HYDROLOGICAL PROCESSES
*Hydrol. Process.* **28**, 905–912 (2014)
Published online 17 December 2012 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/hyp.9596

# Simulation of historical temperatures using a multi-site, multivariate block resampling algorithm with perturbation

Leanna M. King,[1]* A. Ian McLeod[2] and Slobodan P. Simonovic[1]

[1] *Department of Civil and Environmental Engineering, University of Western Ontario, London, Ontario, Canada, N6A 3K7*
[2] *Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, Canada, N6A 3K7*

## Abstract:

Stochastic weather generators have evolved as tools for creating long time series of synthetic meteorological data at a site for risk assessments in hydrologic and agricultural applications. Recently, their use has been extended as downscaling tools for climate change impact assessments. Non-parametric weather generators, which typically use a K-nearest neighbour (K-NN) resampling approach, require no statistical assumptions about probability distributions of variables and can be easily applied for multi-site use. Two characteristics of traditional K-NN models result from resampling daily values: (1) temporal correlation structure of daily temperatures may be lost, and (2) no values less than or exceeding historical observations can be simulated. Temporal correlation in simulated temperature data is important for hydrologic applications. Temperature is a major driver of many processes within the hydrologic cycle (for example, evaporation, snow melt, etc.) that may affect flood levels. As such, a new methodology for simulation of climate data using the K-NN approach is presented (named KnnCAD Version 4). A block resampling scheme is introduced along with perturbation of the reshuffled daily temperature data to create 675 years of synthetic historical daily temperatures for the Upper Thames River basin in Ontario, Canada. The updated KnnCAD model is shown to adequately reproduce observed monthly temperature characteristics as well as temporal and spatial correlations while simulating reasonable values which can exceed the range of observations. 

KEY WORDS weather generator; stochastic; climatology

*Received 25 April 2012; Accepted 8 October 2012*

## INTRODUCTION

Stochastic weather generators are computational algorithms that can be used to produce synthetic meteorological data of any length with the same statistical characteristics as the historical record for a site (Wilks and Wilby, 1999). The advantage of the stochastic approach is that the simulated datasets are long enough to be used in risk assessments of extreme precipitation and temperature events (Semenov and Barrow, 2002). Motivation for the development of such algorithms comes mainly from the water sector— long time series of climate data can be used in computer models to (1) predict the response of crops to extreme heat, (2) assess water availability for the supply of people and industry, (3) understand the change in frequency and magnitude of extreme flood and/or drought events and (4) quantify the risk to water infrastructure due to changing conditions (Wilks and Wilby, 1999; Semenov and Barrow, 2002; Soltani and Hoogenboom, 2003; Dibike and Coulibaly, 2005; Sharif and Burn, 2007; Eum *et al.*, 2010). In recent years, the use of weather generators has been extended for downscaling global climate models to assess the local impacts of climate change (Dubrovsky, 1997; Semenov and Barrow, 1997; Soltani and Hoogenboom, 2003; Kuchar, 2004; Dibike and Coulibaly, 2005; Eum and Simonovic, 2011; Eum *et al.*, 2011). Much of the current research involves the simulation of future climate data for impact assessments dealing with the management of water resources systems.

There are three major categories of weather generators in the literature: parametric, semi-parametric and non-parametric. Some examples of parametric models that have been employed to simulate temperature for single-site applications are WGEN (Soltani and Hoogenboom, 2003; Kuchar, 2004; Craigmile and Guttorp, 2011), WGENK (Kuchar, 2004), AAFC-WG (Qian *et al.*, 2004), SIMMETEO (Geng *et al.*, 1988; Soltani and Hoogenboom, 2003; Elshamy *et al.*, 2006) and GEM (Hanson and Johnson, 1998). Parametric models require the user to make stronger assumptions than the non-parametric approach and require careful statistical diagnostic checks in order to ensure that the statistical characteristics of the historical time series are adequately captured in the resulting synthetic series.

There are a variety of different semi-parametric weather generators that have been used to simulate temperature and precipitation (Semenov and Barrow, 2002; Apipattanavis *et al.*, 2007). Two widely used algorithms are SDSM and LARS-WG (Semenov and Barrow, 2002; Wilby and Dawson, 2007). SDSM is a single-site, regression-based model with a stochastic component where large-scale atmospheric variables are used to linearly condition local temperature or precipitation data (Wilby and Dawson, 2007). A drawback of SDSM is that each variable is simulated independently so the relationships between them are not preserved. LARS-WG uses a semi-empirical wet and dry spell distribution to simulate precipitation occurrence where chosen the amounts are chosen conditional on

*Correspondence to: Leanna M. King, Department of Civil and Environmental Engineering, University of Western Ontario, London, Ontario, Canada, N6A 3K7.
E-mail: leannamichelleking@gmail.com

spell length (Semenov and Barrow, 2002). Temperatures are simulated conditional on the day's wet or dry status (Semenov and Barrow, 2002); however, some studies have found LARS-WG underestimates the occurrence of extreme temperature events (Mavromatis and Hansen, 2001; Qian *et al*., 2004; Semenov, 2008).

A drawback associated with most of the popular parametric and semi-parametric models is the underlying statistical assumption about the probability distribution of the weather variables, which can be highly site dependent. Another issue is that spatial correlations must be assumed for multi-site applications. Non-parametric weather generators have evolved as a way around these limitations and have been employed successfully for multi-site simulation in various regions (Young, 1994; Yates *et al*., 2003; Mehrotra *et al*., 2006; Sharif and Burn, 2006; Eum *et al*., 2010). Many of these algorithms are an extension of the Young (1994) approach where a nearest-neighbour resampling scheme is used to select the next days' weather from a subset of days with similar characteristics to the current day. Sampling is done from the historical record, with replacement (Sharif and Burn, 2006).

