Infilling Missing Daily Precipitation Data at Multiple Sites Using the Multivariate Truncated Normal Distribution Model for Weather Generation

W. W. Ng[1] and U. S. Panu[2]

[1]Department of Civil Engineering, University of Manitoba, Winnipeg, Manitoba, Canada, umngw@cc.umanitoba.ca

[2]Department of Civil Engineering, Lakehead University, Thunder Bay, Ontario, Canada, uspanu@lakeheadu.ca

Abstract

Stochastic weather modeling of daily precipitation amounts is frequently subject to a number of challenges, e.g., the difficulty in modeling the unique statistical characteristics of observations. The difficulty could further be complicated when the varied spatial-dependency of observations at multiple sites and the uncertainty induced by the existence of missing observations are considered. A multivariate truncated Normal distribution model is proposed to transform the skewed distribution of precipitation amounts at multiple sites into a multivariate Normal distribution model. The missing observations are then estimated through the conditional simulation using the parameters obtained from the multivariate Normal distribution model. Historical daily precipitation records from 10 Canadian meteorological stations in the Winnipeg region were utilized to evaluate the efficacy of the model. The evaluation results show that the model can reasonably preserve the statistical characteristics of the historical records while estimating the missing records at multiple sites.

Keywords—precipitation, simulation, multivariate, truncated Normal distribution

1. Introduction

Statistically, spatial-dependency can conventionally be modeled through the correlation parameter of a multivariate Normal distribution model. To acquire normalized data from skewed precipitation data, the multivariate truncated Normal distribution has been commonly used in hydrologic literature to estimate the parameters of the Normalized population of a skewed data set (Bardossy, 1992). The advantages of this approach are well recognized in weather generation (Hutchinson, 1995). To facilitate the parametric modeling of daily precipitation in this paper, the entire records are transformed into a normally distributed data set using a multivariate truncated Normal distribution approach. The parameters of the transformed Normal distribution can be estimated from the original skewed precipitation records based on the concept illustrated in Figure 1 wherein the zero values are transformed into synthetic negative values (such that the area under the probability curve corresponding to the negative values equals the probability of zero values), which in turn can be imagined to be distributed to the left hand side of the truncated Normal distribution. Such a transformation serves two purposes: [1] removal of the high frequency of zero records from the distribution, and [2] contribution of the synthetic negative data set to fill up the truncated part (left hand side) of the distribution. Thus, a complete Normal distribution can be constructed by joining the normalized above-zero records (i.e., positive side of the truncated distribution) and the transformed zero records (i.e., negative side of the truncated distribution).

Besides zero records, an ordinary precipitation data set frequently contains missing data that hinder the formation of a Normal distribution. Assessment of incomplete data sets invariably raises the level of complexity in modeling of precipitation data. Thus, this paper aims to generate daily precipitation amounts concurrently at various sites using the multivariate truncated Normal distribution approach and also illustrates an application of the model for infilling of missing data and weather generation.

2. Methodology

The fitting of a truncated Normal distribution is specifically accomplished through the following six steps (Figure 2). Steps 1-5 estimate the multivariate parameters based on the above-zero and zero records while excluding all of the missing records. Step 6 substitutes the zero and missing records based on the conditional parameters of the truncated distribution (estimated through steps 1-5) when a complete data set is required. In this step, the parameters are estimated individually for a specific month based on the observed records.

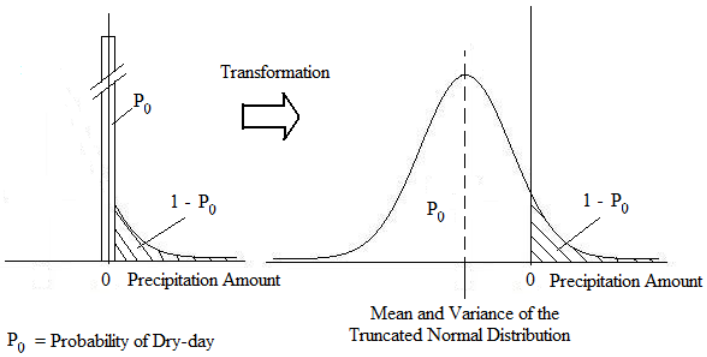

Figure 1. Transformation of a Truncated Distribution into a Normal Distribution.

