A nonparametric approach for representing interannual dependence in monthly streamflow sequences

Ashish Sharma

School of Civil and Environmental Engineering, University of New South Wales, Sydney, New South Wales, Australia

Robert O'Neill

Department of Land and Water Conservation, Sydney, New South Wales, Australia

Received 18 September 2001; revised 21 January 2002; accepted 21 January 2002; published 3 July 2002.

[1] The estimation of risks associated with water management plans requires generation of synthetic streamflow sequences. The mathematical algorithms used to generate these sequences at monthly timescales are found lacking in two main respects: inability in preserving dependence attributes particularly at large (seasonal to interannual) time lags and a poor representation of observed distributional characteristics, in particular, representation of strong asymmetry or multimodality in the probability density function. Proposed here is an alternative that naturally incorporates both observed dependence and distributional attributes in the generated sequences. Use of a nonparametric framework provides an effective means for representing the observed probability distribution, while the use of a "variable kernel" ensures accurate modeling of streamflow data sets that contain a substantial number of zero-flow values. A careful selection of prior flows imparts the appropriate short-term memory, while use of an "aggregate" flow variable allows representation of interannual dependence. The nonparametric simulation model is applied to monthly flows from the Beaver River near Beaver, Utah, and the Burrendong dam inflows, New South Wales, Australia. *INDEX TERMS:* 1812 Hydrology: Drought; 1833 Hydrology: Hydroclimatology; 1884 Hydrology: Water supply; 1857 Hydrology: Reservoirs (surface); 1869 Hydrology: Stochastic processes; *KEYWORDS:* stochastic, streamflow, simulation, nonparametric, disaggregation, low-frequency

1. Introduction

[2] An important goal of stochastic hydrology is to generate synthetic streamflow sequences that are statistically similar to the observed flow record. These sequences serve as inputs for Monte Carlo simulation of a reservoir system to help identify plans and policies for efficient management of available water resources. A key requirement in stochastic streamflow simulation is that the generated sequences be "similar" to the observed flows. This implies that the distributional and dependence attributes of observed flows should be accurately reproduced in the simulations. The representation of seasonal to interannual dependencies commonly associated with sustained droughts or periods of large flows is of particular importance for reservoir system management. Absence of such dependencies in simulations can result in an inaccurate representation of the flows that are likely to occur. This can in turn lead to biased reservoir operating policies, causing both loss of revenue in reservoir operation and a possible hazard for users downstream. This paper presents an approach for stochastic simulation of seasonal streamflow sequences that attempts to reproduce such longer-term dependence characteristics and the observed distributional attributes in the generated flow sequences. The approach is developed within a nonparametric density estimation framework that ensures accurate representation of the distributional attributes present in the historical flow record. Use of an aggregate streamflow variable, details of which are presented later, ensures an accurate characterization of the seasonal to interannual dependencies in the model simulations.

[3] Stochastic simulation of seasonal flows has traditionally been approached using two different perspectives. Autoregressive moving average (ARMA) models have been commonly used to model both seasonal and annual streamflow sequences. These models assume that the current flow is linearly related to previous observations. Many times the actual flow values need to be transformed to an alternate variable that conforms well with the assumptions of linearity (or a Gaussian probability density) implicit in the model structure. Use of such a framework offers an accurate representation of the dependence between the current and a few past flow values but does not necessarily ensure that longer-term (seasonal to interannual) dependencies are accurately reproduced.

[4] An alternative to the ARMA models discussed above are stochastic disaggregation approaches [*Mejia and Rouselle*, 1976; *Stedinger and Vogel*, 1984; *Koutsoyiannis and Manetas*, 1996; *Tarboton et al.*, 1998]. Here the stochastic simulation proceeds in two stages. First, an annual flow sequence is generated using an appropriately chosen model, using previous years' flows as the basis to prescribe the observed annual dependence structure. Next, the generated annual or aggregate flow for each year is disaggregated or divided into the various seasonal components. This ensures

0043-1397/02/2001WR000953$09.00

that if the annual flow corresponds to a low flow year, the associated seasonal flows will also represent the same. While this offers a reasonable alternative to ARMA models and also ensures that some measure of interannual dependence is translated to the seasonal flow simulations, the resulting flow sequences offer only an approximate representation of the processes observed in the historical flow record. Some of the disadvantages in the use of ARMA and stochastic disaggregation models for seasonal streamflow simulation are the following.

1. First is representation of over-year dependence. A disaggregation approach is designed to reproduce the dependence structure between the aggregate annual flow and the seasonal components as well as the dependence among the seasonal flow values. However, the dependence between the seasonal flows from one year to the next is not modeled. Of particular concern here is that the first season in each new year bears little resemblance to the flow in the preceding season. Hence, if the pattern at the end of the year indicates the development of a drought, this may be completely reversed in the flow values for the next year. While modifications to the conventional disaggregation model described above have been proposed that reduce the effect of the discontinuity between adjacent years [see *Mejia and Rouselle*, 1976], the problem of ensuring that an accurate dependence structure is maintained in seasonal flows across years is not fully resolved [see *Stedinger and Vogel*, 1984]. An alternative that could solve the problem of representing the dependence across annual to seasonal timescales must involve either a coupled seasonal and annual streamflow simulation approach (such as is proposed by *Koutsoyiannis* [2001]) or focusing solely on the generation of seasonal flows, the generation procedure being altered to impart the interannual dependence structure. While the coupled procedure has advantages, it would still have limitations because of the use of a "water year" in representing the annual flow value.

2. Second is misrepresentation of interannual dependence in seasonal flows due to the water year formulation. Disaggregation models use the water year as the basis for simulating an aggregate annual flow value. While such an approach is essential in the disaggregation modeling framework, the assumption that seasonal flows in the current year are dependent on the aggregate flow in the current water year and weakly dependent on the aggregate flow in the previous water year, may not always be a realistic one. If the aim is to ensure dependence is maintained between annual and seasonal timescales, a more realistic way to ensure this would be to use a moving aggregate flow variable as compared to the static water year aggregate flow that is used to simulate seasonal streamflow in disaggregation models. In other words, one could maintain dependence between the previous year's flow and the current month's flow by formulating a variable that represents the summation of flows over the past 12 months, instead of using a variable that will not change for the simulation of all 12 months of flows as is the case in a disaggregation approach. The approach proposed here uses the above logic and aims to ensure an appropriate representation of annual to seasonal flow dependence, irrespective of the 12 month period being used to estimate the annual flow value. It must, however, be pointed out that use of the past 12 months flow as the basis for representing persistence in monthly flows is simply a convenient assumption, as persistence in flows is introduced through a combination of factors that include variability in rainfall and the resulting variations in catchment storage [*Salas et al.*, 1980].

