Simulating climate change scenarios using an improved K-nearest neighbor model

Mohammed Sharif, Donald H Burn *

Department of Civil Engineering, University of Waterloo, 200 University Avenue West, Waterloo, Ont., Canada N2L 3G1

Received 30 September 2004; revised 6 October 2005; accepted 11 October 2005

Abstract

An improved weather-generating model that allows nearest neighbour resampling with perturbation of the historic data is applied to generate weather data based upon plausible climate scenarios. The intent is to create an ensemble of climate scenarios that can be used for the assessment of the vulnerability of a watershed to extreme events, including both floods and droughts. Analysis of the results clearly indicates that the model adequately simulated extreme unprecedented events for five different climate change scenarios. Based on the simulation results, the increasing precipitation scenario is identified as the critical scenario for the assessment of risks associated with the occurrence of floods in the basin. The increasing temperature scenario appears to be the critical scenario for the analysis of droughts in the basin. Frequency analysis was carried out to determine the impact of potential climatic change on the occurrence of storm depths of any given magnitude. A promising potential application of the model is in rainfall-runoff modelling where the storms depths could be related to the occurrence of extreme events in the basin. The proposed model, in conjunction with a rainfall-runoff model, has the potential of providing valuable aid in developing efficient management strategies for a watershed. The model produces spatially correlated data, which is crucial for accurate runoff estimation. Although the model is applied to the Upper Thames River Basin in the Canadian province of Ontario, it is generic and transportable to any other watershed with minimal changes.

© 2005 Elsevier B.V. All rights reserved.

Keywords: Weather generator; Climate change; K-NN; Perturbation; Stochastic; frequency

1. Introduction

Global climate is expected to change significantly due to the continuously increasing levels of carbon dioxide and other greenhouse gases. Many aspects of the natural environment, including water resources, are anticipated to experience potentially serious climatic impacts. Reservoir operations, crop production, erosion processes, runoff production and many other hydrological processes are likely to be impacted by climate change. Revelle and Waggoner (1983); Gleick (1987) have indicted that climate change can adversely affect the availability of water supply. Climate change impacts are expected to result in not only changes in the average availability of

* Corresponding author. Tel.: +1 519 888 4567; fax: +1 519 888 4349.
E-mail addresses: msharif@uwaterloo.ca (M. Sharif), dhburn@uwaterloo.ca (D.H. Burn).
water but also changes in the extremes (Burn, 1998; Simonovic and Li, 2003). Other impacts of climate change that have been identified include changes in the quantity of runoff produced (Gleick, 1986; Lettenmaier and Gan, 1990) and changes in the timings of hydrological events (Lettenmaier and Gan, 1990; Kite, 1993; Burn, 1994). Burn and Simonovic (1996) investigated the potential impacts of climate change on the performance of reservoir operations. Scenarios representing diverse sets of climatic conditions were generated and used as input to a reservoir operation model. It was concluded that reservoir performance is sensitive to climate change. Westmacott and Burn (1997) evaluated the effects of climatic changes on hydrological variables pertaining to the magnitude and timing of hydrological events in the Churchill–Nelson River Basin in West-central Canada. The magnitude of hydrological events was found to decrease over time while snowmelt runoff events occurred earlier. Mortsch et al. (2000) studied the impact of changing climatic conditions on the Great Lakes region under the scenario of doubling of CO₂ concentration. Climate change scenarios considered by Mortsch et al. (2000) indicated declines in runoff and lake levels that could lead to potential water allocation problems in the region. Southam et al. (1999) evaluated the impact of climate change in Ontario’s Grand River basin under 21 scenarios of future surface water supply, streamflow regulation, population and water use. They concluded that climate change might have serious impacts on the capability of the Grand River to assimilate wastewater and yield a reliable supply of water for municipal purposes while maintaining existing water quality standards.

A means of generating future climate scenarios is required for developing efficient climate adaptation strategies. At present, there is no ideal method for generating future climate scenarios (Gleick, 1989). Existing methods can be classified into three categories: empirically based, process based and linked methods that combine empirically based and process based concepts. Empirical models make use of historical observations to identify trends in important climatic variables such as temperature and rainfall and changes in important weather patterns such as the El Nino-southern oscillation (ENSO) cycle. Process based models use mathematical representations of the processes that govern atmospheric and oceanic circulation to estimate future climate variables and seasonal changes in climate. Global circulation models (GCMs) are the most sophisticated process based models that simulate the climate system. Empirical models can be applied for estimating climate change impacts using two different approaches. In the first approach, models can be built using observed data and the impacts estimated using GCM projections as input to the model (Ehrman et al., 2000). The accuracy of GCMs is, however, questionable due to their relatively large spatial resolution, which is of the order of 2 × 2.5⁰ (latitude × longitude). The second approach involves relating regional-scale variables to global scale atmospheric fields using empirical downscaling models (Hewitson, 1996; Landman et al., 2001). Typically, the global scale atmospheric variables are better resolved by GCMs than are the regional-scale temperature and precipitation fields.