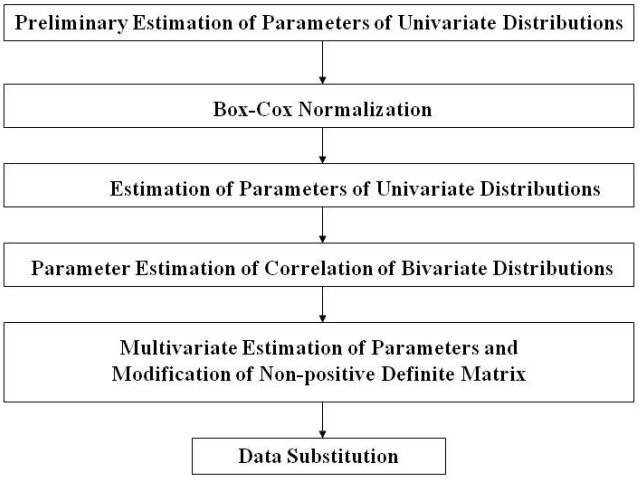

Figure 2. Steps in the Multivariate Truncated Normal Distribution Approach.

2.1 Step 1: Preliminary Estimation of Parameters of Univariate Distributions

The parameters of Normal populations from truncated samples can be estimated by maximum likelihood equations (Cohen, 1950), however, an alternative flexible approach is proposed. The parameters of a univariate distribution can be estimated by minimizing the differences between the empirical distributions and theoretical density functions. Two objective functions are formulated for the estimation of mean and variance based on the minimization of error between: [1] density of a standardized histogram of above-zero records and the corresponding pdf, and [2] P_0 (probability of dry days) and the corresponding cumulative distribution function (CDF). For the formation of the first objective function, the daily above-zero precipitation observations (**x**) of a specific month at a specific site are allocated to "n" bins of the histogram and a probabilistic histogram is obtained. Equation 1 describes the first error equation corresponding to the difference between the probability y_i of an empirical distribution and the theoretical $f(x_i)$. The pdf of a Normal distribution function is used, although Equation 1, in general is applicable to any function. The mean and variance can be estimated by minimizing the least square error:

$$Error_{pdf} = \sum_{i=1}^{n} [\, f(x_i) - y_i \,]^2 \qquad (1)$$

Alternatively, another objective function based on the CDF can be used for the estimation of parameters. For formation of the second objective function, Equation 2 describes the second error corresponding as the difference between the empirical P_0 and the theoretical F(x=0) values of a CDF. Again, the parameters can be optimized solely using Equation 2. However, to enhance the accuracy of the overall estimation, Equations 1 and 2 are integrated into Equation 3 for the estimation of parameters.

$$Error_{CDF} = [F(x=0) - P_0]^2 \qquad (2)$$

$$Error_{Total} = wkError_{pdf} + (1-w)Error_{CDF} \qquad (3)$$

Where, w refers to the assigned weight to each of the error terms and can be adjusted depending on the reliability of

the relevant data. Since both of the error terms are deemed to be equally important in this paper, w = 0.5. The k^{th} term (intervals between bins in the histogram) revises the $Error_{pdf}$ term into a comparable manner to the $Error_{CDF}$ term. The mean and variance in Equation 3 are iteratively evaluated by minimizing the $Error_{Total}$.

2.2 Step 2: Box-Cox Normalization

The estimated parameters in step 1 may be biased by the presence of non-normally distributed above-zero records. Thus, Box-Cox normalization (Box and Cox, 1964) is applied to ameliorate the fitness of a data set to a Normal distribution through exponentially reducing the magnitude of all records. It is noted that parameters of the Box-Cox normalization cannot be estimated without a priori knowledge of parameters of the underlying Normal distribution since the Normal probability plot are established based on the previously estimated parameters of distribution. The selection of Box-Cox parameters rely on the criterion of either minimizing the absolute difference or maximizing the absolute correlation between best estimates of the normal probability plot and the transformed data plot.

