Improved *K*-Nearest Neighbor Weather Generating Model

Mohammed Sharif[1] and Donald H. Burn[2]

Abstract: A major limitation of *K*-nearest neighbor based weather generators is that they do not produce new values but merely reshuffle the historical data to generate realistic weather sequences. In this paper, a modified approach is developed that allows nearest neighbor resampling with perturbation of the historic data. A strategy is introduced that resamples the historical data with perturbations while preserving the prominent statistical characteristics, including the interstation correlations. The approach is similar in spirit to traditional autoregressive models except that the new values are obtained by adding a random component to the individual resampled data points. An advantage of the approach is that unprecedented precipitation amounts are generated that are important for the simulation of extreme events. The approach is demonstrated through application to the Upper Thames River Basin in Ontario. Daily weather variables (maximum temperature, minimum temperature, and precipitation) were simulated at multiple stations in and around the basin. Analysis of the simulated data demonstrated the ability of the model to reproduce important statistical parameters of the observed data series while allowing perturbations to the observed data points. Additionally, no site-specific assumptions regarding the probability distribution of variables are required.

DOI: 10.1061/(ASCE)1084-0699(2007)12:1(42)

CE Database subject headings: Stochastic processes; River basins; Canada; Weather.

Introduction

Development of stochastic weather models is an important task that has many practical applications in hydrology and water resources management. These models are often employed for impact assessment studies that require stochastically generated weather sequences as input. Traditionally, parametric weather generators (Nicks and Harp 1980) have first focused on independent generation of precipitation while the remaining variables are modeled conditionally on precipitation occurrence. Daily precipitation amounts are generated using a two-state first-order Markov model from an assumed probability distribution fit to the observed values. Todorovic and Woolhiser (1975) combined the first-order Markov model for daily precipitation occurrence with a statistical model for daily nonzero precipitation amounts described by an exponential distribution. Katz (1977), Buishand (1978), and Stern and Coe (1984) used the two-parameter gamma distribution to describe the occurrence of precipitation amount on wet days. Smith and Schreiber (1974), Woolhiser and Roldan (1982), and Wilks (1999) fit the three-parameter mixed exponential distribution to describe precipitation amounts on wet days. Richardson (1981) describes a Markov-chain exponential model for generating other meteorological variables in addition to precipitation. Stochastic weather generators of the type proposed by Richardson (1981) are commonly referred to as WGEN (for "weather generator") as in Richardson and Wright (1984). Nicks et al. (1990) describe an extended version of WGEN, called WXGEN, that takes into account the non-normal distribution of wind speed and relative humidity. Wind speed and dewpoint (from which relative humidity can be derived) are included in the weather generator GEM (generation of weather elements for multiple applications) developed by Hanson and Johnson (1998). Parlange and Katz (2000) further extended WGEN to include daily mean wind speed and dewpoint in the model. Wilks and Wilby (1999) present an excellent review of stochastic weather models.

A major drawback associated with the "Richardson type" weather generators is that persistent events, such as drought or prolonged rainfall, are not well reproduced. To overcome this problem, the serial approach to weather generation has been presented by Rackso et al. (1991), Semenov and Barrow (1997), and Semenov et al. (1998), among others. An example of this approach is LARS-WG in which the sequence of dry and wet series of days is modeled first while the precipitation amounts and other variables are generated conditioned on the wet or dry status. Both WGEN and LARS-WG, however, have difficulty in reproducing the annual variability in monthly means of the variables. Further, they cannot simultaneously simulate weather data at multiple sites. Nevertheless, several applications of weather generators for multisite simulation of variables have been reported in the literature (e.g., Smith 1994; Wilby 1994; Wilks 1998). Parametric methods are indeed very useful but they have several inadequacies. First and most importantly, they do not adequately reproduce various aspects of the spatial and temporal dependence of variables. Second, an assumption has to be made regarding the form of probability distribution of the variables, which is often subjective. Third, non-Gaussian features in the data cannot be adequately captured as multivariate autoregressive (MAR) models implicitly assume a normal distribution, which is difficult to satisfy. Fourth, a large number of parameters are separately fit to

[1] Research Fellow, Dept. of Civil Engineering, Univ. of Waterloo, 200 University Ave. West, Waterloo ON, Canada N2L 3G1. E-mail: msharif@uwaterloo.ca

[2] Professor, Dept. of Civil Engineering, Univ. of Waterloo, 200 University Ave. West, Waterloo ON, Canada N2L 3G1 (corresponding author). E-mail: dhburn@civmail.uwaterloo.ca

Note. Discussion open until June 1, 2007. Separate discussions must be submitted for individual papers. To extend the closing date by one month, a written request must be filed with the ASCE Managing Editor. The manuscript for this paper was submitted for review and possible publication on July 15, 2004; approved on June 7, 2006. This paper is part of the ***Journal of Hydrologic Engineering***, Vol. 12, No. 1, January 1, 2007. ©ASCE, ISSN 1084-0699/2007/1-42–51/$25.00.

each time period and the number further increases if the simulations are to be conditioned. Fifth, the models are not easily transportable to other sites due to the site-specific assumptions made regarding the probability distributions of the variables.

Nonparametric methods can circumvent many problems associated with the parametric methods. The most promising nonparametric technique for generating weather data is the *K*-nearest neighbor (*K*-NN) resampling approach. Recently, interest has emerged in the application of these techniques for generating synthetic weather data. The works of Young (1994), Lall and Sharma (1996), Lall et al. (1996), Rajagopalan and Lall (1999), Buishand and Brandsma (2001), and Yates et al. (2003) describe applications of the *K*-NN resampling scheme for simulation of weather data. Young (1994) employed a *K*-NN model for simulation of weather data that preserves the correlation between the temperature and the precipitation, and the wet or dry spell statistics. Simulated sequences, however, showed reduced persistence and underestimation of the fraction of dry months. Lall et al. (1996) used a *K*-NN resampling scheme with kernel density estimators to represent the probability distributions of dry spell lengths, wet spell lengths, and wet day precipitation amounts. Rajagopalan and Lall (1999) compared nearest neighbor resampling with a parametric time series model and demonstrated the superiority of the nonparametric approach. Sharma et al. (1997) describe a nonparametric method for the simulation of streamflow sequences. Their model was able to reproduce both linear and nonlinear dependence. Brandsma and Buishand (1998) describe the application of a nearest neighbor resampling procedure to single site simulation of daily precipitation and temperature for multiple stations in the Rhine Basin. Conditional simulation of weather variables on atmospheric flow was also considered. Buishand and Brandsma (2001) extended the nearest neighbor resampling to simultaneous simulation of daily precipitation and temperature at multiple stations. Yates et al. (2003) describe a *K*-NN resampling strategy that can be used to generate desired climate change scenarios.