Models that have been used in empirical downscaling studies include multiple linear regression (Klein, 1983), canonical correlation analysis (Landman et al., 2001), nearest neighbour (Zorita and von Storch, 1999) and artificial neural network approaches (Crane and Hewitson, 1998; Cannon and Whitfield, 2002). These models can account for variability in the surface variables to an acceptable level but extrapolation beyond the historical conditions may be unreliable. The ability of the downscaling models to predict climate change impacts is greatly reduced due to the coarse spatial resolution of GCMs. Such a resolution is unsatisfactory for catchment level hydrologic processes and gives rise to uncertainties when downscaling is carried out using the GCM outputs. Promising work addressing the issue of spatial resolution has been carried out by Hughes and Guttorp (1994); Wilby (1994). Still there is a great deal of uncertainty regarding the regional GCM output under future scenarios of increasing CO₂ and aerosol changes. Further, a climate change scenario based on the output from a GCM represents only one of the many future climate change scenarios whereas exploring several alternative climate scenarios would be more useful for effective management of water resource systems. Estimates of weather variables, particularly precipitation, on finer
geographic and temporal scales are needed to predict the potential effects of climate change on a regional scale.

Recently, weather generators have been employed for producing alternative climate data sets based upon prescribed conditioning criteria (e.g. Yates et al., 2003; Semenov et al., 1998). Weather generators can be broadly classified into two categories: parametric and nonparametric. Parametric weather generators have first focussed on independent generation of precipitation while the remaining variables are modelled conditionally on precipitation occurrence (e.g. Richardson, 1981; Nicks and Harp, 1980; Stern and Coe, 1984). 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, which takes into account the non-normal distribution of wind speed and relative humidity. Wind speed and dew point temperature (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 incorporate daily mean wind speed and dew point temperature in the model. A major drawback associated with the ‘Richardson type’ weather generators is that persistent events, such as drought or prolonged rainfall, are not very well reproduced. Models presented by Rackso et al. (1991); Semenov et al. (1998) overcome this problem. Rackso et al. (1991) used predefined distributions for modelling of wet and dry series whereas semi-empirical distributions are used in LARS-WG (Semenov and Barrow, 1997; Semenov et al., 1998). WGEN and LARS-WG are single-site models and therefore cannot simultaneously simulate weather data at multiple sites. Moreover, they require specification of model parameters and generally have difficulty in reproducing the annual variability in monthly means of the variables.

Nonparametric methods avoid the difficult model specification issues and can circumvent many other problems associated with the parametric methods. The most promising non-parametric technique for generating weather data is the $K$-nearest neighbour ($K$-NN) resampling approach. Recently, interest has emerged in the application of $K$-NN based 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); Yates et al. (2003) describe successful applications of $K$-NN resampling for simulation of weather data. A major limitation of these models is that they do not produce new values but merely reshuffle the historical data to generate synthetic weather data. Application of such sequences, in conjunction with hydrological models, to catchment response evaluation could lead to underexploration of the possible effects of climatic variability and to suboptimal policies for system management. A weather generator capable of producing precipitation amounts larger than observed, and consequently simulating more extreme events, would be advantageous in formulating effective flood and drought management strategies at the catchment level. This paper describes the development and application of an improved $K$-NN based weather generator to simulate weather data conditioned upon plausible climate scenarios. The model results are derived from future climate scenarios that are consistent with projections based on GCM results. The model is applied to the Upper Thames River Basin (UTRB) in the Canadian province of Ontario. The intent is to create an ensemble of climate scenarios that can be used for the assessment of the vulnerability of a watershed to extreme events, including both floods and droughts, under future climatic conditions. The model is capable of generating unprecedented precipitation amounts, which is possible due to the perturbation of the observed data points. Consequently, extreme events more severe than observed events can be simulated for different climate change scenarios.