A more recent version of the Young (1994) approach was developed by Yates *et al*. (2003) and later Sharif and Burn (2006) in which the Mahalanobis distance metric is used to retain the closest K neighbours from a subset of days within a temporal window centred on the current day. A probability distribution is then used to choose one of the K-nearest neighbours (K-NN) as the weather for the next day (Sharif and Burn, 2006). Multi-site application is done by simply choosing the corresponding weather of all stations so the spatial correlation are preserved. The K-NN model of Sharif and Burn (2006), hereinafter referred to as KnnCAD Version 1 (V1), was extended further by Prodanovic and Simonovic (2008) to include a leap year modification (KnnCAD V2). Eum and Simonovic (2008) further modified the algorithm to allow for the inclusion of more climate variables, without increasing computational demand, by using only the first principal component in the calculation of Mahalanobis distance (KnnCAD V3).

Our work shows that a limitation of these traditional K-NN models results from the resampling of only one day at a time: the temporal correlation structure of daily temperatures is lost. Daily temperature simulations are highly important for watersheds where winter snow accumulation and spring melt events are the cause of major floods. Accurate modeling of the stream flows requires a time series of temporally correlated temperature data as well as precipitation inputs. Furthermore, evaporation is one of the major drivers of the hydrologic cycle and is directly proportional to the temperature. For downscaling of global climate model data, accurate simulations of temperature data become increasingly important as the global climate models are able to better represent daily temperatures than precipitation amounts.

Another major drawback to the traditional K-NN models is the inability of these algorithms to simulate values outside of the observed record (Young, 1994; Yates *et al*., 2003; Sharif and Burn, 2006; Apipattanavis *et al*., 2007; Furrer and Katz, 2008). Because the simulated data is merely resampled, the three highest temperature values in the observed record will be the same as those in the simulated data; no values lying in between these individual observations can be generated nor can values which exceed the observed maximum and minimum (Furrer and Katz, 2008). Sharif and Burn (2006) modified the K-NN approach of Yates *et al*. (2003) in KnnCAD V1 to include perturbation by adding a noise term to smooth the simulated precipitation values. The noise term is calculated using a non-parametric density estimator or kernel with a Gaussian distribution and a bandwidth according to the Silverman (1998) rule of thumb. To avoid negative values of precipitation, a maximum acceptable bandwidth is used (see Sharif and Burn, 2006).

Eum and Simonovic (2011) used a similar approach in KnnCAD V3 to perturb K-NN resampled temperature data, employing maximum bandwidths corresponding to various significance levels to prevent unrealistic values of temperature from being simulated. For significance levels ranging from 6% to 0.5%, it was found that while the maximum bandwidth did improve the range of temperature values simulated, unrealistic values were still generated by the model (Eum and Simonovic, 2011). Furthermore the determination of a 'reasonable' value of temperature is somewhat ad hoc and would change depending on the region in which the weather generator is employed.

The focus of this study is to develop a new methodology for simulating synthetic climate data using the non-parametric K-NN approach of the KnnCAD algorithm. Long time series of temperature data are highly important for water resource management applications, and preserving the temporal correlation structure of the simulated data is crucial, particularly in areas where snow accumulation and melt can cause major flooding events. The existing models do not have the ability to produce unprecedented values for temperatures, so a smoothing of the simulated data is necessary to produce unique values both within and exceeding the historical range of observations. The next section of this report provides details of KnnCAD V3 (Eum and Simonovic, 2008) and the proposed modifications to the algorithm. Following, descriptions of the study area and datasets used for this study are given. Next the results are presented followed by concluding remarks.

## METHODOLOGY

### *KnnCAD version 3 algorithm*

The weather generator of Eum and Simonovic (2008), KnnCAD V3 is a K-NN algorithm with principal component analysis in order to include multiple variables in the selection of the nearest neighbours. It was developed by Eum and Simonovic (2008) as an extension of the models of Sharif and Burn (2006), and Prodanovic and Simonovic (2008). The KnnCAD V3 algorithm has the following steps:

(1) Compute the regional means of $p$ variables ($x$) across all $q$ stations for each day in the historic record:

 

$$\bar{X}_t = \left[\bar{x}_{1,t}, \bar{x}_{2,t} \ldots, \bar{x}_{p,t}\right] \quad \forall t = \{1, 2, \ldots, T\} \tag{3}$$

where $$\bar{x}_{i,t} = \frac{1}{q}\sum_{j=1}^{q} x_{i,t}^{j} \quad \forall i = \{1, 2, \ldots, p\} \tag{4}$$

(2) Choose a temporal window of length w, and select a subset of potential neighbours $L$ days long for each day in $N$ years of record for all $p$ variables, where $L = N*(w+1) - 1$. Yates *et al.* (2003) used a temporal window of 14 days in the great lakes region, so if January 20th is the current day, the potential neighbours are all days that fall between January 13th and January 27th for all $N$ years, excluding the value of the current day.

(3) Compute the regional means $\bar{X}_l$, of the L potential neighbours for each day across all $q$ stations.

(4) Compute the covariance matrix, $C_t$ for day $t$ using the potential neighbours from (3) with a data block of size $L$ by $p$.

(5) The weather on the first time step (e.g. January 1) consisting of $p$ variables at $q$ stations is randomly chosen from the $N$ current day values.

(6)

(6a) Calculate the eigenvector and eigenvalue from the covariance matrix $C_t$.

(6b) Retain the eigenvector **E** which corresponds to the highest eigenvalue which explains the largest fraction of variance in the p variables.