3. Last is representation of nonlinear dependence and nonstandard probability density functional forms. Traditional approaches for stochastic simulation are often based on rigid assumptions about the form of dependence between the various flow variables or the underlying joint or marginal probability density functions. Such assumptions may not always be valid, as illustrated by *Sharma et al.* [1997]. An added problem is the inability of any of these approaches to represent "unusual" features in the probability density functions of the monthly flow variables. Features difficult to represent include strong asymmetry (or a large positive or negative coefficient of skewness) and multimodality (existence of definite modes that could be representing certain states such as low, medium, and high that the flows could assume). While normalizing transformations can be used to model asymmetry and mixture distributions could be used to model multimodality, these require expert specification and parameterization and can lead to difficulties in representing statistics that are not being directly modeled. Some alternatives that effectively remove the above mentioned problem are suggested by *Lall and Sharma* [1996], *Sharma et al.* [1997], and *Tarboton et al.* [1998]. These approaches are nonparametric and make no prior assumptions about the form of dependence or probability distribution. Use of the data-based framework ensures that resulting simulations have similar dependence and distributional attributes as observed in the historical record, attributes that include representation of zero and very low flow values which are otherwise difficult to model in conventional stochastic models.

[5] Proposed here is a seasonal streamflow generation approach that is free from the disadvantages noted above. Generated sequences reproduce both the short-term as well as interannual dependence present in the historical flows. Use of the nonparametric framework ensures that dependence and distributional attributes in generated flows are similar to those in the historical record. What follows is a brief background on nonparametric methods, their applications in hydrology and water resources, and how they can be used to formulate conditional streamflow simulation models. Next, methodological and algorithmic details on the nonparametric streamflow simulation model proposed here are presented. The model is next applied to two streamflow data sets: 84 years (1914–1998) of monthly streamflows in the Beaver River near Beaver, Utah, and 105 years (1890–1994) of Burrendong dam inflows on the Maquarie River in eastern New South Wales, Australia. We conclude our presentation with a summary of the main features in the proposed approach and the benefits it can bring when used in a water resources planning and management application.

2. Nonparametric Applications for Stochastic Streamflow Generation

[6] The past few years have seen a surge in applications of nonparametric methods for probability density and

regression function estimation to a range of hydrologic problems. Interested readers may refer to *Lall* [1995] for a review. Some of the applications related to the present work are a synthetic streamflow resampling approach using nearest neighbor density estimation principles [*Lall and Sharma*, 1996], a nonparametric alternative to the autoregressive order p model (the NPp, or the nonparametric order p streamflow simulation model) [*Sharma et al.*, 1997], and a nonparametric alternative to traditional disaggregation approaches (the NPD, or the nonparametric disaggregation model) [*Tarboton et al.*, 1998]. Streamflow simulation is an exercise in conditional probability distributions [*Bras and Rodriguez-Iturbe*, 1985]. Simulation of flow X_t conditional to p prior flows $(X_{t-1}, X_{t-2}, \ldots, X_{t-p})$ involves estimation of the conditional probability density function $f(X_t \mid X_{t-1}, X_{t-2}, \ldots, X_{t-p})$. Similarly, disaggregation of an aggregate flow $Z = X_1 + X_2 + \ldots + X_d$ into the d seasonal components $(X_1, X_2, \ldots, X_d)$ requires estimation of the conditional multivariate probability density $f(X_1, X_2, \ldots, X_d|Z)$. Conventional approaches assume certain distributional forms for the joint and marginal probability densities of the flow variables, from which the above conditional probability density functions are derived. These conditional densities are then expressed using parameters such as the mean, variance, and skewness and measures of dependence such as correlation. As these methods rely solely on parameters (mean, variance, skewness, and correlation) of the data to characterize the assumed probability density functions, they are termed parametric. Such methods are useful only if the assumptions about the underlying distributional forms are accurate. One often comes across streamflow records that are not easily characterized by the commonly used probability distributions (see examples of *Lall and Sharma* [1996], *Sharma et al.* [1997], and *Tarboton et al.* [1998]).

[7] Nonparametric methods offer an efficient alternative to traditional parametric approaches. A nonparametric kernel probability density estimate is obtained by considering the cumulative effect of smooth functions called kernels placed over each sample data point. Using a Gaussian kernel function, the multivariate kernel probability density $\hat{f}(\mathbf{X})$ of a d-dimensional variable set $\mathbf{X}$ is estimated as

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

where $\mathbf{x}_i$ is the ith multivariate data point for a sample of size n, $\mathbf{S}$ is the sample covariance of the variable set $\mathbf{X}$, and λ is a smoothing parameter, known as the "bandwidth" of the kernel density estimate.

[8] The bandwidth λ is the key to an accurate estimate of the probability density. A large value of λ results in an oversmoothed probability density, with subdued modes and overenhanced tails. A low value, on the other hand, can lead to density estimates overly influenced by individual data points, with noticeable bumps in the tails of the probability density. Several operational rules for choosing optimal values of the bandwidth λ are available in the literature. This study uses the least squares cross validation (LSCV) approach, the LSCV optimal bandwidth being estimated by a minimization of the following function:

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

where $L_{ij} = 1/\lambda^2(\mathrm{x_i}-\mathrm{x_j})^T\,\mathrm{S}^{-1}(\mathrm{x_i}-\mathrm{x_j})$ and the other variables follow the same notation as before. Readers are referred to *Sharma et al.* [1997] for an application of this method for simulation of monthly flows and *Sharma et al.* [1998] for a discussion and sensitivity analysis of this bandwidth choice.

[9] The nonparametric kernel probability density estimate in equation (1) can be considered to be a mixture of n multivariate Gaussian probability density functions (called kernels), each kernel having a mean equal to the ith sample value $\mathbf{x}_i$, and a covariance matrix equal to $\lambda^2\mathbf{S}$, a scaled version of the overall covariance $\mathbf{S}$ of the sample data set. One may wrongly infer from the above definition of the kernel density estimate that it represents a parametric probability density as Gaussian probability density functions with known means and covariances are being used. This would be an incorrect inference as even though parametric probability density functions are used and parameters such as the bandwith λ and the sample covariance $\mathbf{S}$ need to be estimated, the estimated probability density will represent the relative frequency distribution of the sample and will have little similarity to a Gaussian or any other probability density function unless the sample happens to follow that distribution. In other words, following a definition by *Scott* [1992], the density estimate in equation (1) is nonparametric because the influence of any data point $\mathbf{x}_i$ on the density estimate at $\mathbf{X}$, vanishes asymptotically if $|\mathbf{X} - \mathbf{x}_i| > \varepsilon$ for any $\varepsilon > 0$. This is not true for parametric estimators as a parametrically specified probability density function will depend on the parameters used in specifying the probability density function (sample mean, variance, skewness, or other parameters), which will depend on all data points and not necessarily more so on the ones that are close to $\mathbf{X}$. In simple terms, an estimator is nonparametric if it is asymptotically local, which is not the case with a parametric estimator.

