Stochastic Hydrology and Hydraulics 11: 65-93 © Springer-Verlag 1997

Multivariate nonparametric resampling scheme for generation of daily weather variables

B. Rajagopalan
Lamont-Doherty Earth Observatory, Columbia University, POB 1000, Rt. 9W, Palisades, NY 10964, USA

U. Lall, D. G. Tarboton and **D. S. Bowles**
Utah Water Research Lab., Utah State University, Logan, UT 84322-8200, USA

Abstract: A nonparametric resampling technique for generating daily weather variables at a site is presented. The method samples the original data with replacement while smoothing the empirical conditional distribution function. The technique can be thought of as a smoothed conditional Bootstrap and is equivalent to simulation from a kernel density estimate of the multivariate conditional probability density function. This improves on the classical Bootstrap technique by generating values that have not occurred exactly in the original sample and by alleviating the reproduction of fine spurious details in the data.Precipitation is generated from the nonparametric wet/dry spell model as described in Lall et al. [1995]. A vector of other variables (solar radiation, maximum temperature, minimum temperature, average dew point temperature, and average wind speed) is then simulated by conditioning on the vector of these variables on the preceding day and the precipitation amount on the day of interest. An application of the resampling scheme with 30 years of daily weather data at Salt Lake City, Utah, USA, is provided.

Key words: Nonparametric, Monte Carlo, precipitation, weather.

1 Introduction

Daily weather variations influence agricultural and engineering management decisions. Crop yields and hydrological processes such as runoff and erosion are very sensitive to weather. Recognizing the inherent variability in climate, it is often necessary to assess management scenarios for a number of likely input sequences. Stochastic models are consequently useful for simulating weather scenarios. Such models need to simulate sequences that are representative of the data. While there is substantial literature for rainfall simulation and for other variables one at a time, only a few "multivariate" models have been developed.

In this paper, we develop and exemplify nonparametric procedures for resampling a vector of daily weather variables, such that selected lag 0 and lag 1 dependence characteristics are preserved. Dependence is defined in terms of joint or conditional probabilities, rather than correlation.

This work is an offshoot of the ongoing Water Erosion Prediction Project (WEPP) of the United States Department of Agriculture (USDA). WEPP is a key model for soil and forest conservation studies. WEPP includes a Climate Generator (CLIGEN), and the work presented here intends to improve it. Hill slope erosion is driven largely by precipitation and a suite of other weather variables. Hence, the main objective is to generate weather sequences which will be used by WEPP to estimate hill slope erosion. In this study, we chose a set of five daily variables (total daily solar radiation (SRAD, Langleys), maximum temperature (TMX, °F), minimum temperature (TMN, °F), average wind speed (WSPD, m/sec), and average dew point temperature (DPT, °F) in addition to precipitation (P, inches), that are of interest for erosion prediction. Most of these weather variables are sensitive to precipitation. Solar radiation, dew point temperature, maximum temperature, and minimum temperature are more likely to be below normal on rainy days than on dry days, while the wind speed may be above normal on rainy days than on dry days. Consequently, precipitation is chosen as the driving variable of the models developed so far. Typically [see Jones et al., 1972; Nicks and Harp, 1980; Richardson, 1981], daily precipitation is generated independently and the other variables are generated by conditioning on precipitation events (i.e., whether a day is wet or dry).

Throughout this paper, we denote the historical time series of the five weather variables chosen above as $[z]_{mkj}$ $(m = 1, \cdots, NY, k = 1, \cdots, 366, j = 1, \cdots, NV)$, where NY is the number of years of record, NV(=5) is the number of variables considered (SRAD, TMX,TMN, DPT, and WSPD). Further, define $[\bar{\mathbf{Z}}]_{kj}$ and $[\mathbf{STD}]_{kj}$ as the corresponding mean and standard deviation vector for each calendar day k (k=1,$\cdots$,366) of each variable j (j=1,$\cdots$,5). The historical time series of the precipitation is denoted as $[P]_{mk}$.

We now discuss key attributes of some strategies for resampling or synthesizing vectors of these variables.

Resampling approaches

Multivariate stochastic simulation of weather variables has not been studied as extensively as streamflow or precipitation. Two broad approaches that are possible are:

1. Parametric

2. Nonparametric - Bootstrap (Raw, Conditional, and Smoothed)

1. Parametric. The parametric approach is the traditional method [see Jones et al., 1972; Bruhn et al., 1980; Nicks and Harp, 1980; Lane and Nearing, 1989; Richardson, 1981] for stochastic daily weather simulations. Figure 1 summarizes the general structure of the parametric approach. The general strategy is to generate precipitation independently and the other variables conditioned on the status of precipitation (i.e., rain or no rain on the day). The other variables are generated from either independent statistical distributions fitted separately to each of the variables for each of the two precipitation states (i.e., rain, no rain). Independently or jointly fitted auto regressive models of order 1 (AR-1) are sometimes used.

Usually the year is divided into periods (seasons), and moments (i.e., mean standard deviation and skew) are calculated for each variable for each period for each precipitation state. The moments are used to fit statistical distributions or models.

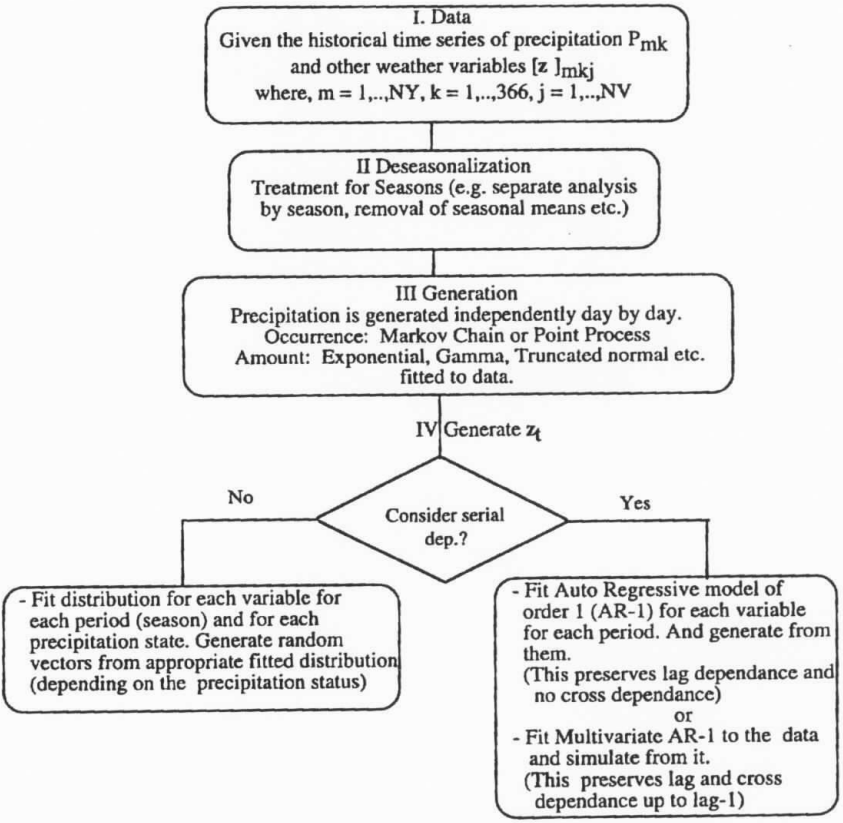

Figure 1. General structure of parametric approach.

Dividing the year into various periods assumes homogeneity within each period and offers a treatment of seasonality. Jones et al. [1972], Bruhn et al. [1980], Nicks and Harp [1980], and CLIGEN [Lane and Nearing, 1989] divide the year into 14-day and one-month periods, respectively, in their work. Richardson [1981] adopted a method wherein the means and standard deviations of each period and each precipitation state are smoothed using the Fourier series. The smoothed daily values of the means and standard deviations are subsequently used for deseasonalization.

Daily precipitation is typically generated from a fitted first-order Markov Chain for precipitation occurrence and by sampling from the distribution (such as Gamma, Exponential, Truncated Normal, etc.) fitted for the daily precipitation amounts for each period.

One approach to generate the other variables is to fit distributions independently for each variable for each period and for each precipitation state. Here, the simulations are made under the assumption that each variable is independent and identically distributed (i.i.d). This approach and its variants are used by Jones et al. [1972], Bruhn et al. [1980], and CLIGEN [Lane and Nearing, 1989]. In CLIGEN, each variable is assumed to be an independent Gaussian variable for each month, with parameters dependent on the precipitation state transition (e.g., wet to wet, dry to wet, etc.). This approach does not consider the dependence between the variables or the serial dependence for each variable.Only the dependence on the precipitation state or the precipitation transition is considered.