It is often necessary to evaluate the response of hydrological models to extreme precipitation events that cause floods or droughts in a basin. A weather generator capable of simulating the occurrence of extreme precipitation events, while preserving the important temporal and spatial correlation of the observed data, is likely to be of immense help in formulating effective flood and drought management strategies at the catchment level. Earlier works of Yates et al. (2003), Buishand and Brandsma (2001), and Lall and Sharma (1996) describe successful applications of the basic *K*-NN approach to the simulation of weather data. However, a limitation of these models is that they do not produce new values but merely reshuffle the historical data to generate new weather sequences. Application of such sequences, in conjunction with the hydrological models, to catchment response evaluation could lead to underexploration of the possible effects of climatic variability, and to suboptimal policies for system management. The principal focus of this study is to develop and evaluate a weather generating model that allows nearest neighbor resampling with perturbation of the historical data. The proposed model is capable of extrapolating beyond the observed record to produce precipitation and temperature values that are different from the observed values. Particular emphasis is placed on the simulation of extreme unprecedented precipitation events that are likely to be beneficial in improving the prediction of hydrologic extremes, including both floods and droughts. Evaluation of the model is through application to data from the Upper Thames River Basin (UTRB) in the Canadian province of Ontario.

The remainder of the paper is organized in the following manner. The next section describes model development and outlines the methodology used to adapt the *K*-NN algorithm for simulating daily weather sequences based upon the modified approach. The subsequent section describes the physical characteristics of the UTRB. Application of the algorithm to the basin along with a description of results is presented in the next section. The paper concludes with a summary of the results.

Model Development

Nearest neighbor methods have been intensively investigated in the field of statistics and in pattern recognition procedures. Despite their inherent simplicity, nearest neighbor algorithms are considered versatile and robust. Although more sophisticated alternative techniques have been developed since their inception, nearest neighbor methods remain very popular. A nearest neighbor algorithm typically involves selecting a specified number of data vectors similar in characteristics to the vector of interest. One of these vectors is randomly resampled to represent the vector of the given time step in the simulation period. In the context of weather data simulation, the nearest neighbor approach involves simultaneous sampling, with replacement, weather variables, like precipitation and temperature, from the observed data. To generate weather variables for a new day, $t+1$, days with similar characteristics to those simulated for the previous day t are first selected from the historical record. One of these nearest neighbors is then selected according to a defined probability distribution or kernel and the observed values for the day subsequent to that nearest neighbor are adopted as the simulated values for day $t+1$. Models based on the *K*-NN approach can easily be extended to multisite simulation of weather data while keeping the spatial correlation structure virtually intact. The spatial dependencies are preserved because the same day's weather is adopted as the weather for all stations. Apart from the spatial dependencies, temporal dependence is likely to be preserved as the simulated values for day $t+1$ are conditioned on the values for the previous day t. Further, the cross correlation among the variables at any given site is automatically preserved as a block of variables, rather than a single variable, is resampled from the observed data.

Consider that the daily historic weather vector consists of p variables. Suppose the number of stations considered in the model is q and data are available for N years. Let X_t^j denote the vector of weather variables for day t and station j, where $t=1,\ldots,T$, and $j=1,\ldots,q$; T being the total number of days in the observed time series. The feature vector for day t can be expressed, in expanded form, as $X_t^j=(x_{1,t}^j,x_{2,t}^j,\ldots,x_{p,t}^j)$ where $x_{i,t}^j$ represents the value of the weather variable i for station j. Suppose that the simulation begins on day t corresponding to January 1. The algorithm cycles through various steps to obtain the weather for day $t+1$. The procedure continues for all 365 days of a given year and the whole procedure is repeated to generate data for as many years as required. The steps of the algorithm are as follows:

1. Compute regional means of the p variables across the q stations for each day of the historical record

$$\bar{X}_t=(\bar{x}_{1,t},\bar{x}_{2,t},\ldots,\bar{x}_{p,t}) \tag{1}$$

where

$$\bar{x}_{i,t}=\frac{1}{q}\sum_{j=1}^{q}x_{i,t}^j,\quad i=1,\ldots,p,\quad \text{and } t=1,\ldots,T \tag{2}$$

2. Determine the size, L, of the data block that includes all potential neighbors to the current feature vector from which the resampling is to be done. A temporal window of width w is chosen and all days within the window are considered as potential candidates to the current feature vector. Yates et al. (2003) used a temporal window of 14 days, which implied that if the current day is January 20 then the window of days consists of all days between January 13 and January 27 for all N years but excluding January 20 for the given year. Thus, the data block of potential neighbors from which to resample consists of $L=(w+1)\times N-1$ days.
3. Compute mean vectors across q stations for each day in the data block consisting of potential neighbors using the expressions given in Step 1.
4. Compute the covariance matrix, C_t, for day t using the data block of size $L\times p$.
5. The weather on the first day t (e.g., January 1) comprising all p variables at q stations is randomly chosen from the set of all January 1 values in the historic record of N years. The algorithm cycles through the following steps to select one of the nearest neighbors to represent the weather for day $t+1$ of the simulation period.
6. Compute Mahalanobis distances (Davis 1986) between the mean vector of the current day's weather $\bar{X}_t$ and the mean vector $\bar{X}_i$ for day i, where $i=1,\ldots,L$. The distance metric can be defined through

$$d_i = \sqrt{(\bar{X}_t - \bar{X}_i)C_t^{-1}(\bar{X}_t - \bar{X}_i)^T} \tag{3}$$

   where T represents the transpose operation; and C_t^{-1}=inverse of the covariance matrix. Yates et al. (2003) used the Mahalanobis distance metric to determine the closeness of any given neighbor to the current vector as it does not require explicit weighting and standardization of the variables.
7. Determine the number of first K nearest neighbors to be retained for resampling out of the total of L neighbors. Lall and Sharma (1996) suggested the use of the generalized cross validation score (GCV) for choosing K. Rajagopalan and Lall (1999) and Yates et al. (2003) recommended the use of a heuristic method for choosing K according to which $K=\sqrt{L}$.
8. Sort the Mahalanobis distances in ascending order and retain the first K nearest neighbors. A discrete probability distribution that gives higher weights to the closer neighbors was used for resampling from the K nearest neigbors. Weights are assigned to each of these j neighbors according to the metric defined by

$$w_j = \frac{1/j}{\sum_{i=1}^{K} 1/i} \tag{4}$$

   The cumulative probabilities, p_j, are given by

$$p_j = \sum_{i=1}^{j} w_i \tag{5}$$

   The neighbor with the smallest distance is assigned the highest weight, while the neighbor with the largest distance (i.e., the Kth neighbor) gets the least weight. Lall and Sharma (1996) developed this function through a local Poisson approximation of the probability density function of state space neighbors.
9. Determine the nearest neighbor of the current day by using the cumulative probability metric given by Eq. (5). Generate a random number, $r\subset(0,1)$ and if $p_1<r<p_K$, then the day j for which r is closest to p_j is selected. If $r\leqslant p_1$, the day corresponding to d_1 is selected and if $r=p_K$, then the day corresponding to d_K is selected. The observed values for the day subsequent to the selected nearest neighbor are adopted to represent the weather for day $t+1$. In the modified approach presented here, the data points resampled using the basic K-NN approach are perturbed by adding a random component as described in Step 10 below.
10. For each station and each variable, a nonparametric distribution is fit to the K nearest neighbors identified in Step 8. This involves estimating the conditional standard deviation, σ, and the bandwidth, λ (Sharma et al. 1997; Sharma and O'Neill 2002). Perturbation of the values of weather variables obtained using the basic K-NN approach is carried out in the following steps:
    a. Let σ_i^j be the conditional standard deviation of variable i for station j computed from the K nearest neighbors. Let z_{t+1} be a random variate for day $t+1$ in the simulation period from a normal distribution with zero mean and unit variance. The new value of weather variables i for day $t+1$ and station j is given by

