Improved *K*-Nearest Neighbor Weather Generating Model

Yiwen Yang

Abstract: A major limitation of the basic *K*-nearest neighbor based weather generating model is that they do not produce new values but merely reshuffle the historical data to generate realistic weather sequences. In this paper, an improved approach is developed that allows nearest neighbor resampling with perturbation of the historic data. The improved 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 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 while allowing perturbations to the observed data points. Additionally, an advantage of the improved approach is that unprecedented precipitation amounts are generated that are important for the simulation of extreme events.

Introduction

Development of weather models is an important task that has many applications in hydrology and water resources management. 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. In this process, daily precipitation amounts were generated with a single parameter using a two-state first-order Markov model from an assumed probability distribution fit to the observed values at first, then Katz (1977), Buishand (1978), and Sternand Coe (1984) used the two-parameter gamma distribution to predict the precipitation amounts. Next, Smith and Schreiber (1974), Woolhiser and Roldan (1982), and Wilks (1999) used three parameters from a mixed exponential distribution to describe precipitation amounts. Afterwards, other meteorological variables in addition to precipitation began to be researched. Richardson (1981) employed a Markov-chain exponential model to generate other meteorological variables at first and this weather generating model proposed by Richardson (1981) are commonly called WGEN, Which is a famous parametric weather model. Then Nicks et al. (1990) described an extended version of WGEN, called WXGEN, which 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 developed by Hanson and Johnson (1998).

Although parametric methods are very useful, they have several inadequacies. First and most importantly, they do not adequately reproduce various aspects of the spatial and temporal correlation 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 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 solve many problems that parametric methods cannot solve. The most famous nonparametric technique is the *K*-nearest neighbor (*K*-NN) approach. 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 approach 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. However, this method underestimated the number of dry months. Lall et al. (1996) used a *K*-NN scheme with kernel density estimators to represent the probability distributions of dry spell lengths, wet spell lengths, and wet day precipitation amounts

It is often necessary to evaluate the performance of K-NN model to extreme precipitation events that cause floods or droughts in a basin. A weather generating model capable of simulating the extreme precipitation events, while preserving the important correlation of the observed data, is likely to be help in formulating effective flood and drought management strategies at the catchment level. Earlier works of Yates et al. (2003), Buishand and Brandsma (2001), andLall and Sharma (1996) used the basic *K*-NN approach to the simulate weather data. However, a limitation of this model is that it does not produce new values but merely reshuffles the historical data to generate new weather values, which lead to underestimate climate variability. Therefore, this basic model needs to be improved to generate data values beyond the observed data in terms of simulating extreme events. 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 so that the extreme events can be simulated very well.

This paper is organized in the following manner. The next section illustrates the K-NN algorithm for simulating daily weather values. The subsequent section describes the application of the K-NN algorithm to the basin along with a description of results. Some important conclusions and ideas about the K-NN algorithm are presented in the next section.

K-NN Model Development

Nearest neighbor methods have been intensively used in the field of statistics and in pattern recognition procedures. Despite they are simple, nearest neighbor algorithms are considered robust. Although many other sophisticated techniques have been developed, nearest neighbor methods remain very popular. A K-NN algorithm involves selecting K data vectors similar to the vector of interest. One of these K vectors is selected to represent the vector of the given time in the simulation period based on the given criterion. In the context of weather data simulation, the K-NN approach involves selecting the k nearest neighbors of current feature vector t according to a defined Mahalanobis distances criterion, then the nearest neighbor of the current feature vector t is chosen from k neighbors according to a defined probability distribution and the nearest neighbor is considered as future feature vector t+1. Models based on the *K*-NN approach can easily reproduce historical statistics and correlation structure of the observed data because the generated simulated data from the model are very similar to the observed data. This is an important reason why the model is very popular and robust.

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 , . . . , *T*, and *j*=1, . . . , *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 $X_t^j = (x_{1,t}^j, x_{2,t}^j, ..., x_{p,t}^j)$ where $x_{i,t}^j$ represents the value of the weather variable i for station j and day t. The algorithm cycles through various steps to obtain the weather for day *t*+1 and 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

$$\overline{X}_t = (\overline{x}_{1,t}, \overline{x}_{2,t}, ..., \overline{x}_{p,t}) \qquad (1)$$

Where

$$\overline{x}_{i,t} = \frac{1}{q}\sum_{j=1}^{q} x_{i,t}^j \qquad \text{i=1,..., p,} \qquad \text{and t = 1,…,T} \qquad (2)$$

For example, for the vector $\overline{x}_{1,t}$, if t =1, 1 represents precipitation, then the $\overline{x}_{1,t}$ represents means of the precipitation across q stations for day 1.