Serial dependence was incorporated by Nicks and Harp [1980] who fit Auto Regressive models of order one (AR-1) independently to each variable for each period. Consideration of dependence across variables is added by Richardson [1981] who used a Multivariate Auto Regressive model of order one (MAR-1). When the cross dependence terms are neglected in MAR-1, it reduces to an AR-1 process. These AR models suffer from the drawback of assuming the data to be normally distributed. As a result, only linear dependence can be reproduced. In practice, changes in the weather variables relative to a change in precipitation or other weather variables are not proportional, and the assumption of linearity is questionable. Transformation of the data to be multivariate normal may be difficult and may lead to biased statistics upon transforming back to the original space.

The parametric approaches discussed have three main drawbacks, which are (i) choice of a model (i.e., a statistical distribution or the order) is often subjective and rarely formally tested on a site by site basis; (ii) reliance on an implicit Gaussian framework (e.g., AR or MAR) which preserves only linear dependence and is not appropriate for bounded variables; and (iii) the fitted models have limited portability in the sense that procedures/distributions used at one site may not be best at other sites. The last point is important where an agency wishes to prescribe a uniform procedure over its domain.

2. Nonparametric. Nonparametric techniques do not require preselected distributions or models to be fit to data. The Bootstrap (or raw Bootstrap) is a nonparametric technique introduced by Efron [1979]. It is often used for constructing a confidence region, attaching a standard error to an estimate, carrying out a test of a hypothesis, or estimating the sampling distribution of some statistic. Historical data are resampled with replacement. Since it is the same data, the simulations by construction

have the same distributional properties as that of the historical data. Since each resampled observation is drawn independently, serial dependence is not preserved. Serial dependence can be accommodated by using the '*block-resampling scheme*' (a conditional Bootstrap) developed by Kunsch [1989] and Liu and Singh [1988]. Here a block of 'k' observations is resampled as opposed to a single observation in the Bootstrap. Serial dependence is preserved within, but not across a block. The block length 'k' determines the order of the serial dependence that can be preserved.

A property of the Bootstrap technique is that the simulated samples will only have values that have occurred in the historical data, and, consequently, the simulations are restricted to the historical set of values. Silverman [1986, p. 142] points out that this behavior may reproduce spurious fine structure in the original data. This is not a desirable feature when applying the technique to simulation of daily weather variables, where we may wish to have simulated values that have not been observed in the historical data and may be also beyond the maximum/minimum of the observed data. This problem can be alleviated by using a 'smoothed Bootstrap.'

In the smoothed Bootstrap [Silverman, 1986, p. 144], each observation $y_i (i = 1, \cdots, n)$ is considered to be representative of a region $(y_i - h, y_i + h)$ around it. The extent of this region h is called the bandwidth and is determined from the data. Intuitively, it is desirable to resample such that the maximum weight is given to the observation y_i and weights decrease when moving towards $y_i - h$ or $y_i + h$. This is accomplished by having a weight function centered at each observation. The weight function is usually chosen to be a valid probability density function, such as the Gaussian (N(0,1)). The simulation proceeds by picking an observation y_i with replacement from $\{y_i, \cdots, y_n\}$ and then generating a value from $N(y_i, h)$ with h specified. Formally, the smoothed Bootstrap is equivalent to resampling from a kernel density estimate (k.d.e.).

In this paper, we develop a smoothed conditional Bootstrap that considers multivariate and serial dependence among the variables of interest. Hereafter, we refer to the scheme presented as the NP model. We first provide the motivation and main ideas of the NP model. The simulation algorithm is outlined next. The utility of the model is then illustrated through application to daily weather data at Salt Lake City, Utah, USA. In related work, Sharma et al. [1995] describe the application of the NP model to simulation of monthly streamflow.

2 Main ideas of the NP model

Our goal is to develop an approach that is driven directly by the observed data with reasonable assumptions, is easy to implement, is readily transferable from site to site, and captures the relative frequencies of the data in a natural manner. We do this by defining the appropriate probability densities that we need to resample from and then discuss their estimation.

2.1 Overview of the NP model

A conceptual flowchart of the model is shown in Figure 2. The historical data of the weather variables other than precipitation are standardized as $[\mathbf{x}]_{lkj} = ([\mathbf{z}]_{lkj} - [\bar{\mathbf{Z}}]_k)/[\mathbf{STD}]_{kj}$, where l, k, and j are the same as defined earlier. This removes the seasonality from each variable. Precipitation for day 't' (P_t) is generated from the wet/dry spell model as described in Lall et al. [1995] that is

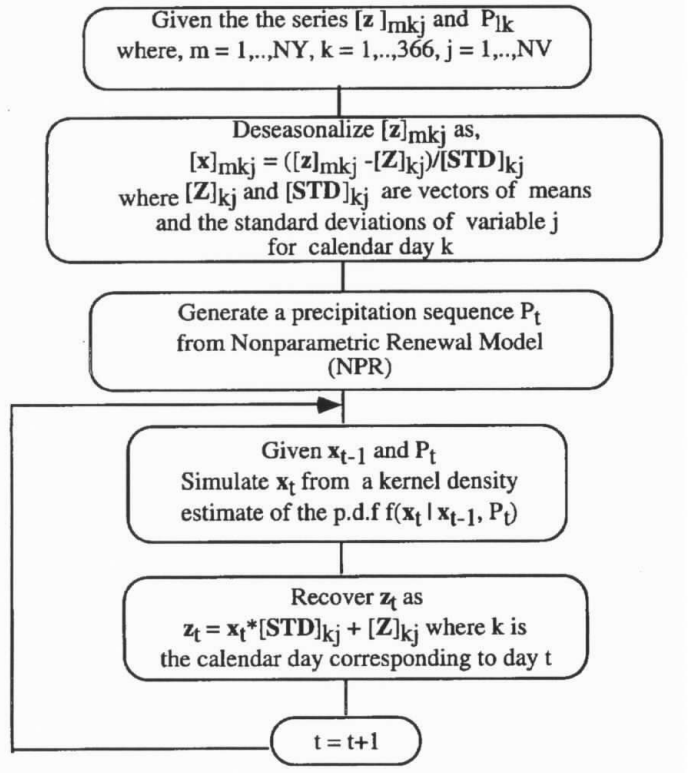

Figure 2. Overview of development of the NP model.

briefly summarized later in this paper. However, the user can generate daily precipitation from any other model that is considered appropriate.

In the NP model, the year is divided into four periods or seasons (for the Salt Lake City example, these are Season 1 (Jan-Mar), Season 2 (Apr- Jun), Season 3 (Jul-Sep), Season 4 (Oct-Dec)). Simulations for days in any particular period are made using the historical data of that period. Subsequently, the comparison between the simulations and the historical data is also made by season. One could choose different periods (e.g., monthly, weekly, etc.). We chose the above four periods so as to be consistent with the wet/dry spell model [Lall et al. 1995] for daily precipitation.

The aim of the model is to capture the day-to-day dependence present between the variables. The standardized vector of variables $\mathbf{x}_t$ for any day 't' is simulated from the multivariate conditional p.d.f. $f(\mathbf{x}_t|\mathbf{V}_t)$. Here, $\mathbf{x}_t$ is a standardized vector of [SRAD,TMX, TMN, WSPD, DPT]$_t$ of length d(=5) that is to be generated for day t, P_t is the generated precipitation for day t from the wet/dry spell model, and $\mathbf{V}_t = [\mathbf{x}_{t-1}, P_t]$ is the conditioning vector of length d'(=6). The joint density is estimated in a space of dimension dg (=d+d').

The conditional density $f(\mathbf{x}_t|\mathbf{V}_t)$ is defined as

$$f(\mathbf{x}_t|\mathbf{V}_t) = \frac{f(\mathbf{x}_t, \mathbf{V}_t)}{\int f(\mathbf{x}_t, \mathbf{V}_t) d\mathbf{x}_t} = \frac{f(\mathbf{x}_t, \mathbf{V}_t)}{f_V(\mathbf{V}_t)} \quad (1)$$

where $f_V(\mathbf{V}_t)$ is the marginal density of $\mathbf{V}_t$.

The standardized sequences $\mathbf{x}_t$ are then transformed to $\mathbf{z}_t = \mathbf{x}_t^*[\mathbf{STD}]_k + [\bar{\mathbf{Z}}]_k$, where k is the calendar day associated with day 't'. Thus, the key idea here is the estimation of this conditional probability density function from the historical data using nonparametric density estimators (kernel estimators) and subsequently simulating or bootstrapping from it. The mechanism of kernel density estimation and the algorithm for simulation from a conditional p.d.f. (as in Equation 1) using kernel density estimators is developed and outlined in later sections.

2.2 Precipitation model

The seasonal wet/dry spell model for daily precipitation described fully in Lall et al. [1995] has three random variables: i) wet spell length, L_w days, ii) dry spell length, L_d days, and iii) wet day precipitation amount, P inches. The periods (seasons) are as defined in the previous section. Variables wsp and dsp are defined through the set of integers between 1 and the season length, and P is defined as a continuous, positive random variable. A mixed set of discrete and continuous random variables is thus considered. The simplified version of the wet/dry spell model described in Lall et al. [1995] that considers successive wet day's precipitation amount and successive wet and dry spell lengths to be independent is adopted in this study. Correlation statistics computed for the data sets analyzed supported these assumptions.

