DEVELOPMENT AND APPLICATION OF K-NEAREST NEIGHBOUR WEATHER GENERATING MODEL

by

Mohammed Sharif[(1)] and Donald H Burn[(2)]

[(1)] *Research Fellow, Department of Civil Engineering, University of Waterloo, ON, Canada.* e-mail: msharif@uwaterloo.ca

[(2)] *Professor, Department of Civil Engineering, University of Waterloo, ON, Canada.* e-mail: dhburn@uwaterloo.ca

Abstract

A generic weather generator, based on the K-nearest neighbour algorithm, is presented for producing synthetic weather sequences that can be used in conjunction with hydrological models. Application of the model to the Upper Thames River basin in Ontario has clearly demonstrated the practicality of the approach in generating plausible climate change scenarios for the basin. Daily weather variables (maximum temperature, minimum temperature, and precipitation) were simulated at multiple stations in the basin. Statistical analysis of the synthetic series generated by the model clearly demonstrated the ability of the model to reproduce important statistical parameters of the observed data series such as the mean, variance, and skewness. A distinct practical advantage of the approach presented here over the traditional Richardson and serial type weather generator is that the spatial correlation of the variables can be adequately reproduced. Cross-correlations of the variables at a station, and autocorrelations of variables, are also strongly preserved. Site-specific assumptions regarding the probability distributions of the variables are not required thereby permitting transportability of the model to other basins with very few modifications.

Keywords: weather generator, K-NN, climatic change, Upper Thames River Basin

INTRODUCTION

'Richardson type' models (Richardson, 1981) are the most widely used weather generators but a major drawback associated with these models is that persistent events, such as drought or prolonged rainfall, are not well reproduced. To overcome this problem, the serial approach to weather generation was presented by Rackso et al. (1991) and Semenov et al. (1998), wherein the sequence of dry and wet series of days is modelled first and the precipitation amounts, and other variables, are generated conditioned on the wet or dry series. Stochastic weather generators have recently been employed for producing future climate change scenarios by Wilks (1999) and Semenov and Barrow (1997). However, the inherent drawbacks of the parametric models have motivated the development of nonparametric methods.

This paper describes the development and application of a nonparametric weather generator, based on the K-nearest neighbour (K-NN) algorithm (Young, 1994; Rajagopalan and Lall, 1999; Buishand and Brandsma, 2001; Yates et al., 2003) for producing weather data based on plausible climate change scenarios for the Upper Thames River Basin (UTRB) in Ontario.

THE K-NN ALGORITHM

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 historical series. The vector consisting of the weather variables for the current day is called the feature vector and can be expressed, in expanded form, as $X_t^j = [x_{1,t}^j, x_{2,t}^j, ..., x_{p,t}^j]$ where $x_{i,t}^j$ is the value of the weather variable i at time step t for station j. The various steps of the algorithm are:

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

2. Determine the size, L, of data block that includes all potential neighbours to the current feature vector from which 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. A fixed length, 14-day temporal window was used in this study. For $w = 14$, and $N = 38$, $L = 569$.

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 current day t using the data block of size $L \times p$.

5. 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 proceeds to select one of the nearest neighbours to represent the weather of the given day in the simulation period.

6. Compute Mahalanobis distances between the mean vector of the current day's weather, $\overline{X}_t$ and the mean vector for day i, $\overline{X}_i$ where $i = 1,...,L$.

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

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

7. Determine the number of nearest neighbours, K, to be retained for resampling out of the total of L neighbours. Rajagoplan and Lall (1999) and Yates et al. (2003) recommended the use of a heuristic method for choosing K according to which $K = \sqrt{L}$. In this study $L = 569$ and hence a value of K equal to 24 was adopted.

8. Sort the Mahalabonis distances in ascending order and retain the first K nearest neighbours. Assign weights to each of the j neighbours according to the probability metric defined as

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

9. The weather on the given day in the simulation period (i.e., the day $t+1$) is selected from amongst the K-nearest neighbours using the probability metric defined by (4).

Steps 6 to 9 are repeated to generate synthetic data for the required number of years. The driving data set for the model can be strategically resampled from the historical data to obtain sequences conditioned upon a desired climate change scenario.