2. Determine the size, *L*, of the data block that includes all potential neighbors to the current feature vector. 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. Although the value of w is 14, actual number of all days between January 13 and January 27 is 15. Therefore every year in N years has (w+1) potential neighbors. And the current day January 20 should not be consider as a neighbor of itself. Thus, the day block of potentials consists of L =(w+1) × N−1 days.

3. Compute mean vectors across *q* stations for each potential day in the data block consisting of potential neighbors using the expressions given in Step 1.

4. Compute the covariance matrix, c_t, for the current day t.

5. Selecting day t in the historic record of N years and considering it as “current day t”:
The weather on the day t (e.g., January 20) comprising all p variables is randomly chosen from the set of all January 20 values in the historic record of N years. For example, if today is December 1, 2014 and tomorrow weather needs to be predicted. In this case, December 1 of a given year should be selected as the current day. However, today should not be selected as the current day because some of its neighbors(e.g,. December 2, 2014, December 3, 2014) have not happened so that these

neighbors cannot be selected as neighbors of today in the simulation period. Therefore, December 1of any year in the historical records is chosen as the current day. If N=10 and all years between 2004 and 2013 are chosen as a sample, then December 1 of any year between 2004 and 2013 can be chosen as the current day. Similarly, if another 10 years are chosen as a sample, then the December 1 needs to be selected from these 10 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 $\overline{X}_t$ and the mean vector $\overline{X}_i$ for day i, where i = 1,……,L. day i means a potential neighbor of the current day t . The distance metric can be given through

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

   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 *K* nearest neighbors from all potential 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 is used for resampling from the *K* nearest neighbors. Weights are assigned to each neighbor of j neighbors according to the metric given by

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

   The cumulative probabilities, p_j , are given by

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

   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.

9. Determine the nearest neighbor of the current day by using the cumulative probability metric given by Eq. (5):
   Firstly, generate a random number, $r \subset (0,1)$.
   If p_1 <r< p_k , then the day *j* for which *r* is closest to p_j is selected.

If $r \le p_1$, then the day corresponding to d_1 is selected.
If $r = p_k$, then the day corresponding to d_k is selected.
The selected neighbor is adopted to represent the day t+1. The step 1 to step 9 of the process is called basic K-NN approach. In the improved approach 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. Perturbation of the values of weather variables obtained using the basic *K*-NN approach is carried out in the following steps:

a. This involve calculating some notations:

σ_i^j : Conditional standard deviation of variable i for station j computed from the K nearest neighbors.

λ : Bandwidth, a function of the number of samples. (Sharma et al. 1997; Sharma and O'Neill 2002).

γ_{t+1} : A random variate for day t+1 in the simulation period from a normal distribution with zero mean and unit variance.

Then the new value of weather variable i for day t+1 and station j is given by

$$y_{i,t+1}^j = x_{i,t+1}^j + \lambda \sigma_i^j \gamma_{t+1} \qquad (6)$$

Where $x_{i,t+1}^j =$ value of the weather variable for day t+1 and station j obtained from the basic K-NN model; $y_{i,t+1}^j =$ corresponding value obtained after perturbation.

b. Since the precipitation values are bounded, there is a possibility that Eq. (6) in the above step could lead to negative precipitation amounts. 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 α =0.06 for which γ =-1.55. The largest value of λ is therefore given by $\lambda^\alpha = x_{3,t+1}^j / (1.55 \times \sigma_3^j)$, where subscript 3 refers to precipitation values and λ^α =acceptable value of λ .If the calculated value of λ is larger than λ^α, then λ^α 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.