The p.d.f.'s of wet day precipitation amount f(P) and the probability mass functions (p.m.f.'s) of wet spell length $f(L_w)$, and dry spell length $f(L_d)$ are estimated for each season using kernel density estimators.

A dry spell is first generated using $f(L_d)$. Then a wet spell is generated using $f(L_w)$. Precipitation for each of the 'L_w' wet days is then generated from f(P). The process is repeated with the generation of another dry spell. If a season boundary is crossed, the p.d.f.'s used for generation are switched to those for the new season. This procedure

continues until a synthetic sequence of the desired length has been generated. The p.d.f.'s $f(L_w)$, $f(L_d)$, and $f(P)$ are estimated using kernel density estimators detailed in Lall et al. [1995] and Rajagopalan et al. [1995] and are described below.

2.3 Kernel density estimation

The kernel density estimator generalizes the frequency histogram as an estimator of the p.d.f. While the histogram is capable of showing some features of the data, it has several drawbacks. It is difficult to manipulate analytically; it is not easy to visualize for multivariate situations; and it allows for no extrapolation beyond the data. The histogram is sensitive to the class width, as well as the origin of each class. Silverman [1986, p. 9-11] illustrates these problems graphically. One can improve the histogram by centering rectangular boxes at each observation (to gain independence from choice of origin). A kernel density estimator, introduced by Rosenblatt [1956], is formed by centering a smooth kernel function at each observation.

An attractive feature of kernel estimators of the p.d.f. is that they are local (use only a neighborhood around the point of estimate) and, hence, are not globally affected by outliers. Since they make weak prior assumptions of the underlying probability density function, they are data driven and robust and are portable across sites/data sets. For details on kernel density estimation, refer to Silverman [1986] and Scott [1992].

Univariate continuous variables.

The kernel density estimator for a continuous variable (such as the wet day precipitation P) is defined as

$$\hat{f}(P) = \sum_{i=1}^{n} \frac{1}{nh} K\left(\frac{P - P_i}{h}\right) \tag{2}$$

where K(.) is a kernel function centered on the observation Pi and can be any valid probability density function, and h is a bandwidth. The bandwidth h controls the amount of smoothing of the data in the density estimate. Bandwidth h may be constant or variable, taking on different values at different locations. An estimator with constant bandwidth h (as in Equation 2) is called a fixed kernel estimator. Commonly used kernel functions are:

Gaussian Kernel $\quad K(t) = (2\pi)^{-1/2} e^{-t^2/2}$ (3a)

Epanechnikov Kernel $\quad K(t) = 0.75(1 - t^2) \quad |t| \leq 1$ (3b)

Bisquare Kernel $\quad K(t) = (15/16)(1 - t^2)^2 \quad |t| \leq 1$ (3c)

The kernel function represents the weight given to the observation P_i based on distance between P and P_i. One can see from Equation 2, that the kernel estimator is a convolution estimator that forms a local weighted average of the relative frequency of observations in the neighborhood of the point of estimate. The kernel function K(.) prescribes the relative weights, while h prescribes the range of data values over which the average is computed. This is illustrated in Figure 3.

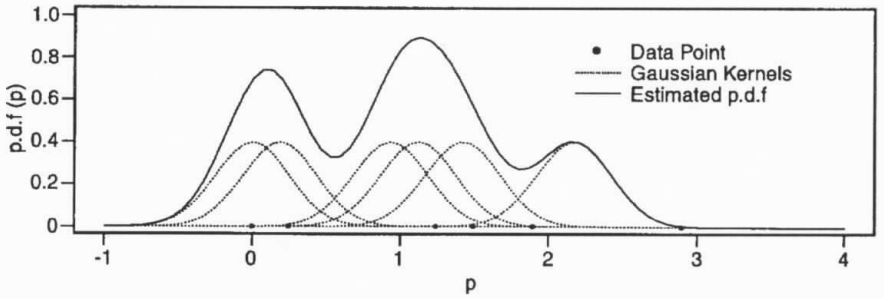

Figure 3. Example of kernel density estimation using five data points with Gaussian Kernel, h = 0.5.

The p.d.f. of wet day precipitation $\hat{f}(P)$ is obtained by applying a kernel density estimator to log transformed data. Note that most of the data of wet day precipitation are concentrated near the lower boundary (i.e., 0). This is a problem for kernel density estimation methods since modifications to kernel density estimate are necessitated within a bandwidth of the boundary. The kernel centered at an observation that is within one bandwidth of the boundary extends past the boundary, thereby leading to leakage of probability mass in the resulting density estimate (i.e., an increase in the bias of the estimate). This boundary problem can be avoided by applying the k.d.e. to logarithmically transformed data. The resulting estimator is given as

$$\hat{f}(P) = \frac{1}{nhP} \sum_{i=1}^{n} K\left(\frac{\log(P) - \log(P_i)}{h}\right) \tag{4}$$

The Epanechnikov kernel is used, and the bandwidth h is chosen for the log transformed data using the recursive approach of Sheather and Jones [1991] to minimize the Mean Integrated Square Error (MISE) of estimate of f(log(P)).

Silverman [1986] points out that, in terms of mean square error of the estimated density, the kernel density estimator is more sensitive to the choice of the bandwidth than to that of the kernel, and the general practice is to choose a kernel and then seek an optimal estimate of the bandwidth h under some criteria.

Univariate discrete variables. In this section, we present procedures for the estimation of the univariate probability mass functions for discrete variables (such as wet spell lengths w, dry spell lengths d). We recommend using the Discrete Kernel (DK) estimator developed in Rajagopalan and Lall [1995]. The DK estimator for the p.m.f. $\hat{f}(L)$, where L is either w or d, and n is the corresponding sample size is given as

$$\hat{f}(L) = \sum_{j=1}^{L_{max}} K_d\left(\frac{L-j}{h}\right) \tilde{\alpha}_j \tag{5}$$

where $\tilde{\alpha}_j$ is the sample relative frequency (n_j/n) of spell length j, n_j is the number of spells of length j, L_{max} is the maximum observed spell length (note that $\sum_{j=1}^{L_{max}} \tilde{\alpha}_j = 1$), $K_d(\cdot)$ is a discrete kernel function, and L, j, and h are positive integers. The kernel function $K_d(\cdot)$ is given as

$$K_d(t) = at_j^2 + b \text{ for } |t| \leq 1 \tag{6}$$

The expressions for a and b for the interior of the domain, L>h+1 and the boundary region L<h are developed in Rajagopalan and Lall [1995].

The bandwidth h is estimated by minimizing a Least Squares Cross Validation (LSCV) function given as

$$LSCV(h) = \sum_{j=1}^{L_{max}} (\hat{f}(j))^2 - 2\sum_{j=1}^{L_{max}} \hat{f}_{-j}(j)\tilde{\alpha}_j \tag{7}$$

where, $\hat{f}_{-j}(j)$ is the estimate of the p.m.f. of spell length j, formed by dropping all the spells of length j from the data. This method has been shown by Hall and Titterington [1987] to automatically adapt the estimator to an extreme range of sparseness types. Monte Carlo results showing the effectiveness of the DK estimator with bandwidth selected by LSCV are presented in Rajagopalan and Lall [1995].

Multivariate continuous variables. Extending the idea of the kernel density estimator for univariate continuous variables, a kernel density estimate of the multivariate p.d.f. of a vector $\mathbf{y}$ is defined as [Silverman, 1986, p. 76-78]

$$\hat{f}(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} K(u) \tag{8}$$

where $u = \frac{(\mathbf{y}-\mathbf{y}_i)^T S^{-1}(\mathbf{y}-\mathbf{y}_i)}{h^2}$, and K(u) is a multivariate Gaussian kernel function. $\mathbf{y} = [y_1, y_2, \cdots, y_d]^T$ denotes the d dimensional random vector whose density is being estimated with, $\mathbf{y}_i = [y_{1i}, y_{2i}, \cdots, y_{di}]^T$ i=1 to n the sample values of $\mathbf{y}$, n is the number of sample vectors; h is a bandwidth; and S the sample covariance matrix. The Gaussian kernel function used is given as

$$K(u) = \frac{1}{(2\pi)^{d/2} det(S)^{1/2} h^d} \exp(-u/2) \tag{9}$$

Just as in the univariate case described in the earlier section, K(u) represents the weight given to an observation $\mathbf{y}_i$ that is based on distance between $\mathbf{y}$, and $\mathbf{y}_i$. The distance used here is the Euclidean distance modified to recognize the covariance of the $\mathbf{y}$. It can be seen that the estimator in Equation 8 is similar to the univariate estimator in Equation 2 since the estimate is a local weighted average of the relative frequency of observations in the neighborhood of the point of estimate. Here also the kernel function, $K(\cdot)$ prescribes the relative weights, h prescribes the range of data values over which the average is computed, and the covariance matrix $\mathbf{S}$ provides the orientation of the weight function.