2.3 Step 3: Estimation of Parameters of Univariate Distributions

As above-zero records are normalized by the Box-Cox Normalization procedure; the parameters of the univariate truncated distribution have to be reevaluated. Steps 1 to 3 are conducted individually and iteratively on data sets of the remaining sites. However, the generated precipitation amounts at each site obtained from the parameters (mean and variance) of the univariate distribution do not take in to consideration the spatial-dependency (correlation) from one site to another. For this purpose, the estimation of a bivariate parameter (i.e., correlation coefficient) is required.

2.4 Step 4: Estimation of Correlation Parameter of Bivariate Distributions

The bivariate parameter is estimated through minimizing the squared errors between the empirical probability of a joint event and its corresponding theoretical total bivariate density. The joint events are considered under four categories: simultaneously observed wet-wet events in the first quadrant, dry-wet events in the second quadrant, dry-dry events in third quadrant, and wet-dry events in the fourth quadrant of a bivariate pdf. The empirical probabilities corresponding to the four quadrants are: P^B_1, P^B_2, P^B_3, and P^B_4. The theoretical total densities corresponding to the four quadrants are: F^B_1, F^B_2, F^B_3, and F^B_4. The empirical probabilities are computed from the joint observations and the theoretical total densities are computed based on the CDF of the bivariate distribution. Assume that the theoretical F^B_1 is restricted in a domain of x_1: $[0, \infty]$ and x_2: $[0, \infty]$ and is equal to the empirical P^B_1, then ρ_{12} can be estimated as other variables are estimated previously. To ameliorate the estimate of correlation being biased because the estimation solely relied on the quadrant 1, the rest of the available information (e.g., P^B_2, P^B_3, and P^B_4) is also taken into consideration. The correlation can be estimated by minimizing the squared errors in Equation 4.

$$Error^B_{CDF} = \sum_{Quadrant=1}^{4} [F^B_{Quadrant} - P^B_{Quadrant}]^2 \quad (4)$$

2.5 Step 5: Multivariate Estimation of Parameters and Modification of Non-positive Definite Matrix

The parameters of a multivariate distribution (mean vector, covariance matrix) corresponding to precipitation stations can be formulated by integrating parameters that were individually estimated in steps 3 and 4 (such as mean, variance or correlation scalars). However, since each element inside the correlation matrix is obtained individually, these elements may not concur with one another. Such a conflict leads to inconsistency of errors in the multi-dimensional generalization of variance and may result in a non-positive definite matrix. The problem of non-positive definite matrix commonly occurs when a large amount of zero records and missing data are involved in the analysis. The consequence of non-positive definite matrix is that the calculated conditional variance values and the determinant of the correlation matrix could become negative, which is statistically infeasible (i.e., negative variance). There are a few methods to handle this problem; however, the methods purports only case specific adjustment procedures (Rasmussen et al., 1996). This research adopts two notions to convert the non-positive to positive definite matrices into a positive or semi-positive definite matrix: [1] by decreasing the non-diagonal elements of the correlation matrix, and [2] by minimizing the resulting change in the correlation matrix subject to the constraint that all eigenvalues be positive.

2.6 Step 6: Data Substitution (Infilling Missing Data)

The data substitution procedures are applied to cases requiring transformation of zero and missing records into a normally distributed data set. The method of substitution relies on conditional sampling to the conditional parameters of the spatial-dependent distribution and the parameters are obtained through steps 1-5. The general form of conditional mean and covariance are presented in Equations 5 and 6 (Johnson and Wichern, 1982).

$$E(X_1 \mid X_2) = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2) \quad (5)$$

$$Cov(X_1 \mid X_2) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \quad (6)$$

Where $\mathbf{X_1}$ and $\mathbf{X_2}$ respectively refer to the conditional mean vectors that are intended to be estimated and the given conditional observations (e.g., above-zero records). $\mathbf{X}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in Equations 7 respectively refer to partitioned multivariate observations for a specific day, mean vector, and covariance matrix.

$$X = \left[\frac{X_1}{X_2}\right]; \qquad \mu = \left[\frac{\mu_1}{\mu_2}\right]; \qquad \Sigma = \left[\frac{\Sigma_{11}}{\Sigma_{21}} \middle| \frac{\Sigma_{12}}{\Sigma_{22}}\right] \quad (7)$$

Since $\mathbf{X_2}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (Equation 7) are either available or previously estimated, the conditional mean and covariance can be subsequently evaluated using Equations 5 and 6.