Model Application

Model Parameters

The performance of the K-NN model depends on the proper parameters whose values need to be determined before the model is carried out. Two important model parameters of the K-NN model are the width of the temporal window, w, and the value of K. It is worth mentioning that Although K and w are parameters, the k-NN model is a nonparametric model 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. For example, if the parametric models also need neighbors and current day t to predict weather and the weather condition of spring is different from winter, then the number of potential neighbors of the current day t in spring is possibly different from that in winter, thus the parameters of the parametric model must be different in this two conditions. However, in the K-NN model, the effect of seasonality is through the moving window w. For example, if the current day t in spring is April 15 and it only needs a window of width 10 because the weather condition of days outside the window are very different from the April 15. And if the current day t in winter is January 15 and it needs a window of width 20 because the weather condition of days inside window are very similar to the January 15. Additionally, the effect of seasonal variation is greatly reduced for the K-NN model because the search for nearest neighbors is restricted to days within a moving window. Therefore, the width of window must be sufficiently large such that the correlation between the days outside the window can be neglected. In other words, if width of the window is small, some nearest neighbors are likely to be lost. 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 but other approaches are also available. Lall and Sharma (1996) recommended a heuristic value of $K = \sqrt{L}$ .Buishand and Brandsma (2001) observed that resampling with a small number of nearest neighbors (i.e, the value of K) might lead to duplication of large parts of the historical record. 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 is adopted.

Reproduction of Historical Statistics

The performance of the proposed model was evaluated through application to data from the Upper Thames River Basin. The goal of simulation was to produce a data series that preserved the statistical attributes of the historic data while involving perturbation. The value of w (w=14), N (N=38), L (L=569), K (K=24) were used in this study. If the current day t is transformed in the historical records or another N years are selected as a sample, then there are a lot of simulated data. A fixed number 800 simulated data were used in this study. The simulation data were compared to the observed data using box plots. Box plots are a preferred method of data analysis as they show the range of variation in the simulation data and provide a straightforward method of comparing the simulations data with the historical data. 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(i.e, N=38).

London was one of the 15 stations and Results are presented below only for London since the results for other stations were similar.

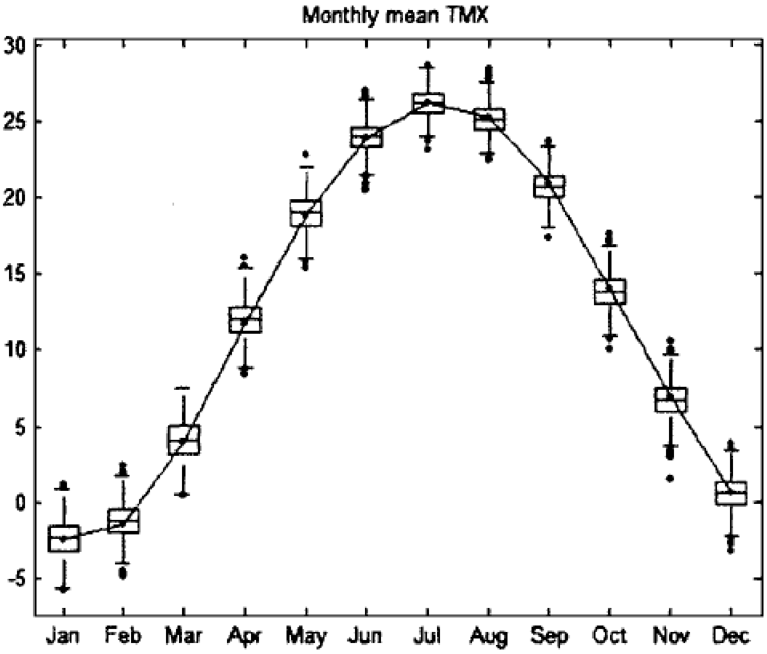
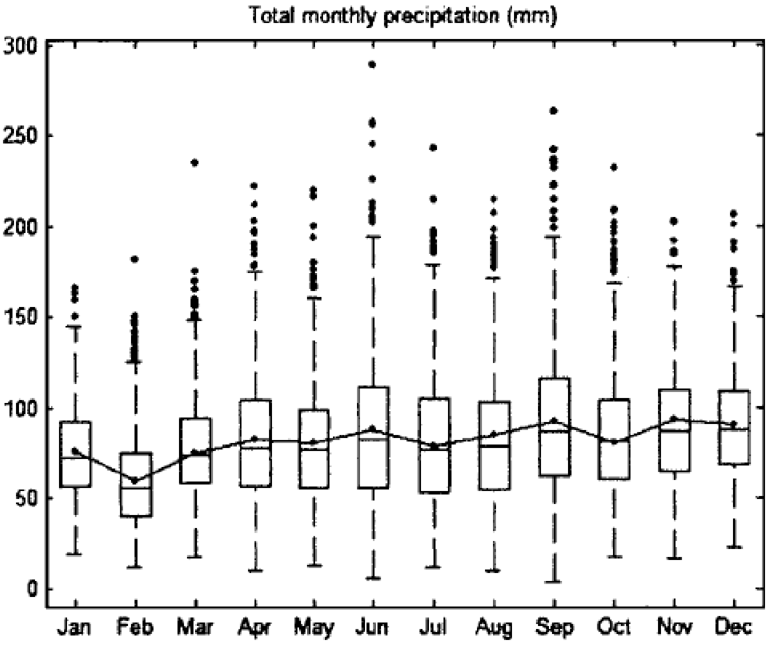