Here, we chose the bandwidth h as the one that minimizes mean integrated square error in $\hat{f}(y)$ if the underlying distribution is assumed to be multivariate Gaussian. Silverman [1986, p. 86-87] gives an appropriate h to use for a multivariate Gaussian p.d.f. using the Gaussian kernel as

$$h = \{(4/(2d+1))^{1/(d+4)}\} n^{-1/(d+4)} \tag{10}$$

Here n is the number of observations and d is the dimension. As the dimension d increases, h also increases. This happens because in higher dimensions large regions of high density may be completely devoid of observations in a sample of moderate size. The bandwidth in such a situation has to be bigger to cover these large regions.

The above choice of bandwidth is optimal for p.d.f.'s that are near Gaussian and is an adequate choice for many cases [Silverman, 1986, p. 45-48]. Cross validation [see Sain et al., 1994] or plug-in methods [see Wand and Jones, 1994] could be used here to choose h as in the wet/dry spell model. However, this increases the computational burden substantially. Recall that the parametric approaches often assume a Gaussian distribution. In a Bayesian context, using this bandwidth can be thought of as developing a posterior kernel density estimate with a Gaussian prior. The resulting tail behavior and degree of smoothing supplied will be consistent with an underlying Gaussian p.d.f., with some adaption to local features.

In the Bootstrap context, we have a region that each observation $\mathbf{y}_i$ represents. The orientation and shape of the region is given by the scaling factor h^2S and the kernel function K(u). Resampling from the kernel density estimate entails picking a point $\mathbf{y}_i$ uniformly in $[\mathbf{y}_1, \cdots, \mathbf{y}_n]$ and then simulating from the kernel K(u), i.e., $N(\mathbf{y}_i, h^2S)$. We extend this approach formally for simulation from a multivariate conditional p.d.f. in the following section.

3 Kernel density estimation of multivariate conditional p.d.f.

Here an estimate of the conditional p.d.f. $f(\mathbf{x}_t|\mathbf{V}_t = \mathbf{V}^*)$ is needed for the simulation of interest. The strategy used here is similar to the one used by Sharma et al. [1995] for streamflow simulation. Applying the estimator in Equation 8 to the conditional p.d.f. in Equation 1 with sample vectors $\mathbf{x}_i = [\mathbf{x}_t, \mathbf{x}_{t-1}, P_t]_i$ denoted as $[\mathbf{x}_i, \mathbf{V}_i]$ we get

$$f(\mathbf{x}_t|\mathbf{V}_t = \mathbf{V}^*) = \frac{1}{nh^d} \frac{1}{f_v(\mathbf{V}^*)} \sum_{i=1}^{n} \frac{1}{\det(S)^{1/2}} K\left(\frac{[\mathbf{x}_t - \mathbf{x}_i; (\mathbf{V}^* - \mathbf{V}_i)^T] S^{-1} \begin{bmatrix} \mathbf{x}_t - \mathbf{x}_i \\ \mathbf{V}^* - \mathbf{V}_i \end{bmatrix}}{h^2} \right) \tag{11}$$

where S is the dg by dg covariance matrix of the vector $(\mathbf{x}_i, \mathbf{V}_i)$ estimated from historical data. Let the matrix S be partitioned as

$$S = \begin{bmatrix} S_X & S_{XV}^T \\ S_{XV} & S_V \end{bmatrix} \tag{12}$$

where S_X is the d by d covariance matrix of $\mathbf{x}$, S_V is the d' by d' covariance matrix of $\mathbf{V}$, and S_{XV} the d by d' cross covariance between $\mathbf{x}$ and $\mathbf{V}$. Using the Gaussian kernel function (i.e., Equation 9), we can reduce Equation 11 to a weighted sum of Gaussian functions,

$$f(\mathbf{x}_t|\mathbf{V}_t = \mathbf{V}^*) = \sum_{i=1}^{n} w_i N(\mathbf{b}_i, \mathbf{c}_i) \tag{13}$$

where

$$w_i = w_i' / \sum_{i=1}^{n} w_i', w_i' = \exp(-a_i/2);\ a_i = \frac{([\mathbf{V}^* - \mathbf{V}_i]^T [S_V]^{-1} [\mathbf{V}^* - \mathbf{V}_i])}{h^2}\ ; \tag{14}$$

$$\mathbf{b}_i = \mathbf{x}_i + ([\mathbf{V}^* - \mathbf{V}_i]^T)[S_V]^{-1}[S_{XV}]\ ;\ \ \mathbf{c} = h^2(S_X - S_{XV}^T S_V^{-1} S_{XV}) \tag{15}$$

Note that $\sum_{i=1}^{n} w_i = 1$.

From Equation 13, we see that the conditional p.d.f. reduces to a weighted sum of Gaussian functions. It can be thought of as a slice through a multivariate density function, estimated as a weighted sum of slices with the same orientation through the kernels placed on each observation. Simulation from the conditional p.d.f. can be achieved by picking a point $\mathbf{x}_i$ with probability w_i, then sampling from $N(\mathbf{b}_i, \mathbf{c})$.

3.1 NP Simulation algorithm

The simulation proceeds as:

1. Simulate precipitation for all the days of the year from the wet/dry spell model described earlier.
2. Estimate the NP model parameters (i.e., bandwidth h and the covariance matrix S) from the data for each season.
3. At the start of each period of interest, initialize t=0, $\mathbf{x}_t$ = one of the historical observations randomly selected.
4. Generate $\mathbf{x}_t$ sequentially (day by day) from $f(\mathbf{x}_t|\mathbf{V}_t)$, where the conditioning vector $\mathbf{V}_t$ consists of the previous day's vector $\mathbf{x}_{t-1}$ and the current day's generated precipitation P_t (i.e., $\mathbf{V}_t = [\mathbf{x}_{t-1}, P_t]$) as:
   i) Estimate weights (w_i) associated with each data point ($\mathbf{x}_i$) (Equation 14).
   ii) Resample an index i using $w_i (i = 1, \cdots, n)$ as probabilities point $\mathbf{x}_i$ and $\mathbf{V}_t$ (Equation 15).
   iii) Generate vector $\mathbf{x}_t = \mathbf{b}_i + \epsilon$ where ϵ is from a multivariate normal distribution with mean $\mathbf{0}$ and variance $\mathbf{c}$ [see Devroye, 1986, p. 565].
5. Recover $\mathbf{z}_t$ as $\mathbf{z}_t = \mathbf{x}_t^*[\mathbf{STD}]_k + [\bar{\mathbf{X}}]_k$ where k is the calendar day corresponding to day t.
6. At the start of a new simulation, go to step 3.

4 Model application

To demonstrate the utility of the resampling model for generation of daily weather variables, the model was applied to daily weather data from the station in Salt Lake

City in Utah. Thirty years of daily weather data were available from the period 1961-1991. Salt Lake City is at 40°46'N latitude, 111°58'W longitude, and at an elevation of 1288 m. Most of the precipitation comes in the form of winter snow. Rainfall occurs mainly in spring, with some in fall.

We shall first outline the experimental design and then use some measures of performance to judge the utility of the model.

4.1 Experimental design

Our purpose here is to test the utility of the NP generation scheme. The main steps involved in accomplishing this are:

1. Daily precipitation is generated from the wet/dry spell model.
2. The other variables are generated following the simulation algorithm described in the previous section.
3. Twenty-five synthetic records of 30 years each (i.e., the historical record length) are simulated using the NP model.
4. The statistics of interest (described below) are computed for each simulated record, for each period, and are compared to statistics of the historical record using boxplots.

4.2 Performance measures

The following statistics were considered to be of interest when comparing the historical record and the NP simulated record of other weather variables.

Moments:

1. Mean of each variable for each season.
2. Standard deviation of each variable for each season.
3. Skew of each variable for each season.
4. Coefficient of variation of each variable for each season.

Relative Frequencies:

5. 25% quantile of each variable for each season.
6. 75% quantile of each variable for each season.

Dependence:

7. Cross correlation on any given day between the variables for each season.
8. Lag-1 daily cross correlation between the variables for each season.
9. Lag-1 daily correlation of each variable for each season.

5 Results

The statistics of interest calculated from the simulations are compared with those for the historical record using boxplots. A box in the boxplots (e.g., Figure 4) indicates the interquartile range of the statistic computed from 25 simulations, the line in the middle of the box indicates the median simulated value. The solid lines correspond to the statistic of the historical record. The boxplots show the range of variation in the statistics from the simulations and also show the capability of the simulations to reproduce historical statistics.

Figures 4 through 7 show the boxplots of moments and relative frequency measures of solar radiation, maximum temperature, minimum temperature, and average dew point temperature, respectively. It can be seen that the historical values of mean, and the quantiles are well-reproduced, while standard deviation, coefficient of skew, and coefficient of variation are not quite well-reproduced. This is to be expected as the kernel methods inflate the variance by a factor equal to $(1 + h^2)$ [see Silverman, 1986, p. 143] which, in turn, affects the skew and the coefficient of variation. This inflation can be corrected through an appropriate scaling of the random terms during simulation [see Silverman, 1986, p. 143]. However, it may be desirable to have to have a slight increase in the variance of the simulations as compared to that of the historical.

