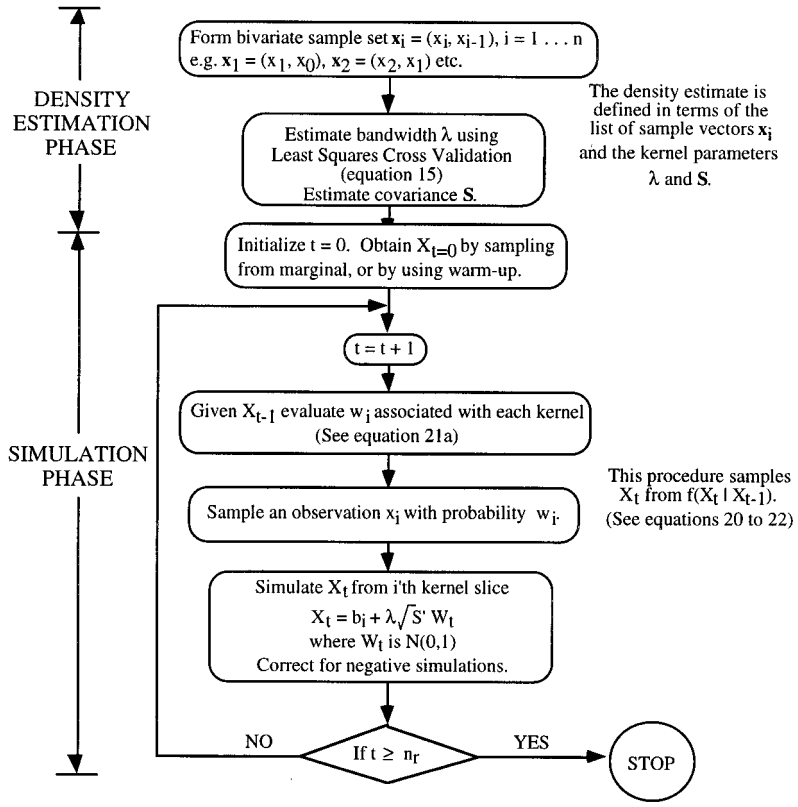


Figure 3. Flowchart of NP1 model.

There are two alternatives for initializing the nonparametric simulation procedure. The first is to sample X_t at $t = 0$ from the appropriate marginal distribution, which is a univariate kernel density function given by (4) with bandwidth $h = \lambda(S_{11})^{1/2}$. Each prior data point contributes equal weight $(1/n)$ to this kernel estimate. Therefore the initial variate may be obtained by picking one of the prior values x_{i-1} at random with probability $1/n$ and then using

$$X_t = x_{i-1} + \lambda(S_{11})^{1/2} W_t \tag{23}$$

where W_t is $N(0, 1)$.

The second initialization alternative is to specify X_t at $t = 0$ arbitrarily (e.g., equal to the mean) and provide a suitably long "warm-up" period, discarding the first several values simulated to avoid any initialization bias.

The nonparametric simulation model has been presented from the perspective of formally estimating the underlying probability density function and then sampling from it. However, when viewed operationally one sees that it has close ties to the bootstrap [*Efron*, 1979; *Efron and Tibishirani*, 1993]. In fact, it is a smoothed bootstrap. Each kernel slice that contributes weight w_i is centered over a prior data pair (x_i, x_{i-1}), so picking a kernel slice amounts to picking a prior data pair with probability w_i. The bootstrap is a statistical method that involves resampling the original data (with replacement) that has applications in estimation of confidence intervals and quantification of parameter uncertainty [*Hardle and Bowman*, 1988; *Woo*, 1989; *Tasker*, 1987; *Zucchini and Adamson*, 1989]. The classic bootstrap assumes data are independent and identically distributed and resamples from each prior data point with equal probability. The nearest neighbor bootstrap method presented by *Lall and Sharma* [1996] was designed for bootstrapping dependent data. It is similar to the approach here in that a data pair nearby to the conditioning vector is picked, and its successor is chosen as the simulated data value. However, it uses a conditional probability density represented by a discrete kernel which is based on the assumption of a local Poisson distribution in the neighborhood of the conditioning vector. The nearest neighbor bootstrap also differs in that there is no perturbation of the selected point. Consequently, it only reproduces streamflow values that have been observed. The approach presented here amounts to picking a prior data pair (x_i, x_{i-1}) that is nearby, that is, x_{i-1} near to the conditioning value X_{t-1}, according to weights based on Gaussian kernels, then through (22), adding a perturbation. The perturbations in our approach serve to smooth over the gaps between data points in the density estimate and provide alternative streamflow realizations that are different but are stochastically similar to the historical record.

Simulations from this nonparametric approach retain the marginal and joint density structure of the historical time series including nonlinearities and state dependence. One can also analytically calculate the marginal distribution and the values of the NP1 model mean, standard deviation, skewness, and lag 1 correlation from the kernel density estimate (equation (18)). These are given in Appendix B and compared in the results below to sample statistics from the historical data.

5. Testing With Synthetic Data

In order to evaluate the ability of our model to recover structure from known linear and nonlinear parametric models, we conducted tests using two synthetic models. The purpose of these experiments was to verify the performance of the NP1 model where the true model is known. The first model used was a linear autoregressive order 1 (AR1) model of the type commonly used to model streamflow. The second was a self-exciting threshold autoregressive (SETAR) model [*Tong*, 1990]. A two-level Monte Carlo experiment was used in both cases: (1) generate 100 level 1 sample records of length 80 from the true model (AR1 or SETAR); and (2) for each sample record, generate 100 level 2 realizations, each of length 80, from the NP1 model.

Note that we refer to the 100 samples first generated as level 1 samples. These are representative of the observed data. We refer to the 10,000 realizations generated (100 NP1 realizations from each of the 100 level 1 sample records) as level 2 realizations. These are representative of what our model would simulate given the level 1 sample records and were used to evaluate how well the NP1 model reproduces statistics of the samples it is based on as well as the underlying population statistics. Since all realizations are the same length as the record from which they are generated, the 100 realizations from each record provide estimates of the natural sampling variability associated with that record length. Statistics such as the mean, standard deviation, lag 1 correlation, and skewness were estimated, as well as marginal and joint kernel density estimates from the sample records and realizations.

Box plots are used for the graphical comparisons. These consist of a box that extends over the interquartile range of the quantity being plotted, estimated from the 100 realizations. The line in the center of this box is the median, and whiskers extend to the 5% and 95% quantiles of the compared statistic.

5.1. Tests With AR1 Data

The AR1 model used was

$$X_t = 0.5\, X_{t-1} + 0.866\, W_t \tag{24}$$

where W_t was a Gaussian random variate with mean zero and standard deviation one. For brevity comparisons for the standard statistics are not given. The mean, variance, lag 1 correlation, and skewness of each AR1 sample were well reproduced in the simulations based on it. These values were also close to the corresponding model statistic.

Figure 4 shows the marginal density estimates from one of the sample records and its corresponding 100 NP1 simulations. Shown are the true Gaussian density function, the NP1 model marginal density function (from equation (B1)), and a univariate kernel density estimate based on the sample records, with the boxes giving the univariate kernel density estimates for the 100 NP1 simulations. To ensure that these univariate kernel density estimates are comparable, we used the same bandwidth for each of them, namely, the median bandwidth from the set of bandwidths obtained by applying the LSCV procedure to each simulation. Figure 4 shows that the marginal density of the data is reproduced quite well by the simulations.