The remainder of the paper is organized in the following manner. The next section gives the description of the model and outlines the methodology used to adapt the $K$-NN algorithm for simulating daily weather sequences based upon the improved model. The subsequent section describes the physical characteristics of the UTRB. Application of the model to produce weather data conditioned upon the prescribed climate change criteria is described in the next section. Analysis of results obtained from simulation of extreme events under
different climate change scenarios is then presented. The subsequent section presents the frequency analysis of selected storm depths. The paper concludes with a summary of results and the identification of possible avenues for further research.

2. Model description

A K-NN algorithm typically involves selecting a specified number of days similar in characteristics to the day of interest. One of these days is randomly resampled to represent the weather of the next day in the simulation period. Despite their inherent simplicity, nearest neighbour algorithms are considered versatile and robust. These methods have been intensively investigated in the field of statistics and in pattern recognition procedures that aim at distinguishing between different patterns. The nearest neighbour approach involves simultaneous sampling of the weather variables, such as precipitation and temperature. The sampling is carried out from the observed data, with replacement. To simulate weather variables for a new day \( t + 1 \), days with similar characteristics as those simulated for day \( t \) are first selected from the historical record. One of these nearest neighbours is then selected according to a defined probability distribution or kernel and the observed values for the day subsequent to that nearest neighbour are adopted as the simulated values for day \( t + 1 \). Models based on the K-NN approach can be easily extended to multisite simulation of weather data while keeping the spatial correlation structure virtually intact. This feature of the K-NN model is not available in parametric models. 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 day \( t \). The cross-correlation among the variables at any given site is automatically preserved as a block of variables is resampled from the observed data.

Consider that the daily historic weather vector consists of \( p \) variables. Here \( p = 3 \) including maximum temperature (TMX), minimum temperature (TMN) and precipitation (PPT). Suppose the number of stations considered in the model is \( q \) and data are available for \( N \) years. Let \( X_t \) 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 \( X_t = [x_{1,t}, x_{2,t}, \ldots, x_{p,t}] \) where \( x_{i,t} \) represents the value of 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 year and the 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

   \[
   \bar{X}_t = [\bar{x}_{1,t}, \bar{x}_{2,t}, \ldots, \bar{x}_{p,t}]
   \]  

   where

   \[
   \bar{x}_{i,t} = \frac{1}{q} \sum_{j=1}^{q} x_{i,j,t},
   \]  

   \( i = 1, \ldots, p \), and \( t = 1, \ldots, T \)

2. Determine the size, \( L \), of data block that includes all potential neighbours 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 consist of all days between 13 January and 27 January for all \( N \) years but excluding January 20 for the given year. Thus, the data block of potential neighbours 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 neighbours using the expressions given in Step 1.

4. Compute the covariance matrix, \( C_t \), for the current day \( t \) using the data block of size \( L \times p \).

5. Determine the number of first \( K \) nearest neighbours to be retained for resampling out of
the total of $L$ neighbours. Lall and Sharma (1996) suggested choosing $K$ using the generalized cross validation score (GCV), which is similar to the Akaike information criteria (AIC) used in traditional autoregressive models. Rajagopalan and Lall (1999); Yates et al. (2003) recommended the use of a heuristic method for choosing $K$ according to which $K = \sqrt{L}$. The performance of the algorithm with this value of $K$ was found to be good. In this study $L = 569$ and hence a value of $K$ equal to 24 has been adopted.

(6) The weather on the first day $t$ (e.g. 1 January) 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 neighbours to represent the weather for day $t+1$ of the simulation period.

(7) Compute Mahalanobis distances between the mean vector of the current day’s weather, $\bar{X}_i$, and the mean vector $\bar{X}_i$ for day $i$, where $i = 1, \ldots, L$. The distance metric can be defined through

$$d_i = \sqrt{(\bar{X}_i - \bar{X}_j)C^{-1}(\bar{X}_i - \bar{X}_j)^T} \quad (3)$$

where $T$ represents the transpose operation and $C^{-1}$ is the inverse of the covariance matrix.