Illustrative statistics of wet spell lengths, dry spell lengths, and wet day precipitation for the simulations from the wet/dry spell model are also estimated and are shown in Figures 8, 9, and 10, respectively. Figure 8 shows the boxplots of average wet spell length, standard deviation of wet spell length, fraction of wet days, and length of longest wet spell length for each season. Figure 9 shows the boxplots of these statistics of the dry spell length. Figure 10 shows the boxplots of average wet day precipitation, standard deviation of wet day precipitation, and percentage of yearly precipitation in each season. The boxplots in Figures 8, 9, and 10 show that the historical statistics are reproduced well by the simulations.

Figures 11 and 12 show the boxplots of the lag-0 cross correlation and lag-1 cross correlation between the variables. Figure 13 shows the lag-1 auto correlation of each variable for each of the four seasons. The correlations from the simulations and the historical correlations seem to be different in a number of cases. The correlations that are reproduced most poorly are the ones with precipitation. While the correlations of the variables with precipitation are very small as can be seen from these figures and in many cases seem insignificant.

One reason for this mismatch of the correlations is that the precipitation is supplied externally from the wet/dry spell model. As a result, the covariance between $\mathbf{x}_{t-1}$ and P_t need not correspond to that of the historical covariance between them. This introduces a bias in the conditioning plane from which $\mathbf{x}_t$ is generated and results in a mismatch of the correlations. To verify this, we made 25 simulations without conditioning on precipitation (i.e., simulated $\mathbf{x}_t$ from $f(\mathbf{x}_t|\mathbf{x}_{t-1})$ where both $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$ are of Dimension 5). The correlations from this simulation are shown in Figures 14, 15, and 16, respectively. It can be seen from these three figures that the correlations are well-reproduced, which strongly suggests that the conditioning on the precipitation is the reason for mismatch of correlations in Figures 11, 12, and 13.

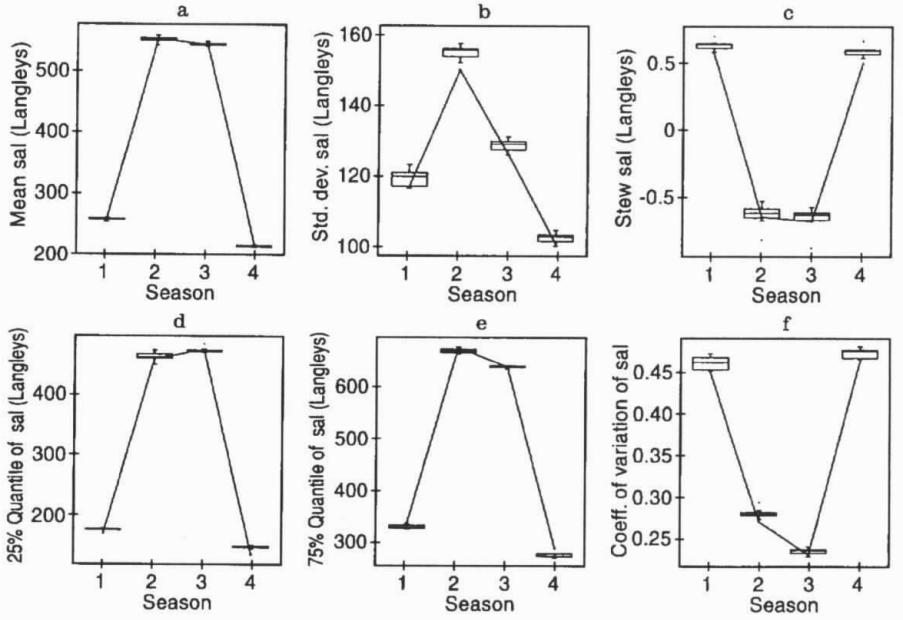

Figure 4. Boxplots of statistics of total daily solar radiation, SRAD (a) mean SRAD, (b) standard deviation of SRAD, (c) skew of SRAD, (d) 25% quantile of SRAD, (e) 75% quantile of SRAD, and (f) coefficient of variation of SRAD for model simulations, along with the historical values for the four seasons.

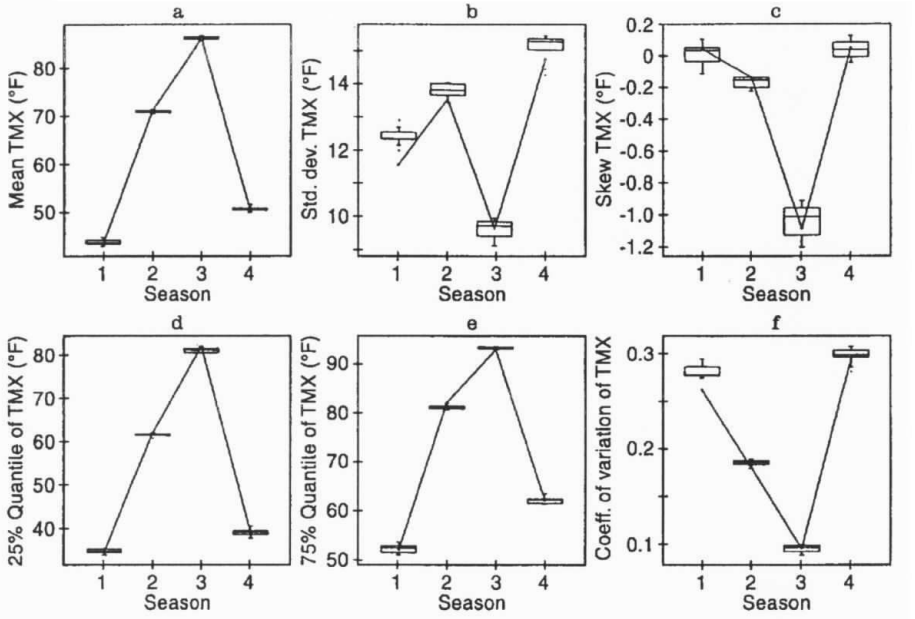

Figure 5. Boxplots of statistics of daily maximum temperature, TMX (a) mean TMX, (b) standard deviation of TMX, (c) skew of TMX, (d) 25% quantile of TMX, (e) 75% quantile of TMX, and (f) coefficient of variation of TMX for model simulations, along with the historical values for the four seasons.

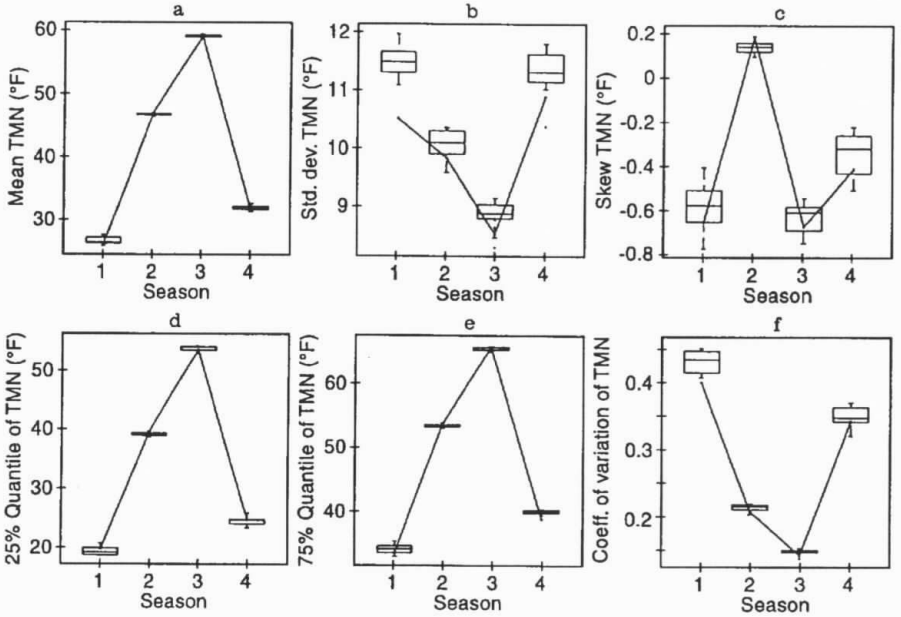

Figure 6. Boxplots of statistics of daily minimum temperature, TMN (a) mean TMN, (b) standard deviation of TMN, (c) skew of TMN, (d) 25% quantile of TMN, (e) 75% quantile of TMN, and (f) coefficient of variation of TMN for model simulations, along with the historical values for the four seasons.