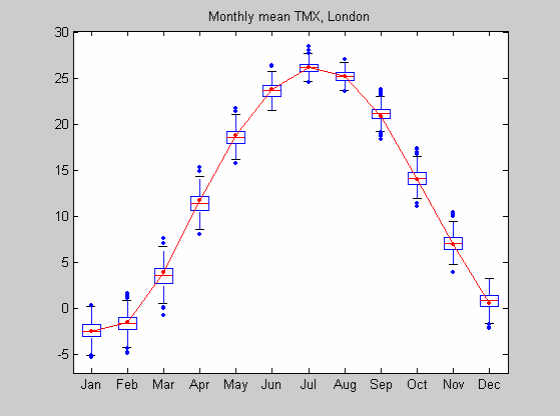
APPLICATION TO UPPER THAMES RIVER BASIN

The K-NN model described above was applied to data from the UTRB in Ontario. Daily maximum temperature (TMX), minimum temperature (TMN) and precipitation (PPT) data from nine stations in the basin were used for the period 1964-2001. The model was used to generate 800 years of synthetic weather data for simulating the observed weather and for simulating potential climate change scenarios.

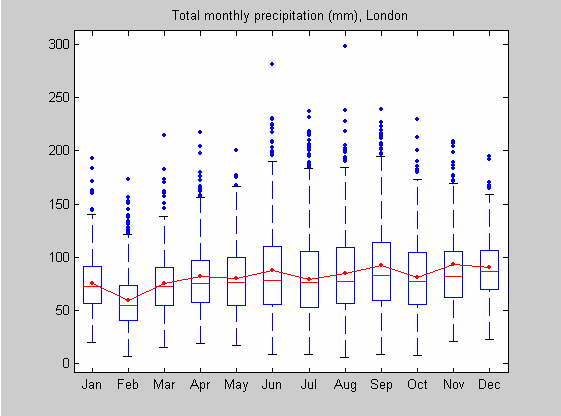
Case 1: Reproduction of Historical Data Statistics

The first simulation was conducted to reproduce the statistical characteristics of the historical data. *Figure 1(a)* shows box plots of simulated values of mean TMX values for London station. The model was able to adequately reproduce the historical values which is highly satisfactory given that monthly statistics are not explicitly specified in fitting the K-NN model. *Figure 1(b)* provides box plots of total monthly precipitation. It can be seen that the historical mean of the total precipitation is close to the median of the simulated data for all months. *Figure 1(c)* shows box plots of the total number of wet days. The model reproduced the historical statistics very well, although there was a slight overestimation for the months of April and September, and underestimation for June.

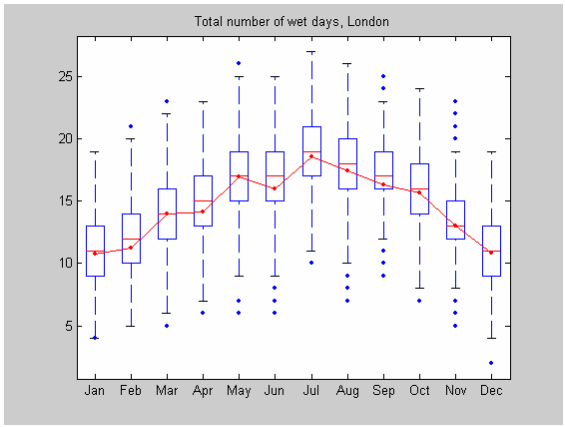
The performance of the K-NN model for the reproduction of the correlation structure was also investigated. Box plots of correlation between TMX and PPT and autocorrelations of PPT are shown in *Figure 1(d)* and *1(e)* respectively. The model adequately reproduced the historical correlation structure. The interstation correlations for mean monthly PPT are shown in *Figure 1(f)*. For q stations, there are $q(q-1)/2$ pairwise correlations resulting in 36 such correlation coefficients for each month. The performance of the model with regards to the reproduction of interstation correlations is extremely good. Although parametric models such as LARS-WG (Semenov et al., 1998) and WGEN (Richardson and Wright, 1984) can be effectively used 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 the stations.



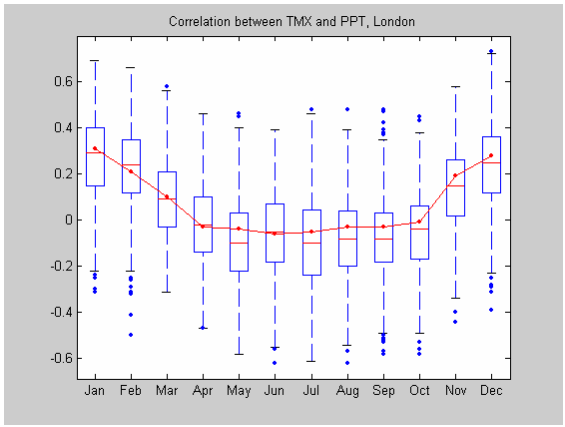
(a)