The zero and missing records are sequentially and individually substituted based on the available observations on the same day and the conditional parameters, $E(\mathbf{X_1}|\mathbf{X_2})$ and $Cov(\mathbf{X_1}|\mathbf{X_2})$, of the multivariate Normal distribution. First, for a specific day, site by site zero records are substituted and/or infilled through drawing sample from the conditional distribution. Only negative samples are accepted as the zero records are restricted to the replacement of synthetic negative values. Second, the missing records are substituted simultaneously through drawing multivariate samples from the conditional distribution and the previously substituted and/or infilled records of zero (negative values) are also taken into consideration as part of $\mathbf{X_2}$. The missing data can be substituted by positive or negative values. The procedure for the simulation of a multivariate Normal distribution can be found literature (Scheuer and Stoller, 1962).

3. Case Study

Daily precipitation data sets from the years 1961 to 1993 corresponding to 10 stations located in the Winnipeg region of Canada have been utilized for the analysis. Because the region of Winnipeg city continues to experience recurring flooding events, a number of meteorological stations have been established. The availability of sufficient and reliable daily precipitation data is one of the reasons for the selection of this region. The total data used for the evaluation of parameters of each month is approximately 900 days (30 years * 30 days). Although the data size is large, approximately 73.5% of the records are zero and/or missing records. Precipitation records were first divided into a training data set from January 1, 1961 to December 31, 1990 and a testing data set from January 1, 1991 to December 31, 1993. The training data set was further separated into three streams of data (i.e., untransformed Control-1, untransformed Control-2, and transformed) in Figure 3. The untransformed data sets did not involve the parameters estimation using the multivariate truncated Normal distribution, except data sets were normalized using Box-Cox Normalization. The parameters of Control-1 and Control-2 can be directly calculated from the data sets. The parameters of these three data sets were subsequently used in the analysis of the two cases studies.

3.1 Case Study 1 (Infilling of Missing Data)

For case study 1, the testing data set was chosen such that it included a missing data period. A total number of 1095 daily precipitation records (3 years * 365 days) per station were organized in a manner as shown in Figure 4. In this figure, rows and columns respectively correspond to stations (1 to 10) and daily precipitation records (1 to 1095). For operational purposes, daily precipitation records (station-1 through station-10 in a column) are assumed to be missing at one station at a time. For example, element-1 in column-1 is assumed to be missing for infilling purpose. This process is continued until all elements of column-1 were assumed to be missing and subsequently infilled in successive order. A similar process was followed for column-2 and the remainder of the columns. The infilling procedure is the same as described in step-6 and the parameters estimated from the training data sets are used.

3.2 Case Study 2 (Generation of Data)

The parameters obtained from the training data set were used to generate a total number of 27000 realizations. The 27000 realizations represent 900 times of 30 days in a month. Such a large size of data set is considered statistically sufficient and is expected to provide reasonable estimates for the evaluation measures used in this study.

4. Results and Discussion

For ease of illustration, the results and discussion of the two cases studies are presented and divided into three portions: [1] parameters estimation of the mean, variance, and correlation of the multivariate truncated distribution

approach using the training data set from the years 1961 to 1990, [2] infilling of "presumed to be missing data" from the years 1991 to 1993, and [3] generation of data.