$$y^j_{i,t+1} = x^j_{i,t+1} + \lambda\sigma^j_i z_{t+1} \tag{6}$$

       where $x^j_{i,t+1}$=value of the weather variable for day $t+1$ and station j obtained from the basic K-NN model; $y^j_{i,t+1}$=corresponding value obtained after perturbation; and λ=bandwidth (a function of the number of samples) determined following Sharma et al. (1997). The perturbed resampled value obtained at time step t is not used in the neighbor calculation for day $t+1$, but rather the original resampled value is used.
    b. Since the precipitation values are bounded, there is a possibility that Eq. (6) in the above step could lead to negative precipitation amounts. Setting these negative values to zero would lead to bias that might produce monthly totals higher than the observed values, which is unacceptable. To overcome this problem, the bandwidth is transformed if the probability of generating a negative value is too large. A threshold probability, α, for generating a negative value is selected. Sharma and O'Neill (2002) use $\alpha=0.06$ for which $z=-1.55$. The largest value of λ corresponding to the probability of generating a negative value of exactly α is therefore given by $\lambda^a=x^j_{3,t+1}/(1.55\times\sigma^j_3)$, where subscript 3 refers to precipitation values and λ^a=acceptable (largest) value of λ. If the calculated value of λ is larger than λ^a, then λ^a is used instead of λ.
    c. If the precipitation computed in Step 10b is still negative, a new value of the random variate is generated and the value of precipitation recomputed from Eq. (6).
    d. Step 10c is repeated until the generated value of precipitation becomes non-negative.

Steps 6–10 are repeated to generate as many years of synthetic data as required. If multiple sequences of data are required, then the algorithm starts at Step 5. The modified approach presented here recognizes that the variability associated with low precipita-

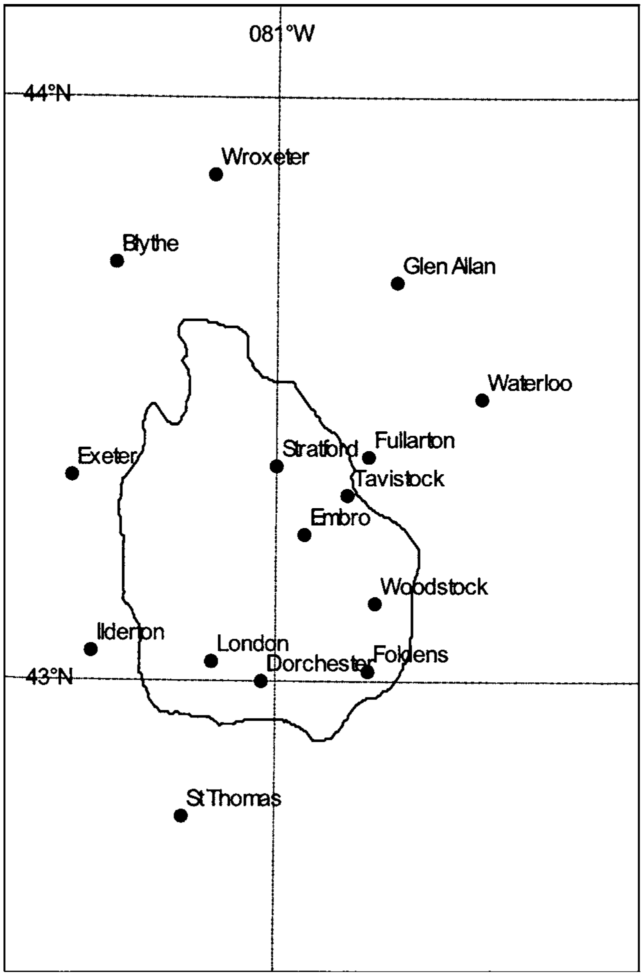

Fig. 1. Geographical location of different stations in basin

tion values is significantly smaller than that associated with higher precipitation values. A certain amount of bias is introduced due to the use of a new value of the random variate in case the computed value of precipitation is negative. However, the overestimation of precipitation amounts caused by this bias is insignificant as can be seen from the model results presented in the following sections of the paper.

Upper Thames River Basin

The Thames River Basin is located in the agricultural heartland of the southwestern region of the Canadian province of Ontario. The Thames River is the major river of the basin. It is 273 km long and has a catchment area of around 5,825 km^2, making it the second largest watershed in southwestern Ontario. Southwestern Ontario is a highly developed region and as such, the basin faces pressures from urban and rural land uses. Most of the precipitation comes in the form of winter snow. Rainfall occurs mainly in spring, with some in fall.

Daily maximum temperature (TMX), minimum temperature (TMN), and precipitation (PPT) data from 15 stations in and around the basin were used for the period 1964–2001. Hence, for our model, $p=3$. The geographical location of the stations is shown in Fig. 1. The data set used in this study is Environment Canada corrected. The mean annual values of weather variables and the latitude and longitude of each meteorological station are presented in Table 1. The meteorological stations in the basin are distributed across an area of approximate dimension 80 km (east–west) by 120 km (north–south). The interstation distances range from approximately 10 to 120 km.

Table 1. Station Characteristics

Station	Latitude (degrees N)	Longitude (degrees W)	Mean annual TMX (°C)	Mean annual TMN (°C)	Mean annual PPT (mm)
Blythe	43°43′	81°23′	11.3	2.2	1,159
Dorchester	43°0′	81°2′	12.3	2.6	1,034
Embro	43°15′	80°56′	11.9	2.5	984
Exeter	43°21′	81°29′	12.0	2.8	1,008
Foldens	43°1′	80°47′	11.9	3.2	945
Fullarton	43°23′	80°47′	11.8	2.5	1,012
Glen Allan	43°41′	80°43′	10.9	1.7	989
Ilderton	43°3′	81°26′	12.7	3.3	1,008
London	43°2′	81°9′	12.4	2.4	980
Stratford	43°22′	81°0′	11.4	2.4	1,056
St. Thomas	42°46′	81°13′	12.9	3.0	985
Tavistock	43°19′	80°50′	11.8	2.5	1,048
Waterloo	43°29′	80°31′	11.6	1.6	915
Woodstock	43°8′	80°46′	12.45	2.52	941
Wroxeter	43°52′	81°9′	11.25	2.18	995

There were a large number of missing temperature records in the available data which were infilled using a two-step procedure. In the first step, missing records for London, Ilderton, Foldens, Stratford, and Woodstock were infilled with the mean daily values. Once the data set for these five stations was complete, the missing records for Embro, Dorchester, Tavistock, and Fullarton were infilled. At these four stations, the precipitation records were available but maximum temperature and minimum temperature records were missing for the entire period. The weighted average inverse distance square method was used to estimate the missing temperature data for these four stations. The method has the advantage that the estimated values will always be less than the greatest and greater than the smallest value at the surrounding stations. The calculation of distances between various stations was based on the latitude and longitude of the stations that are shown in Table 1. It was not possible to use the weighted average inverse distance square method for London, Foldens, Stratford, and Woodstock data as there were many days in the record for which the data are missing for either one or more of the remaining stations. The observed record for St. Thomas was also incomplete. Missing values for St. Thomas were computed from the available records of Ilderton, London, and Dorchester.