(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 neighbors. Weights are assigned to each of these $j$ neighbors according to the metric defined by

$$w_j = \frac{1/j}{\sum_{i=1}^{K} 1/i} \quad (4)$$

The cumulative probabilities, $p_j$, are given by

$$p_j = \sum_{i=1}^{j} w_i \quad (5)$$

The neighbor with the smallest distance is assigned the highest weight, while the neighbor with the largest distance (i.e. the $K$th 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, $r \in (0,1)$ and if $p_1 < r < p_K$, then the day $j$ for which $r$ is closest to $p_j$ is selected. If $r \leq p_1$, the day corresponding to $d_1$ is selected and if $r \geq p_K$, 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 fitted to the $K$ nearest neighbors identified in step 8. This involves estimating the conditional standard deviation, $\sigma$ and the bandwidth, $\lambda$ (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 $\sigma_i^t$ 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

$$y_{i,t+1} = x_{i,t+1} + \lambda \sigma_i^t z_{i,t+1} \quad (6)$$

where $x_{i,t+1}$ is the value of the weather variable for day $t+1$ and station $j$ obtained from the basic $K$-NN model, $y_{i,t+1}$ is the corresponding value obtained after perturbation and $\lambda$ is the bandwidth (a function of the number of samples) determined following Sharma et al. (1997).

(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 far 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, \( \alpha \), 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 \( \lambda \) corresponding to the probability of generating a negative value of exactly \( \lambda^a \), is therefore given by:

\[
\lambda^a = \left( \lambda_{3j+1}^3 / 1.55 \sigma^2 \right),
\]

where subscript 3 refers to precipitation values and \( \lambda^a \) is the acceptable (largest) value of \( \lambda \). If the calculated value of \( \lambda \) is larger than \( \lambda^a \), then \( \lambda^a \) is used instead of \( \lambda \).

(c) If the precipitation computed in step 10(b) is still negative, a new value of the random variate is generated and the value of precipitation recomputed from Eq. (6).

(d) Step 10(c) 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 precipitation 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.

3. The Upper Thames River Basin

The Upper Thames River Basin is located in the South-western region of the Canadian province of Ontario. The Thames River, which is 273 km long, is the major river of the basin. The basin has a catchment area of around 5825 km², making it the second largest basin in South-western Ontario. Since South-western Ontario is a highly developed region, the basin faces pressure from urban and rural land uses. Most of the forest cover in the watershed has been cleared for agricultural fields or urban development. Despite these pressures, the Thames River remains one of the most biologically diverse rivers in Canada.

Daily TMX, TMN and PPT data from 15 stations in and around the basin were used for the period from 1964 to 2001. The geographical location of stations is shown in Fig. 1. The data set used in this study has been corrected by Environment Canada using procedures discussed in Mekis and Hogg (1999); Vincent et al. (2002). The mean annual values of different 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 10–120 km.

4. Model application

The improved model was applied to simulate weather data conditioned upon plausible climate change scenarios for the UTRB. For each simulated sequence, box plots have been used to present the statistics of interest. Box plots are a favoured method of data analysis in many hydrological applications as they show the range of variation in statistics of simulations and provide a straightforward method of comparing the statistics of simulations with historical data. The bottom and top horizontal lines in the box in a box plot indicate the 25th and 75th percentile, respectively, of the statistics computed from the simulated data. The horizontal line within the box represents the median. The whiskers are lines extending from each end of the box to show the extent of the rest of the data. The whisker extends to the most extreme data value within 1.5 times the inter-quartile range of the data. The values beyond the ends of the whiskers are called outliers and are shown by dots. The statistics of the historical record are represented by dots and joined by solid lines. Results are presented for the London station only since the results for other stations are similar. Five different simulations were carried out, each involving simulation of 800 years of data. Simulation 1 was carried out to reproduce the statistics of the observed data. Simulation 2 represents an increasing temperature scenario. The increasing precipitation scenario is represented by simulation 3. Simulations 4 and 5 represent a wetter spring and a warmer spring scenario, respectively. Based on prescribed conditioning criteria, strategic resampling was used to obtain the driving data set for each climate change scenario. Strategic resampling implies that a
certain subset of years will be used as a base data set in the $K$-NN algorithm. Model runs were then carried out to simulate 800 years of data based on the driving data set for each scenario. The results of the simulations are described below.

5. Simulation 1: reproduction of historical data statistics

The first simulation was carried out to reproduce the statistical characteristics of the historical data. The purpose of simulation 1 was to analyse the performance of the improved model in reproducing various statistical attributes of the observed data, particularly the correlation structure, while perturbing the observed data points. A new subset of years that constitute the driving data for the model was obtained by using an integer function that returned integers between the specified upper and lower bounds. To obtain the driving data set for the model consisting of $N$ (here $N=38$) years, the integer function was queried $N$ times. With this method, each year has an equal probability of being selected but some years may be selected more than once, while other years may not be selected at all.