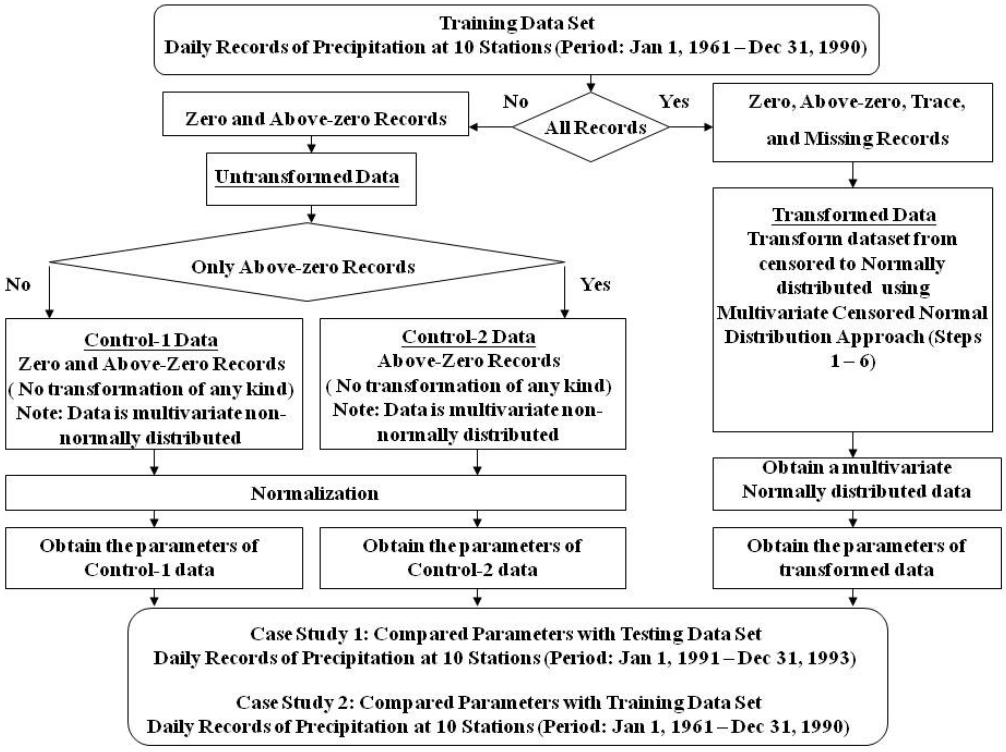

Figure 3. Schematic Representation of the Transformed and Untransformed Data Sets.

$$Daily\ Precipitation\ Amounts\ (p_{station,day}) = \begin{bmatrix} \textcircled{p_{1,1}} & p_{1,2} & \cdots & p_{1,1095} \\ \boxed{p_{2,1}} & p_{2,2} & \cdots & p_{2,1095} \\ \boxed{\vdots} & \vdots & \ddots & \vdots \\ \boxed{p_{10,1}} & p_{10,2} & \cdots & p_{10,1095} \end{bmatrix}$$

◯ Missing Data ($\mathbf{X_1}$); ▭ Conditional Observed Data ($\mathbf{X_2}$)

Figure 4. Schematic Table for Infilling of Missing Daily Precipitation Records.

4.1 Parameters Estimation Using the Multivariate Truncated Distribution Approach

The training data set of daily precipitation records (above-zero and the number of zero records) at each station corresponding to each month was used for the estimation of parameters of the respective probability distribution.

Step 1: The estimates of mean and variance were found to converge asymptotically before 400 iterations as shown in Figures 5 and 6. Such a convergence suggested that the differences in parameters between the theoretical density functions and the standardized histograms were minimized.

Step 2: An improvement in normality was found after the use of the Box-Cox transformation on the observed data.

Step 3: Figure 7 illustrates that the pdf at each station adequately describes the properties of the observed histogram.

Step 4: To express spatial association between any two sites, bivariate relationships were developed in the form of bivariate distributions. For example, a bivariate distribution (Figure 8) between stations 1 and 2 in the month of January shows that contours of the estimated bivariate distribution reasonably describe the distribution of above-zero observed data. Table 1 summarizes the differences between the observed and estimated probability densities occupied in each of the quadrants of a bivariate distribution. Each number in the table denotes the average differences (%) between the observed and estimated densities of the four quadrants of a bivariate probability density

distribution. The average difference is approximately 8.58% in the month of January. Therefore, the estimated bivariate distribution depicts: [1] a reasonable fit of the marginal distributions of stations 1 and 2 as demonstrated by the presence of minor differences between the histogram and theoretical pdf in Figure 7, [2] a reasonable match in above-zero bivariate distribution which is illustrated by the closeness between the bivariate above-zero data and the theoretical contours at the right upper corner for a joint distribution in Figure 8, and [3] a reasonable match also exists in terms of bivariate probability distributions occupied in each quadrant of the bivariate distribution (Table 1).
Step 5: Correlation matrices corresponding to all months were found to be a non-positive definite; therefore, each matrix was individually modified according to the procedure described in step 5. The changes (%) in elements of various correlation matrices to transform them into positive definite matrices are summarized in Table 2. The averaged out minimum and maximum changes of elements inside the correlation matrix are 0.60% and 13.51% for the monthly group of April and November respectively. The average change is 4.75%.