Model Application

Model Parameters

The performance of the *K*-NN model depends upon the proper setting of various parameters whose values must be determined before the evaluation of the model can be carried out. Two important model parameters in a *K*-NN based resampling approach are the width of the temporal window, *w*, and the number, *K*, of nearest neighbors. It is worth mentioning that although *K* and *w* are parameters the method itself is nevertheless nonparametric and does not require specifying model parameters from the observed data (Karlson and Yakowitz 1987). In parametric models, the effect of seasonality is taken into account by fitting different model parameters to each season, whereas it is through the moving window that the seasonal characteristics of the observed data are reproduced in a K-NN resampling technique. Because the search for nearest neighbors is restricted to days within a moving window, the effect of seasonal variation is greatly reduced. A subset of days within a moving window, centered on the Julian day of interest, is selected and all days within the temporal window are potential candidates for the weather on the given day in the simulation period. Therefore, the width of the moving window must be sufficiently large such that the dependence between the observations outside the moving window can be neglected. A fixed length 14-day temporal window was used in this study. For $w=14$ and $N=38$, the total number of potential candidates consists of L days ($L=569$) as described in Step 2.

The choice of K is vital for good performance of the model. The value of K depends on the type of kernel used for resampling, the number, L, of days from which the nearest neighbors are selected, and the dimension of the feature vector (Buishand and Brandsma 2001). A simple approach to determining K is to try many values and obtain a satisfactory value by trial and error but other approaches are also available. Lall and Sharma (1996) recommended a heuristic value of $K=\sqrt{L}$. Buishand and Brandsma (2001) observed that for a decreasing kernel, the reproduction of autocorrelation coefficients gradually deteriorates with increasing K. On the other hand, resampling with a small number of nearest neighbors might lead to duplication of large parts of the historical record and to repeated sampling of the same historical values. Rajagopalan and Lall (1999) and Yates et al. (2003) found that the heuristic method of choosing K led to good model performance. In our case $L=569$ and hence a value of $K=24$ has been adopted. With the parameters of the model defined, it is possible to generate weather data based upon the modified approach and evaluate the performance of the model.

Reproduction of Historical Statistics

The performance of the proposed model was evaluated through application to data from the Upper Thames River Basin. A new subset of years that constitute the driving data for the model was obtained by using an integer function that returned integers between specified upper and lower bounds. To generate N years of data, the integer function was called N times. With this method, each year has an equal probability of being selected but some years may be selected more than once. A new data set was thus obtained and the K-NN algorithm was used to generate 800 years of synthetic data with this data set. The goal of simulation was to produce a data series that preserved the statistical attributes of the historic data while perturbing the data points. The statistics of interest were computed from the simulated sequence and compared to the statistics of the observed record using box plots. Box plots are a preferred method of data analysis as they show the range of variation in the statistics of simulations and provide a straightforward method of comparing the statistics of simulations with the historical data. Results are presented below only for London since the results for other stations are similar.

Fig. 2 shows the box plots of 800 simulated values of mean TMX values for London. Although the model was applied on daily data, the statistics from the daily data have been aggregated to a monthly time scale to facilitate presentation of the results. The statistics of simulations are shown by box plots while the solid lines with dots represent the mean of the monthly values from the historical data. Comparison of historical monthly values with the simulated values clearly showed that the model was able to adequately reproduce the historical values. This is highly satisfactory given that monthly statistics were not explicitly specified in fitting the K-NN model.

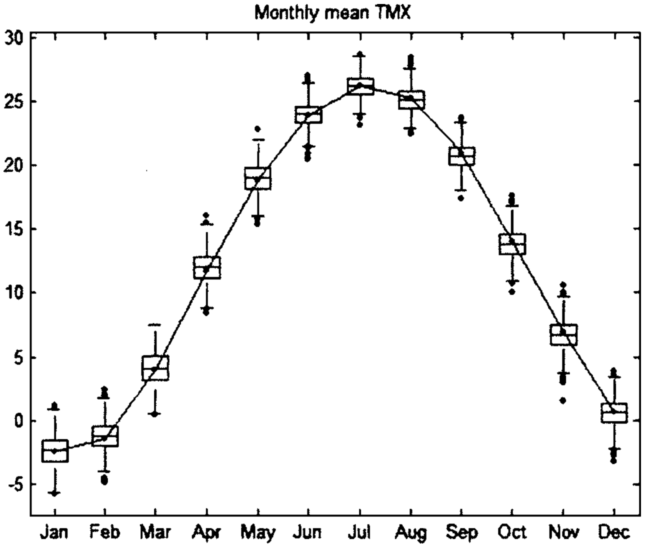

Fig. 2. Box plots of monthly mean maximum temperature

Fig. 3 provides box plots of total monthly precipitation for London. It can be seen from the box plots that the historical mean of the total precipitation is close to the median of the simulated data for all the months. A number of values were found to lie beyond the whiskers but these outliers are indicative of the variability in the simulated data. The total annual precipitation simulated by the model (987 mm) matched very closely the historical value (980 mm). The model slightly overestimated the monthly totals for February, August, September, and November. For the rest of the months, model results are very close to the observed values. Among all weather variables, precipitation has the greatest variability in time and space and therefore the performance of the model in simulating the total monthly precipitation may be considered to be very good. Since kernel based perturbation tends to