(6c) Calculate the first principal component using **E** from Equation (6b):

$$PC_t = \bar{X}_t E \tag{5}$$

$$PC_l = \bar{X}_l E, \quad \forall l = \{1, 2, \ldots, L\} \tag{6}$$

Where $PC_t$ and $PC_k$ are one-dimensional values for the current day, $t$ and the kth neighbour transferred from the eigenvector in Equation (6b).

(6d) Calculation of the Mahalanobis distance using the values obtained in Equations (5) and (6) as well as the variance, Var(**PC**), between all L values of $PC_k$.

$$d_k = \sqrt{\frac{(PC_t - PC_k)^2}{Var(PC)}} \quad \forall k = \{1, 2, \ldots, L\} \tag{7}$$

(7) Select the number $K$ of nearest neighbours to retain out of the $L$ potential values. Rajagopalan and Lall (1999) and Yates *et al.* (2003) recommend taking $K = \sqrt{L}$.

(8) Sort the Mahalanobis distance metric from smallest to largest, and retain the closest $K$ neighbours on the list. Use a discrete probability distribution weighting closest neighbours highest for resampling one of the $K$ values, following Equations (8) and (9).

$$w_k = \frac{1/k}{\sum_{i=1}^{k} 1/i} \quad \forall k = \{1, 2, \ldots, K\} \tag{8}$$

$$p_j = \sum_{i=1}^{j} w_i \tag{9}$$

(9) Generate a random number, $u(0,1)$ and compare this to the cumulative probability, $p_j$, to determine the current day's neighbour. The day $j$ for which $u$ is closes to $p_j$ is selected as the neighbour, and the corresponding weather is used for all stations in the region. Through this step, spatial correlation among the variables is preserved.

(10) Steps 6 through 9 are repeated for each day in the observed record to produce a synthetic output file of the same length. Multiple simulations can be run to produce long datasets of synthetic climate data for a site.

*Proposed modifications to KnnCAD model*

In order to preserve the temporal characteristics of temperature data, a block resampling scheme is proposed which resamples a specified number of days at a time. The model, hereinafter referred to as KnnCAD V4, follows the same steps as the KnnCAD V3 above, but resamples a block of days of length $B$ which follow the selected day in the historical record. The choice of $B$ depends on the ability of the model to reproduce temperature autocorrelations. It can change depending on how correlated the day-to-day temperature values are in the historical record (McLeod and King, 2012).

In order to ensure sequences of days are not repeated in the record and to produce unprecedented values for temperature, a perturbation scheme is also introduced. While some successful attempts have been made for the perturbation of precipitation data (Sharif and Burn, 2006; Eum *et al.*, 2011), results have been less satisfactory for the perturbation of temperatures (Eum and Simonovic, 2011). The proposed modifications for KnnCAD V4 are outlined below. Steps (1) through (9) from KnnCAD V3 remain the same.

(10a) Resample $B$ days from the historical record which follow the selected day ($j$) from step (9). For example, if the selected neighbour to the current day is January 21st, 1979 from the historical record and $B = 10$ days, the days January 21st, 1979 to January 30th, 1979 are resampled from the historical record. The days in $B$ may extend outside of the temporal window described in Step (2).

(10b) Perturbation of the reshuffled temperature values $\left(x_{i,t+b}^{j}\right)$ for temperature variable $i$, station $j$ and day $b$ (where $b = 1,2,\ldots,B$), following Equation (10):

$$y_{i,t+b}^{j} = \lambda x_{i,t+b}^{j} + (1 - \lambda)Z \tag{10}$$

(10c) Where $y_{i,t+b}^{j}$ is the simulated perturbed value for day $b$ of the block, and $\lambda$ is chosen between 0 and 1 (1 gives an unperturbed result and 0 yields a result based entirely on perturbation). $Z$ is a normally distributed value with a mean of $x_{i,t+b}^{j}$ and a standard deviation of $\sigma_{i,t}^{j}$, where $\sigma_{i,t}^{j}$ is the standard deviation of the K-NN for day $t$, station $j$ and temperature variable $i$. To prevent minimum temperature (TMIN) from exceeding maximum temperature (TMAX), the same random normal variable is used for both TMAX and TMIN across all stations, along with the variables' corresponding $x^j_{i,t+b}$ and $\sigma^j_{i,t}$ values. In order to adequately preserve the inter-site spatial correlations and the variable autocorrelations, $\lambda$ should be chosen as large as possible but less than 1 (which would result in no perturbation). A value of $\lambda = 0.9$ will produce a result that is 10% based on perturbation and contains 90% of the original resampled value.

(11) Repeat steps 6 through 10 until the end of the historical record is reached. Multiple simulations can be done to produce long synthetic datasets.

### *Evaluation of weather generator performance*

In order to evaluate the performance of KnnCAD V4 with the proposed block resampling and perturbation scheme, the ability of the model to simulate historical climate characteristics is investigated. This is done using 25 independent simulations, each with the same length as the historical record. The ability of the updated model to reproduce temporal correlations of temperatures is investigated using monthly boxplots, and results are compared to KnnCAD V3 to demonstrate the improvement made by the block resampling procedure. KnnCAD V4-simulated monthly mean temperatures are compared to historical means through the use of boxplots. Line plots of simulated and observed monthly standard deviations, 99$^{th}$ percentile values, 1$^{st}$ percentile values and absolute maxima and minima are plotted. The ability of the updated KnnCAD V4 to preserve inter-site spatial correlation with the new perturbation component is also evaluated as this is a key advantage of the non-parametric models.

## APPLICATION

### *The Upper Thames River basin*

The Upper Thames River basin (UTRB), shown in Figure 1 (Census of Canada, 2006a,2006b) is located in southwestern Ontario, between the great lakes of Erie and Huron. The River runs in two main channels, the North Thames and the South Thames, which meet in London, Ontario and flow as a single channel through the Lower Thames into Lake St. Clair (Wilcox *et al*., 1998). The basin drains an area of 3482 km$^2$, which is mainly agricultural land with some heavily urbanized regions and a few remaining forested areas (Wilcox *et al*., 1998).