[10] This above difference between parametric and nonparametric methods should, however, not be interpreted as suggesting that nonparametric methods are mutually exclusive competitors to the parametric methods used conventionally in hydrology. In many cases a nonparametric probability density or regression function estimate may suggest that a parametric probability density function or regression equation is appropriate to use. In many cases, again, one would prefer to adopt nonparametric methods where an obvious parametric function does not exist. The authors firmly believe that nonparametric methods should be used in representing uncertainty in physical variables such as streamflow, for which a conventional parametric probability density function cannot be theoretically specified, or found to take the same shape and form irrespective of time step, catchment characteristics, rainfall characteristics, and any other factors that introduce variability. However, as nonparametric methods are sample estimators, one needs to be careful when they are used in situations where there are few sample observations or that the observations are not repre-

sentative of the population and an expert judgment about the population needs to supersede what the data suggests. Care must also be taken to not rely overly on nonparametric methods in situations where many variables are being modeled simultaneously as the effective sample size of a multidimensional variable set needed to maintain the same accuracy as the kernel density estimate of a single variable sample is proportional to the number of data points in the univariate sample raised to the dimension of the multidimensional variable set, the so-called "curse of dimensionality" [*Scott*, 1992]. While the same limitation is valid for parametric methods as well, where the high-dimensional multivariate probability density is assumed instead of estimated from the sample, it is less visible as compared to nonparametric methods, mainly because the latter depend explicitly on the individual observations that constitute the sample. One of the applications in stochastic hydrology where such high dimensional variable sets need to be modeled is disaggregation of annual flows into their monthly components. Our proposed approach for representing the interannual persistence in monthly flows, described in the next section, stems from a need for a method that can accomplish the same result as disaggregation approaches, without requiring the use of excessively high dimension multivariate variable sets. As will be clear after our presentation of the proposed method in section 3, the approach presented manages to remove many other deficiencies associated with conventional disaggregation techniques and offers an alternative that can be used in hydrology whenever representation of interannual or higher-level persistence is important in monthly (or subannual timescale) hydrologic variables.

3. Proposed Approach

[11] This approach is aimed at accurately representing both short-term (month to month) as well as interannual (month to year and year to year) dependence in simulated flows. Consider the flow at time t to be X_t, where t could represent annual, seasonal, or monthly time steps. For example, for monthly flows, X_1, X_2, ..., X_{12} would be the flows for the first 12 months, X_{13}, ..., X_{24} the flows for the next 12 months, and so on. The aggregate flow variable Z_t can then be defined as

$$Z_t = \sum_{j=1}^{m} X_{t-j}, \tag{2}$$

where m is the number of prior flows included in the aggregate variable. This study uses monthly flows and an annual aggregate level ($m = 12$) to formulate the simulation model. The variable Z_t thus represents the annual flow during the past 12 months for the month being simulated, and its use as a conditioning variable enables proper representation of interannual dependence features. Simulation proceeds from the following conditional probability density:

$$\begin{aligned} f\left(X_t \mid X_{t-1}, X_{t-2}, \ldots, X_{t-p}, Z_t\right) &= \frac{f\left(X_t, X_{t-1}, X_{t-2}, \ldots, X_{t-p}, Z_t\right)}{\int f\left(X_t, X_{t-1}, X_{t-2}, \ldots, X_{t-p}, Z_t\right) dX_t} \\ &= \frac{f\left(X_t, X_{t-1}, X_{t-2}, \ldots, X_{t-p}, Z_t\right)}{f_m\left(X_{t-1}, X_{t-2}, \ldots, X_{t-p}, Z_t\right)}, \end{aligned} \tag{3}$$

where $f_m(\,)$ represent the marginal probability density of the variable set. Note that the above conditional probability density is a function of $(p + 1)$ variables: the aggregate flow over the past ($m = 12$) months, Z_t, and the past p months of flows, (X_{t-1}, X_{t-2}, ..., X_{t-p}). While use of the variables (X_{t-1}, X_{t-2}, ..., X_{t-p}) enforces a short-term (till lag p) dependence structure in the simulated flow value, the aggregate variable Z_t ensures that the month to annual dependence pattern is represented more accurately than would be the case with the use of short-term dependence variables alone. Note that in a general application the aggregation level m must be chosen such that it is greater than p, or else the aggregate variable Z_t would become an explicit function of the other predictors (X_{t-1}, X_{t-2}, ..., X_{t-p}). Also note that the conditional probability density in equation (3) has been specified as a function of p prior lags of X_t. One needs to estimate the appropriate value for p in case of a real application using an order selection scheme such as the Akaike information criterion (AIC) [*Akaike*, 1974] or generalized cross validation (GCV) [*Craven and Wahba*, 1979] or a nonparametric measure of partial dependence, the partial mutual information (PMI) criterion [*Sharma*, 2000]. The authors recommend the use of PMI for estimation of the optimal short-term dependence predictors. In regard to the choice of the aggregate variable (or the aggregation period m) we recommend that the aggregate variable be selected by trial and error, with a period m being selected that can result in an acceptable representation of the persistence observed in the historical record. The present application assumes p equal to 1 and m equal to 12 for the sake of simplicity. The conditional density used for simulation then becomes

$$f(X_t \mid X_{t-1}, Z_t) = \frac{f(X_t, X_{t-1}, Z_t)}{f_m(X_{t-1}, Z_t)}. \tag{4}$$

Using the kernel density estimator in equation (1), the conditional density in equation (4) is estimated as

$$\hat{f}(X_t \mid X_{t-1}, Z_t) = \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi\lambda^2 S'}} w_i \exp\left(-\frac{(X_t - b_i)^2}{2\lambda^2 S'}\right), \tag{5}$$

where $\hat{f}(X_t \mid X_{t-1}, Z_t)$ is the conditional probability density estimate, S' is a measure of spread of the conditional probability density, expressed as

$$S' = S_{11} - \begin{bmatrix} S_{12} \\ S_{1z} \end{bmatrix}^T \begin{bmatrix} S_{22} & S_{2z} \\ S_{2z} & S_{zz} \end{bmatrix}^{-1} \begin{bmatrix} S_{12} \\ S_{1z} \end{bmatrix},$$

where the covariance matrix of the variable set (X_t, X_{t-1}, Z_t) is written as

$$\mathrm{Cov}(X_t, X_{t-1}, Z_t) = \begin{bmatrix} S_{11} & S_{12} & S_{1z} \\ S_{12} & S_{22} & S_{2z} \\ S_{1z} & S_{2z} & S_{zz} \end{bmatrix},$$

w_i is the weight associated with each kernel that constitutes the conditional probability density,