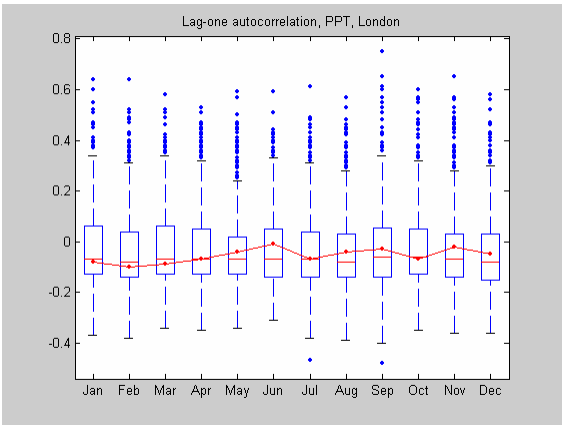
(b)



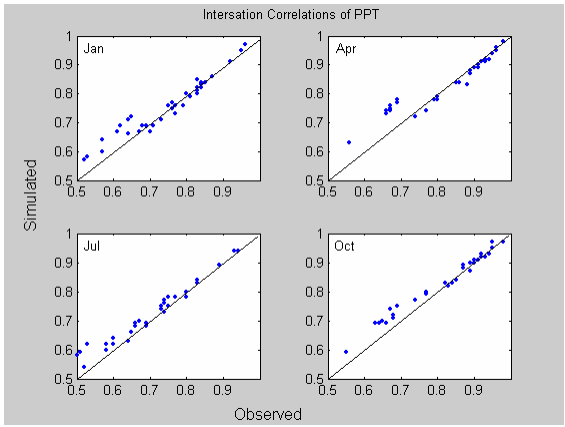
(c)



(d)



(e)

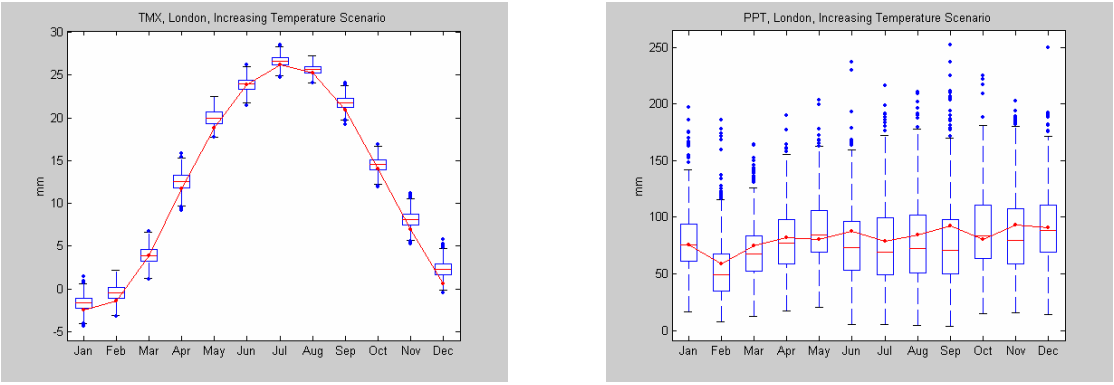


(f)

Figure 1 Box plots of (a) monthly mean maximum temperature, (b) total monthly precipitation (c) total number of wet days, (d) correlation between TMX and PPT (e) lag-one autocorrelations of PPT (f) interstation correlations

Case 2: Increasing Average Temperature Scenario

The driving data set for the model comprising years with increased temperatures is obtained by resampling strategically from a ranked list generated on the basis of the deviations of mean annual average temperature from the long term historical mean. To compute the deviation for each year, the overall long term mean is subtracted from the mean yearly value for that particular year. On the basis of deviations, a ranked list of years is generated with the first rank corresponding to the year with the lowest deviation and the last rank corresponding to the year with the highest deviation. Biasing of certain years over others was carried out using an index function.