Fig. 2(a) shows box plots of simulated values of mean TMX. The model adequately reproduced the
historical values, which is highly satisfactory given that monthly statistics are not explicitly specified in fitting the \( K \)-NN model. Fig. 2(b) provides box plots of total monthly precipitation. It can be seen that the historical mean of the total precipitation is well preserved in the simulated data. The inter-annual variability in the simulated data is quite evident from the box plots presented in Fig. 2(b). Overall, the performance of the model in reproducing the monthly precipitation totals is satisfactory. It is important to note that the perturbations made using a normal random variate did not greatly alter the mean values. Fig. 2(c) shows the distribution of the total number of wet days for different months computed from the simulated data. Analysis of statistics of wet days is important as it gives an indication about the ability of the model to reproduce the persistence structure of the underlying data. The model reproduced the historical statistics very well, although there was a slight overestimation for the months of April, June and September.

With the basic \( K \)-NN model, the correlation structure of the underlying data is bound to be preserved. Since the improved model perturbs the observed data points, it was necessary to investigate the performance of the model with respect to reproduction of the correlation structure. Box plots of the correlation between TMX and PPT and lag-one autocorrelations of PPT are shown in Fig. 2(d) and (e), respectively. The model adequately reproduced the historical correlation structure, including the interstation correlations that are shown in Fig. 2(f). For \( q \) stations, there are \( q(q-1)/2 \) pairwise correlations resulting in 105 such correlation coefficients for each Julian day. The performance of the model concerning the reproduction of interstation correlations is satisfactory as can be seen from the scatter plots presented in Fig. 2(f). It is not clear why the model seems to slightly over-predict the correlation for the lower values of correlation. This remains an area for future investigation. Although parametric models such as LARS-WG (Semenov et al., 1998) and WGEN (Richardson and Wright, 1984) can be easily applied to generate weather data for any number of stations independently, they cannot be expected to preserve important interstation correlations of the variables. With the \( K \)-NN model, the spatial dependence is preserved by resampling simultaneously the same day’s weather as the weather for all stations. It is worth noting that the correlation structure is practically unaffected by the perturbations made to the data points.

6. Simulation 2: increasing average temperature scenario

To assess the vulnerability of the basin under an increasing average temperature scenario, a new data
Fig. 2. Box plots of simulation 1: (a) monthly mean TMX; (b) total monthly precipitation; (c) total number of wet days; (d) correlation between TMX and PPT; (e) lag-one autocorrelations of PPT; and (f) daily interstation correlations.
set comprising years with increased average temperatures is required. The amount of increase in temperature is obtained based on output of the Canadian GCM projections available on the University of Victoria’s climate change website (www.cics.uvic.ca) for a number of emission scenarios. For this study, the scenario considered was ‘greenhouse gases with aerosol’, which projects an approximately 2 °C increase in temperature by the year 2050. The driving data set comprising of years with increased temperatures was obtained using strategic resampling that involves generating a ranked list of years on the basis of the deviations of mean annual average temperature from the long-term historical mean. To compute the deviation for a given year, the overall long-term mean is subtracted from the mean yearly value for that particular year. In the ranked list of years, the first rank corresponds to the year with the lowest deviation (coldest year) and the last rank corresponds to the year with the highest deviation (warmest year). An index function (Yates et al., 2003) of the following form was used to cause biasing of certain years over others

\[ I = \text{INT}[N(1 - r^S)] + 1 \]  

where \( I \) refers to the index of the year in the ranked list, \( r \subset (0,1) \) is a random number, \( S \) is a shape factor, and \( N \) is the number of years in the historical data set. The desired amount of biasing can be obtained by suitably adjusting the shape parameter. No biasing is introduced when \( S = 1 \). To bias selection of warm years, shape parameter values greater than 1 must be used while shape parameter values less than 1 tend to bias the selection of cold years. Once the new data set with increased temperature is obtained, the model is executed to simulate 800 years of data. The results of simulation are presented through box plots shown in Fig. 3.

The box plots in Fig. 3(a) show that the model did produce weather data with the desired statistical attributes (i.e. increased TMX values for all months). The results suggest that with strategic resampling, the model is capable of producing alternative climate scenarios with desired statistical attributes. The effect of increasing average temperature on total monthly precipitation can be seen in Fig. 3(b). There appears to be a decrease in precipitation for most of the months except for January, March, and October when the median of the simulated total precipitation amounts did not deviate substantially from the mean historical values. The simulation results are in accordance with the observed climate relationships in the basin as reflected in the correlation structure between TMX and precipitation shown in Fig. 2(d). Although not
shown here, the correlation structure of the resampled data was adequately reproduced in the simulations largely due to the manner in which the $K$-NN algorithm works.