$$w_i = \frac{\exp\left(-\frac{1}{2\lambda^2}\begin{bmatrix} (X_{t-1} - x_{i-1}) \\ (Z_t - z_i) \end{bmatrix}^T \begin{bmatrix} S_{22} & S_{2z} \\ S_{2z} & S_{zz} \end{bmatrix}^{-1} \begin{bmatrix} (X_{t-1} - x_{i-1}) \\ (Z_t - z_i) \end{bmatrix}\right)}{\sum_{j=1}^{n} \exp\left(-\frac{1}{2\lambda^2}\begin{bmatrix} (X_{t-1} - x_{j-1}) \\ (Z_t - z_j) \end{bmatrix}^T \begin{bmatrix} S_{22} & S_{2z} \\ S_{2z} & S_{zz} \end{bmatrix}^{-1} \begin{bmatrix} (X_{t-1} - x_{j-1}) \\ (Z_t - z_j) \end{bmatrix}\right)},$$

b_i is the conditional mean associated with each kernel,

$$b_i = x_i + \begin{bmatrix} S_{12} \\ S_{1z} \end{bmatrix}^T \begin{bmatrix} S_{22} & S_{2z} \\ S_{2z} & S_{zz} \end{bmatrix}^{-1} \begin{bmatrix} (X_{t-1} - x_{i-1}) \\ (Z_t - z_i) \end{bmatrix},$$

and x_i and z_i represent observations, z_i being estimated using the 12 prior flows as expressed in equation (2).

[12] The derivation of equation (5) from the multivariate kernel estimator of equation (2) and the conditional probability density in equation (4) is presented in Appendix A. The conditional probability density estimate in equation (5) can be viewed as consisting of n kernels having relative areas equal to weight w_i, centered at b_i, and having a spread proportional to S'. Each of these are slices of the trivariate kernels that constitute the joint probability density of (X_t, X_{t-1}, Z_t), along the conditioning plane specified by (X_{t-1}, Z_t). The weight w_i depends directly on how far the kernel is from the conditioning plane. A smaller weight implies that the kernel is far from the conditioning plane and does not make up a significant proportion of the conditional density estimate. On the other hand, a large w_i implies that kernel i is close to the conditioning plane and constitutes a significant portion of the conditional density estimate. Consequently, simulation will proceed with more emphasis given to the observed data points lying closer to the conditioning plane and lesser emphasis given to the data points that lie farther away.

[13] Synthetic streamflow generation from the conditional density in equation (4) proceeds as follows:

1. Estimate the bandwidth λ and the covariances S_{11}, S_{12}, S_{1z}, S_{22}, S_{2z}, S_{zz}.
2. Start the simulation by arbitrarily assigning values to X_{t-1} and Z_t.
3. Given X_{t-1} and Z_t, estimate the weight w_i associated with each kernel.
4. Pick a data point i with probability w_i.
5. The new value of X_t can now be obtained as $X_t = b_i + \lambda(S')^{1/2} W_t$, where W_t is a Gaussian random variate with zero mean and unit standard deviation.
6. Increment time step t, X_{t-1}, and Z_t.
7. Repeat steps 3–6 as many times as required.

It should be noted that step 4, identifying an observation to sample from, ensures that the sampled value is chosen randomly and allows the simulation to generate a new sequence of observations and that step 5, generating a new realization from the sampled conditional kernel, ensures that the value that is simulated is different than observed data points. The bandwidth λ is important in both defining the weights w_i (a smaller bandwidth would result in a greater weight for observations that are close to the conditioning plane) and in generating the new value (a low bandwidth here would mean that the new realization bears a closer resemblance to the associated data point x_i). In practice, the first few values simulated are discarded to reduce the effect of the arbitrary initialization used. In all results reported, the number of values discarded was set to 120 (the first 10 years of the simulation).

[14] Because of the smooth and symmetric nature of the kernel function used, if an excessive percentage of observed data points lie on or close to the zero flow boundary, it can result in a significant amount of probability to be associated with negative (hence infeasible) values of flow. To get around this problem, a "variable kernel" [*Scott*, 1992] has been used for data points close to the boundary. The bandwidth or the spread of the conditioned kernel slice used for simulating the new flow value (step 5 of the algorithm) is reduced depending on the distance of its center (b_i) from the zero-flow boundary. The modified step 5 of the above algorithm is as follows.

[15] For step 5a, estimate a transformed bandwidth λ' such that

$$\begin{aligned} \lambda' &= \lambda \qquad \text{if } F_{N\left(b_i, \lambda^2 S'\right)}(X_t \leq 0) \leq \alpha \\ &= \lambda' \qquad \text{if } F_{N\left(b_i, \lambda^2 S'\right)}(X_t \leq 0) > \alpha, \text{such that} \\ &\qquad\qquad F_{N\left(b_i, \lambda'^2 S'\right)}(X_t \leq 0) = \alpha, \end{aligned}$$

where $F_{N(\mu, \sigma 2)}$ is the cumulative probability of a normal distribution with mean μ and variance σ^2, with the bandwidth being transformed to λ' if for the selected normal kernel, the probability of the flow X_t being less than or equal to zero is estimated to be greater than a specified threshold α.

[16] For step 5b, sample a new value of X_t as $X_t = b_i + \lambda'(S')^{1/2} W_t$, where W_t is a Gaussian random variate with zero mean and unit standard deviation.

[17] For step 5c, repeat step 5b if the sampled X_t is less than zero until a positive value results.

[18] The rationale behind the use of the above transformation is to leave the bandwidth unaltered if the kernel is far away from the zero-flow boundary but to reduce the bandwidth if that is not the case. If the kernel is far from the boundary, it is unlikely that the cumulative probability for a zero-flow level will be significant and greater than the threshold probability α. If, on the other hand, the kernel is lying close to the boundary, the zero-flow cumulative probability may well exceed the threshold, in which case, λ will have to be modified such that the zero-flow cumulative probability is exactly equal to the threshold α. In the present study, a threshold probability α equal to 0.06 has been used. If a negative flow value is simulated (as would happen in 6% of all cases for kernels lying close to the zero-flow boundary), a new value is sampled from the same kernel until a positive flow results. Note that the use of such a variable kernel ensures that if a significant number of observed data points represent low-flow conditions, the nature of dependence that leads to such low flows in the historical record is naturally enforced in the simulations. If such a procedure were not used, the simulation would proceed in the same manner for both low- and high-flow observations, without recognizing that the variability associated with low-flow values is significantly smaller than that associated with the higher-flow values.