Fig.1. Box plots of monthly mean maximum temperature

Fig. 1 shows the box plots of 800 simulated values of mean TMX values for London. Although the model is 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 simulated data 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 can adequately reproduce the historical data.



\

Fig.2. Box plots of total monthly precipitation

Fig. 2 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. Some values are found to be beyond the acceptable values, but these outliers indicate variability of the simulated data. The model slightly overestimates the total monthly precipitation for February, August, September, and November, but for the rest of the months, model results are very close to the observed data values. Among all weather variables, precipitation has the greatest variability and therefore the performance of the model in terms of simulating the total monthly precipitation may be considered to be very good. Because the perturbation tends to increase the variance of the simulated values, the monthly standard deviations of 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. Thus, the simulations also adequately reproduce the probability distributions of historical values.

Preservation of Correlation Structure

Parametric models cannot reproduce the correlation structure of the observed data. However, the basic K-NN model resamples from the observed data by conditioning on the weather for the previous day, and is therefore likely to preserve the correlation structure. And because of perturbation, the correlation structure of the observed data may be tempered with the improved K-NN model. To keep the correlation structure complete, it was decided to use a constant value of the random normal variate(i.e., the value of γ_{t+1} was constant) 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.

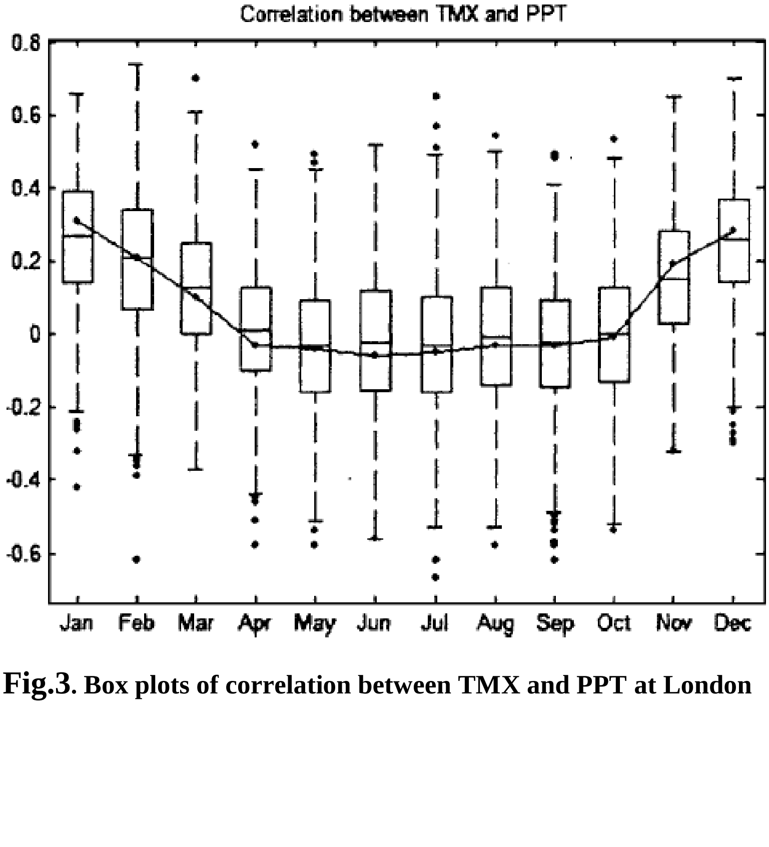