The UTRB has a history of major floods, which typically occur in the early spring from the combination of rainfall and snowmelt, or in summer after major rainfall events. Because snow accumulation and melt play an important role in the prediction of runoff, long time series of both temperature and precipitation data are required for flood risk assessment purposes. Simulated temperature data must adequately preserve the observed temporal correlation in order to provide accurate hydrologic models that account for snow accumulation and melt events. Many studies have indicated that the UTRB is vulnerable to climate change in terms of extreme precipitation and flooding (Sharif and Burn, 2006; Prodanovic and Simonovic, 2007; Simonovic, 2010; Solaiman *et al*., 2010; Eum *et al*., 2011). As such, climate impact assessments for water resources management in the UTRB will play an important role for future development of this region.

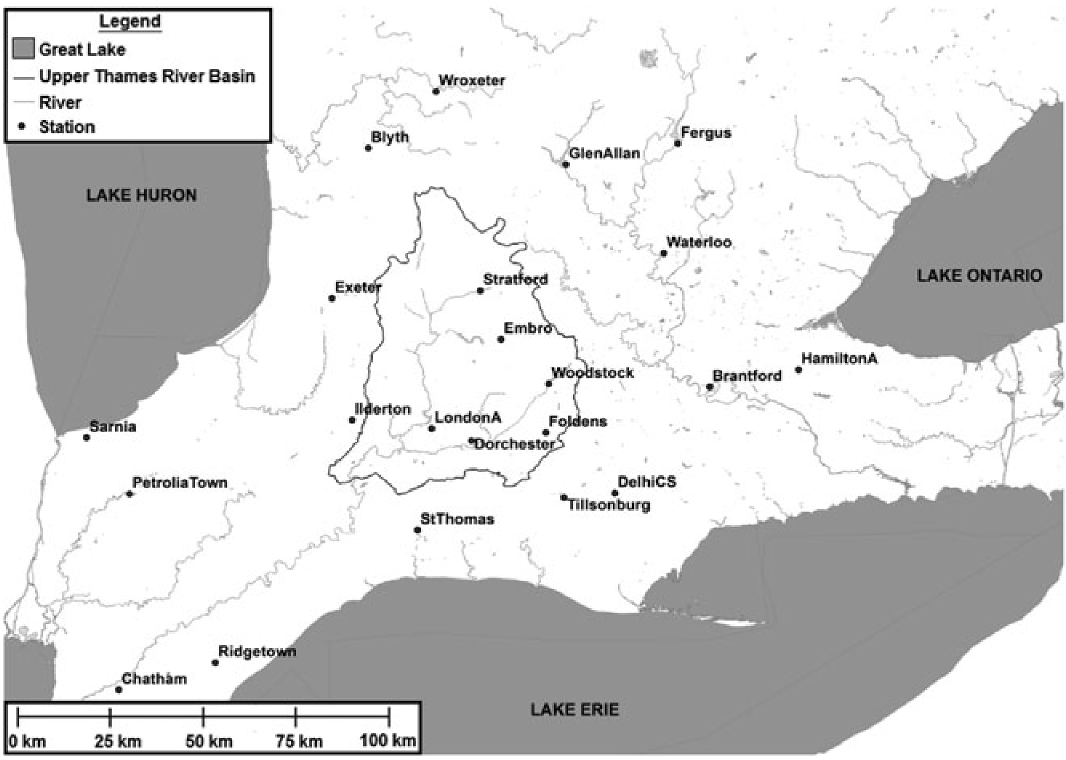


Figure 1. The Upper Thames River Basin

 

*Data*

Historical climate data for Canada is available from Environment Canada's Canadian Daily Climate Data archives. Table I provides a list of the stations used in this study along with the latitudes, longitudes and elevations of each location. Figure 1 shows the locations of the stations in the watershed. Stations are chosen based on the length and completeness of the historical record. Based on the availability of data, 27 years of records from 1979–2005 are used in this study. This dataset provides three time series each of length 9862 corresponding to the daily values of precipitation, TMAX and TMIN over the period 1979–2005. The variables used are precipitation, TMAX and TMIN. Our focus is on the two temperature variables.

## RESULTS

The updated KnnCAD algorithm (KnnCAD V4) with block resampling and perturbation is used to produce 25 independent simulations of 27-year length based on the observed record, for a total of 675 years of synthetic climate data. The ability of KnnCAD V4 to adequately reproduce historical temperature characteristics was investigated through sensitivity analysis using several values of $\lambda$ for perturbation (see Equation (10)). A value of $\lambda = 0.9$ which yields a result that is 10% based on perturbation is found to adequately preserve inter-site correlations and historical characteristics while still producing temperature values outside of the observed range. As such, $\lambda = 0.9$ is employed in KnnCAD V4 for the remainder of this study.

The choice of block length B was investigated by examining the effect of the block bootstrap for autocorrelations likely to be found in daily temperature series (McLeod and King, 2012). Various choices of $B$ were tested, and for temporal correlations in the temperature series of up to 0.7, it was found that $B = 10$ days is an adequate block length (McLeod and King, 2012). This was confirmed in the application of KnnCAD V4. In the selection of $B$, there is a tradeoff between choosing a value large enough to preserve temporal correlation but small enough to ensure diversity between the values in subsequent simulations. While perturbation does reduce this effect to an extent, $B$ should be a value large enough to preserve temporal correlation but as small as possible. The lag-1 autocorrelation within each block is inherently preserved, but at the interface between subsequent blocks, autocorrelation is not preserved. With a sufficiently large block length such as $B = 10$ days, used in this study, the effect of this is very small (one instance every 10 days) and thus is undetectable, as seen in the results. The use of block resampling also inherently helps to preserve wet and dry spell lengths.