[19] A related outcome of this procedure is that if the historical record contains any zero-flow values and a kernel representing such a zero-flow is selected in step 4 of the algorithm, the transformed bandwidth λ' will be set equal to zero or the simulated value will be set equal to b_i, which in this case, reduces to x_i, which is equal to zero. Hence simulations will contain zero-flow values if the historical data set contains zero flows. On the other hand, the simulation will not contain values exactly equal to zero if

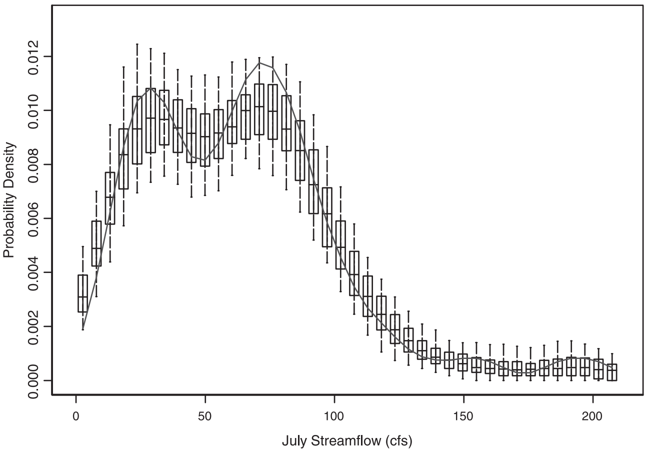

Figure 1. An averaged shifted histogram [*Scott*, 1992] probability density estimate of the July month flows in Beaver River near Beaver, Utah. Box plots represent the probability density function estimates for 100 NPL1 model simulations.

there are no zero flows in the historical record. In our application of this procedure, we have found that using the transformed bandwidth offers an acceptable solution to historical flows containing a significant number of low- and zero-flow values, a solution that is better than the use of a skewness stabilizing transformation, as is used conventionally in most stochastic hydrology applications. Such a procedure is especially important and useful when working with flows from arid regions such as Australia, where one seldom comes across long monthly flow records without any zero-flow occurrences.

[20] Readers should note that the model proposed here is similar to the NP1 model of *Sharma et al.* [1997], except that the proposed model uses an aggregate flow variable in addition to the previous month's flow as the two model predictors. The use of the aggregate flow variable is to impose a longer-term dependence in the simulated flows. Such dependence is missing in the NP1 or any other models that assume a Markov order 1 dependence (the assumption of a Markovian process implying that the variable being modeled is dependent on only a finite set of prior values, which is assumed to equal 1 in case of a Markov order 1 process). To distinguish between the NP1 model of *Sharma et al.* [1997] and the nonparametric simulation model proposed here, the following convention will be used: the NP1 model of *Sharma et al.* [1997] with no long-term dependence will be denoted as before (NP1), whereas the nonparametric model proposed here will be denoted as NPL1 in the discussions that follow.

4. Application to Monthly Streamflow From Beaver River, Utah

[21] Eighty-four years (October 1914 to September 1998) of monthly streamflow data from the Beaver River near Beaver, Utah, (U.S. Geological Survey station number 10234500) was used to test the applicability of the NPL1 simulation model. This station is at 1889.76 m above mean sea level and represents a total catchment area of 235.59 km^2. These data have been used in earlier studies [*Sharma et al.*, 1997; *Tarboton et al.*, 1998] illustrating the use and applicability of nonparametric techniques for reasons evident in the probability density functions for the observed and simulated flows for the month of July shown in Figure 1. The July month streamflow has a clearly bimodal probability density function, which is difficult to model using conventional stochastic techniques. The box plots in Figure 1 represent the variability in the probability density function of 100 flow sequences, each of an 84 year length, simulated using the nonparametric simulation model described earlier. As used here, a box plot consists of a

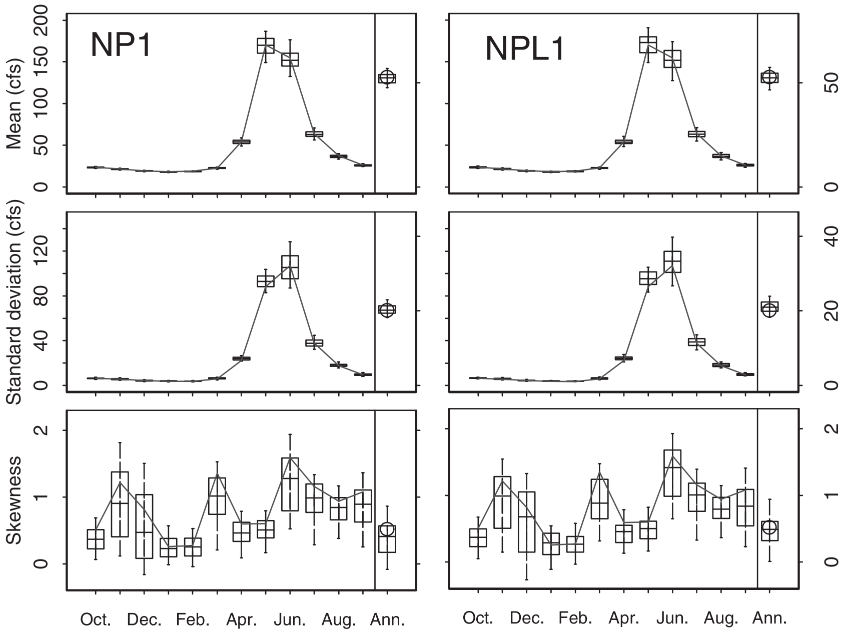

Figure 2. Summary statistics of NP1 and NPL1 model simulations of the Beaver River monthly flow record.

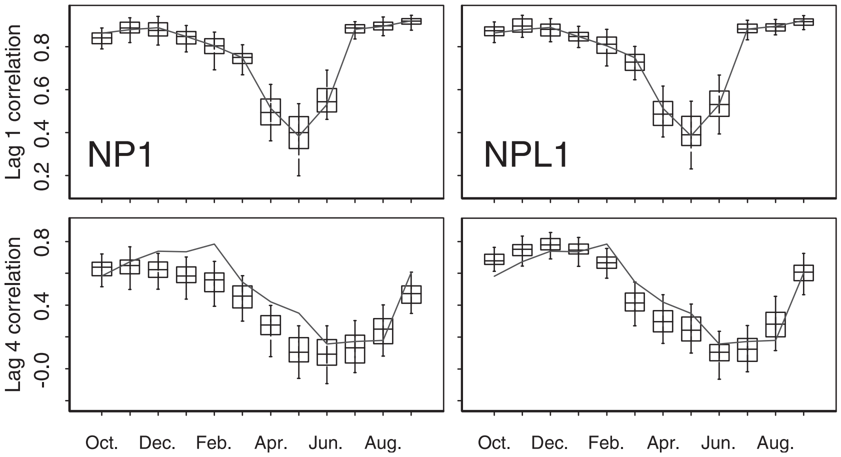

Figure 3. Lag 1 and 4 serial correlations for Beaver River monthly flows.

box that represents the 25th and 75th percentiles, a horizontal line within the box that represents the median, and whiskers that extend to the 5th and 95th percentiles of the simulated statistic. As one can infer from Figure 1, the nonparametric model is able to represent the bimodality is a reasonably accurate manner. This would not have been possible through the use of conventional parametric approaches unless a specific assumption about the nature of the bimodality (specified possibly as a mixture of two Gaussian probability density functions with appropriately estimated means, variances, and mixture weights) was made. On the other hand, one may argue that the bimodality may not be statistically significant; hence not representing it in a parametric framework would be appropriate. Such an argument does not have a meaning in a nonparametric approach: the nonparametric simulation model "letting the data speak for themselves" [*Wand and Jones*, 1995] rather than an expert intervention being used.

