A Stochastic Daily Weather Generation Model at Multiple Sites

by

Wai Wah Ng

A Thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfillment of the requirements of the degree of

DOCTOR OF PHILOSOPHY

Department of Civil Engineering
University of Manitoba
Winnipeg

Contents

List of Tables

List of Figures

ABSTRACT

Stochastic generation of daily precipitation at multiple sites is frequently needed to evaluate the long-term effects of hydrologic and climate-change in design and operation of water resources systems. Capturing the spatial dependence of precipitation at multiple sites into a stochastic model presents a great challenge because of the non-normal bivariate distributions of precipitation-amounts. Without normalizing the precipitation amounts, many models have attempted to establish spatial dependence through alternative methods that tended to be cumbersome and wieldy. In contrast, representing precipitation in Gaussian fields provides a generic structure that is well-amenable to statistical analyses facilitating easy implementation of models. The thrust of this thesis is to generate normalized precipitation data and transform them back into the original domain for applications and analyses.

A multivariate censored distribution (MCD) and a multivariate autoregressive censored process (MACP) are developed to formulate two weather generation (WG) models. Parameters of censored distributions were estimated by using the maximum likelihood method. To reduce the magnanimity in the number of parameters and their temporal variation, elements of covariance matrices of models were represented by periodic functions.

The performance of models was evaluated by comparing discrepancies in attributes. Three performance measures (i.e., the coefficient of determination, the coefficient efficiency and the root mean square error) suggested that simulated data to be indistinguishable from the historical precipitation sequences. The models were implemented with other techniques to address the three most common problems encountered in daily precipitation records.

The first implementation is related to simulation of precipitation at un-gauged sites using the WG-MACP model with general regression neural networks or Kriging methods. Simulations of these methods were conducted by projecting the output of the WG-MACP model or by expanding the dimension of the WG-MACP model. Simulations of these methods were evaluated by the K-fold cross-validation technique. These methods were found to perform satisfactorily in the generation of precipitation at un-gauged sites.

The second implementation was related to infilling of missing observations a using the WG-MCD and WG-MACP models with Gibbs sampling. Attributes obtained before and after data infilling were compared. Infilling with the WG-MCD model was found to be superior in preserving the mean and covariance of the historical precipitation. Infilling with the WG-MACP model was found less effective in preserving the temporal dependence in the historical precipitation sequences.

The third implementation was related to downscaling of monthly and daily output of the Canadian regional climate model (CRCM) using traditional and parametric Delta change methods. These methods modified the output and parameters of the WG-MACP model to capture the future changes in statistics as reflected in the CRCM data. The Delta change method was found to adequately downscale the future change in monthly precipitation from a synoptic to local scale. For the future climate change study on daily precipitation, the parameters of WG-MACP model were found to increase. The results indicate likelihood of frequent occurrences of wet days coupled with deluge of extreme precipitation in the future.

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my advisor, Professor Umed Panu, who provided me with many constructive suggestions, constant support, and heartfelt encouragement during these years. Without his guidance and persistent help, this thesis would not have been possible.

I would also like to thank my co-advisor, Professor Peter Rasmussen, who shared with me his expert knowledge and insights to my research area, and provided me with many useful guidance through out my graduate education.

My appreciation also extends to my examining committee. I am thankful to Professor Ahmed Shalaby's insightful critique and kind support through out these years. I am highly appreciated for his generous help. Thanks also go to Professor Brad Johnson, who generously supplied me with his professional opinion and suggestion for the improvement of the thesis which gave me a better perspective on my own research.

In addition, a big thank to Dr. Tribeni Sharma for his support and insightful critique that helped me to make a significant improvement to my thesis.

Last but not the least, I am grateful to my parents for their patience and love. I also wish to thank to my sibling: Kim, Sharon, Carmen, Alan, Polly, and Linda for their constant help and support. Especially, I owe Carmen a big thank for her endless support and encouragement.