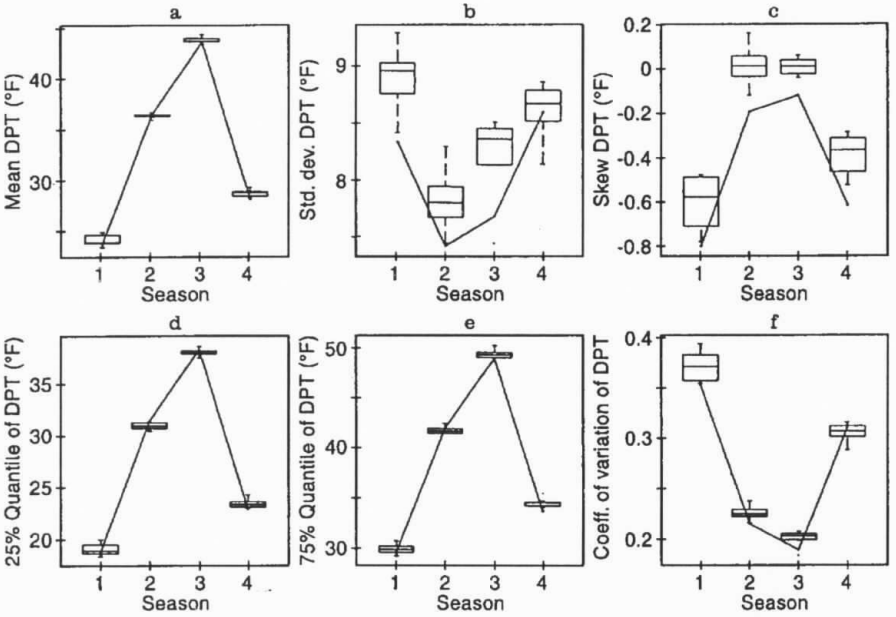

Figure 7. Boxplots of statistics of dew point temperature, DPT (a) mean DPT, (b) standard deviation of DPT, (c) skew of DPT, (d) 25% quantile of DPT, (e) 75% quantile of DPT, and (f) coefficient of variation of DPT for model simulations, along with the historical values for the four seasons.

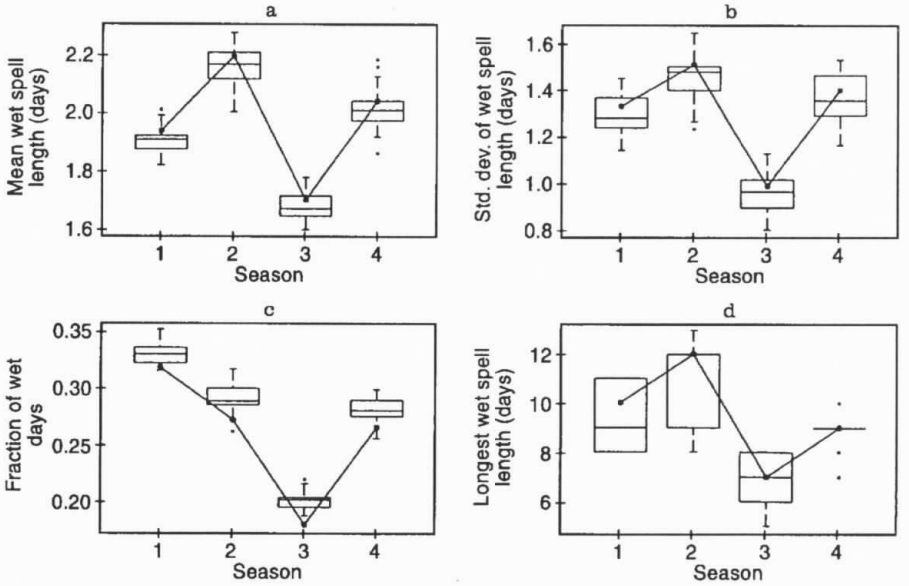

Figure 8. Boxplots of statistics of wet spell length (a) mean wet spell length, (b) standard deviation of wet spell length, (c) fraction of wet days, and (d) longest wet spell length for simulations from wet/dry spell model, along with the historical values for the four seasons.

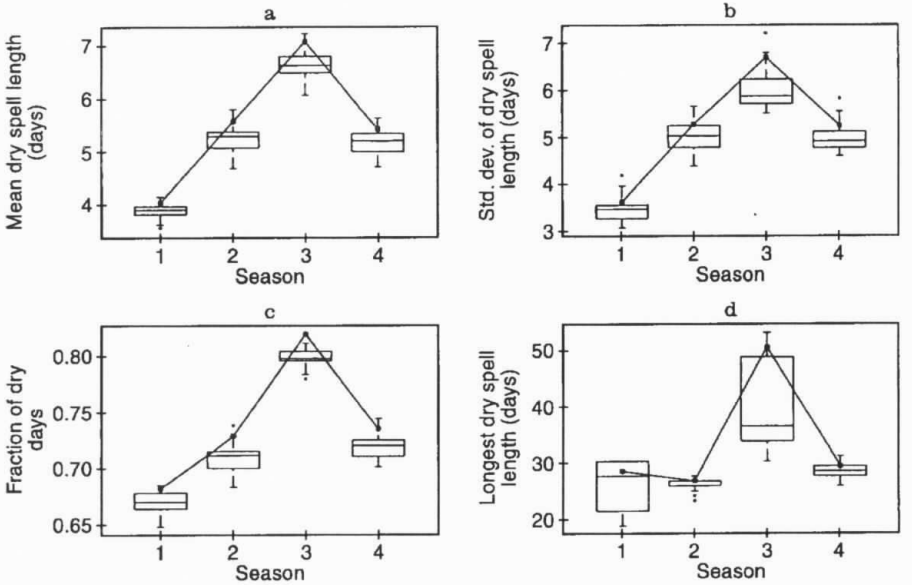

Figure 9. Boxplots of statistics of dry spell length (a) mean wet spell length, (b) standard deviation of wet spell length, (c) fraction of wet days, and (d) longest wet spell length for simulations from wet/dry spell model, along with the historical values for the four seasons.

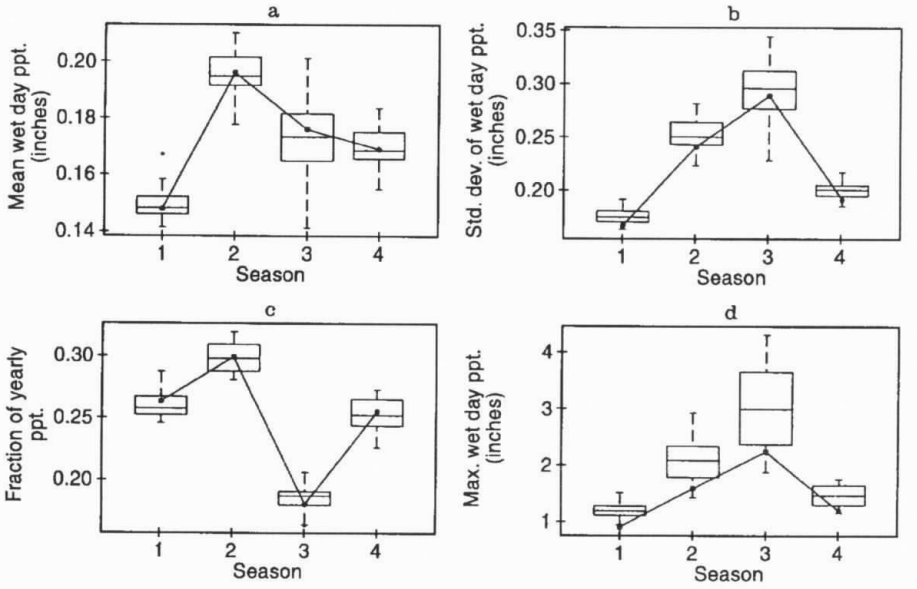

Figure 10. Boxplots of statistics of wet day precipitation (a) mean wet day precipitation, (b) standard deviation of wet day precipitation, (c) fraction of yearly wet day precipitation, and (d) maximum wet day precipitation for simulations from wet/dry spell model, along with the historical values for the four seasons.