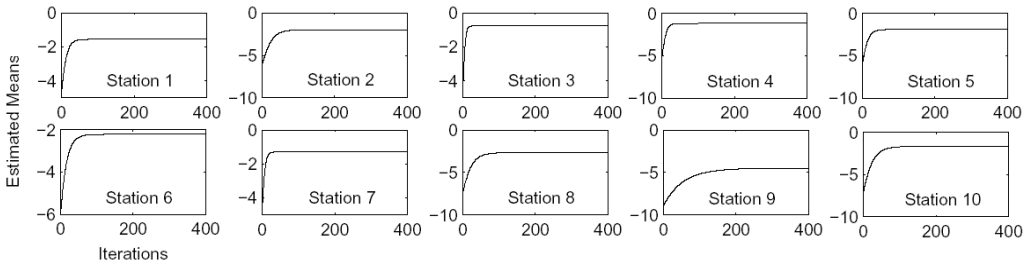

Figure 5. Convergence of Estimates of Mean Based on Truncated Univariate Gaussian Distribution.

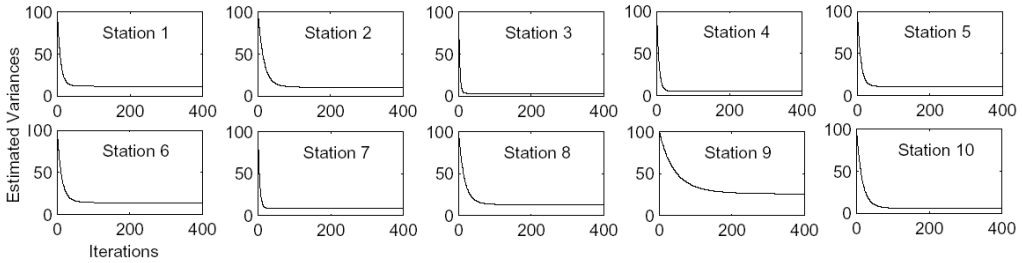

Figure 6. Convergence of Estimates of Variance Based on Truncated Univariate Gaussian Distribution.

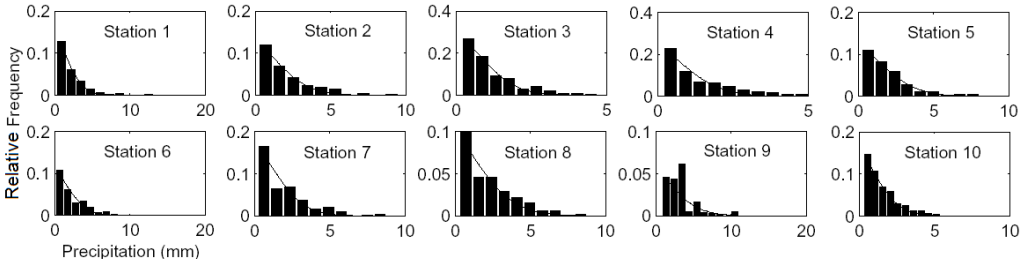

Figure 7. Empirical (Normalized) and Estimated Distributions for the Month of January.

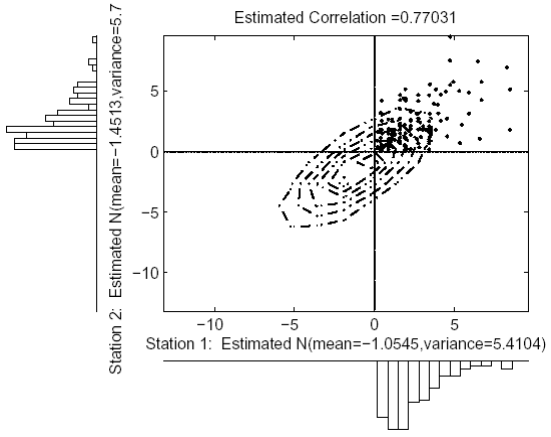

Figure 8. A Plot Showing Bivariate Distribution between Station 1 and 2 for the Month of January.