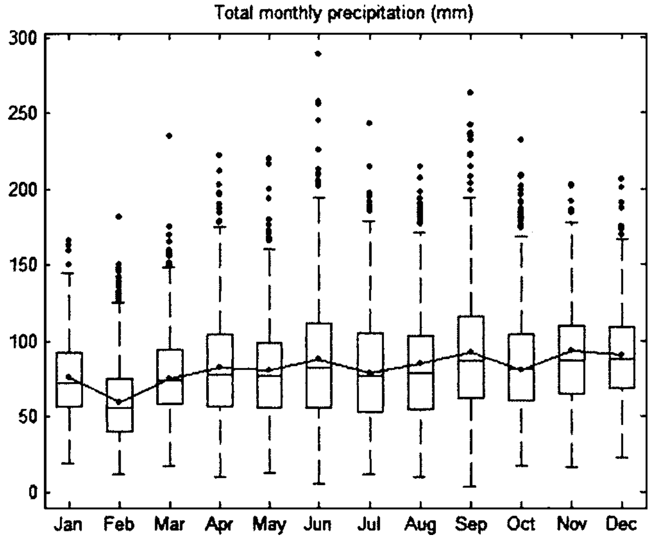

Fig. 3. Box plots of total monthly precipitation

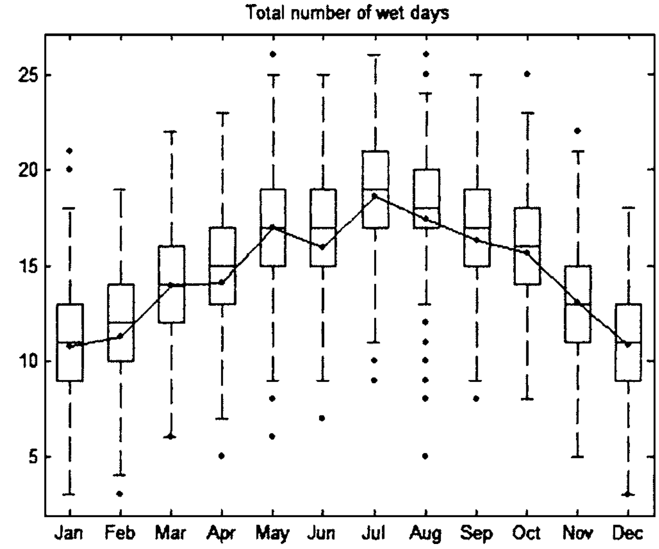

Fig. 4. Box plots of total number of wet days

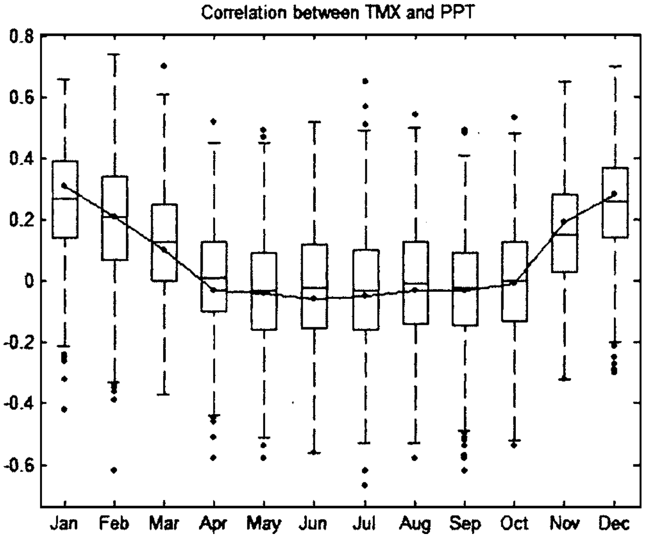

Fig. 5. Box plots of correlation between TMX and PPT

increase the variance of the simulated values, the distributions of monthly standard deviations for TMX and PPT were investigated. The standard deviations of the simulated values of TMX were found to be larger than the historical values. Interestingly, for PPT the simulated standard deviations were in good agreement with the historical values. The simulations also adequately reproduced the probability distributions of historical values.

Fig. 4 shows box plots of total number of wet days for London. This statistic is important for sequences that are generated for use in crop production and flood management models. Box plots in Fig. 4 indicate that the model adequately reproduced the total number of wet days in different months of the year. There was a slight overestimation for the months of April, June, and August, but the overall results are satisfactory.

Preservation of Correlation Structure

Parametric models often fail to reproduce the correlation structure of the observed data. Due to the inherent structure of the basic *K*-NN model, there is a strong likelihood of the correlation structure being preserved. With the modified model presented here, the correlation structure of the observed data might be tempered. To keep the correlation structure intact, it was decided to use a constant value of the random normal variate for all the variables and all the stations at any given time step. The extent to which correlation structure might change with the approach presented here was then investigated. Box plots for correlation between TMX and precipitation and autocorrelation of PPT are shown in Figs. 5 and 6, respectively. It can be observed from the box plots shown in Fig. 5 that the observed data have a positive correlation between TMX and PPT during the winter months while the correlation is very close to zero during the summer months, thus indicating a statistically insignificant correlation for these months. These seasonal correlation characteristics are adequately reproduced by the *K*-NN model, as shown by the box plots.

Simple resampling schemes tend to destroy prominent time correlations of the observed data but the *K*-NN scheme resamples from the observed data by conditioning on the weather for the previous day, and is therefore more likely to preserve important time correlations. The model proposed here involves perturbations of the observed data points and as such does not explicitly reproduce daily correlations. Clark et al. (2004) present a method to preserve daily correlations that involves reordering the ensemble members to reconstruct the spatial and temporal correlation statistics of the observed data. The performance of the proposed model in reproducing autocorrelation of precipitation data at monthly time scale was investigated, and is presented through box plots of autocorrelation of precipitation shown in Fig. 6. It can be seen from the box plots that the mean values of autocorrelation coefficients of the historical record for different months are close to zero, which implies a very weak lag-1 autocorrelation of PPT, and the model adequately captured these characteristic of the observed data.

For agricultural models, weather data can be generated separately at different sites without taking into account spatial correlations because the interaction between processes at different sites is often weak. In hydrological models, especially those dealing with flood prediction, the spatial distribution of the generated precipitation amounts is crucial. Many studies have shown that the lack of spatially distributed precipitation amounts can have a serious impact on basin runoff generation (Shah et al. 1996; Yang et

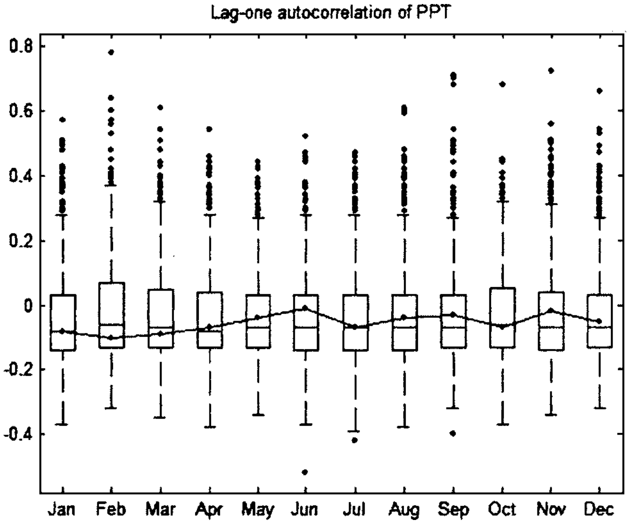

Fig. 6. Lag-one autocorrelation of PPT at London

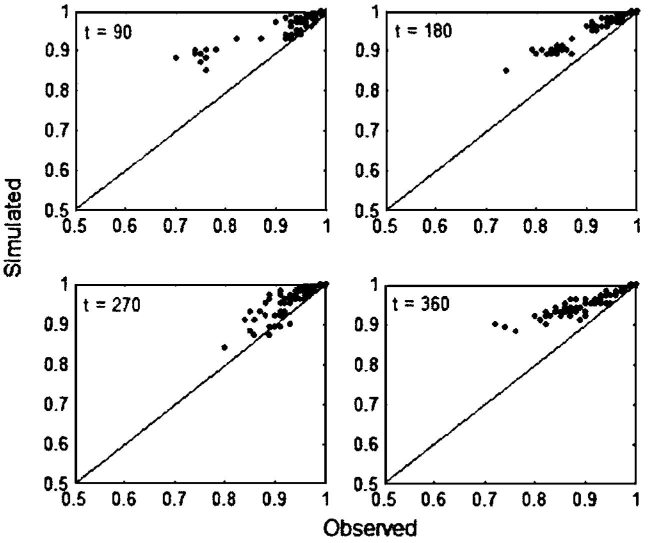

Fig. 7. Comparison of observed versus simulated interstation correlations for daily TMX values between all station pairs for 4 representative days