7. Simulation 3: increasing precipitation scenario

A resampling procedure similar to the one used for the increasing average temperature scenario is followed except that the deviations are computed for the precipitation instead of the temperature. A shape parameter value greater than 1 tends to bias the selection of wetter years. Therefore, a new data set comprising years with increased annual precipitation is obtained and used as the driving data set for the $K$-NN model. Fig. 4(a) shows the impact of an increase in precipitation on TMX while the simulated total monthly precipitation is shown in Fig. 4(b). It can be seen from Fig. 4(a) that the model produced some increase in TMX values for the months of February, March and April. As seen earlier in Fig. 2(d), there is a positive correlation between the historical TMX and PPT values during the winter months. Owing to this positive correlation, an increase in TMX values accompanies an increase in precipitation in some winter months. The maximum increase in PPT was obtained for August (nearly 30 mm) but a similar increase in TMX is not visible, which is due to the lack of correlation (see Fig. 2(d)) between TMX and PPT for August.

8. Simulation 4: wetter spring scenario

The purpose of simulation 4 was to simulate weather data with increased precipitation during the spring months. A resampling procedure similar to that for the increasing precipitation scenario was employed with the modification that the biasing was carried out for the spring months only. Fig. 5(a) and (b) includes box plots of monthly TMX and PPT, respectively. The effect of increased spring precipitation on monthly TMX values is clearly evident from Fig. 5(a). The median of the simulated TMX is higher than the historical mean value for February and May while it is slightly lower for July, August, September and December. These deviations of the simulated median values from the mean historical values are likely due to the interdependence of TMX and precipitation, which is also reflected in the observed correlation statistics. For the remaining months, changes in the observed mean TMX values are insignificant.

As expected, the median of simulated total precipitation was higher than the mean historical
values for February and March. Interestingly, no change in the January value was observed although the effect of increased spring precipitation appears to have been propagated to other months where deviations of simulated data from the historical could be seen. For example, August had a higher median value while April, June, September, and December exhibit a lower median value compared to the historical mean values. However, as can be observed from the box plots shown in Fig. 5(b), these deviations are insignificant. The inter-annual variability, which is a desirable feature, is quite evident in the simulated data. Overall, the performance of the model in simulating a wetter spring scenario can be considered acceptable.

9. Simulation 5: warmer spring scenario

The driving data set for the model was obtained through strategic resampling of the observed data in a manner similar to the wetter spring scenario. The objective was to derive a data set that has increased temperatures for the spring months. The desired increase in temperature was obtained by suitably adjusting the shape parameter for the spring season. Fig. 6(a) shows the box plots of mean TMX. The median of the simulated TMX values for the months of January, February and March are higher than the historical mean values, with the highest increase achieved for February thereby suggesting that the model was able to adequately simulate a warmer spring scenario for the basin. Increased temperatures were also observed for April, August, September and November. This may be attributed to the presence of certain years in the resampled data in which warmer springs are accompanied by warmer falls. For the remaining months, the simulated values are close to the median values.

The effect of warmer springs on the monthly precipitation is shown in Fig. 6(b). Historically, there is a small positive correlation between TMX and precipitation for January and February but there is no correlation for March (Fig. 2(d)). As a result, the effect of increasing TMX on precipitation is not very well defined. There is some decrease in the simulated precipitation for January and February while the March values remain unaffected. For September, the simulated total precipitation was found to be somewhat lower than the historical mean value. For the remaining months, the simulated precipitation totals were very close to the mean values. As in other simulations, the inter-annual variability in the simulated data is quite evident.
10. Extreme precipitation events simulation

A major objective of this research is to assess the vulnerability of watersheds to floods and droughts under changing climatic conditions. The capability 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 unprecedented events for the basin. Five potential climate change scenarios in the basin were simulated as described earlier. Fig. 7 shows the box plots of total precipitation that occurred during the most extreme precipitation event in each year of the simulated record for different simulations. An extreme precipitation event is defined as consisting of a continuous sequence of wet days. For each year of the data, a single multi-day extreme event was determined along with the total precipitation that occurred during the event. The first box plot in Fig. 7 shows the total precipitation during the most extreme event in each year of the observed data while the remaining box plots show the same statistics for five different simulations.