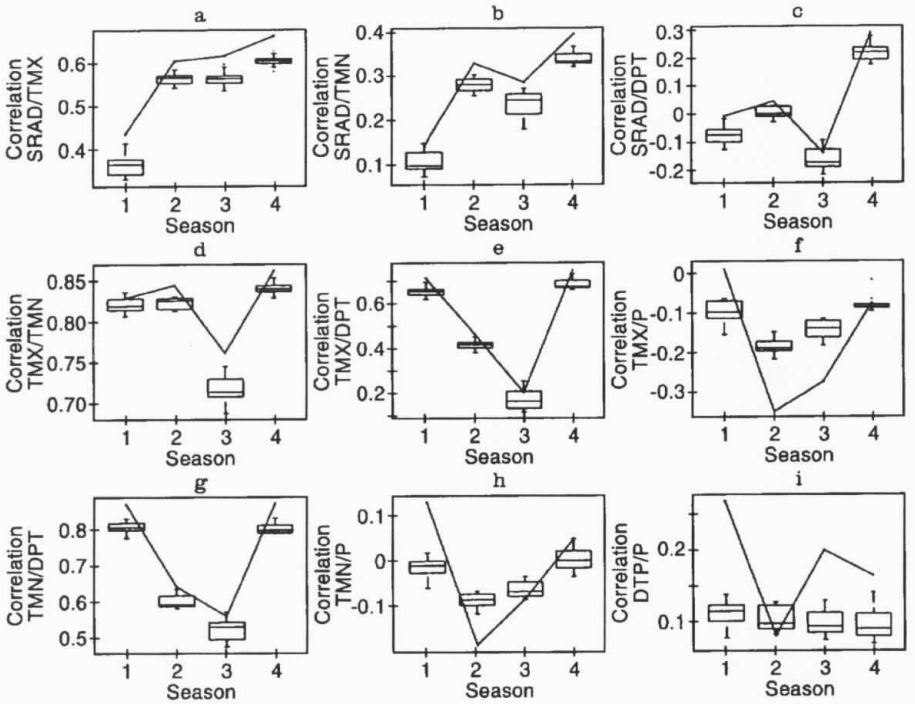

Figure 11. Boxplots of Lag-0 cross correlation between (a) SRAD and TMX, (b) SRAD and TMN, (c) SRAD and DPT, (d) TMX and TMN, (e) TMX and DPT, (f) TMX and P, (g) TMN and DPT, (h) TMN and P, and (i) DPT and P for model simulations, along with the historical values for the four seasons.

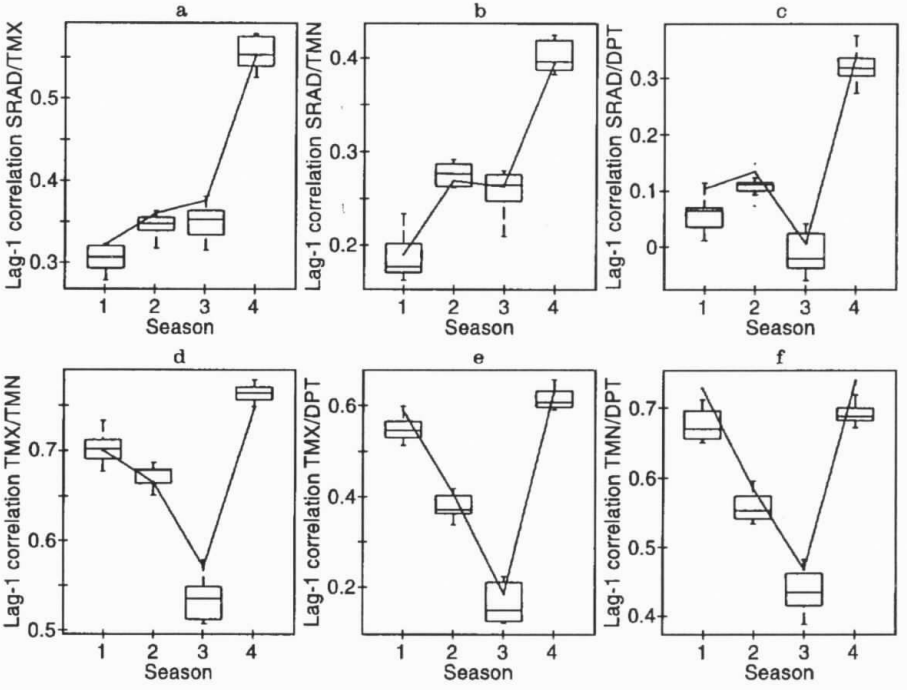

Figure 12. Boxplots of Lag-1 cross correlation between (a) SRAD and TMX, (b) SRAD and TMN, (c) SRAD and DPT, (d) TMX and TMN, (e) TMX and DPT, and (f) TMN and DPT for model simulations, along with the historical values for the four seasons.

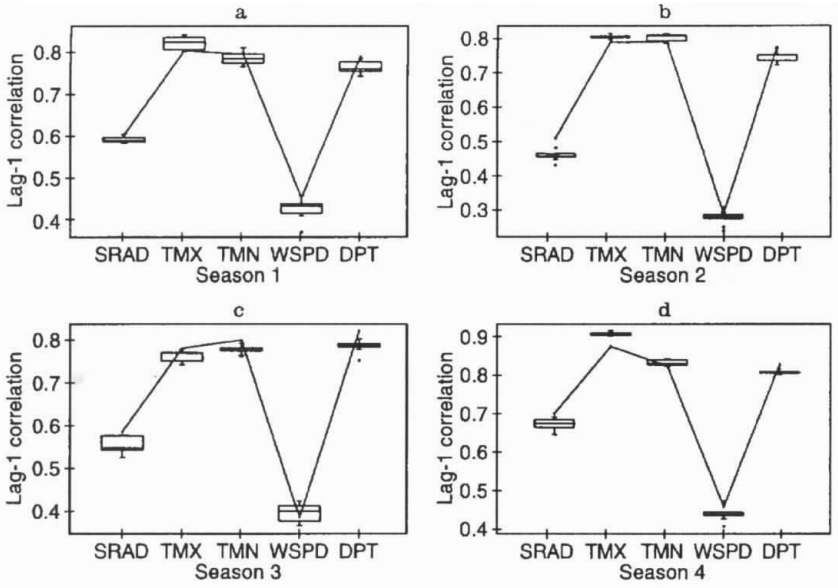

Figure 13. Boxplots of Lag-1 Auto Correlation of SRAD, TMX, TMN, WSPD, and DPT for (a) Season 1, (b) Season 2, (c) Season 3, and (d) Season 4 for model simulations, along with the historical values.

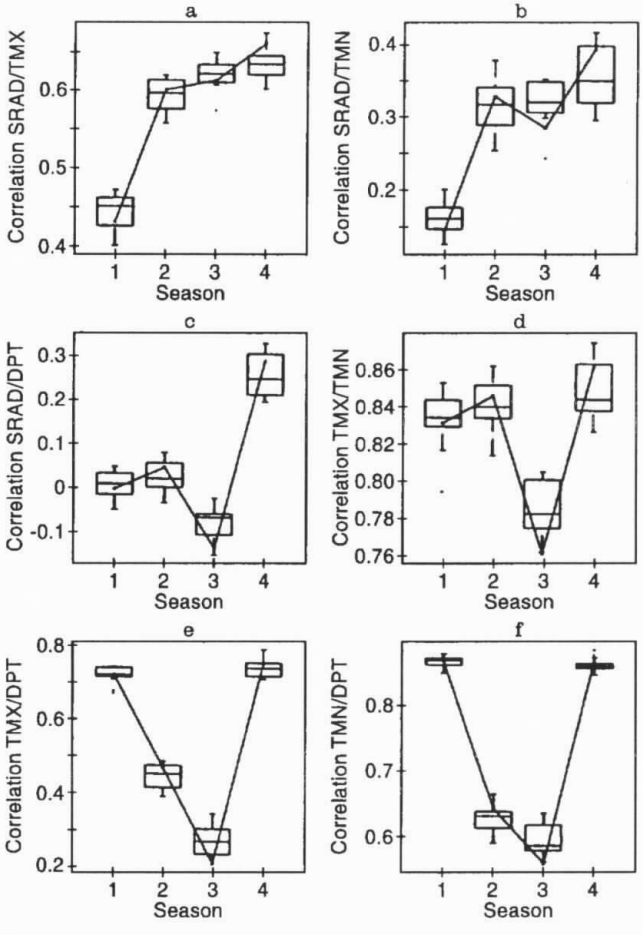

Figure 14. Boxplots of Lag-0 cross correlation between (a) SRAD and TMX, (b) SRAD and TMN, (c) SRAD and DPT, (d) TMX and TMN, (e) TMX and DPT, and (f) TMN and DPT for model simulations (without conditioning on precipitation), along with the historical values for the four seasons.

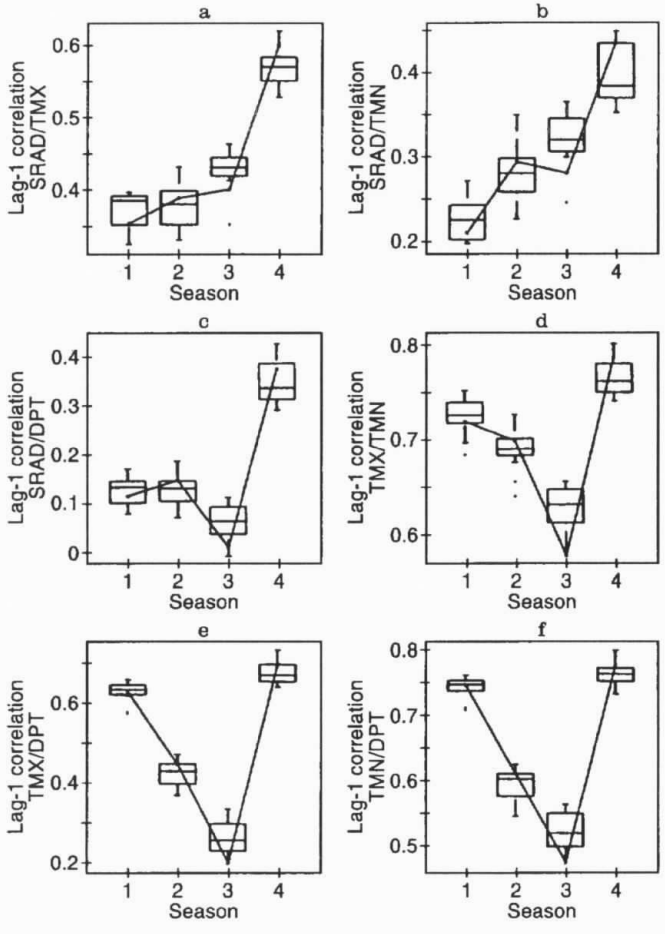

Figure 15. Boxplots of Lag-1 cross correlation between (a) SRAD and TMX, (b) SRAD and TMN, (c) SRAD and DPT, (d) TMX and TMN, (e) TMX and DPT, and (f) TMN and DPT for model simulations (without conditioning on precipitation), along with the historical values for the four seasons.