The mean integrated square error (MISE) of the joint density was estimated by averaging the integrated square error (ISE; see (9)) between the kernel estimate on the basis of each level 1 sample record and the true distribution. This provides a measure of model error which was 0.0093. The corresponding MISE from fitting an AR1 model joint density to each level 1 sample is 0.0046. This is lower than the MISE from kernel density estimates because the assumed joint density (in this

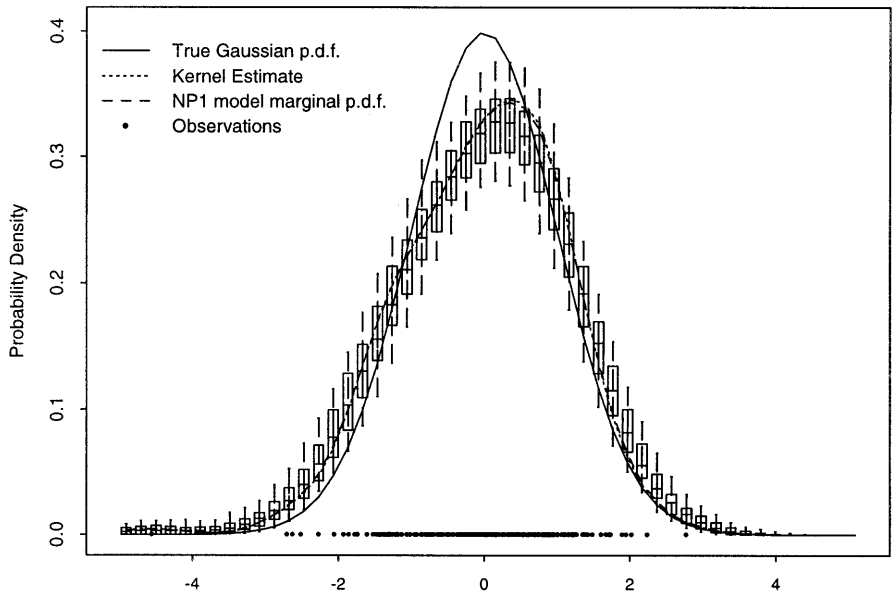


Figure 4. Marginal density estimates for NP1 simulations of an AR1 data set. The NP1 underlying density is estimated using (B1).

case a bivariate Gaussian) is the same as the underlying density.

The level 2 Monte Carlo experiment involved calculating bivariate density estimates using the procedure given in section 3 for each of the 10,000 level 2 NP1 realizations. The ISE for each of these was calculated, and the average is 0.0161, which is greater than the 0.0093 given above. This reflects the additional error introduced by reestimating the density function from simulated values. These experiments serve to illustrate that although the nonparametric approach cannot match a parametric approach when the true density is known, it is still able to reasonably approximate the properties of the underlying AR1 process.

5.2. Tests With SETAR Data

The SETAR [*Tong*, 1990, section 3.3.1.1] model used was

$$\begin{aligned} X_t &= 0.4 + 0.8X_{t-1} + W_t \qquad X_{t-1} \le 0.0 \\ X_t &= -1.5 - 0.5X_{t-1} + W_t \qquad X_{t-1} > 0.0 \end{aligned} \tag{25}$$

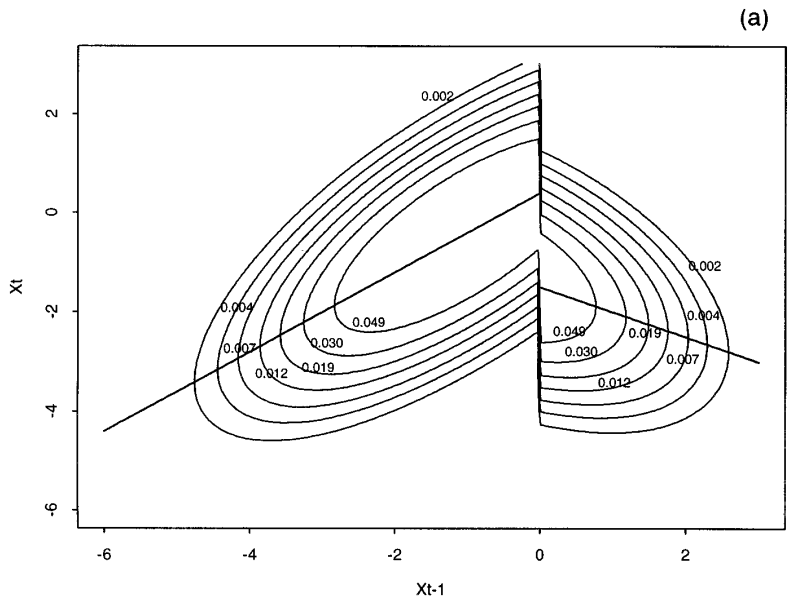
where W_t was $N(0, 1)$. This is a state-dependent time series model with parameters that depend on the system state as determined by a threshold. This model may be representative of the monthly streamflow time series one could get from threshold-driven hydrologic processes such as snowmelt and evapotranspiration.

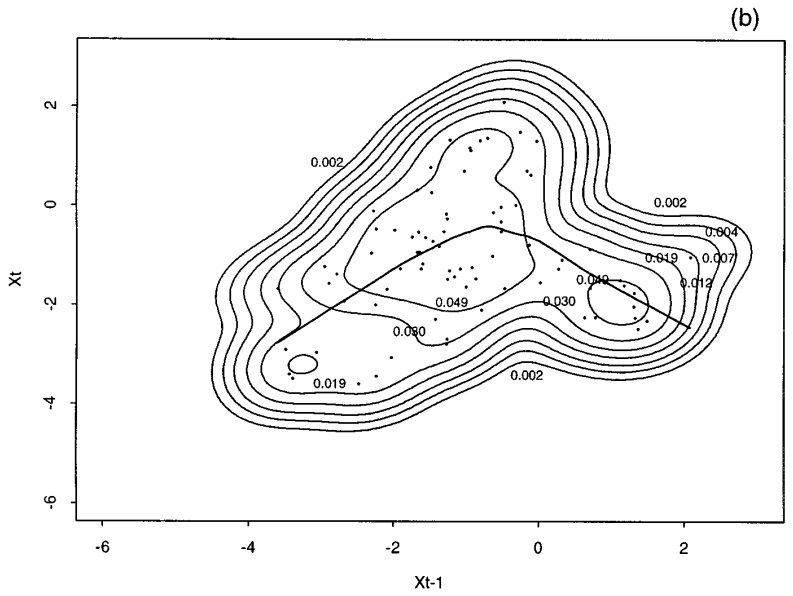
As with the AR1 case, mean, variance, lag 1 correlation, and skewness of each SETAR sample were well reproduced in the simulations based on it. These values were also close to the corresponding model statistic.

Figure 5 shows the underlying true joint density $f(X_t, X_{t-1})$ for the SETAR model in (25), the bivariate kernel density estimate for one SETAR sample record, and the density estimate of the NP1 simulations averaged over all 10,000 realizations. The line in Figure 5a shows the true conditional mean from (25) with W_t set to 0. Figure 5b shows an estimate of the conditional mean based on the sample record obtained using LOESS [*Cleveland and Devlin*, 1988]. LOESS is a locally weighted regression smoother that calculates a weighted least squares fit (assigning weights using a tricubic weight function centered at the point of estimation) at each data point on the basis of a fixed number of nearest neighbors. The number of nearest neighbors (expressed as a fraction of the total number of data points, called "span") used to compute the LOESS smooth was chosen as the one that resulted in an optimal value of Mallow's Cp. The function "loess," available in the software package Splus [*Chambers and Hastie*, 1992], was used in our calculations. The LOESS smooth is plotted to show that the sample record and nonparametric density function based on it reproduce the change in conditioning structure (with some smoothing) as the threshold is crossed. The illustrated fit based on an optimal Mallow's Cp has a span of 0.75. The particular kernel density estimate shown in Figure 5b has an integrated square error (ISE; see (9)) of 0.0084 that was evaluated by integrating the squared differences between Figures 5a and 5b. By averaging across the 100 level 1 sample records, we obtain an estimate of the NP1 model fitting MISE as 0.0082. The corresponding MISE from fitting an AR1 joint density to each level 1 sample is 0.0131.

As for the AR1 example, bivariate density estimates were calculated using the procedure given in section 3 for each of the 10,000 level 2 realizations. The ISE for each of these was calculated, and the average is 0.010, which is greater than the NP1 model MISE (0.0082), owing to the additional error added by reestimating the density function from simulated values. It is again representative of the difference between NP1 simulations and the underlying model. Figure 5c shows the

Figure 5. (opposite) Bivariate joint density. (a) SETAR model in (5). The straight lines denote the conditional mean of the model. (b) An example SETAR level 1 sample. Dots represent individual observations and contours represent the NP1 kernel density estimate. The line is a LOESS smooth through the data. (c) Average of the kernel density estimates from all 10,000 level 2 realizations from the NP1 model.