[22] The monthly mean, standard deviation and coefficient of skewness of the observed and simulated flows are illustrated in Figure 2. Statistics for both models (NP1 and NPL1) are shown. All statistics were reproduced well by both models, as is expected in theory [see *Sharma et al.*, 1997, Appendix B].

[23] Figure 3 illustrates the lag 1 and lag 4 serial (or auto) correlations of the observed and simulated samples from both NP1 and NPL1 models. The use of the aggregate flow variable (the running sum of the last 12 months of flow) in NPL1 enables higher lag correlations to be represented better than the NP1 model simulations. This is all the more notable as the model was not designed to ensure accurate reproduction of these higher lag correlations. One would need to have a higher-order Markov dependence structure in a conventional stochastic simulation model (an autoregressive lag 4 (AR4) or its nonparametric equivalent, the NP4 model) to achieve the same results.

[24] Figure 4 illustrates the correlation between 1 month's flow and the sum of the previous 12 months flows, for lags of 1 and 4 months. As would be expected, the NPL1 model reproduces these correlations significantly better than the NP1 model for a lag of 1 month. A similar result is observed at a lag of 4 months. The NP1 model results are not as encouraging as would be expected for any Markov order 1 dependence model. This is an important result as it indicates that longer-term dependence is being properly represented in the NPL1 model simulations.

[25] To understand better how well the model reproduces dependence statistics at an annual level, annual flows were estimated by adding the monthly simulated flow values for each water year. Annual flow lag 1 autocorrelations for the historical and simulated flow values are presented in Table 1. Note how well the annual lag 1 correlations are simulated

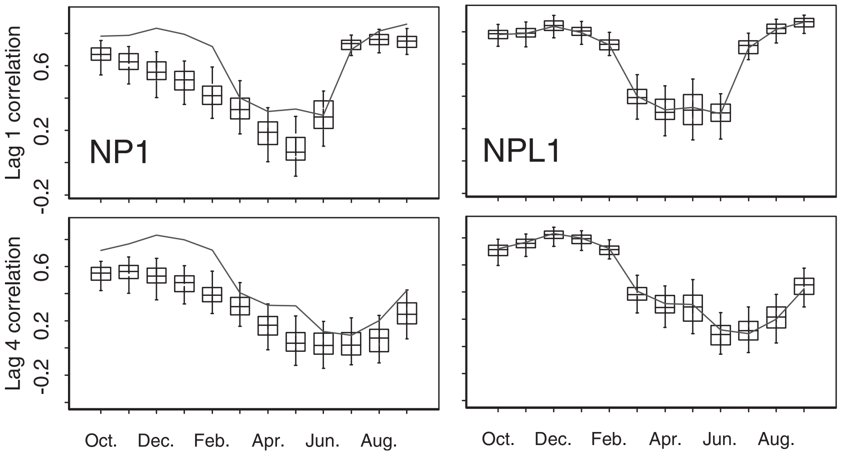

Figure 4. Lag 1 and 4 month to 12 month aggregate flow correlations for Beaver River monthly flows.

Table 1. Lag 1 Autocorrelations for Observed and Simulated Beaver River Annual Flow Values

Historical Flows	Percentile	NP1	NPL1
0.307	25th	0.030	0.209
0.307	median	0.101	0.309
0.307	75th	0.172	0.360

by the NPL1 model. This is an important result as the model has not been designed to ensure the proper representation of dependence at the annual timescale. It should also be noted that while the water year was used to estimate annual flow values in presenting the annual lag 1 correlation results, the results are likely to be as good if a different sequence of 12 months had been used instead. This is an important aspect of NPL1 model simulations that would be difficult to represent in conventional monthly simulation or annual-to-monthly disaggregation models.

[26] The reservoir storage volumes required to sustain constant monthly demands were estimated using the sequent peak algorithm. These storages are illustrated in Figure 5 for specified fractions of the demand and the mean annual flow. While both models perform reasonably well, the bias and variance of the storages calculated for the NPL1 model simulations is smaller than that for the NP1 model simulations.

5. Application to Burrendong Dam Inflows, New South Wales, Australia

[27] The nonparametric simulation model was next applied to 105 years (1890–1994) of reservoir inflows to the Burrendong dam in eastern New South Wales, Australia. The Burrendong dam is located on the Maquarie River and has an approximate catchment area of 7500 km^2. While flow data have been measured since the opening of the dam in 1967, the earlier periods of record have been estimated by the New South Wales Department of Land and Water Conservation using the observed rainfall record and a rainfall-runoff model. It was assumed that both the pre-1967 estimated streamflow and the post-1967 observed streamflow are representative of the underlying flow series, even though it is well known that the use of a rainfall-runoff model in estimating the flows is likely to reduce the variability that would normally be present. This streamflow data poses many problems to the stochastic modeler. First, there are several instances where the flow has stayed at fairly low levels for 6–10 months at a stretch. Second, there are several “zeroes” in the flow record, which always pose a few challenges when prescribing a continuous probability density function. Last, this river is known to be susceptible to prolonged droughts, leading to long periods of very low flows (the minimum 11 and 12 month average flows are 0.8 and 2.2% of the mean annual flow, respectively).

[28] One hundred realizations each 105 years long were simulated using the two nonparametric stochastic streamflow generation models. A comparison of the monthly flow statistics of both NP1 and NPL1 model simulations indicated results similar to what have been reported for the case of the Beaver River data. These results are not presented here but are available on request from the authors. A comparison of some of the statistics of the annual flow volumes (summation of monthly flows over the water year) is presented in Table 2. While both models are able to model the annual mean flow reasonably well, the NP1 simulations are unable to reproduce the observed annual flow standard deviation, coefficient of skewness, or lag 1 correlation. This is an important result that illustrates the ability of the NPL1 model to ensure that simulated flow values preserve distributional and dependence attributes at both monthly and annual scales. We wish to reemphasize that the results would be just as good if a different sequence of 12 months were used in estimating the annual flow values.

[29] Figure 6 illustrates the reservoir storages estimated on the basis of NP1 and NPL1 model simulations using the sequent peak algorithm. The NPL1 model simulations lead to reservoir storage volumes that are close to those estimated on the basis of the historical flow record. NP1 model simulations, however, lead to substantially smaller storage volumes than those suggested by the historical flows. These smaller storage volumes are due to the lack of a longer-term dependence structure in the NP1 model simulations and have obvious implications for any water management applications they might be used for.