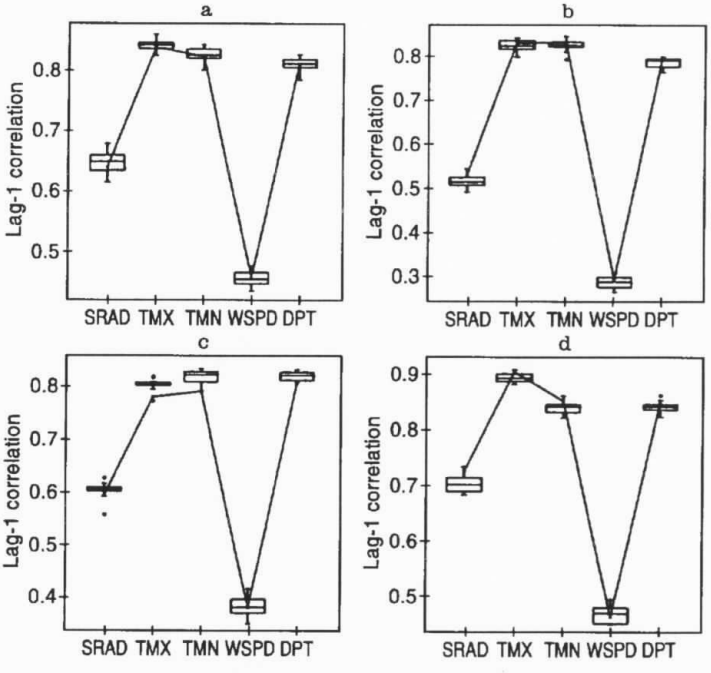

Figure 16. Boxplots of Lag-1 Auto Correlation of SRAD, TMX, TMN, WSPD, and DPT for (a) Season 1, (b) Season 2, (c) Season 3, and (d) Season 4 for model simulations (without conditioning on precipitation), along with the historical values.

One way to get around this problem would be to include precipitation in the multivariate model (i.e., simulate $\mathbf{x}_t$ from $f(\mathbf{x}_t|\mathbf{x}_{t-1})$ where both $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$ are of Dimension 6 and include precipitation). This may reproduce the correlation statistics. However, the fixed bandwidth Gaussian kernel framework used here is inappropriate for a highly skewed variable such as precipitation with a concentration at 0. There would be considerable leakage at the boundary, that lead to negative precipitation values being simulated. It was consequential of these problems and the desire to represent a more general dependence structure in rainfall occurrences that motivated the development of the precipitation model used [Lall et al., 1995].

6 Summary and conclusions

A multivariate nonparametric model NP that aims at capturing dependence up to lag-1 was presented and illustrated. The simulations were made from the conditional p.d.f. estimated from the data using kernel density estimators. The kernel estimators being local average estimators have the advantage of readily admitting arbitrary probability densities without requiring that they be hypothesized or formally identified. Broader dependence structures can be consequently considered. The need to choose/justify and fit the best p.d.f. is side stepped.

The bandwidth is the key parameter in the NP model, as it determines the degree of smoothness that will be imparted to the p.d.f. The larger the bandwidth, the smoother the p.d.f. and vice versa. Choosing h automatically using cross-validation [see Sain et al., 1994] or plug-in approaches [see Wand and Jones, 1994] from the data would be more appropriate than the choice used here. However, the additional variance in the choice of h induced by such an estimation process may detract from its use where the primary purpose is to resample the data. Bandwidth selection methods are undergoing continuous improvement. We expect to implement more formal selection procedures in due course. One could also use a local covariance matrix estimated at each data point using a few neighbors of that point (i.e., S_i instead of S in Equation 8). Sharma et al. [1995] use this method for streamflow simulation.

Another problem with simulations is the boundary effect. For the variables that are bounded (e.g., solar radiation and precipitation), values that violate the bounds could be generated. Typically, these are censored to the bound. This may introduce a bias in the simulations. Procedures to better address this problem in univariate situations are described in Rajagopalan et al. [1995], but for multivariate situations, effective methods are yet to be developed.

We chose to apply the NP model on a seasonal time scale because the precipitation model that was used to drive the NP model is a seasonal model. However, we checked the results of the seasonal NP model at monthly time scale and found the performance to be similar (results are not presented here).

The NP model developed here underscores our growing conviction that nonparametric techniques have an important role to play in improving the synthesis of hydrologic time series. They can capture dependence structure present in the data without imposing arbitrary distributional assumptions and produce synthetic sequences that are statistically similar to the historic sequence. The idea of resampling the data with appropriate perturbation of each value while maintaining selected dependence characteristics (or data sequencing) is easy to accept as a practical matter.

Acknowledgments

Partial support of this work by the U.S. Forest Service under contract numbers INT-915550-RJVA and INT-92660-RJVA, Amend #1 is acknowledged.

References

Bruhn, J.A.; Fry, W.E.; Fick, G.W. 1980: Simulation of daily weather data using theoretical probability distributions, J. Appl. Meteorology 19(9), 1029-1036

Devroye, L. 1986: Non-Uniform Random Variate Generation. Springer-Verlag, New York

Jianping, D.; Simonoff, J. 1994: The construction and properties of boundary kernels for sparse multinomials, J. Comp. Graph. Statis. 3, 1-10

Efron, B. 1979: Bootstrap methods: Another look at the Jackknife, Ann. Stat. 7, 1-26

Hall, P.; Titterington, D.M. 1987: On smoothing sparse multinomial data, Australian J. Stat. 29(1), 19-37

Jones, W.; Rex, R.C.; Threadgill, D.E. 1972: A simulated environmental model of temperature, evaporation, rainfall, and soil moisture, Trans. of the ASAE, 366-372

Kunsch, H.R. 1989: The Jackknife and the Bootstrap for general stationary observations, Ann. Stat. 17, 1217-1241

Lall, U.; Rajagopalan, B.; Tarboton, D.G. 1995: A nonparametric wet/dry spell model for resampling daily precipitation, Water Resour. Res.

Lane, L.J.; Nearing, M.A. 1989: USDA - Water Erosion Prediction Project: Hill Slope Profile Model Documentation, NSERL Report No. 2, National Soil Erosion Research Laboratory, USDA-Agricultural Research Service, W. Lafayette, Indiana 47907

Liu, R.Y.; Singh, K. 1988: Using iid Bootstrap Inference for Some Non-iid Models, Preprint. Department of Statistics, Rutgers University

Nicks, A.D.; Harp, J.F. 1980: Stochastic generation of temperature and solar radiation data, J. Hydr. 48, 1-7

Rajagopalan, B.; Lall, U. 1995: A kernel estimator for discrete distributions, J. Nonparametric Stat. 4, 409-426

Rajagopalan, B.; Lall, U.; Tarboton, D.G. 1995: Evaluation of kernel density estimation methods for daily precipitation resampling, Water Resour. Res., Sep. 1996

Richardson, C.W. 1981: Stochastic simulation of daily precipitation, temperature, and solar radiation, Water Resour. Res. 17(1), 182-190

Rosenblatt, M. 1956: Remarks on some nonparametric estimates of a density function, Ann. Math. Stat. 27, 832-837

Sain, S.R.; Baggerly, K.A.; Scott, D.W. 1994: Cross-validation of multivariate densities, J. Amer. Stat. Assoc. 89(427), 807-817

Scott, D.W. 1992: Multivariate Density Estimation, John Wiley and Sons, Inc., New York

Sharma, A.; Tarboton, D.G.; Lall, U. 1995: Streamflow simulation: A nonparametric approach, Water Resour. Res., in press

Sheather, S.J.; Jones, M.C. 1991: A reliable data-based bandwidth selection method for kernel density estimation, J. Royal Stat. Soc. B53, 683-690

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

Wand, M.P.; Jones, M.C. 1994: Multivariate plug-in bandwidth selection, Comp. Stat, 9, 97-116