(a)
Xt
Xt-1
-6
-4
-2
0
2
0.002
0.004
0.007
0.012
0.019
0.030
0.049


(b)
Xt
Xt-1
-6
-4
-2
0
2
0.002
0.004
0.007
0.012
0.019
0.030
0.049

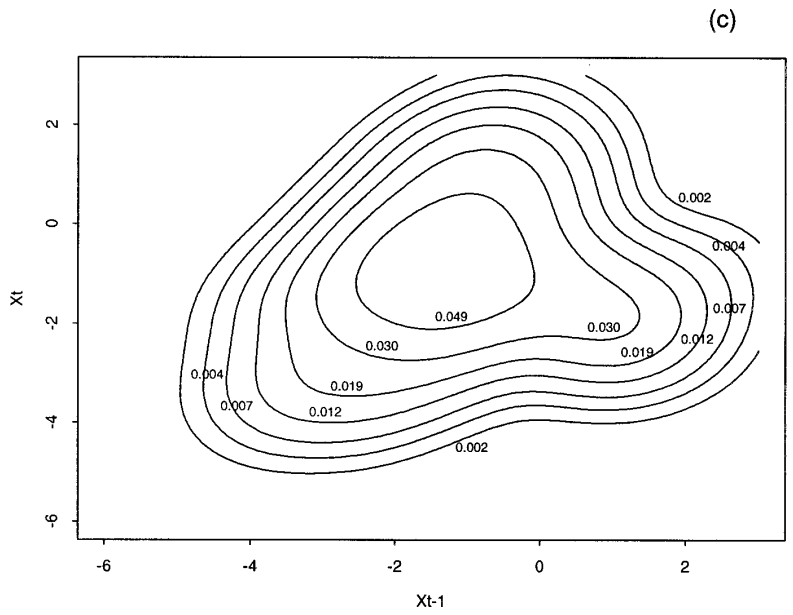

(c)
Xt
Xt-1
-6
-4
-2
0
2
0.002
0.004
0.007
0.012
0.019
0.030
0.049

Table 1. State-Dependent Lag 1 Correlations for NP1 Simulations From a SETAR Model

Statistic	Single Level 1 Sample Record	Level 2 Simulation Statistics From NP1 Model Fit to Single Level 1 Record		
		5% Quantile	Median	95% Quantile
r	0.164	−0.094	0.160	0.329
r_{af}*	−0.528	−0.595	−0.373	−0.202
r_{bf}†	0.504	0.167	0.425	0.620
r_{ab}‡	0.172	−0.056	0.161	0.380
r_{bb}§	0.310	−0.215	0.239	0.490
	Average Over 100 Level 1 Sample Records and 10,000 Level 2 Realizations			
		NP1 Simulation Statistics		
Statistic	100 Records	5% Quantile	Median	95% Quantile
r	0.074	−0.105	0.041	0.226
r_{af}*	−0.555	−0.557	−0.396	−0.164
r_{bf}†	0.494	0.187	0.377	0.558
r_{ab}‡	0.165	0.000	0.131	0.305
r_{bb}§	−0.020	−0.219	−0.039	0.169

*Above and forward.
†Below and forward.
‡Above and back.
§Below and back.

average density function estimated from the 10,000 realizations. This captures the essential nonlinearity of the SETAR model despite smoothing over the discontinuity. No model from the MGTM class of models is able to reproduce samples which exhibit such nonlinear structure. A bivariate normal distribution with the true mean and covariance of the model in (25), that is, without fitting errors, has an ISE relative to Figure 5a of 0.0119, larger than that obtained from the NP1 model fits with 80 data points. Asymptotically, the NP1 kernel density estimate will converge to the underlying SETAR model exactly, that is, with no fitting errors. This reiterates the point that model mis-specification, for example, by selection of the bivariate normal distribution, precludes a model from convergence to whatever the underlying distribution may be.

Table 1 shows the state dependent correlation statistics (described in Appendix A) for NP1 model simulations based on SETAR data. Note how well the NP1 model reproduces the big difference between above median and forward correlations and below median and forward correlations.

It is clear from the synthetic examples presented that the NP1 model is able to (1) approximate the underlying joint distribution of the data, (2) reproduce the nonlinear structure suggested by the data in model simulations, and (3) approximate both linear and nonlinear dependence between the variables involved. No assumptions about marginal distributions or normalizing transforms are required.

6. Application of NP1 to Simulation of Monthly Streamflow

This section describes the application of the NP1 model to simulate seasonal streamflow sequences. Assume that the year is divided into s periods (seasons or months, in which case $s = 12$) and there are n years of data (total $n \times s$ data values). The model applied to seasonal sequences then consists of s (one for each period) bivariate density functions estimated directly from the historical data. For all periods except the first the random vector (X_t, X_{t-1}) is replaced by $(X_{t,j}, X_{t,j-1})$, where subscript t denotes the year and j denotes the period. For the first period the conditioning flow is the flow in the last period of the previous year and the vector is $(X_{t,1}, X_{t-1,s})$. Simulations proceed sequentially from density estimates for one period pair to the next.

Results from an AR1 model representative of current hydrologic practice are also presented for comparison to NP1 simulations. SPIGOT, a synthetic streamflow generation software package developed by *Grygier and Stedinger* [1990], uses four choices for monthly marginal probability densities: (1) Gaussian, (2) two-parameter lognormal, (3) three-parameter lognormal, and (4) approximate three-parameter gamma distributions. 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]). Here we used the same procedure as SPIGOT to fit a marginal probability distribution and to obtain a normalizing transformation for each month. Then the AR1 model with seasonally varying coefficients given in (2) was applied to the transformed monthly flows.

Both the NP1 and AR1 (with normalizing transformations) models were applied to the 79-year (1914–1992) record of monthly streamflow in the Beaver River, used earlier for Figure 1. This data set is one of many streamflow data sets that we have tested the model with, all with satisfactory results. We chose to present results from the Beaver River because it illustrates well some of the points we want to emphasize.

Figure 6 shows the joint densities for the April-May and the June-July month pairs, using the kernel estimator in section 3. Also shown are the conditional expectations $E(X_t|X_{t-1})$ estimated using LOESS. The NP1 model simulates streamflow from such joint density functions. A notable aspect of both