[30] Both models were also tested for their ability to simulate flows that would be likely in the event of a sustained drought. Figure 7 presents the lowest monthly flow rate as a function of duration. It is interesting that both models are able to properly simulate observed low-flow sequences for durations longer than 11 months. Neither model performs as well at simulating the worst 11 month drought on record. We feel this could be due to a number of factors, which include the simplistic model order (p = 1 and m = 12) that has been assumed in this study.

6. Summary

[31] A synthetic streamflow generation model was presented that was capable of modeling both short and longer-

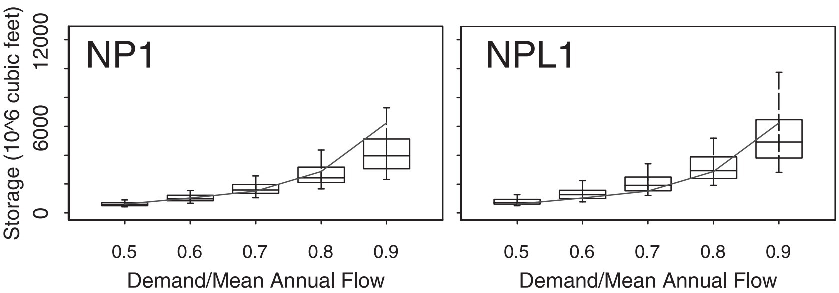

Figure 5. Beaver River reservoir storage volume estimates for meeting seasonally nonvarying demands.

Table 2. Comparison of Observed and Simulated Burrendong Dam Annual Inflow Statistics

Statistic	Observed	Percentile	NP1	NPL1
Mean, ML	1,096,697	25th	1,049,526	953,064
Mean, ML	1,096,697	median	1,086,588	1,022,806
Mean, ML	1,096,697	75th	1,159,303	1,102,851
Standard deviation, ML	1130,626	25th	830,305	881,642
Standard deviation, ML	1,130,626	median	928,528	995,120
Standard deviation, ML	1,130,626	75th	1,007,665	1,170,716
Skewness	3.01	25th	1.56	2.18
Skewness	3.01	median	1.91	2.91
Skewness	3.01	75th	2.13	3.31
Lag 1 correlation	0.114	25th	−0.040	0.049
Lag 1 correlation	0.114	median	0.015	0.090
Lag 1 correlation	0.114	75th	0.078	0.134

term dependencies as well as nonstandard probability density functional forms. This model was tested on two monthly streamflow data sets representing very different climatological and topographical regimes. The results indicated that the proposed model was able to represent longer-term dependence in a better way as compared to conventional streamflow simulation approaches. The improved representation of the longer-term dependence leads to significant improvements in the representation of reservoir storage characteristics. It should be noted that the results presented here were based on arbitrary choices for the number and type of variables used to represent the short- and long-term dependence. It is likely that results will improve further if these variables are chosen on the basis of nonparametric partial dependence measures such as the partial mutual information [*Sharma*, 2000] coupled with an elaborate sensitivity analysis. It should also be noted that it is not necessary to couple the rationale behind the proposed approach (using aggregate flow in addition to prior flows for conditional simulation of monthly streamflow) with nonparametric approaches alone. Even though the authors recommend that nonparametric methods be used because of their natural simplicity and their faithfulness to the observed flow record, a similar logic could well be used in formulating an appropriately structured parametric streamflow simulation model.

Appendix A. Derivation of Kernel Estimate of the Conditional Probability Density Function

[33] Here we derive the kernel estimate of the conditional probability density function given in equation (5) from the joint probability density estimate in equation (2). For the variable set (X_t, X_{t-1} and Z_t) the joint probability density estimate of equation (2) can be expressed as

$$\hat{f}(X_t, X_{t-1}, Z_t) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(2\pi)^{3/2}\lambda^3 \det(\mathbf{S})^{1/2}} \cdot \exp\left(- \frac{\begin{pmatrix} X_t - x_i \\ X_{t-1} - x_{i-1} \\ Z_t - z_i \end{pmatrix}^T \mathbf{S}^{-1} \begin{pmatrix} X_t - x_i \\ X_{t-1} - x_{i-1} \\ Z_t - z_i \end{pmatrix}}{2\lambda^2} \right),$$

where

$$\mathbf{S} = \begin{bmatrix} S_{11} & S_{12} & S_{1z} \\ S_{12} & S_{22} & S_{2z} \\ S_{1z} & S_{2z} & S_{zz} \end{bmatrix},$$

which is equivalent to the scaled sum of n normal probability density functions:

$$\hat{f}(X_t, X_{t-1}, Z_t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{N}_3\left(\begin{pmatrix} x_i \\ x_{i-1} \\ z_i \end{pmatrix}^T, \lambda^2\mathbf{S} \right), \qquad \text{(A1)}$$

where $\mathbf{N}_3(\mu, \Sigma)$ represents a trivariate normal probability density function with mean μ and covariance matrix Σ.

[34] Using known relationships of a multivariate normal distribution [*Mardia et al.*, 1988, theorems 3.2.3 and 3.2.4,

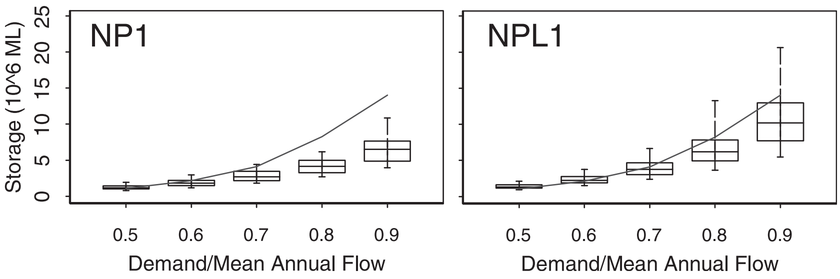

Figure 6. Burrendong dam reservoir storage volume estimates for meeting seasonally nonvarying demands.