Fig.3. Box plots of correlation between TMX and PPT at London

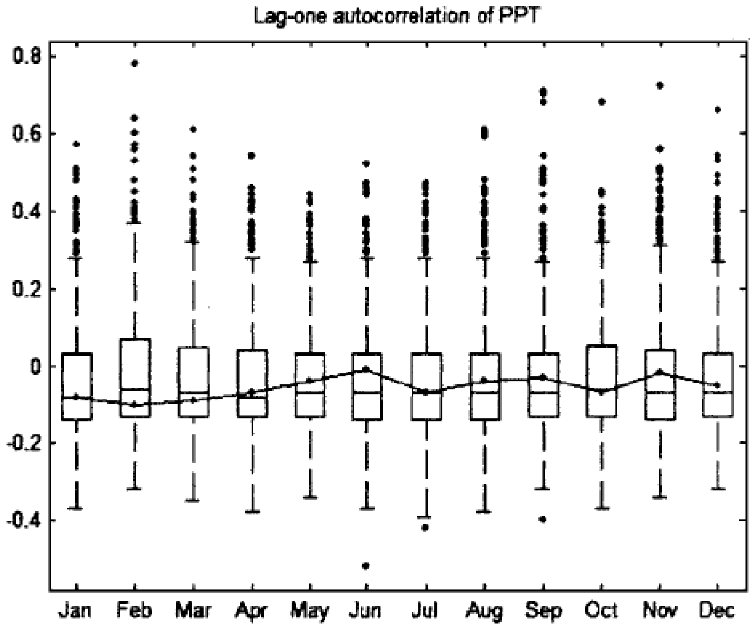


Fig.4. Lag-one autocorrelation of PPT at London

Box plots for correlation between TMX and precipitation and autocorrelation of PPT are shown in Fig.3 and 4 respectively. It can be seen from the box plots shown in Fig. 3 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. And It can be seem From the Fig.3 that the correlations are adequately preserved by the K-NN model. However, the improved K-NN model cannot clearly preserve daily correlations. Clark et al. (2004) presented a method to preserve daily correlations that involves reordering the ensemble members to reconstruct the spatial and temporal correlation statistics of the observed data.

It can be seen from the box plots shown in Fig. 4 that autocorrelation of PPT of the observed data for all months are close to zero, which implies a very weak autocorrelation of PPT, and the K-NN model adequately preserve the correlations.

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 However, 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 al.1998). The hypothesis of uniform spatial distribution of precipitation is invalid, even for small basins, because runoff simulation in the basin is significantly influenced by the distribution of precipitation. Thus, it is important to evaluate the performance of the K-NN model in preserving the spatial correlations of the observed data. Scatter plots are also a method of data analysis. And scatter plots of station correlations for daily TMX and precipitation values are presented in Figs. 5 and 6, respectively. It is worth mentioning that the spatial locations of the 15 stations that were used in this paper were different.

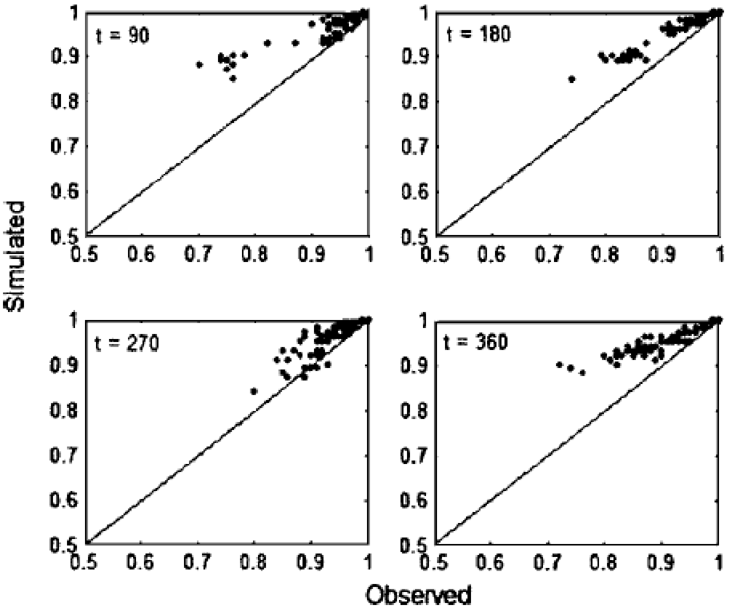


Fig.5. Comparison of observed versus simulated station correlations for daily TMX values between all station pairs for 4 representative days.