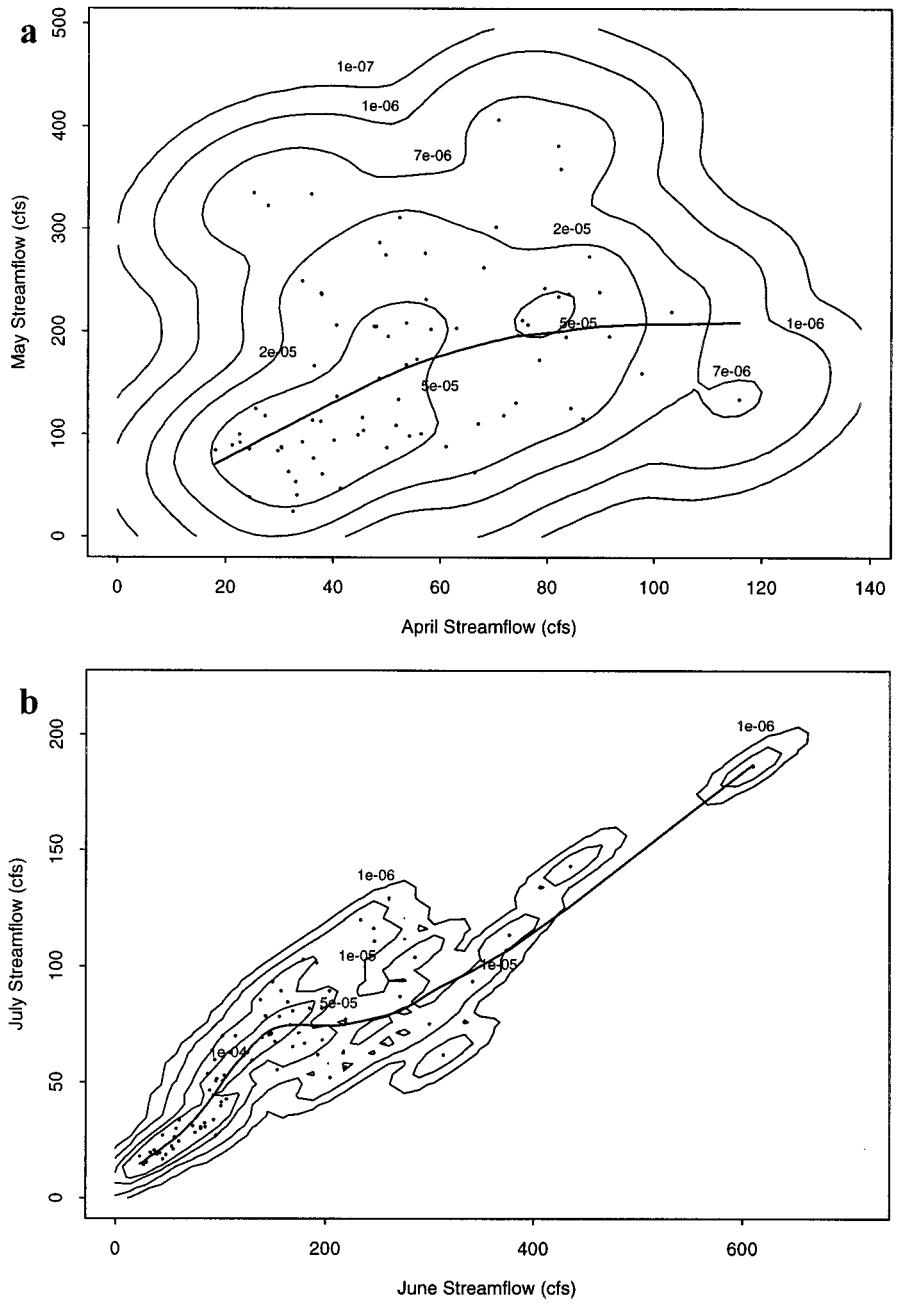


Figure 6. Underlying bivariate densities in NP1 model simulations for selected month pairs. The dots represent observations. (a) April–May flows. The line represents a LOESS smooth with a span of 0.85 corresponding to an optimal Mallow's Cp. (b) June–July flows. The line represents a LOESS smooth with a span of 0.45 corresponding to an optimal Mallow's Cp.

figures is that the LOESS fits exhibit a certain degree of nonlinearity. It is possible that both bivariate samples could have originated from a threshold-driven process of the type illustrated in the synthetic example of section 5.2.

One hundred simulations, each with a length of 79 years (initialized with the average flow of the first month and with a warm-up period of 1 year), were made using both the NP1 and AR1 models. Results comparing the simulations from NP1 and AR1 are presented below.

Box plots of selected monthly statistics are shown in Figures 7–10. The mean flows of the AR1 and NP1 simulations (Figure 7) match well those of the streamflow record. The annual means also match well. Figure 8 shows standard deviations of flows for the AR1 and NP1 simulations. The standard deviations of the NP1 simulations are slightly inflated with respect to the historical record, as expected from (B5). Annual standard deviations, though not modeled directly by either approach, compare well with the historical value. Figure 9 shows box plots of the correlation between sequential month pairs. The NP1 model reproduces monthly lag 1 correlations without any bias, as proved in (B7). The AR1 simulations approximate the monthly lag 1 correlation well, although some bias is present depending on which transformation (or which marginal distribution) is used. Annual lag 1 correlations are relatively small

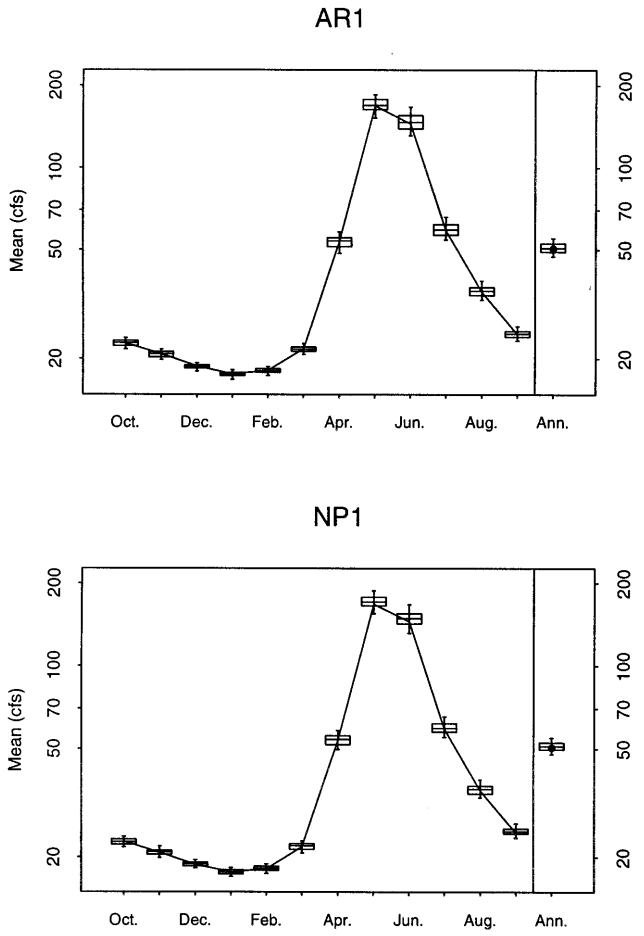


Figure 7. Comparison of means of simulated and historical flows. The continuous line represents monthly means of the historical record. The dot in the right panel is the observed annual mean.

but not reproduced by either model. This is a deficiency requiring higher-order or multiscale (such as disaggregation) models. Skewness is reproduced well in NP1 simulations (Figure 10), although a small downward bias (equation (B12)) is evident. Some bias is apparent in AR1 simulations, too, which is another indication of the difficulty in fitting marginal distributions.

The marginal distributions for each month were also compared. Selected marginal distributions are shown in Figure 11. In these figures the model underlying density function (equation (B1) in the case of NP1 or one of the SPIGOT [*Grygier and Stedinger*, 1990] densities in the AR1 case) is shown as a dashed line. The solid line is a univariate kernel density estimate applied to the original data, and the boxes represent the range of univariate kernel density estimates applied to the 100 simulations. For these univariate kernel density estimates the same bandwidth, chosen as the median of the set obtained by minimizing LSCV over the 100 simulations, is used for all. Here the univariate kernel density estimator is being used as a plotting tool to compare observed and simulated data. The dots on the axis represent the historical data. These figures show that for some months, the best fitting SPIGOT marginal distribution is inadequate. Note how the NP1 model is able to reproduce the bimodality in the July marginal distribution (also shown earlier in Figure 1), whereas the fitted three-parameter gamma distribution does not. Overall, we find that the common normalizing transformations are not able to capture all the structure, in particular bimodality, sometimes present in data. This structure is captured by the kernel density estimates.

Recall that the joint densities illustrated in Figure 6 indicated nonlinear conditional expectations. For the April-May month pair the slope of the conditional expectation for flows less than 70 feet^3 s^{-1} (1982 L s^{-1}) appeared different to the slope for flows greater than 70 feet^3 s^{-1}. To quantify the dependence of autocorrelation on the magnitude of flow, we split each series into flows above and below the median and then calculated the state-dependent correlation statistic described in Appendix A. These results are illustrated in Figure 12. The historical data (solid line) have significant differences (at the 95% level by a hypothesis test for equality of two sample correlations; see Appendix A for details) between forward above- and below-median correlations for the following three month pairs: October-November, July-August, and September-October. These state-dependent correlations are modeled effectively by the NP1 approach. Some bias is apparent in AR1 simulations for month pairs exhibiting significant differences between the above- and below-median correlations.