Table 2. Average and Largest Precipitation Values at Various Stations

	Average annual PPT (mm)		Largest PPT value (mm)	
Station	Observed	Simulated	Observed	Simulated
Blythe	1,159	1,161	137	153
Dorchester	1,034	1,042	94	112
Embro	984	990	107	122
Exeter	1,008	1,018	159	179
Foldens	945	952	110	126
Fullarton	1,012	1,017	106	118
Glen Allan	989	994	104	118
Ilderton	1,008	1,010	99	107
London	980	987	89	98
Stratford	1,056	1,068	137	155
St. Thomas	985	992	89	105
Tavistock	1,048	1,051	94	109
Waterloo	915	923	90	101
Woodstock	941	954	114	127
Wroxeter	995	998	166	187

al. 1998). The assumption of uniform spatial precipitation distribution is often invalid, even for small basins, because runoff simulation is significantly impacted by the distribution of precipitation over the basin. Therefore, it was considered important to evaluate the performance of the model in reproducing the spatial dependencies of the observed data, especially since the proposed approach involves perturbing the observed data points. Scatter plots of interstation correlations for daily TMX and precipitation values are presented in Figs. 7 and 8, respectively.

Fig. 7 shows scatter plots of interstation correlation coefficients for daily TMX values in the simulated and the observed data. For 15 stations, there are 105 pairwise correlation coefficients for each day. The scatter plots have been shown for 4 representative days. As can be seen from Fig. 7, there are pronounced interstation correlations between TMX values across the basin, mostly in the range of 0.8–1.0. Most data points lie in the close vicinity of the 45° sloping solid line shown in the scatter plots. However, the stronger correlations appear to be preserved better than the weaker correlations, which is a desirable outcome. Additionally, the simulated spatial correlations are higher than the observed ones possibly due to the same random number being used for all variables and all sites to perturb the resampled points.

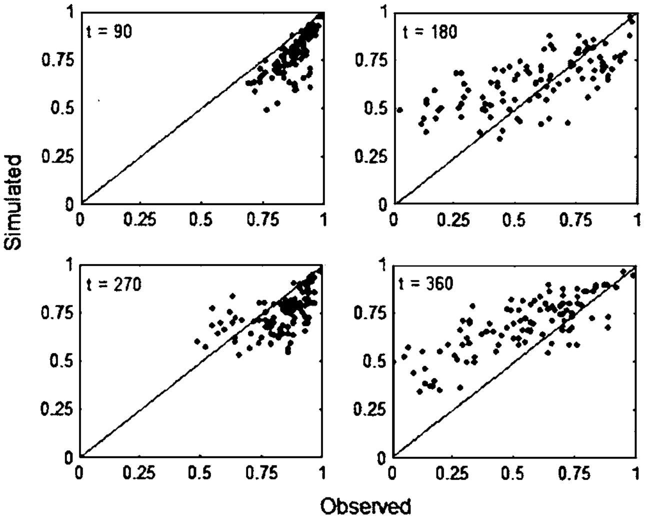

Fig. 8. Comparison of observed versus simulated correlations for daily PPT values between all station pairs for 4 representative days

Given the structure of the basic *K*-NN model, the spatial characteristics of the observed data are bound to be preserved on a daily time scale. Small deviations of data points from the 45° line may be attributed to the perturbations made to the observed values. However, the overall performance of the *K*-NN model in reproducing the historical interstation correlation structure is satisfactory.

The scatter plots of interstation correlations of daily PPT values between the observed and the simulated data are shown in Fig. 8. Although the correlations in the observed data are not as strong as in the case of TMX, the model reproduced the historical structure very well. It was observed that the standard and modified *K*-NN models give similar results in terms of the statistics reported in Figs. 2–8. Well known parametric models such as LARS-WG and WGEN can be effectively used to generate independent weather data for any number of stations but they cannot be expected to preserve important interstation correlations of the variables. However, some parametric models are available that are capable of preserving spatial correlations (e.g., Hughes and Guttorp 1994; Wilks 1998). With the *K*-NN model, the spatial dependence is preserved by resampling simultaneously the same day's weather as the weather for all the stations. This feature of the *K*-NN model makes it an attractive option for use in conjunction with hydrological models where the spatial dependencies may be crucial for the accuracy of runoff predictions.

Extreme Precipitation Events Simulation

A major focus of this study was to evaluate the performance of the proposed model in simulating precipitation amounts larger than the observed amounts. In addition, the effect of perturbations on the reproduction of annual average precipitation needs to be investigated. Table 2 summarizes the results of simulation with respect to reproduction of long-term average annual precipitation.

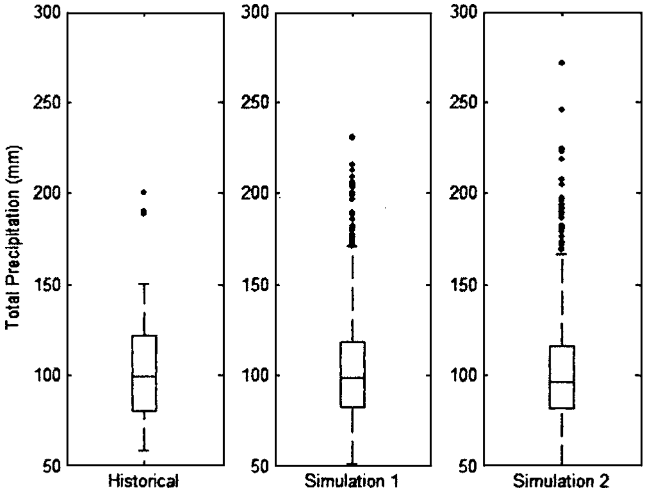

Fig. 9. Box plots of total precipitation during extreme events in each year of historical and simulated data (Simulation 1 refers to basic *K*-NN model, Simulation 2 refers to modified model)

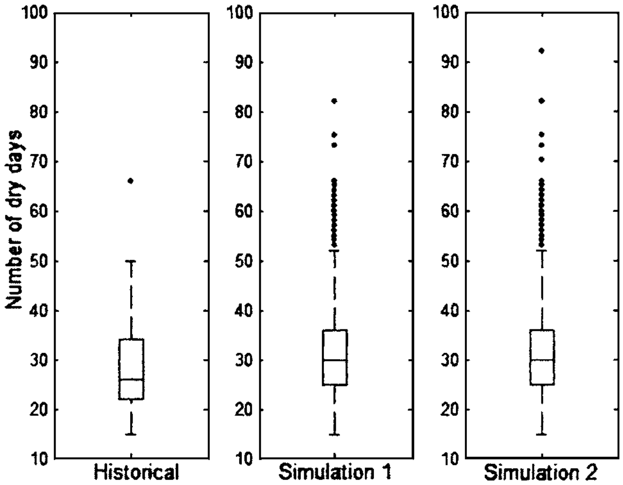

Fig. 10. Box plots of dry days during extreme events in each year of historical and simulated data (Simulation 1 refers to basic *K*-NN model, Simulation 2 refers to modified model)