Table I. Location of stations in the UTRB

| Station | Latitude (deg N) | Longitude (deg W) | Elevation (m) |
|---|---|---|---|
| Blyth | 43.72 | 81.38 | 350.5 |
| Brantford | 43.13 | 80.23 | 196.0 |
| Chatham | 42.38 | 82.2 | 198.0 |
| Delhi CS | 42.87 | 80.55 | 255.1 |
| Dorchester | 43.00 | 81.03 | 271.3 |
| Embro | 43.25 | 80.93 | 358.1 |
| Exeter | 43.35 | 81.50 | 262.1 |
| Fergus | 43.73 | 80.33 | 410.0 |
| Foldens | 43.02 | 80.78 | 328.0 |
| Glen Allan | 43.68 | 80.71 | 404.0 |
| Hamilton A | 43.17 | 79.93 | 238.0 |
| Ilderton | 43.05 | 81.43 | 266.7 |
| London A | 43.03 | 81.16 | 278.0 |
| Petrolia Town | 42.86 | 82.17 | 201.2 |
| Ridgetown | 42.45 | 81.88 | 210.3 |
| Sarnia | 43.00 | 82.32 | 191.0 |
| Stratford | 43.37 | 81.00 | 354.0 |
| St. Thomas | 42.78 | 81.21 | 209.0 |
| Tillsonburg | 42.86 | 80.72 | 270.0 |
| Waterloo Wellington | 43.46 | 80.38 | 317.0 |
| Woodstock | 43.14 | 80.77 | 282.0 |
| Wroxeter | 43.86 | 81.15 | 355.0 |

Figure 2 shows the simulated and observed values for monthly lag-1 autocorrelations from the London A station. Results from the other stations and other lags are similar. The boxplots represent simulated correlations, and the solid line shows the observed median. The left column contains KnnCAD V4 results for TMAX and TMIN, and the right column contains results from KnnCAD V3 of Eum and Simonovic (2010). It is clear from the figure that the updated algorithm provides a major advantage over the one-day resampling approach which greatly underestimates lag-1 autocorrelations. While there is still a slight underestimation in the medians for some months, the medians of the observed data all lie within the interquartile ranges of the simulated data from KnnCAD V4. Because 10 days are resampled at a time, the proposed algorithm inherently preserves the temporal characteristics of daily temperatures.

Figure 3 shows the simulated and observed monthly temperature characteristics for the London A station from KnnCAD V4. The previous version KnnCAD V3 of Eum and Simonovic (2010) produced similar results, as did the outputs from different stations. The first column contains boxplots of simulated TMAX (top) and TMIN (bottom) with the historical medians shown as a line plot. From the figure, it is clear that KnnCAD V4 is able to adequately simulate monthly average temperatures with very little deviation from the observed values.

The second column in Figure 3 shows the monthly standard deviations in TMAX (top) and TMIN (bottom). While there are some very slight over and underestimations in certain months, KnnCAD V4 performed very well for standard deviations. The third column of Figure 2 shows the 99$^{th}$ percentile (top) and 1$^{st}$ percentile (bottom) of simulated monthly TMAX and TMIN values. There is a very close agreement between observed and simulated values.

The last column in the figure shows the absolute maximum (top) and minimum (bottom) monthly TMAX

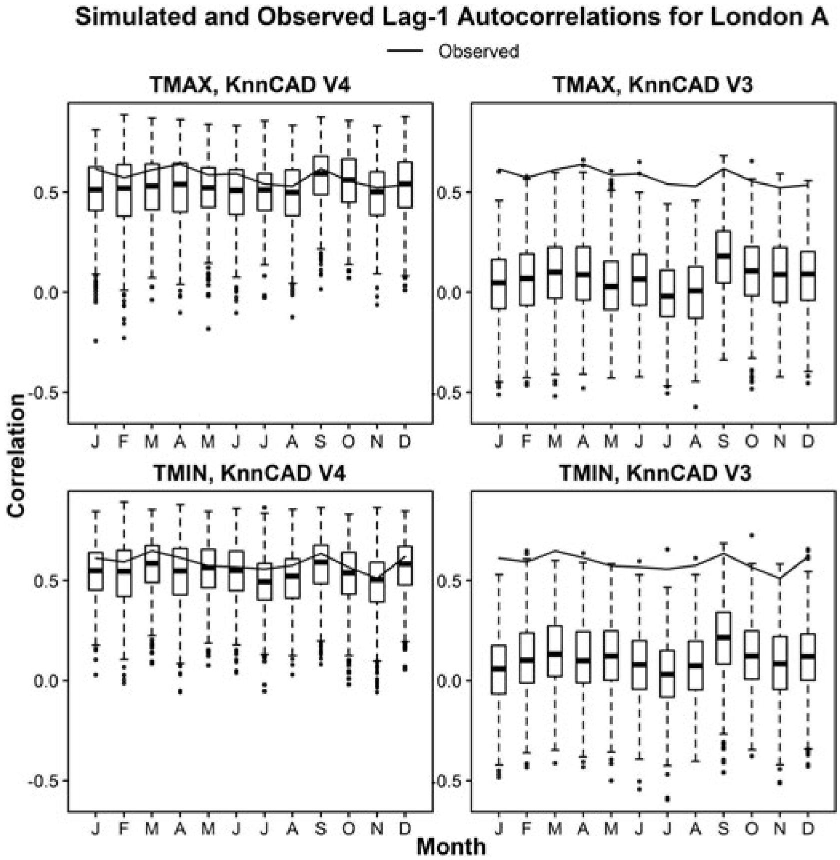


Figure 2. Observed and simulated monthly lag-1 autocorrelations of TMAX and TMIN from KnnCAD Versions 3 and 4