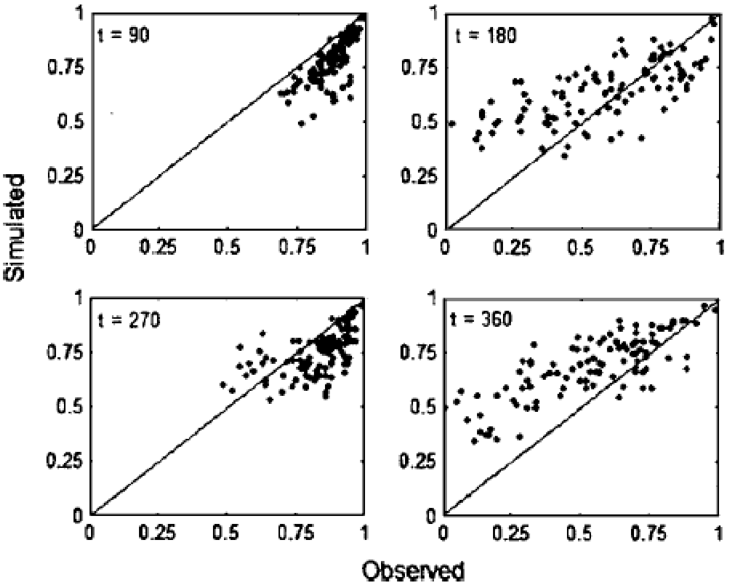


Fig.6. Comparison of observed versus simulated correlations for daily PPT values between all station pairs for 4 representative days

Fig. 5 shows scatter plots of station correlation coefficients for daily TMX values in the simulated and the observed data. The horizontal axis represents station correlation coefficients of the observed data and the vertical axis represents station correlation coefficients of the simulated data. For 15 stations, there are 105(i.e (14+13+…., +1) × 14/2)) pair correlation coefficients for each day. The scatter plots are shown for 4 representative days. It can be seen from Fig.5 that most values of station correlations coefficients are in the range of 0.8-1.0. And most data points lie in the close vicinity of the 45 °sloping solid line that is shown in the scatter, which implies the simulated station correlation coefficients are higher than the observed station correlation coefficients due to effect of the perturbation.

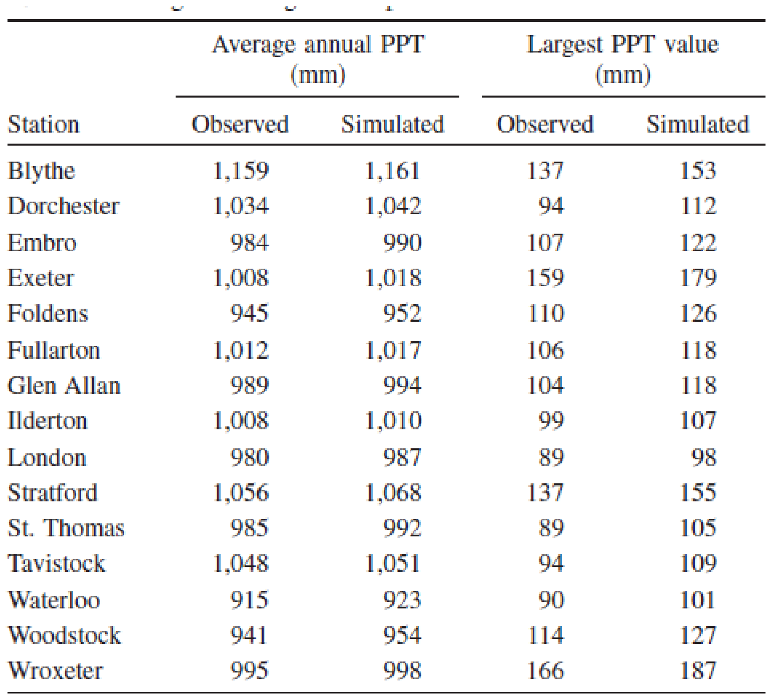
Given the structure of the basic *K*-NN model, the observed station correlation characteristics are definitely preserved on a daily time scale. Therefore, the overall performances of the K-NN models are satisfactory.