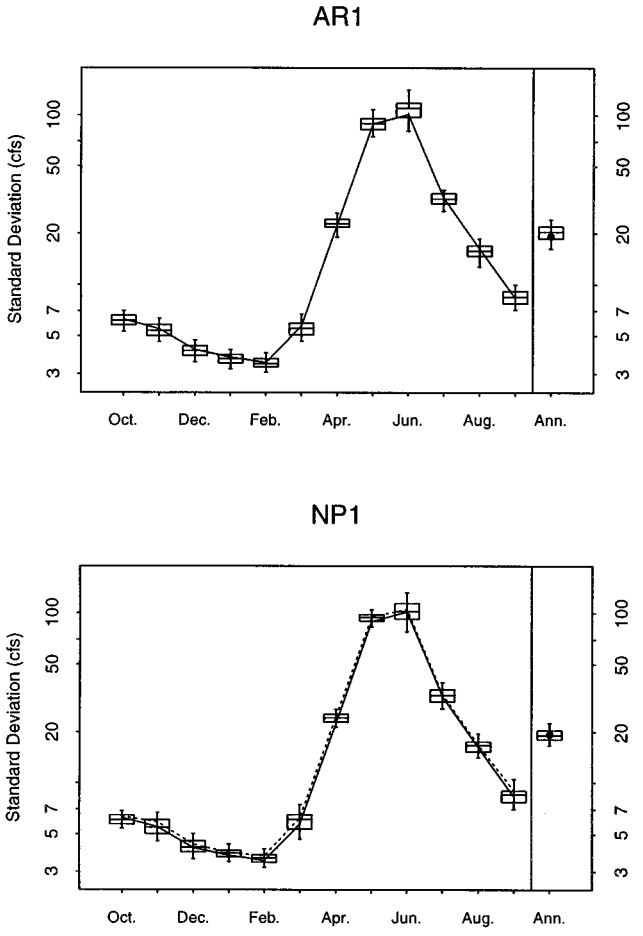


Figure 8. Comparison of standard deviations of simulated and historical flows. The continuous line represents monthly standard deviations of the historical record. The dot in the right panel is the observed annual standard deviation. The dashed line in the NP1 figure shows the NP1 model standard deviations (equation (B5)).

The practical use of synthetic streamflow simulations is often the evaluation of the storage capacity of reservoirs required to support a certain yield. For a given streamflow sequence (observed or simulated) the storage required to support a specified yield can be obtained using the sequent peak algorithm [*Loucks et al.*, 1981, p. 235]. *Vogel and Stedinger* [1988] compared the root-mean-square error (RMSE) and bias of this storage statistic computed directly from data and showed the improvements in precision that result from using stochastic streamflow models. Here reservoir storages required to support firm yields that are specified percentages of the mean annual flow were estimated for the 100 AR1 and NP1 model realizations of Beaver River streamflow. Monthly demand fractions given by *Lall and Miller* [1988] were used. Standardized bias and RMSE estimates for both models, relative to the storage required to support a given yield for the historical record, are given in Table 2.

$$\text{bias}/S_h = \left(S_h - \frac{1}{n_r} \sum_{i=1}^{n_r} S_{s_i} \right) \Big/ S_h \tag{26}$$

$$\text{RMSE}/S_h = \frac{\left[\frac{1}{n} \sum_{i=1}^{n_r} (S_h - S_{s_j})^2 \right]^{1/2}}{S_h} \tag{27}$$

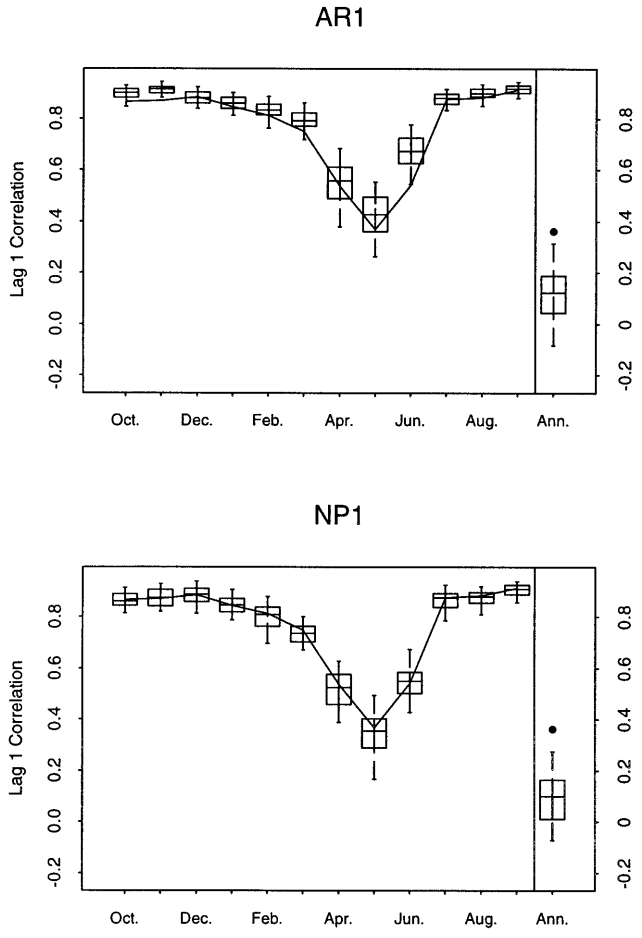


Figure 9. Comparison of lag 1 correlations of simulated and historical flows. The continuous line (and dots for annual flows) represents the lag 1 correlations in the historical record.

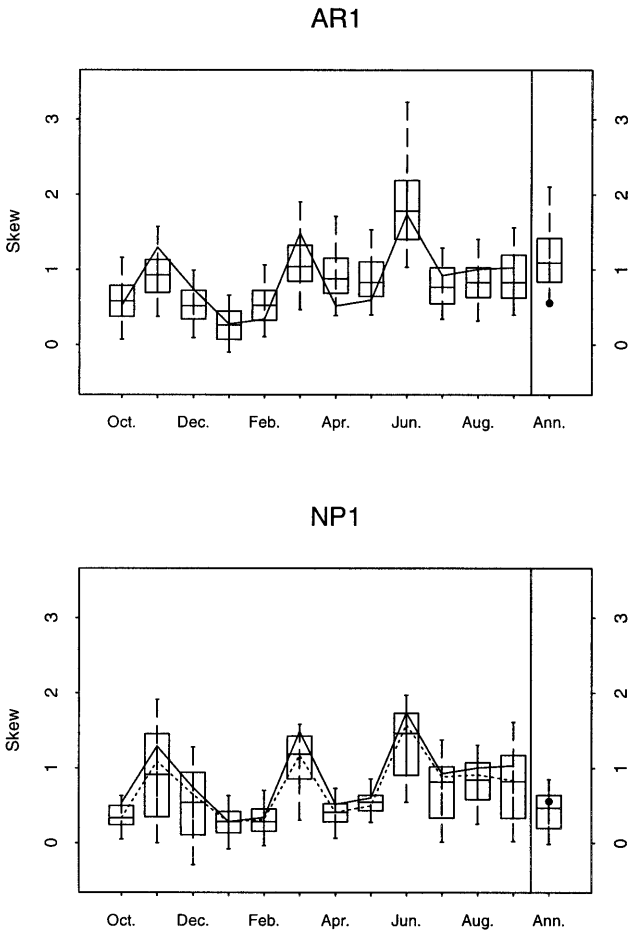


Figure 10. Comparison of skewness of simulated and historical flows. The continuous line represents the monthly skewness of the historical record. The dot in the right panel is observed skewness of the annual flows. The dashed line (in NP1 result) shows model skewness (equation (B12)).

where S_h denotes the storage required with the historical record, S_{s_i} is the storage estimated from the ith AR1 or NP1 realization and n_r is the number of realizations.

The bias and RMSE statistics reported in Table 2 indicate that the NP1 model is better at providing simulations with storage statistics comparable to the historical data.