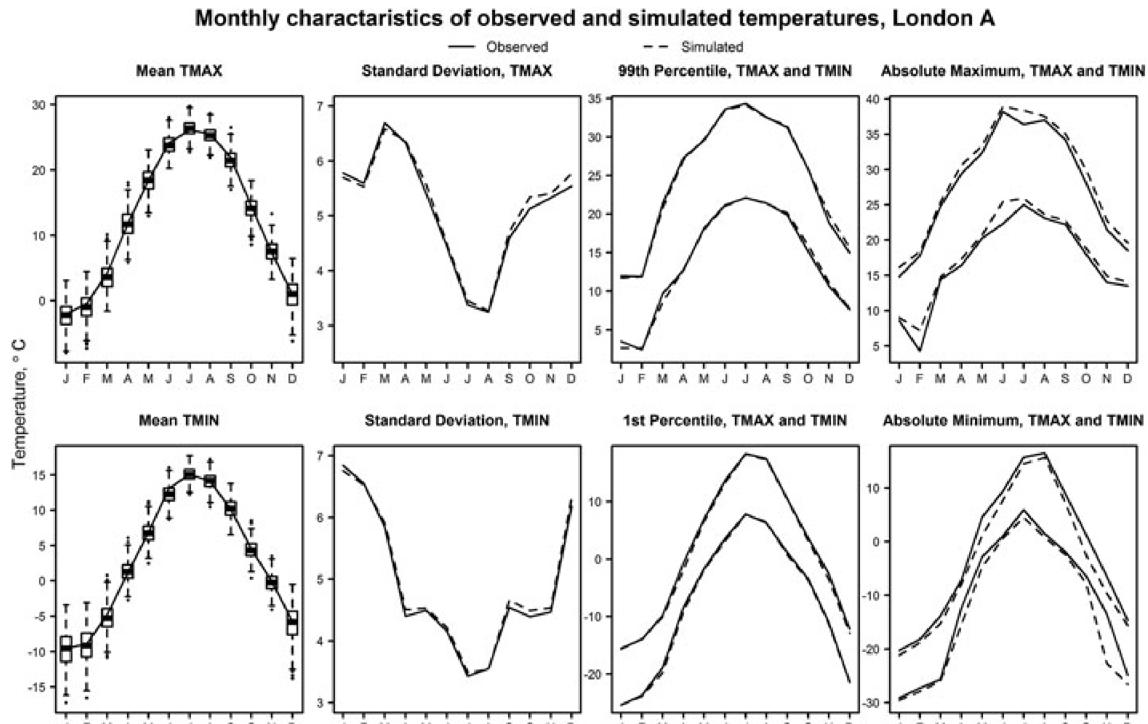


Figure 3. Monthly characteristics of observed and simulated temperatures for London Airport station from KnnCAD V4. The first column shows boxplots of the monthly averages for maximum (top) and minimum (bottom) temperatures with the historical median shown as a solid line. The second column shows the standard deviation of observed and simulated maximum (top) and minimum (bottom) temperatures. The third column shows the 99$^{th}$ (top) and 1$^{st}$ (bottom) percentiles of simulated and observed maximum and minimum temperatures. The remaining column shows the absolute maximum (top) and minimum (bottom) of the observed and simulated data

 

and TMIN values. While the simulated extreme temperatures are close to the observed records, there are some values which lie reasonably outside of the historical range, indicating satisfactory performance of the perturbation component.

Table II shows the observed and simulated results for selected climate indices from CLIMDEX (WCRP, 2009). Results are the average yearly occurrence of each index from the 27 year historical dataset and the averaged results from all ensembles of simulated data. There is a slight underestimation, just under 3 days, in the simulated annual number of frost days but all other indices are simulated quite well, indicating good performance of the perturbation mechanism in preserving the occurrence of extreme temperature conditions.

Figures 4 and 5 show observed and simulated spatial correlations between 21 stations and London A for TMAX in January and July, respectively, from KnnCAD V4. Results from KnnCAD V3 are similar. The solid line with X marks show the historical average correlations, and the points show the simulated correlation results from each of the 25 independent simulations. In both figures, it is clear that the spread of the simulated correlation values is centred on the historical observation for most station pairs, indicating good performance of the algorithm for spatial correlation. Simulations for the station pairs with historically lower correlations (due to inter-site distance or prevailing wind tendencies), such as Blyth-London A, generally had a greater spread than those with higher historical correlations. Overall the simulations reproduced station correlations quite well. Similar results are found for all other months using both TMAX and TMIN. As such, the performance of KnnCAD V4 in preservation of inter-site spatial correlations is deemed satisfactory.

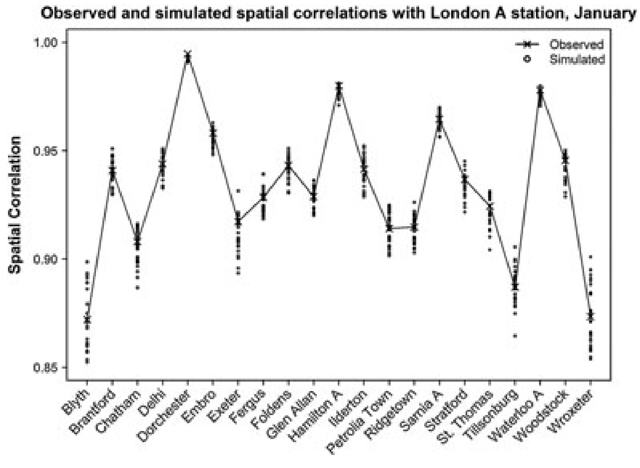


Figure 4. Observed and simulated spatial correlations between London A and all other stations for January