4.2 Case Study 1: Infilling of Missing Data

Missing data as described above in the transformed and untransformed data cases were infilled using their respective parameters obtained from the training data sets. The results thus obtained in each case were evaluated to assess the efficacy of the procedure for infilling of missing data. The infilled data corresponding to these cases were employed for the calculation of the mean squared errors (MSE) in respect to five statistical descriptions (e.g., the mean, mode, median, variance, and percentage of wet day) and are summarized in Table 3. From the table, the MSEs corresponding to the transformed data (case-1) were generally a close match with the observation compared with the untransformed data of Control-1 and Control-2.

A visual inspection of Figure 9a illustrates that the spike pattern of the historical data is approximately infilled by the data substitution in the case of the transformed data. In general, the infilled observations by the data substitution for case of the transformed data reasonably mimic the corresponding historical observations in the entire three years. Figure 9b shows an enlarged view of Figure 9a, where several spikes between Julian days 20-120 were infilled reasonably well.

4.3 Case Study 2: Generation of Data

The results summarized in Table 4 indicate that the values of MSE in the majority of the statistical descriptions were found to be lower for the transformed data when compared to those of the untransformed Control-1 and 2 data sets. The value of MSE related to the mode of Control-1 and the value of MSE related to the variance of Control-2 were found to be less than their corresponding values of MSE in the transformed data. Besides these two special cases, in general the statistical characteristics of the transformed and observed data were found to be reasonably replicated.

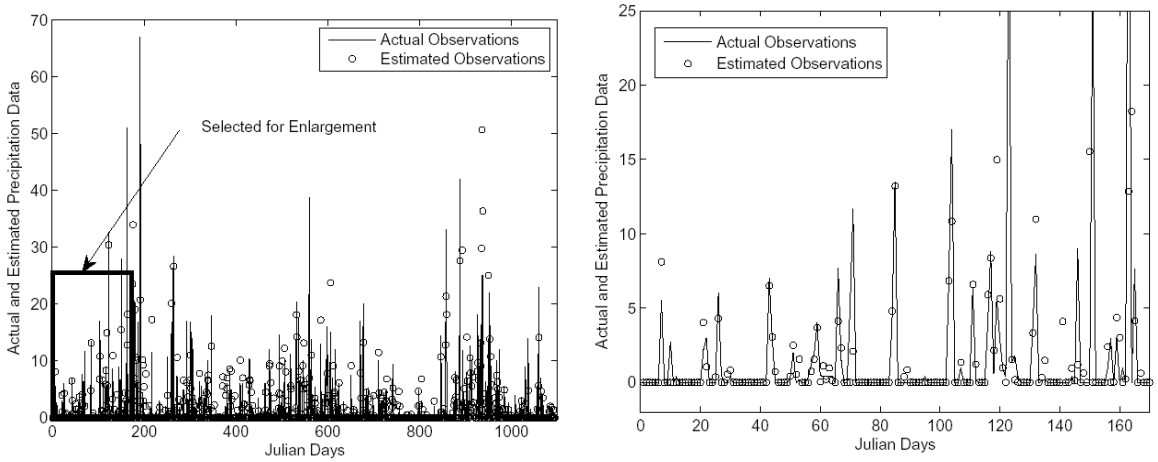

Figure 9a. A Plot of Infilling Missing Data at Station 1. Figure 9b. An Enlargement Block.

Table 1. Summary of Average Differences (%) between Monthly Estimated and Observed Densities to Each of the Quadrants of the Bivariate Probability Density Distributions for the Month of January.