It can be seen from Table 2 that the model consistently overestimates the annual average precipitation but the amount of overestimation is negligibly small. The reason for this overestimation is the bias due to recomputation of the normal random variate whenever the precipitation amounts become negative. Overall, the model yielded practically an exact reproduction of the observed long-term average annual precipitation. A comparison of the largest daily precipitation amounts simulated by the model with the observed values is also presented in Table 2. It can be observed that the simulated amounts are significantly higher than the observed amounts. The model was able to generate a precipitation amount of 187 mm compared to the historical largest value of 166 mm. In addition, the model produced several storm depths greater than the observed values. The results presented in Table 2 clearly show that the model was able to produce unprecedented, but realistic, precipitation amounts throughout the basin.

The output from the *K*-NN model is intended for use in conjunction with a hydrological model of the basin with the objective of assessing the vulnerability of the basin to floods and droughts. Prolonged precipitation events during the winter season combined with heavy rainfall during summer are the most probable cause of flooding in the basin. Particular attention is therefore given to the simulation of extreme precipitation events that are responsible for floods. Similarly, it is important to determine dry spell characteristics of the simulated data in order to gain insight into the possibility of drought in the basin. The ability of the model to simulate the occurrence of extreme events, both high precipitation and low precipitation, was therefore investigated with particular emphasis on generating realistic, but unprecedented, events for the basin. For each year of the historical and simulated record, the single most extreme multiday precipitation event was determined. Fig. 9 shows the box plots of total precipitation that occurred during the most extreme precipitation event in each year of the historical and the simulated records. The results are shown for both the basic *K*-NN model (Simulation 1) and the modified model (Simulation 2). It can be seen from the box plots that in both the simulations the median of the simulated data matches very closely the median of the historical data. Due to the perturbations made to the observed data points, total precipitation on the order of 280 mm in the most extreme precipitation event was observed in Simulation 1. The modified model simulated around five precipitation events that were more severe than the most extreme precipitation event in the observed record. The total precipitation amount of the simulated largest event in Simulation 1 is around 40% higher than the corresponding amount in the extreme observed event while it is only 15% higher in Simulation 2. The interannual variability in the simulated data is quite prominent with a highest total precipitation of about 280 mm compared to a corresponding value of around 200 mm in the historical record. The basic *K*-NN model simulated a most extreme precipitation event with a total precipitation of 230 mm.

Fig. 10 shows the duration of dry spells during extreme events in each year of the historical and the simulated records. As can be seen from the box plots, the median of dry spell durations for the observed data is 25 and the corresponding value in the simulated data is 30. The difference may be attributed to the nature of the modified model, which tends to produce events more severe than in the observed data. Again, the median of the simulated data matched the observed data well although the median value for the simulated data was slightly higher. This clearly indicates that the tendency of dry days to exhibit persistence is adequately represented by the model. The interannual variability in the simulated sequences is also quite evident with the model producing several severe low precipitation events other than those in the historical record. The longest dry spell in the historical data lasted for 67 days but the *K*-NN model was able to simulate several dry spells having a duration more than that observed in the historical data. The largest dry spell in Simulation 1 has a duration of 93 days, which is much more severe than seen in the observed record. With the basic *K*-NN model (Simulation 2), the most extreme dry spell has a duration of 82 days.

Fig. 11 shows the box plots of extreme wet spells in the historical and the simulated sequences. The most severe wet spell simulated by the modified model was on the order of 52 days, while the corresponding value obtained from the basic model was 48. The strength of the perturbation model presented here lies in the simulation of such extreme dry and wet spells that are important for the evaluation of effective drought and flood management policies for the basin. The results presented in Figs. 10 and 11 clearly show that the modified model provides greater variability associated with sustained periods of precipitation and dry days than is possible with the basic *K*-NN model. Comparison of re-

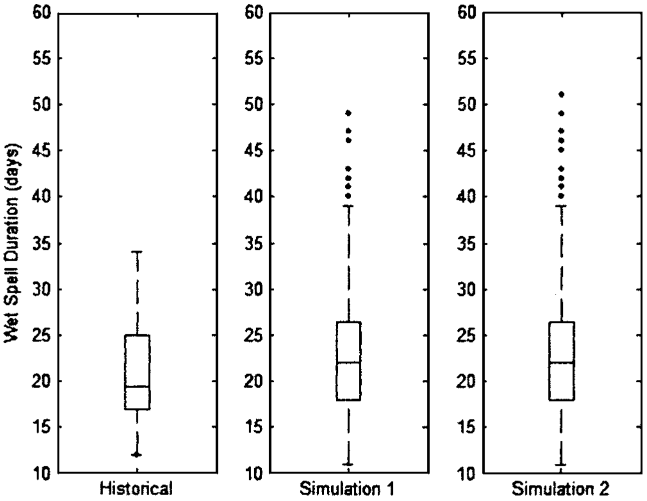

Fig. 11. Box plots of extreme wet spell duration in each year of historical and simulated data (Simulation 1 refers to basic *K*-NN model, Simulation 2 refers to modified model)

sults presented in Figs. 9–11 clearly indicates that the modified model was able to simulate more extreme events than is possible with the basic model.

Conclusions

The development and evaluation of a modified version of the basic *K*-NN weather generator has been presented. The modified model was shown to produce precipitation amounts different from those observed in the historical record, thereby alleviating a common problem associated with the basic *K*-NN approach. A strategy has been devised that allows perturbation of the observed data points by adding a random component to the values suggested by the basic approach while preserving the important statistical characteristics of the observed data. This approach is similar in spirit to traditional autoregressive models where a random component is added to the conditional expectation of the variable to obtain the new values. However, in the proposed approach, the random component is added to the individual resampled data values. The practicality of the approach was demonstrated through application to data from the Upper Thames River Basin. Comparison of observed and simulated data clearly indicated that the model performance was very good with regards to reproduction of various statistics of interest to a hydrologist. Important properties of precipitation spell structure and amounts were preserved. Cross correlation among the variables was preserved, which is particularly important for erosion, crop production, and rainfall– runoff models. An important output of the model is the spatially correlated data for the basin, which is important for evaluating the response of hydrological models to watershed-level processes. Unlike well known models such as LARS-WG and WGEN, which cannot be expected to preserve the spatial dependencies of the variables, the proposed model adequately reproduced the spatial correlation of the observed data.