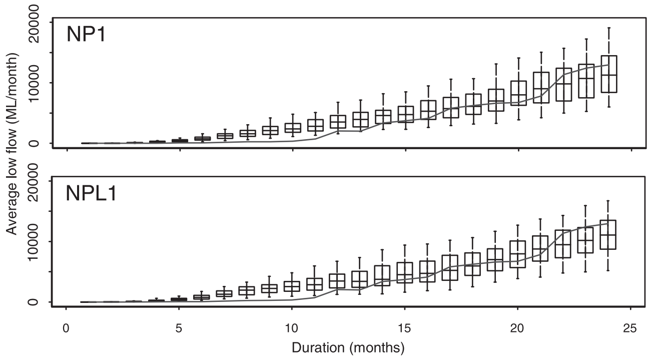

Figure 7. Variation of observed and simulated low-flow values as a function of duration for the Burrendong dam monthly inflow record.

p. 63], the joint probability density estimate in equation (A1) can be further simplified to the following:

$$\hat{f}(X_t, X_{t-1}, Z_t) = \frac{1}{n}\sum_{i=1}^{n}\left[N\left(b_i, \lambda^2 S'\right) \times \mathbf{N}_2\left(\begin{pmatrix} x_{i-1} \\ z_i \end{pmatrix}^T, \lambda^2 \begin{bmatrix} S_{22} & S_{2z} \\ S_{2z} & S_{zz} \end{bmatrix}\right)\right], \qquad (A2)$$

where

$$b_i = x_i + \begin{bmatrix} S_{12} \\ S_{1z} \end{bmatrix}^T \begin{bmatrix} S_{22} & S_{2z} \\ S_{2z} & S_{zz} \end{bmatrix}^{-1} \begin{bmatrix} (X_{t-1} - x_{i-1}) \\ (Z_t - z_i) \end{bmatrix}$$

$$S' = S_{11} - \begin{bmatrix} S_{12} \\ S_{1z} \end{bmatrix}^T \begin{bmatrix} S_{22} & S_{2z} \\ S_{2z} & S_{zz} \end{bmatrix}^{-1} \begin{bmatrix} S_{12} \\ S_{1z} \end{bmatrix}.$$

Similar to equation (A1), the joint probability density estimate of (X_{t-1}, Z_t) can be written as

$$\hat{f}(X_{t-1}, Z_t) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{N}_2\left(\begin{pmatrix} x_{j-1} \\ z_j \end{pmatrix}^T, \lambda^2 \begin{bmatrix} S_{22} & S_{2z} \\ S_{2z} & S_{zz} \end{bmatrix}\right). \qquad (A3)$$

The conditional probability density in equation (4) can now be stated as the ratio of equations (A2) and (A3):

$$f(X_t \mid X_{t-1}, Z_t) = \frac{f(X_t, X_{t-1}, Z_t)}{f_m(X_{t-1}, Z_t)}$$

or

$$\hat{f}(X_t \mid X_{t-1}, Z_t) = \frac{\sum_{i=1}^{n}\left[N\left(b_i, \lambda^2 S'\right) \times \mathbf{N}_2\left(\begin{pmatrix} x_{i-1} \\ z_i \end{pmatrix}^T, \lambda^2 \begin{bmatrix} S_{22} & S_{2z} \\ S_{2z} & S_{zz} \end{bmatrix}\right)\right]}{\sum_{j=1}^{n} \mathbf{N}_2\left(\begin{pmatrix} x_{-1} \\ z_j \end{pmatrix}^T, \lambda^2 \begin{bmatrix} S_{22} & S_{2z} \\ S_{2z} & S_{zz} \end{bmatrix}\right)}$$

$$= \sum_{i=1}^{n} w_i \, N\left(b_i, \lambda^2 S'\right)$$

or

$$\hat{f}(X_t \mid X_{t-1}, Z_t) = \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi\lambda^2 S'}} w_i \exp\left(-\frac{(X_t - b_i)^2}{2\lambda^2 S'}\right), \qquad (A4)$$

where

$$w_i = \frac{\mathbf{N}_2\left(\begin{pmatrix} x_{i-1} \\ z_i \end{pmatrix}^T, \lambda^2 \begin{bmatrix} S_{22} & S_{2z} \\ S_{2z} & S_{zz} \end{bmatrix}\right)}{\sum_{j=1}^{n} \mathbf{N}_2\left(\begin{pmatrix} x_{j-1} \\ z_j \end{pmatrix}^T, \lambda^2 \begin{bmatrix} S_{22} & S_{2z} \\ S_{2z} & S_{zz} \end{bmatrix}\right)}$$

$$= \frac{\exp\left(-\frac{1}{2\lambda^2}\begin{bmatrix} (X_{t-1} - x_{i-1}) \\ (Z_t - z_i) \end{bmatrix}^T \begin{bmatrix} S_{22} & S_{2z} \\ S_{2z} & S_{zz} \end{bmatrix}^{-1} \begin{bmatrix} (X_{t-1} - x_{i-1}) \\ (Z_t - z_i) \end{bmatrix}\right)}{\sum_{j=1}^{n} \exp\left(-\frac{1}{2\lambda^2}\begin{bmatrix} (X_{t-1} - x_{j-1}) \\ (Z_t - z_j) \end{bmatrix}^T \begin{bmatrix} S_{22} & S_{2z} \\ S_{2z} & S_{zz} \end{bmatrix}^{-1} \begin{bmatrix} (X_{t-1} - x_{j-1}) \\ (Z_t - z_j) \end{bmatrix}\right)}.$$

[36] **Acknowledgments.** The authors wish to acknowledge the constructive comments of the two anonymous Water Resources Research reviewers and a useful suggestion by Sri Srikanthan.

References

Akaike, H., A new look at the statistical model identification, *IEEE Trans. Autom. Control*, *AS-19*, 716–723, 1974.

Bras, R. L., and I. Rodriguez-Iturbe, *Random Functions and Hydrology*, Dover, Mineola, New York, 1985.

Craven, P., and G. Wahba, Smoothing noisy data with spline functions, *Numer. Math.*, *31*, 377–403, 1979.

Koutsoyiannis, D., Coupling stochastic models of different time scales, *Water Resour. Res.*, *37*, 379–391, 2001.

Koutsoyiannis, D., and A. Manetas, Simple disaggregation by accurate adjusting procedures, *Water Resour. Res.*, *32*, 2105–2117, 1996.

Lall, U., Recent advances in nonparametric function estimation, *Rev. Geophys.*, *33*, 1093–1102, 1995.

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

Mardia, K. V., J. T. Kent, and J. M. Bibby, *Multivariate Analysis*, Academic, San Diego, Calif., 1988.

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

Salas, J. D., J. W. Delleur, V. Yevjevick, 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 Visualisation*, John Wiley, New York, 1992.

Sharma, A., Seasonal to interannual rainfall probabilistic forecasts for improved water supply management, 1, A strategy for system predictor identification, *J. Hydrol.*, *239*, 232–239, 2000.

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

Sharma, A., U. Lall, and D. G. Tarboton, Kernel bandwidth selection for a first order nonparametric streamflow simulation model, *Stochastic Hydrol. Hydraul.*, *12*, 33–52, 1998.

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

Tarboton, D. G., A. Sharma, and U. Lall, Disaggregation procedures for stochastic hydrology based on nonparametric density estimation, *Water Resour. Res.*, *34*, 107–119, 1998.

Wand, M. P., and M. C. Jones, *Kernel Smoothing*, Chapman and Hall, New York, 1995.

R. O'Neill, Department of Land and Water Conservation, Parramatta, New South Wales 2150, Australia.

A. Sharma, School of Civil and Environmental Engineering, University of New South Wales, Sydney, New South Wales 2052, Australia. (a.sharma@unsw.edu.au)