7. Discussion and Conclusions

We also computed and checked many other statistical attributes of the NP1 and AR1 simulations, but space limitations prevent presentation of the results. The simulated autocorrelation function (acf) for each month (not shown) showed that both models do not model correlations higher than lag 1 very well. In some months lag 2 and 3 correlation was preserved though both NP1 and AR1 model correlations decrease to essentially 0 by lag 7. Longer-range dependence quantified in terms of the annual correlation coefficient or Hurst coefficient [*Hurst*, 1951] was not preserved by either model. Although the bias and errors in the reservoir storages given in Table 2 are generally smaller for the NP1 simulations than for the AR1 simulations, they are still relatively large, especially for high-yield fractions, indicating that the order 1 dependence assumed in both models is possibly inadequate to model reser-

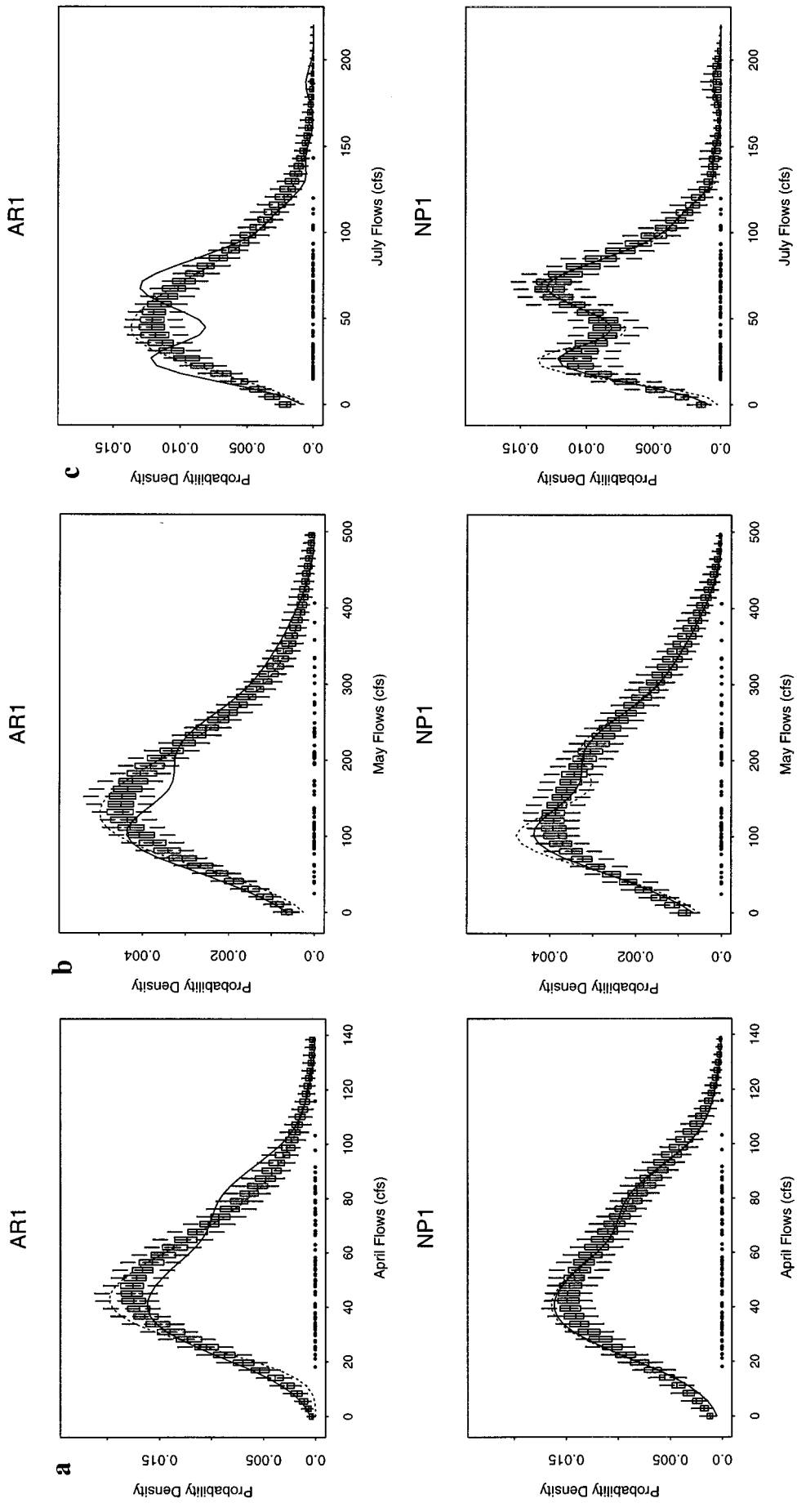


Figure 11. Marginal density estimates for selected months streamflow. The solid line is a univariate kernel density estimate applied to the original data. The model underlying density function ((B1) in the case of NP1 or one of the SPIGOT densities in the AR1 case) is shown as a dashed line. The boxes depict the range of univariate kernel density estimates applied to the 100 simulations. (a) April AR1 and NP1 marginal density estimates. A three-parameter lognormal distribution is used for the AR1 model fit. (b) May AR1 and NP1 marginal density estimates. A three-parameter lognormal distribution is used for the AR1 model fit. (c) July AR1 and NP1 marginal density estimates. A three-parameter gamma distribution is used in the AR1 model fit.

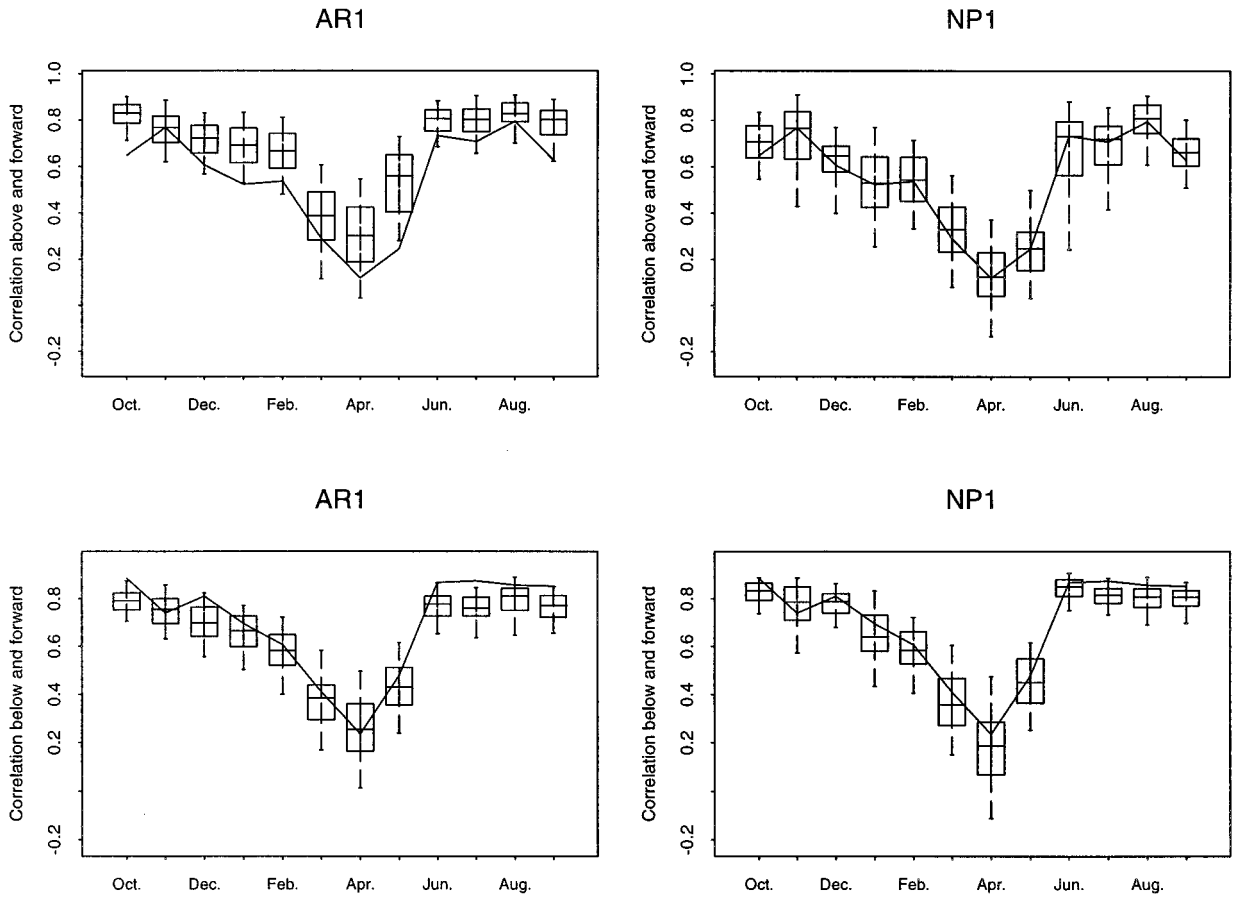


Figure 12. Monthly state-dependent correlations (refer to Appendix A) for simulated and historical flows. Continuous lines represent respective correlations in the historical record.

voir storage for this data. These all indicate the need for models that capture higher-order dependence, such as multivariate or disaggregation models.