The box plot for simulation 1 clearly shows that the median of the simulated data matches closely with the median of the observed data. Greater variability in the simulation of extreme events is achieved owing to the improved nature of the $K$-NN model. A total precipitation of the order of 270 mm in the most extreme precipitation event was observed in the simulated data compared to a corresponding value of 200 mm in the observed record. Simulation 3, which represents an increasing precipitation scenario, is the critical scenario for the analysis of floods as the total precipitation amount of the largest event is the highest among all the scenarios. A total precipitation amount of around 280 mm was simulated, which is substantially higher than the highest value simulated in any scenario and about 1.4 times the largest value in the observed data. Simulation 4 was carried out for the wetter spring scenario with the driving data that comprised years with increased precipitation amounts for the spring months only. Therefore, the median of the simulated data is smaller than that obtained for the increasing precipitation scenario where the driving data set comprised years that had higher precipitation amounts year round and not just for the spring months. Simulation 5 represents a warmer spring scenario and as can be seen for the box plot for simulation 5, the distribution of the simulated data is quite similar to that for simulation 2, which represents the increasing temperature scenario. The inter annual variability is quite high as is the case with all other simulations thus providing a variety of extreme events that can be used as input to hydrological models.

It is important to determine dry spell characteristics of the simulated data in order to assess the risks associated with drought in the basin under future
climatic conditions. Fig. 8 presents box plots of the total number of dry days during the extreme dry spell in each year of the observed as well as simulated record. The first box plot in Fig. 8 presents the distribution of extreme dry spells computed from 38 years of observed data. Simulation 1 provides the basis for comparing the results of all other simulations as it reflects the effects of strategic resampling more accurately. As can be seen from the box plot of simulation 1, the median dry spell duration for the observed data is 25 and the corresponding statistic in the simulated data is 30. The higher median value
produced in simulation 1 may be attributed to the nature of the improved model, which tends to produce more severe events than observed in the historical data. For instance, five extreme events in simulation 1 are more severe than the observed events. For simulation 2, the median of the simulated data was slightly higher than the corresponding value for simulation 1. The model simulated several events that are more severe than produced in simulation 1 thus providing a wider range of events as input to a hydrologic model. It can be seen from the box plot that the longest dry spell in simulation 2 lasted around 92 days, which is quite high compared to the observed value of 67 days. These results clearly indicate that the increasing temperature scenario represented by simulation 2 is the critical scenario for the analysis of droughts in the basin. As expected, simulation 3 and 4 produced results that are not as extreme as in simulation 2. For simulation 5, several extreme events similar to those in simulation 2 were produced. As described earlier, the similarity in the distributions observed in simulation 2 and simulation 5 is primarily due to the similarity in the driving data set used for these simulations. The strength of the improved model presented here lies in the simulation of extreme dry spells for potential climate change scenarios that are important for evaluation of effective drought management policies for the basin.

11. Frequency analysis

A key focus of this study was to evaluate the impact of climate change on the occurrence of floods and droughts in the basin. Since the occurrence of these extreme events is intrinsically linked to extreme storm depths, it is important to determine the probabilities of exceedence of different storms depths. Results from the simulations clearly indicated that the improved model produced storm depths that are more extreme than those observed in the historical record. For simulation 2, the largest simulated storm depth in the basin was 196 mm while the observed historical value is only 165 mm. For each simulation, the model was able to produce several storm depths that are larger than the observed maximum. A total of 10 storm depths were selected and a return period was determined for each of these for the various scenarios considered. There are two basic approaches to determining the return periods of extreme values. The first is the graphical approach based upon the Weibull distribution that involves plotting various storm depths against the recurrence interval or exceedence probability and fitting a curve to the resulting points. The second approach involves mathematically fitting the data to a theoretical probability distribution. The Gumbel distribution is particularly convenient for extreme value distribution purposes and has been commonly used for the estimation of precipitation quantiles. Therefore, the Gumbel probability distribution is assumed as the underlying probability distribution for the data available here. Moreover, the emphasis is on comparing the return period associated with a given storm depth for different scenarios, and therefore the use of the Gumbel distribution should suffice.

For computing return periods associated with different storm depths, a single largest daily precipitation amount was extracted for each year of the simulated data. Thus, the data set for frequency analysis consisted of 800 values of precipitation amounts from which the parameters of the Gumbel distribution were estimated. Table 2 shows return periods associated with ten selected storm depths. Results of frequency analysis clearly indicate that simulation 3, representing the increasing precipitation scenario, is the critical scenario. For example, consider a storm depth of 165 mm. The return period is the highest for simulation 2 and the lowest for simulation 3. Similar trends were observed for other storm depths. A promising potential application of the frequency analysis carried out here is in rainfall-runoff modelling where the storm depths could be related to the frequency of flood and drought events in the basin. The approach would involve identification of critical flood and drought events through the rainfall-runoff model of the basin under each potential climate change scenario. Alternatively, an already identified critical scenario may be used in the analysis. The output of the $K$-NN perturbation model can then be analyzed to determine the probability of occurrence of the corresponding weather data that is likely to cause these events. Because flood events depend upon the precipitation and catchment wetness, some of the weather data identified as being critical by the weather model may not actually be critical. Therefore,
it might be necessary to run the critical weather events through the rainfall-runoff model to verify the severity of such events. A similar procedure would be required for the low flow events as well.