(a) (b)

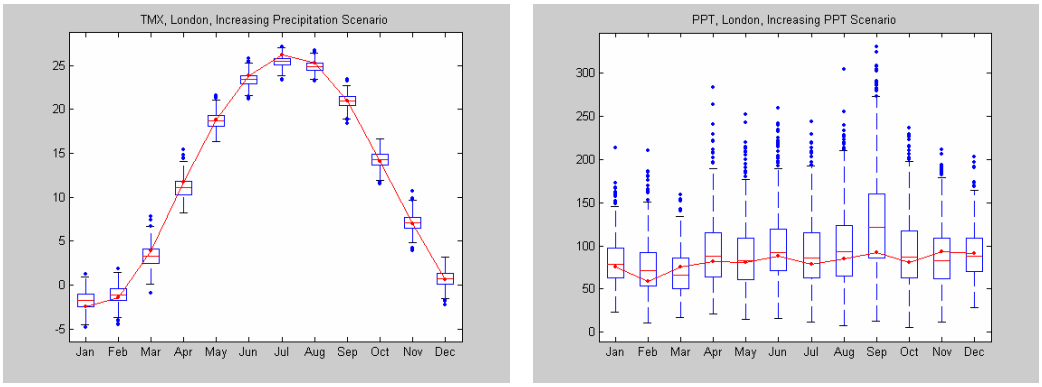
Figure 2 Box plots for increasing TAV scenario: (a) monthly TMX, (b) total monthly PPT

The box plots in *Figure 2(a)* show that the model produces increases in TMX over the historical values for all months. This suggests that with strategic resampling, the model is capable of producing alternate climate scenarios with desired attributes. The effect of increasing TAV on total monthly precipitation is shown in *Figure 2(b)*. It appears that there has been a decrease in precipitation for most of the months except for January, May, October and December when the precipitation remained nearly the same as the historical. The correlation structure of the observed data was preserved in the simulations.

Case 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. A new data set comprising years with increased annual precipitation is obtained and used as the driving data set for the K-NN model.

Figure 3(a) reveals that the model produced some increase in TMX values for the months of January and February. As seen earlier in *Figure 1(c),* there is a positive correlation between the historical TMX and PPT values during the winter months (November, December, January and February). Owing to this positive correlation, an increase in TMX values is accompanied with an increase in precipitation in January and February. There appears to be a slight increase in PPT for November and December but the corresponding increase in TMX is insignificant for these months.



(a) (b)

Figure 3 Box plots for increasing precipitation scenario: (a) Monthly TMX, (b) Total Monthly PPT

SUMMARY AND CONCLUSIONS

Application of the K-NN weather generator to the data from UTRB has clearly demonstrated the practicality of the approach in generating plausible climate change scenarios for the basin. A major advantage of the approach is that no site-specific assumptions regarding the probability distribution of the variables are required in the algorithm. The model is therefore fairly generic and easily transportable to any other basin with minimal changes. Important properties of precipitation spell structure and amounts were preserved in the simulated sequences. Cross correlation among the variables are preserved, which is particularly important for erosion, crop production and rainfall runoff models. The model adequately reproduced the spatial correlation of the observed data unlike most parametric models, which cannot be expected to preserve the spatial dependencies of the variables.

The K-NN model does not produce values not seen in the historical record. To alleviate this limitation, perturbation of the data in the spirit of traditional autoregressive methods was carried out. Preliminary runs suggest that the bounds on variables are indeed exceeded resulting in negative values of precipitation amounts. Setting these negative precipitation values to zero led to a bias that produced monthly totals higher than the observed values, which is unacceptable. A strategy needs to be devised that would preserve the historical totals while producing precipitation values not seen in the historic record. Further research is therefore required in this direction. As expected, the perturbations made to temperature values did not significantly affect the overall mean temperature.

REFERENCES

Buishand, T. A. and T. Brandsma, (2001) Multisite simulation of daily precipitation and temperature in the Rhine Basin by nearest-neighbor resampling, *Water Resources Research*, 37(11): 2761-2776.

Rackso, P., L. Szeidi and M. Semenov, (1991) A serial approach to local stochastic weather models, *Ecol. Model.*, 57: 27-41.

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

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

Richardson, C.W. and D.A. Wright, (1984) WGEN: A model for generating daily weather variables, U.S. Department of Agriculture, Agricultural Research Service, Washington, D.C., ARS-8, 88p.

Semenov, M. A. and E.M. Barrow, (1997) Use of a stochastic weather generator in the development of climate change scenarios, *Climate Change*, 35: 397-414.

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

Wilks, D. S., (1999) Multisite downscaling of daily precipitation with a stochastic weather generator, *Climate Research*, 11: 125-136.

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

Young, K. C., (1994) A multivariate chain model for simulating climatic parameters with daily data, *Journal of Applied Meteorology*, 33: 661-671.