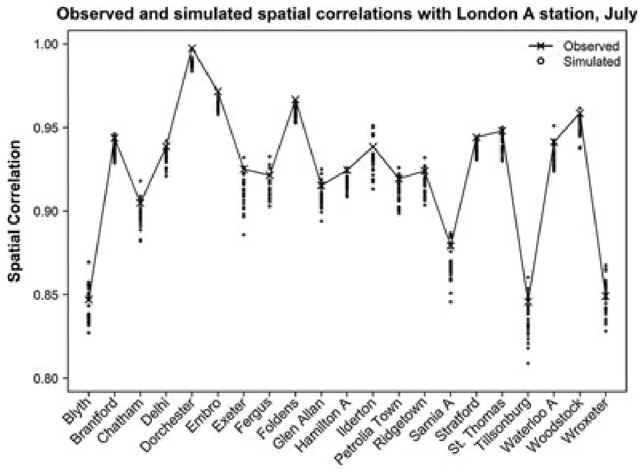


Figure 5. Observed and simulated spatial correlations between London A and all other stations for July

## CONCLUSIONS

In this study, a new methodology for K-NN weather generation is presented (KnnCAD V4). The weather generator is applied to the UTRB in southwestern Ontario, Canada. Meteorological data from 22 stations and several variables are used as inputs to the algorithm to produce 25 simulations of synthetic climate data for a total of 675 years.

The proposed algorithm is found to adequately reproduce historical monthly temperature characteristics and simulates reasonable values that lie outside of the historical range. Spatial correlations are also adequately preserved by the model. The modifications to the KnnCAD V3 of Eum and Simonovic (2010) are found to significantly improve simulation of temporal autocorrelations. This is of crucial importance for hydrologic modeling in the study area where snow accumulation and sudden melt events have historically resulted in high flood levels.

KnnCAD V4 can be easily applied to any site and has excellent potential for use as a downscaling tool in climate change impact assessments for agricultural and hydrological applications. Future work should focus on the development of a perturbation scheme for the resampled precipitation data. Downscaling of data from Atmosphere-Ocean Coupled Global Circulation Models with KnnCAD V4 to develop synthetic temperature and precipitation series for climate change impact assessments is another important research topic.

Table II. Observed and simulated CLIMDEX indices for London A

| Index | **Frost days** | **Summer days** | **Icing days** | **Tropical nights** |
|---|---|---|---|---|
| Definition | Min. temperature $<0\,^{\circ}$C | Max. temperature $>25\,^{\circ}$C | Max. temperature $<0\,^{\circ}$C | Min. temperature $>20\,^{\circ}$C |
| Observed | 140.8 days/year | 62.6 days/year | 57.9 days/year | 5.3 days/year |
| Simulated | 143.3 days/year | 63.0 days/year | 58.9 days/year | 5.4 days/year |

ACKNOWLEDGEMENTS

The authors would like to thank Environment Canada for providing the climate data used in this research, as well as Ontario Graduate Scholarships and the Canadian Federation for Climate and Atmospheric Sciences for funding the work presented in this study. Ian McLeod and Slobodan Simonovic acknowledge the support from the NSERC Discovery Grant Program.

REFERENCES

Apipattanavis S, Podesta G, Rajagopalan B, Katz RW. 2007. A semiparametric multivariate and multisite weather generator. *Water Resources Research* **43**. DOI: 10.1029/2006WR005714

Census of Canada. 2006a. *Ontario Cartographic Boundary File*, *Coast*. Statistics Canada: Ottawa, 2007.

Census of Canada. 2006b. *Ontario Cartographic Boundary File*, *Hydrography*. Statistics Canada: Ottawa, 2007.

Craigmile PF, Guttorp P. 2011. Space-time modelling of trends in temperature series. *Journal of Time Series Analysis* **32**: 378–395. DOI: 10.1111/j.1467-9892.2011.00733.x

Dibike YB, Coulibaly P. 2005. Hydrologic impact of climate change in the Saguenay watershed: comparison of downscaling methods and hydrologic models. *Journal of Hydrology* **307**: 144–163. DOI: 10.1016/j.jhydrol.2004.10.012

Dubrovsky M. 1997. Creating daily weather series with use of the weather generator. *Environmetrics* **8**:409–424. DOI: 10.1002/(SICI)1099-095X(199709/10)8:5<409::AID-ENV261>3.0.CO;2-0

Elshamy ME, Wheater HS, Gedney N, Huntingford C. 2006. Evaluation of the rainfall component of a weather generator for climate impact studies. *Journal of Hydrology* **326**: 1–24. DOI: 10.1016/j.jhydrol.2005.09.017

Eum HI, Simonovic SP. 2008. Selecting variables for the weather generator model. In *Proc., 4th Int. Symposium on Flood Defense*. Institute for Catastrophic Loss Reduction (ICLR), Toronto, Canada.

Eum HI, Simonovic SP. 2011. Assessment on variability of extreme climate events for the Upper Thames basin in Canada. *Hydrological Processes*. DOI: 10.1002/hyp.8145

Eum HI, Simonovic SP, Kim YO. 2010. Climate change impact assessment using k-nearest neighbour weather generator: Case study of the Nakdong River basin in Korea. *Journal of Hydrologic Engineering* **15**: 772–785. DOI: 10.1061/_ASCE_HE.1943-5584.0000251

Eum HI, Sredojevic D, Simonovic SP. 2011. Engineering Input for the Assessment of Flood Risk due to the Climate Change in the Upper Thames River Basin. *Journal of Hydrologic Engineering* **16**(7): 608–612. DOI: 10.1061/(ASCE)HE.1943-5584.0000346.