12. Summary and conclusions

The potential impact of climatic change on the occurrence of extreme precipitation events in the UTRB has been investigated. Weather sequences based upon several plausible climate change scenarios for the basin have been simulated using an improved K-NN model that allows nearest neighbour resampling with perturbation of the historic data. The improved model has been shown to be effective in producing precipitation amounts larger than those observed in the historical record thereby alleviating a common problem associated with the standard K-NN approach. Perturbation of the observed data points was carried out by adding a random component to the values suggested by the basic K-NN model. The proposed approach is philosophically similar in spirit to traditional autoregressive models except that the new values are obtained by adding a random component to the individual re-sampled data points. A remarkable advantage of the model described here is that unprecedented precipitation amounts can be simulated for any given climate change scenario. Consequently, extreme events, both high precipitation and low precipitation, that are more severe than the observed ones can be easily simulated, which is important for assessing the vulnerability of the basin to floods and droughts under changing climatic conditions. At present, the model considers changes in temperature and precipitation separately. Future work will explore approaches for creating scenarios with specified changes in both temperature and precipitation simultaneously.

Five potential climate change scenarios were simulated. The intent was to create an ensemble of scenarios that can be used for the evaluation of the response of a rainfall-runoff model for a variety of simulated data, especially the extremes. Frequency analysis carried out to determine the impact of potential climate change on the occurrence of storm depths of any given magnitude revealed that for the increasing temperature scenario, the return periods of storm depths were found to be the highest while the lowest return periods were seen for the increasing precipitation scenario. Of the five scenarios considered, the results obtained indicate that the increasing precipitation scenario is the critical scenario for the assessment of risks associated with the occurrence of floods in the basin. For the analysis of drought situation in the basin, the increasing temperature scenario is the critical scenario. Once the critical scenarios have been identified, the proposed model, in conjunction with a rainfall-runoff model, can be used to determine the frequency of floods and droughts in the basin. A distinct advantage of the model is that it produces spatially correlated data for the basin, which is crucial for evaluating the response of hydrological models to watershed-level processes.

<table>
<thead>
<tr>
<th>Storm depth (mm)</th>
<th>Return period (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Simulation 1</td>
</tr>
<tr>
<td>59</td>
<td>1.14</td>
</tr>
<tr>
<td>95</td>
<td>3.1</td>
</tr>
<tr>
<td>105</td>
<td>4.64</td>
</tr>
<tr>
<td>130</td>
<td>13.93</td>
</tr>
<tr>
<td>143</td>
<td>25.29</td>
</tr>
<tr>
<td>165</td>
<td>70.53</td>
</tr>
<tr>
<td>170</td>
<td>89.18</td>
</tr>
<tr>
<td>190</td>
<td>228.48</td>
</tr>
<tr>
<td>200</td>
<td>366.03</td>
</tr>
<tr>
<td>220</td>
<td>940.16</td>
</tr>
</tbody>
</table>
On the other hand, WGEN and LARS-WG cannot be expected to preserve the spatial correlation structure of the data.

An interesting application of the model could be in inverse modelling (Cunderlik and Simonovic, 2004), which could involve transforming the critical hydrological exposures (e.g. flood levels) into corresponding critical meteorological conditions, such as extreme precipitation events. The local meteorological conditions could then possibly be linked to largescale climatic outputs available from GCMs. Another application of the proposed model could be in the simulation of scenarios by considering changes in one or more variables as predicted by the GCMs. This would involve modifying the historical data set by applying change fields to one or more variables simultaneously and then using the modified data set as the driving data set for the model. This will be pursued in future work.

Acknowledgements

The research presented in this paper was funded by the Canadian Foundation for Climatic and Atmospheric Sciences. The authors gratefully acknowledge the cooperation received from Prof. S.P. Simonovic and Dr. J.M. Cunderlik, University of Western Ontario. The authors also thank Mr. Mark Helsten for providing meteorological data for the Upper Thames River Basin.

References