The scatter plots of station correlations of daily PPT values between the observed and the simulated data are shown in Fig. 6. Although the correlations in the observed data are not as strong as in the case of TMX, the model preserved the historical structure very well. And it was observed that the basic and improved K-NN models give similar results in terms of simulating station PPT correlations. Thus, the performances of the K-NN models are satisfactory. With the *K*-NN model, the spatial correlation 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 correlation 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 1 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

Table 1 summarizes the results of simulation with respect to reproduction of long-term average annual precipitation. It can be seen from Table 1 that the model overestimates the annual average precipitation but the amount of overestimation is very small. The reason for this overestimation is the bias due to recomputation of the normal random variate(i.e., γ_{t+1}) whenever the precipitation amounts become negative. Overall, the performance of the model is satisfactory in terms of simulating annual average precipitation. A comparison of the largest daily precipitation amounts simulated by the model with the observed data is also presented in Table 1. It can be observed that the simulated amounts are significantly higher than the observed amounts. The model can generate a amount of 187 mm precipitation compared to the historical largest value of 166 mm. Therefore, it can be seen clearly that the model can produce unprecedented, but realistic, precipitation amounts throughout the basin.

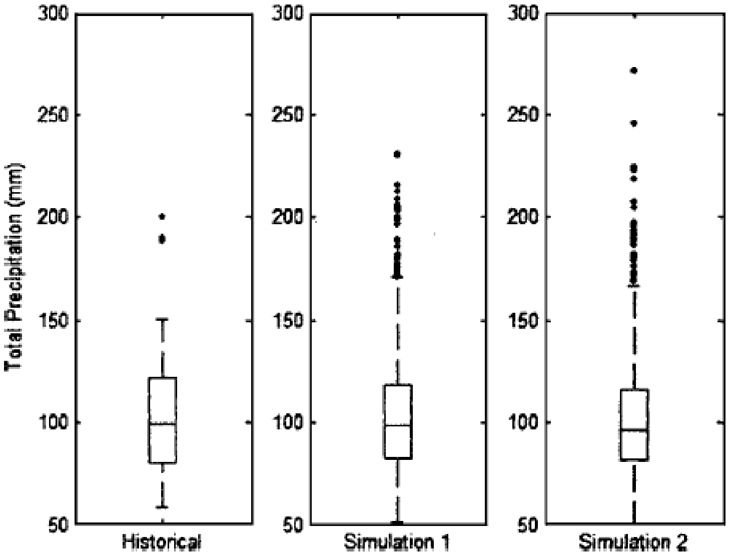


Fig.7. Box plots of total precipitation during extreme events in each year of historical and simulated data (Simulation 1 refers to basic *K*-NN model, Simulation 2 refers to improved model)

Fig.7 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 improved 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. Because of the perturbation, a amount of 280mm extreme precipitation event is observed in Simulation 2 and the improved model simulates around five precipitation events that are more severe than the most extreme precipitation event in the historical record. And the basic *K*-NN model simulates a most extreme precipitation event with a amount of 230mm precipitation.

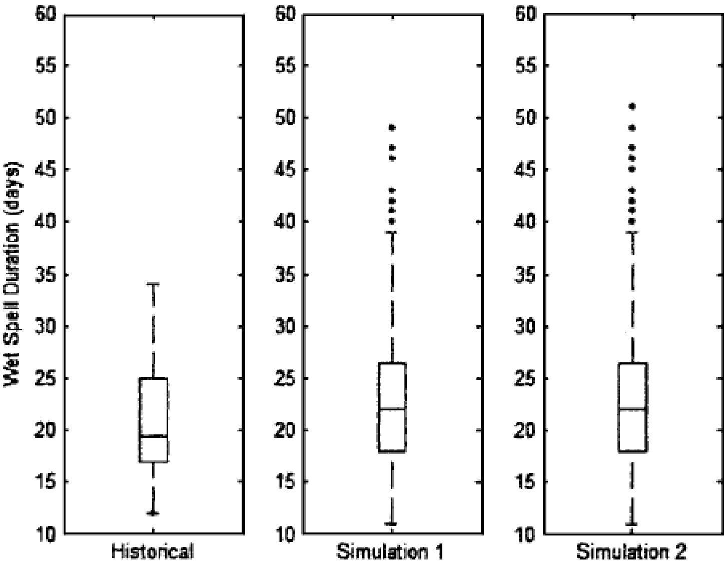


Fig.8. Box plots of extreme wet spell duration in each year of historical and simulated data (Simulation 1 refers to basic *K*-NN model, Simulation 2 refers to improved model)

Fig.8 shows the box plots of extreme wet spells in the historical and the simulated records. It can be seen from the fig.8 that the most severe wet spell simulated by the improved model is 52 days. While the corresponding value obtained from the basic model is 48 and the historical record is 34.