Furrer EM, Katz RW. 2008. Improving the simulation of extreme precipitation events by stochastic weather generators. *Water Resources Research* **44**: 13pp. DOI: 10.1029/2008WR007316

Geng S, Auburn JS, Brandstetter E, Li B. 1988. A program to simulate meteorological variables: Documentation for SIMMETEO. Agronomy Progress Rep 204, Department of Agronomy and Range Science, University of California, Davis CA

Hanson CL, Johnson GL. 1998. GEM (Generation of weather Elements for Multiple applications): its application in areas of complex terrain. *Hydrology, Water Resources and Ecology in Headwaters IAHS* **248**: 27–32.

Kuchar L. 2004. Using WGENK to generate synthetic daily weather data for modelling of agricultural processes. *Mathematics and Computers in Simulation* **65**: 69–75. DOI: 10.1016/j.matcom.2003.09.009

Mavromatis T, Hansen JW. 2001. Interannual variability characteristics and simulated crop response of four stochastic weather generators. *Agricultural and Forest Meteorology* **109**: 283–296. DOI: 10.1016/S0168-1923(01)00272-6 DOI:10.1016/S0168-1923%2801%2900272-6

McLeod AI, King L. 2012. Block Bootstrap for Time Series. *Wolfram Demonstrations Project*. http://demonstrations.wolfram.com/BlockBootstrapForTimeSeries/[accessed 24 March 2012]

Mehrotra R, Srikanthan R, Sharma A. 2006. A comparison of three stochastic multi-site precipitation occurrence generators. *Journal of Hydrology* **331**: 280–292. DOI: 10.1016/j.jhydrol.2006.05.016

Prodanovic P, Simonovic SP. 2007. Development of rainfall intensity duration frequency curves for the City of London under the changing climate, *Water Resources Research Report* no. 58, Facility for Intelligent Decision Support, Department of Civil and Environmental Engineering, London, Ontario, Canada.

Prodanovic P, Simonovic SP. 2008. Intensity duration frequency analysis under changing climatic conditions. *CD Proceedings*, 4th International Symposium on Flood Defense: Managing Flood Risk, Reliability and Vulnerability, edited by SP Simonovic and P Bourget, Paper 142, 8 pages.

Qian B, Gameda S, Hayhoe H, De Jong R, Bootsma A. 2004. Comparison of LARS-WG and AAFC-WG stochastic weather generators for diverse Canadian climates. *Climate Research* **26**: 175–191. DOI: 10.3354/cr026175

Rajagopalan B, Lall U. 1999. A k-nearest neighbour simulator for daily precipitation and other weather variables. *Water Resources Research* **35**(10): 3089–3101.

Semenov M. 2008. Simulation of extreme events by a stochastic weather generator. *Climate Research* **35**: 203–212. DOI: 10.3354/cr00731

Semenov MA, Barrow EM. 1997. Use of a stochastic weather generator in the development of climate change scenarios. *Climate Change* **35**(4): 397–414. DOI: 10.1007/BF01091621

Semenov MA, Barrow EM. 2002. LARS-WG: A stochastic weather generator for use in climate impact studies. *Version 3.0 User Manual*.

Sharif M, Burn DH. 2006. Simulating climate change scenarios using an improved K-nearest neighbor model. *Journal of Hydrology* **325**: 179–196. DOI: 10.1016/j.jhydrol.2005.10.015

Sharif M, Burn DH. 2007. Improved K-nearest neighbour weather generating model. *Journal of Hydrologic Engineering* **12**(1): 42–51. DOI: 10.1061/(ASCE) 1084-0699(2007) 12:1(42)

Silverman BW. 1998. *Density estimation for statistics and data analysis*. Chapman & Hall: Boca Raton, FL; 44–46.

Simonovic SP. 2010. A new methodology for the assessment of climate change impacts on a watershed scale. *Current Science* **98**(8): 1047–1055.

Solaiman T, King L, Simonovic SP. 2010. Extreme precipitation vulnerability in the Upper Thames River basin: uncertainty in climate model projections. *International Journal of Climatology* **31**(15): 2350–2364.

Soltani A, Hoogenboom G. 2003. A statistical comparison of the stochastic weather generators WGEN and SIMMETEO. *Climate Research* **24**: 215–230. DOI: 10.3354/cr024215

Wilby RL, Dawson CW. 2007. SDSM 4.2 – A decision support tool for the assessment of regional climate change impacts. *Version 4.2 User Manual*.

Wilcox I, Quinlan C, Rogers C, Troughton M, McCallum I, Quenneville A, Heagy E, Dool D. 1998. The Thames River watershed: Background study for nomination under the Canadian heritage rivers system. *Upper Thames Conservation Authority*. ISBN 1-894329-00-7.

Wilks DS, Wilby RL. 1999. The weather generation game: a review of stochastic weather models. *Progress in Physical Geography* **23**(3): 329–357. DOI: 10.1177/030913339902300302

World Climate Research Programme (WCRP). 2009. ETCCDI/CRD Climate Change Indices: Definitions of the 27 core indices. Accessed September 4, 2012, from http://cccma.seos.uvic.ca/ETCCDI/list_27_indices.shtml

Yates D, Gangopadhyay S, Rajagopalan B, Strzepek K. 2003. A technique for generating regional climate scenarios using a nearest-neighbour algorithm. *Water Resources Research* **39**(7). DOI: 10.1029/2002WR001769

Young K. 1994. A Multivariate chain model for simulating climatic parameters from daily data. *Journal of Applied Meteorology* **33**: 661–671. DOI: 10.1175/1520-0450(1994)033<0661:AMCMFS>2.0.CO;2

 