The results presented here support the nonparametric approach as a feasible alternative to parametric approaches used to model streamflow. The nonparametric approach presented here is consistent and robust and reproduces not only the linear statistics modeled by the AR1 model but also a broader set of properties based on additional distributional information. The skewness, bimodality, and dependence of correlations on the flow magnitude, when present in the data, can be adequately modeled. One could no doubt find better marginal distributions to use with the AR1 model and improve on some of the AR1 simulations. However, the NP1 approach is effective in sidestepping these difficult model and distribution selection issues that are often somewhat arbitrarily resolved and provides a method that is easy to use and adapts well to the data.

Table 2. Reservoir Capacities Evaluated for 100 AR1 and NP1 Model Realizations

Yield/Mean Annual Flow	AR1		NP1	
	Bias/S_h	RMSE/S_h	Bias/S_h	RMSE/S_h
0.5	0.275	0.312	0.095	0.229
0.6	0.170	0.260	−0.010	0.262
0.7	0.111	0.250	−0.029	0.287
0.8	0.200	0.310	0.137	0.281
0.9	0.373	0.435	0.373	0.426

Although the examples presented here used an order 1 dependence structure, it is easy to extend the model to higher-order dependence. Cross-validatory procedures [*Eubank*, 1988] can be applied to evaluate the benefit gleaned from including additional lags in the model dependence structure. These are somewhat analogous to use of Akiake's information criterion in linear models. We intend to evaluate this further in future work. Future work will also apply the nonparametric approach to multivariate problems in stochastic hydrology, specifically, to the development of nonparametric analogs to multivariate ARMA and disaggregation models. The purpose here was to introduce this approach in a simple univariate setting with order 1 dependence and show that results are satisfactory when compared to current hydrologic practice.

We are convinced that nonparametric techniques 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 historical data without imposing arbitrary linearity or distributional assumptions. They have the capability to reproduce nonlinearity, state dependence, and multimodality while remaining faithful to the historical data and producing synthesized sequences statistically indistinguishable from the historical sequence.

Appendix A: State-Dependent Correlation Coefficients

This appendix describes the measures we used to quantify nonlinear dependence in data. The usual estimator of lag 1 correlation is

$$r = \frac{1}{(n-1)s_x^2} \sum_{t=1}^{n-1} (x_t - \bar{x})(x_{t+1} - \bar{x}) \tag{A1}$$

where $\bar{x}$ and s_x^2 are the mean and variance of x_t, $t = 1, \cdots, n$.

1. Forward, above median correlation (r_{af}) is defined as the correlation between above median flows and flows in the subsequent time step. This is calculated by replacing the sum over all t in the expression above by the sum over those t for which x_t is greater than the median flow x_{median}, replacing the s_x^2 by the product of the standard deviations of the above median flows and flows one time step ahead of above median flows, replacing $\bar{x}$ by the mean of the above median and one time step ahead of above median flows, and adjusting n accordingly.
2. Forward, below median correlation (r_{bf}) is the correlation between all below median flows and the flow in the subsequent time steps, calculated in a similar manner with the sum over those t for which $x_t < x_{\text{median}}$.
3. Backward, above median correlation (r_{ab}) is the correlation between above median flows and the preceding time step's flow, calculated in a similar manner with the sum over those t for which $x_{t+1} > x_{\text{median}}$.
4. Backward, below median correlation (r_{bb}) is the correlation between below median flows and the preceding time step's flow, calculated in a similar manner with the sum over those t for which $x_{t+1} < x_{\text{median}}$.

For a linear Gaussian process the above and below pair of correlations in either the forward or backward direction should be the same. Differences indicate nonlinearity or state dependence in the dependence structure. To test the significance of differences between sample correlation coefficients r_1 and r_2, the following test from *Kendall and Stuart* [1979, p. 315] was used. The test is based on the transformation of the correlation coefficient r as

$$z = \frac{1}{2} \log \left(\frac{1+r}{1-r} \right) \tag{A2}$$

The quantity $z_1 - z_2$ is closely normally distributed with zero mean and variance $1/(n_1 - 3) + 1/(n_2 - 3)$, where n_1 and n_2 are the sample sizes, under the null hypothesis that z_1 and z_2 are calculated from sample correlation coefficients from populations with the same correlation coefficient. Therefore the significance test compares $(z_1 - z_2)/[1/(n_1 - 3) + 1/(n_2 - 3)]^{1/2}$ to the standard normal distribution. This test is approximate unless the samples are from independent bivariate normal populations. In section 6 we used this test to investigate the significance of the difference between r_{af} $(= r_1)$ and r_{bf} $(= r_2)$. Sets of above and below median flows are effectively censored samples, inconsistent with the independence assumptions. Nevertheless, an approximate measure of whether these quantities are significantly different can be obtained by use of this test.

Appendix B: Derivation of Model Statistics

We derive here the expected values of selected statistics of the NP1 model. These depend on the observed data x_i, kernel parameters λ and $\mathbf{S}$, and the Gaussian kernel function.

Marginal Distribution of X_t

The marginal density of X_t (denoted $\hat{f}_m(X_t)$) is estimated as

$$\hat{f}_m(X_t) = \int \hat{f}(X_t, X_{t-1})\, dX_{t-1} = \frac{1}{n} \sum_{i=1}^{n} f_G(X_t - x_i,\ \lambda^2 S_{11}) \tag{B1}$$

where

$$f_G(X_t - x_i,\ \lambda^2 S_{11}) = \frac{1}{(2\pi\lambda^2 S_{11})^{1/2}} \exp\left(-\frac{(X_t - x_i)^2}{2\lambda^2 S_{11}} \right) \tag{B2}$$

denotes a Gaussian density function with mean x_i and variance $\lambda^2 S_{11}$. This follows from (6) with $\mathbf{H}$ from (7) and $\mathbf{S}$ expressed as (19). Equation (6) is the sum of n multivariate Gaussians, each of which when integrated over X_{t-1} results in the univariate Gaussian given above. This marginal distribution is used to calculate model mean, covariance, and skewness.

Mean μ' of X_t

This can be evaluated using the marginal distribution in (B1). Since each kernel is symmetric and centered at a data point, the NP1 model mean (μ') is the sample mean:

$$\mu' = E[X_t] = \int X_t \hat{f}_m(X_t)\, dX_t = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{B3}$$

Standard Deviation of X_t

The variance under the NP1 model can be written as

$$\sigma'^2 = E[(X_t - \mu')^2] = \int (X_t - \mu')^2 \hat{f}_m(X_t)\, dt \tag{B4}$$

where the expectation is over the marginal distribution, (B1). Since $\hat{f}_m(X_t)$ from (B1) is a sum of Gaussian probability density functions (pdf's) with individual means x_i, and variances $\lambda^2 S_{11}$ and the x_i have sample variance S_{11}, we get

$$\sigma'^2 = S_{11}(1 + \lambda^2) \tag{B5}$$

Note the inflation in the underlying variance by the factor $(1 + \lambda^2)$.