Station	2	3	4	5	6	7	8	9	10
1	7.58	5.84	5.84	8.34	10.95	6.17	8.03	19.01	6.52
2	-	7.54	7.54	8.28	10.58	7.54	7.80	19.01	7.63
3	-	-	4.33	8.35	10.96	6.17	8.04	20.00	6.53
4	-	-	-	8.35	10.96	6.18	8.03	19.92	6.53
5	-	-	-	-	11.13	8.35	8.73	19.67	8.45
6	-	-	-	-	-	10.96	11.13	19.01	11.06
7	-	-	-	-	-	-	8.04	19.59	6.53
8	-	-	-	-	-	-	-	19.01	8.09
9	-	-	-	-	-	-	-	-	19.28

Table 2. Summary of Average Correlation Change (%) in Monthly Correlation-matrices.

Month (Season)	1	2	3	4	5	6
Change (%)	5.10	4.00	6.50	0.60	3.97	2.92
Month (Season)	7	8	9	10	11	12
Change (%)	2.26	3.95	4.62	3.84	13.51	5.73

Table 3. Summary of the mean squared error (MSE) with respect to the five statistical criteria for Infilling of Missing Data

MSE	Mean	Mode	Median	Variance	Probability of Wet day
Control 1	9.24	0.75	5.40	3127.17	0.33
Control 2	7.56	77.42	31.91	3244.07	0.41
Transformed Data	2.01	0.89	1.09	2754.87	0.0044

Table 4. Summary of the (MSE) with respect to the five statistical parameters of the generated data.

MSE	Mean	Mode	Median	Variance	Probability of Wet day
Control 1	1.87	2.16	1.99	2645.75	0.13
Control 2	46.85	9.09	74.73	665.15	0.36
Transformed Data	0.49	2.37	0.92	1594.13	0.02

5. Conclusions

For considerations of the spatio-dependency, the parametric approach with multivariate truncated Normal distribution is capable of providing a flexible yet robust approach of generation of precipitation amounts at multiple sites. The non-Normal precipitation records can be transformed into normally distributed records via the use of the multivariate truncated Normal distribution. The paper discusses a simple and general approach for the estimation of parameters. Such an approach in association with a data substitution procedure was found to be capable of infilling missing data and also generating reliable synthetic realizations at multiple sites.

Several advantages of the method are: [1] it facilitates the availability of a Normally distributed data set thus making further statistical analysis possible and tractable, such as time series analysis using autoregressive model, [2] its complex mathematical operations (e.g., derivative for maximizing the likelihood function with multivariate truncated distributed data set) for the parameters estimation can be substituted by a simple estimation method that minimizes the differences between the empirical distribution and probability density function (Steps 4), [3] its method of estimation of parameters is comparatively flexible and does not restrict the application of the Normal distribution function and, [4] since the estimation of its parameters of the multivariate distribution is primarily conducted by dealing with univariate cases one by one in terms of computing mean, variance and bivariate correlation, the model requires significantly less computational resources to handle a large-scale multivariate problem. The involvement of a large number of stations is expected to proportionally increase the time of operation and yet the calculation can still be performed on an ordinary computer system.

Acknowledgement

The partial financial support of the Natural and Engineering Research Council of Canada is gratefully for this project greatfully acknowledged.

References

Bardossy, A. (1992). "Space-time Model for Daily Rainfall Using Atmospheric Circulation Patterns," Water Resources Research 28(5):1247-1259.

Box C. E. and Cox D. R. (1964). "An analysis of transformations," Journal of the Royal Statistical Society Series B 26: 211-246.

Cohen, A. C. (1950), "Estimating the Mean and Variance of Normal Populations from Singly Truncated and Doubly Truncated Samples," The Annals of Mathematical Statistics 21(4): 557-569.

Hutchinson M. F. (1995), "Stochastic Space-time Weather Models from Ground-based Data," Agricultural and Forest Meteorology 73: 237-264.

Johnson, R.A. and Wichern, D.W. (1982), Applied Multivariate Statistical Analysis, Prentice hall, New Jersey, USA.

Rasmussen, P. F., Salas, J. D., Fagherazzi, L., Rassam, J., and Bobee, B. (1996) "Estimation and Validation of Contemporaneous PARMA Models for Streamflow Simulation," Water Resources Research 32(10): 3151-3460.

Scheuer, E. M. and Stoller, D. S. (1962), "On the Generation of Normal Random Vectors," Ecological Modelling 4(2): 278-281.