In a conclusion, it can be seen from Figs. 7-8 that the improved K-NN model can simulate more extreme events than the basic K-NN model.

Conclusions

Main results

The development and evaluation of a improved version of the basic *K*-NN weather generating model has been presented. It can be seen that the improved 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. 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 historical statistics. Important properties of precipitation were preserved. Correlation among the variables was preserved, which is particularly important for erosion, crop production, and rainfall– runoff models. An important advantage 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 preserve the spatial correlations of the variables, the proposed model adequately reproduced the spatial correlation of the observed data.

Furthermore, an encouraging aspect of the proposed model is that extreme unprecedented events, both low precipitation and high precipitation, can be simulated. And the proposed model can predict extreme flood and drought for the basin. Although the *K*-NN algorithm was designed to model daily statistics, the monthly statistics are also adequately reproduced for the application presented here.

My ideas about the paper

Predicting average daily weather

1. The first step of the K-NN model in this paper is that compute regional means of the p variables across the q stations for each day of the historical record. The expressions are $\overline{X_t} = (\bar{x}_{1,t}, \bar{x}_{2,t}, ..., \bar{x}_{p,t})$ (1), where $\bar{x}_{i,t}$ = 1/q $\sum_{j=1}^{q} X_{i,t}^{j}$, i=1,...,p, and t = 1,…,T. But this expressions can be transformed to the form as $\overline{X_i} = (\overline{x_1}, \overline{x_2}, ..., \overline{x_p})$ (2), where, $\bar{x}_i = 1/qT\sum_{t=1}^{T}\sum_{j=1}^{q} x_{i,t}^{j}$, i=1,.....,p and j=1,........q, t=1,...,T which means regional means of the p variables across the T days and q stations of the historical record. If T=1, this expressions are identical with Eq(1).

2. The second step is that determine the value of T. If the means of the P variables across the whole year needs to be investigated, then the value of T is set to 365.Similarly, the value of T should be set to 30 if the means of the p variables across a month needs to be investigated.

3. Similarly, the third step is that determine the size, L, of the data block that includes all potential neighbors to the current feature vector. A window of width w, is chosen and all means of the p variables across days within the window are considered as potential candidates to the current feature vector. It is worth mentioning that width of the window is according to the values of T. For example, if means of the p variables across the whole January of 2015 needs to be examined, then a window of width 3 is selected, which means the window consists of all means of the p variables across the whole January, the whole February and the whole march for all N years but excluding "the current day" that is randomly chosen in advance. Because the average daily weather of January, February and March are similar. The data block of potential neighbors consists of L= $w \times N - 1$.

4. The following steps are similar to the K-NN model steps illustrated by the paper. However, it is important to note that how to select "the current day t". The weather on the day t(e.g., means of p variables across January) is randomly chosen from the set of all means of p variables across January in the historical record of N years. For example, if the means of the p variables across the whole January of 2015 needs to be investigated, N=10 and all years between 2004 and 2014 are selected as a sample, then means of the p variables across the whole January of any year between 2004 and 2014 can be selected as "the current day t".

5. Therefore, average daily weather of any given time can be predicted. And if the value of T is

smaller, the prediction has more research value.

Another method to determining the value of K

The step 8 of the K-NN model gives a discrete probability distribution that gives higher weights to the closer neighbors and weights can be defined by $w_j = \frac{1/j}{\sum_{i=1}^{k} 1/i}$. If the k neatest neighbors are selected as a sample, then the MLE method can be used to determine the value of K. But I am not sure whether this method is correct.

Improved method to determining the value of N and the current day

When the size of L, the data block that includes all potential neighbors to the current feature vector, is determined, w days' weather condition of every year within N years should be examined. If w days' weather condition of any year is very different from the other years, it should be removed and replaced by a new year. Similarly, if the selected current day t (e.g., January 1) is different from the other January 1 values in the historic record of N years, it should be replaced by another January 1 value in the historic record of N years. This improved method can generate less outliers and reduce simulation times but it does not matter if enough simulation times involve. Therefore, the choice of w, N, K and the current day are worth studying.

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.

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 rainfalldata.” *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(120, 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.

,