Lag 1 Correlation

The lag 1 correlation (ρ_1') under the NP1 model is expressed as the ratio

$$\rho_1' = \frac{E[(X_t - \mu')(X_{t-1} - \mu')]}{\sigma'^2} \tag{B6}$$

where expectation is over the joint density estimate in (6). This expression simplifies to

$$\rho_1' = \frac{(1 + \lambda)^2 S_{12}}{(1 + \lambda^2)(S_{11} S_{22})^{1/2}} = r \tag{B7}$$

where r denotes the sample lag 1 correlation:

$$r = \frac{S_{12}}{(S_{11}S_{22})^{1/2}} \tag{B8}$$

Skewness

The coefficient of skewness (γ') under the NP1 model is defined as the ratio

$$\gamma' = \frac{E[(X_t - \mu')^3]}{\sigma'^3} = \frac{\int (X_t - \mu')^3 \hat{f}_m(X_t)\, dt}{\sigma'^3} \tag{B9}$$

where the expectation is over the marginal distribution in (B1). By integrating over the marginal distribution, the numerator can be evaluated as

$$E[(X_t - \mu')^3] = \frac{1}{n}\sum_{i=1}^{n} x_i^3 + 3\lambda^2 S_{11}\frac{1}{n}\sum_{i=1}^{n} x_i - 3\mu'\lambda^2 S_{11} - 3\mu'\frac{1}{n}\sum_{i=1}^{n} x_i^2 + 3\mu'^2\frac{1}{n}\sum_{i=1}^{n} x_i - \mu'^3 \tag{B10}$$

Now recognizing (B3), the second and third terms cancel, and this is equivalent to

$$E[(X_t - \mu')^3] = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu')^3 \tag{B11}$$

The expression for γ' then becomes

$$\gamma' = \frac{(1/n)\sum_{i=1}^{n} (x_i - \mu')^3}{\sigma'^3} = \frac{g}{(1 + \lambda^2)^{3/2}} \tag{B12}$$

where g is the skewness estimator

$$g = \frac{(1/n)\sum_{i=1}^{n} (x_i - \mu')^3}{S_{11}^{3/2}} \tag{B13}$$

and S_{11} is the sample variance. The decrease in skewness is due to the inflation of variance by (B5). The results derived here do not account for the cut and normalize boundary corrections applied.

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 thank Jery Stedinger, Dennis Lettenmaier, and three anonymous reviewers for insightful comments which have improved this work.

References

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

Beard, L. R., *Monthly Streamflow Simulation*, Hydrol. Eng. Cent., U.S. Corps of Eng., Washington, D. C., 1967.

Bendat, J. S., and A. G. Piersol, *Random Data: Analysis and Measurement Procedures*, 2nd ed., John Wiley, New York, 1986.

Benjamin, J. R., and C. A. Cornell, *Probability, Statistics, and Decision for Civil Engineers*, McGraw-Hill, New York, 1970.

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

Chambers, J. M., and T. J. Hastie, *Statistical Models in S*, Wadsworth and Brooks/Cole, Pacific Grove, Calif., 1992.

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.

Efron, B., Bootstrap methods: Another look at the jackknife, *Ann. Stat.*, *7*, 1–26, 1979.

Efron, B., and R. Tibishirani, *An Introduction to the Bootstrap*, Chapman and Hall, New York, 1993.

Eubank, R. L., *Spline Smoothing and Non-Parametric Regression*, Marcel-Dekker, New York, 1988.

Fiering, M. B., *Streamflow Synthesis*, Harvard Univ. Press, Cambridge, Mass., 1967.

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

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.

Hardle, W., and A. W. Bowman, Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bands, *J. Am. Stat. Assoc.*, *83*(401), 102–110, 1988.

Hipel, K. W., A. I. Mcleod, and W. C. Lennox, Advances in Box-Jenkins modeling, 1, Model construction, *Water Resour. Res.*, *13*(3), 567–575, 1977.

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

Kendall, D. R., and J. A. Dracup, A comparison of index-sequential and AR(1) generated hydrologic sequences, *J. Hydrol.*, *122*, 335–352, 1991.

Kendall, M., and A. Stuart, *The Advanced Theory of Statistics*, vol. 2, *Inference and Relationship*, Macmillan, New York, 1979.

Kite, G. W., *Frequency and Risk Analysis in Hydrology*, Water Resour. Publ., Fort Collins, Colo., 1977.

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

Lall, U., and C. W. Miller, An optimization model for screening multipurpose reservoir systems, *Water Resour. Res.*, *24*(7), 953–968, 1988.

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

Lettenmaier, D. P., and S. J. Burges, An operational approach to preserving skew in hydrologic models of long-term persistence, *Water Resour. Res.*, *13*(2), 281–290, 1977.

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

McLeod, A. I., K. W. Hipel, and W. C. Lennox, Advances in Box-Jenkins modeling, 2, Applications, *Water Resour. Res.*, *13*(3), 577–585, 1977.

Pegram, G. G. S., J. D. Salas, D. C. Boes, and V. Yevjevich, *Stochastic Properties of Water Storage*, Colo. State Univ., Fort Collins, 1980.

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., Analysis and modeling of hydrologic time series, in *Handbook of Hydrology*, edited by D. R. Maidment, McGraw-Hill, New York, 1993.

Salas, J. D., and R. A. Smith, Physical basis of stochastic models of annual flows, *Water Resour. Res.*, *17*(2), 428–430, 1981.

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

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

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

Slack, J. R., and J. M. Landwehr, Hydro-climate data network: A U.S. Geological Survey streamflow data set for the United States for the study of climate variations, 1874–1988, *U.S. Geol. Surv. Open File Rep. 92-129*, 1992.

Smith, J. A., Long-range streamflow forecasting using nonparametric regression, *Water Resour. Bull.*, *27*(1), 39–46, 1991.

Smith, L. A., Identification and prediction of low dimensional dynamics, *Physica D*, *58*, 50–76, 1992.
Stedinger, J. R., Estimating correlations in multivariate streamflow models, *Water Resour. Res*., *17*(1), 200–208, 1981.
Stedinger, J. R., and M. R. Taylor, Synthetic streamflow generation, 1, Model verification and validation, *Water Resour. Res*., *18*(4), 909–918, 1982.
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. P. Lettenmaier, and R. M. Vogel, Multisite ARMA(1,1) and disaggregation models for annual streamflow generation, *Water Resour. Res*., *21*(4), 497–509, 1985a.
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, 1985b.
Tasker, G. D., Comparison of methods for estimating low flow characteristics of streams, *Water Resour. Bull*., *23*(6), 1077–1083, 1987.
Thomas, H. A., and M. B. Fiering, Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation, in *Design of Water Resource Systems*, edited by A. Maass et al., pp. 459–493, Harvard Univ. Press, Cambridge, Mass., 1962.
Todini, J., The preservation of skewness in linear disaggregation schemes, *J. Hydrol*., *47*, 199–214, 1980.
Tong, H., *Nonlinear Time Series Analysis: A Dynamical Systems Perspective*, Academic, San Diego, Calif., 1990.
Vogel, R. M., and J. R. Stedinger, The value of stochastic streamflow models in overyear reservoir design applications, *Water Resour. Res*., *24*(9), 1483–1490, 1988.
Wand, M. P., and M. C. Jones, Multivariate plug-in bandwidth selection, *Comput. Stat*., *9*, 97–116, 1994.
Woo, M. K., Confidence intervals of optimal risk-based hydraulic design parameters, *Can. Water Resour. J*., *14*(1), 10–16, 1989.
Yakowitz, S., Nonparametric density estimation, prediction, and regression for markov sequences, *J. Am. Stat. Assoc*., *80*(389), 215–221, 1985.
Yevjevich, V. M., *Stochastic Processes in Hydrology*, Water Resour. Publ., Fort Collins, Colo., 1972.
Zucchini, W., and P. T. Adamson, Bootstrap confidence intervals for design storms from exceedance series, *Hydrol. Sci. J*., *34*(1), 41–48, 1989.

U. Lall and D. G. Tarboton, Utah Water Research Laboratory, Utah State University, Logan, UT 84322-4110. (e-mail: dtarb@cc.usu.edu)

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

(Received October 16, 1995; revised August 12, 1996;
accepted September 18, 1996.)