An encouraging aspect of the proposed model is that extreme unprecedented events, both low precipitation and high precipitation, can be simulated. This allows for evaluation of the response of rainfall–runoff models for a wide variety of simulated data, especially the extremes. The proposed model has the potential for providing valuable aid in developing efficient flood and drought management strategies for the basin because of the ability of the model to simulate extreme dry and wet spells. It may be concluded that the utility of flood prediction models in estimating the probability of extreme events may be greatly enhanced if their performance is evaluated based on synthetic sequences generated by the type of model proposed here. Although the *K*-NN algorithm was designed to model daily statistics, the monthly statistics are also adequately reproduced for the application presented here. An additional practical advantage of the model is that it does not require site-specific assumptions regarding the probability distribution of the variables.

Acknowledgments

The research presented in this paper was funded by the Canadian Foundation for Climatic and Atmospheric Sciences. The writers gratefully acknowledge the cooperation received from Professor S. P. Simonovic and Dr. J. M. Cunderlik, University of Western Ontario. The writers also thank Mr. Mark Helsten for providing meteorological data for the Upper Thames River Basin. This paper was improved by helpful comments from two anonymous reviewers. The writers are particularly grateful for useful suggestions that led to revisions to the approach used to perturb the data from the basic *K*-NN model.

References

Brandsma, T., and Buishand, T. A. (1998). “Simulation of extreme precipitation in the Rhine basin by nearest neighbor resampling.” *Hydrology Earth Syst. Sci.*, 2(2–3), 195–209.

Buishand, T. A. (1978). “Some remarks on the use of daily rainfall models.” *J. Hydrol.*, 36(3–4), 295–308.

Buishand, T. A., and Brandsma, T. (2001). “Multisite simulation of daily precipitation and temperature in the Rhine Basin by nearest-neighbor resampling.” *Water Resour. Res.*, 37(11), 2761–2776.

Clark, M. P., Gangopadhyay, S., Brandon, D., Werner, K., Hay, L., Rajagopalan, B., and Yates, D. (2004). “A resampling procedure for generating conditioned daily weather sequences.” *Water Resour. Res.*, 40(4), 1–15.

Davis, J. (1986). *Statistics and data analysis in geology*, Wiley, New York.

Hanson, C. L., and Johnson, G. L. (1998). “GEM (generation of weather elements for multiple applications): Its application in areas of complex terrain.” *Hydrology water resources and ecology in headwaters*, K. Kovar, U. Tappeiner, N. E. Peters, and R. G. Craig, eds., International Association of Hydrological Sciences Press, Wallingford, U.K., 27–32.

Hughes, J. P., and Guttorp, P. (1994). “Incorporating spatial dependence and atmospheric data in a model of precipitation.” *J. Appl. Meteorol.*, 33(12), 1503–1515.

Karlson, M., and Yakowitz, S. (1987). “Nearest neighbor methods for non-parametric rainfall-runoff forecasting.” *Water Resour. Res.*, 23(7), 1300–1308.

Katz, R. W. (1977). “Precipitation as a chain-dependent process.” *J. Appl. Meteorol.*, 16(7), 671–676.

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

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

Nicks, A. D., Richardson, C. W., and Williams, J. R. (1990). “Evaluation of EPIC model weather generator: Erosion/productivity impact calculator. 1: Model documentation.” *USDA—ARS tech. bull. 1768*, A. N.

Sharpley and J. R. Williams, eds., Washington, D.C.

Nicks, A. D., and Harp, J. F. (1980). “Stochastic generation of temperature and solar radiation data.” *J. Hydrol.*, 48(1–2), 1–7.

Parlange, M. B., and Katz, R. W. (2000). “An extended version of the Richardson model for simulating daily weather variables.” *J. Appl. Meteorol.*, 39(5), 610–622.

Rackso, P., Szeidi, L., and Semenov, M. (1991). “A serial approach to local stochastic weather models.” *Ecol. Modell.*, 57(1–2), 27–41.

Rajagopalan, B., and Lall, U. (1999). “A k-nearest neighbour simulator for daily precipitation and other variables.” *Water Resour. Res.*, 35(10), 3089–3101.

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

Richardson, C. W., and Wright, D. A. (1984). “WGEN: A model for generating daily weather variables.” *ARS-8*, U.S. Dept. of Agriculture, Agricultural Research Service, Washington, D.C.

Semenov, M. A., and Barrow, E. M. (1997). “Use of a stochastic weather generator in the development of climate change scenarios.” *Clim. Change*, 35(4), 397–414.

Semenov, M. A., Brooks, R. J., Barrow, E. M., and Richardson, C. W. (1998). “Comparison of WGEN and LARS-WG stochastic weather generators for diverse climates.” *Climate Res.*, 10(2), 95–107.

Shah, S. M. S., O’Connell, P. E., and Hosking, J. R. M. (1996). “Modelling the effect of spatial variability in rainfall on catchment response. 2: Experiments with distributed and lumped models.” *J. Hydrol.*, 175(1–4), 89–111.

Sharma, A., and O’Neill, R. (2002). “A nonparametric approach for representing interannual dependence in monthly streamflow sequences.” *Water Resour. Res.*, 38(7), 5-1–5-10.

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

Smith, R. L. (1994). “Spatial modelling of rainfall data.” *Statistics for the environment. 2: Water related issues*, V. Barnett and K. F. Turkman, eds., Wiley, New York, 19–41.

Smith, R. E., and Schreiber, H. A. (1974). “Point processes of seasonal thunderstorm rainfall. 2: Rainfall depth probabilities.” *Water Resour. Res.*, 10(3), 418–423.

Stern, R. D., and Coe, R. (1984). “A model fitting analysis of rainfall data.” *Stat. Soc., Ser. A.*, 147, 1–34.

Todorovic, P., and Woolhiser, D. A. (1975). “A stochastic model of n-day precipitation.” *J. Appl. Meteorol.*, 14(1), 17–24.

Wilby, R. L. (1994). “Stochastic weather type simulation for regional climate change impact.” *Water Resour. Res.*, 30(12), 3395–3403.

Wilks, D. S. (1998). “Multisite generalization of a daily stochastic precipitation generation model.” *J. Hydrol.*, 210(1–4), 178–191.

Wilks, D. S. (1999). “Interannual variability and extreme value characteristics of several stochastic daily precipitation models.” *Agric. Forest Meteorol.*, 93(3), 153–169.

Wilks, D. S., and Wilby, R. L. (1999). “The weather generation game: A review of stochastic weather models.” *Prog. Phys. Geogr.*, 23(3), 329–357.

Woolhiser, D. A., and Roldan, J. (1982). “Stochastic daily precipitation models. 2: A comparison of distribution of amounts.” *Water Resour. Res.*, 18(5), 1461–1468.

Yang, D. Q., Goodison, B. E., and Ishida, S. (1998). “Adjustment of daily precipitation data at 10 climate stations in Alaska: Application of World Meteorological Organization intercomparison results.” *Water Resour. Res.*, 34(2), 241–256.

Yates, D., Gangopadhyay, S., Rajagopalan, B., and Strzepek, K. (2003). “A technique for generating regional climate scenarios using a nearest-neighbor algorithm.” *Water Resour. Res.*, 39(7), 7-1–7-14.

Young, K. C. (1994). “A multivariate chain model for simulating climatic parameters with daily data.” *J. Appl. Meteorol.*, 33(6), 661–671.