1. INTRODUCTION

1.1 Background

Significant changes in weather patterns have been observed throughout the world over the last few decades. The changes in weather patterns, in terms of the frequency and intensity of its elements (e.g., precipitation, maximum/minimum temperature, solar radiation, humidity, or wind speed etc.) have caused damages to physical infrastructure worth billions of dollars and immeasurable adverse impacts on the global environment. The changes in weather patterns are ramified in climate change, which are attributed to have caused the deaths of nearly 400,000 people a year and cost the world more than $1.2 trillion, equivalent to 1.6% of the annual global GDP [*DARA and the Climate Vulnerable Forum*, 2012]. To address the problem of climate change and its impacts, the Intergovernmental Panel on Climate Change (IPCC) was established in 1988. The wide ranging probable impacts of climate change are well documented in the IPCC reports [*IPCC*, 2007]. To conduct a thorough impact assessment of climate change, the need for comprehensive weather records is crucial. However, weather records often have insufficient length for the desired analysis and are not always available for specific sites due to the lack of weather monitoring stations.

Weather generation is employed when the historical records are insufficient for the desired analysis. The main purpose of weather generation is to synthesize sequences

of elements (e.g., precipitation, maximum/minimum temperature, solar radiation, humidity, or wind speed) of weather that are statistically indistinguishable from the observed records. Although all elements of weather stated above are important, precipitation assumes a special significance in view of its crucial role in hydrological, agricultural, and water resources systems analysis and design. Precipitation sequences, on a daily time scale, are likely to be desirable from the consideration of weather analysis, prediction and forecasting, while being central elements in hydrological activities and analyses. Furthermore, daily data values can be suitably aggregated into weekly, monthly, seasonal and annual data values for necessary applications.

For example, the generated precipitation data commonly serves as an input to the hydrological systems that are of importance in local precipitation-runoff studies, the modeling of weather systems, and impact assessment of scenarios resulting due to the change in components of the hydrologic cycle. The results of these analyses have direct impacts on the design of water resources systems, such as those related to drinking water supply, hydropower generation, irrigation, rural/urban drainage, and flood mitigation.

Precipitation is a stochastic process in the sense that its occurrence in time and space domains cannot be predicted with certainty using a deterministic formulation. Its modeling therefore requires stochastic considerations in the development of suitable models for data generation (i.e., data simulation), data forecasting, data infilling, etc. The challenges in stochastic modeling of daily precipitation sequences are enhanced due to three primary factors: [1] the records form intermittent time series; [2] the precipitation-amounts tend to be highly skewed and must be modeled by mixed probability distributions, i.e., distribution involving a discrete probability mass at

zero plus a continuous part on the positive axis; and [3] the records frequently contain extreme and/or missing observations that tend to increase the difficulty in fitting a probability distribution function.

In view of the importance and complexity in modeling of daily precipitation, this thesis focuses on the stochastic generation of daily precipitation. Generated precipitation at a weekly, monthly, or annual basis can subsequently be derived through aggregation of daily precipitation, if required.

1.1.1 Precipitation Generation at a Single Site

To generate a daily precipitation sequence at a single site, the most common approach, as stated earlier, separates the modeling tasks into two stages [*Rolda'n and Woolhiser*, 1982; *Woolhiser and Rolda'n*, 1982]. In the first stage, the model simulates a sequence of daily precipitation occurrences as a binary sequence consisting of digits 0 and 1. For the purpose, Markov chain processes have been extensively invoked for the generation of precipitation occurrences. In the second stage, for each wet-day occurrence, a precipitation amount will be assigned by drawing a sample from a selected probability distribution fitted to historical precipitation amounts.

The two-stage approach intends to generate a series of daily precipitation records that preserves the statistical characteristics corresponding to the empirical probability distribution of precipitation amounts and the temporal dependence of precipitation occurrences in a historically observed precipitation series. This modeling approach has been found to be adequate for the generation of daily precipitation at a single site [*Richardson and Wright*, 1984]. Occasionally, single-site models can be adopted for the generation of precipitation at multiple sites when sites of interest are either located

close to each other (i.e., dependence case) or far apart (i.e., independence case). In dependence case, the precipitation amounts observed at closely located sites are highly spatially dependent and precipitation generation at multiple sites can be deemed as a single-site problem. In independence case, the precipitation amounts observed at sites far from each other may be considered spatially independent and precipitation generation at multiple sites can be conducted through several site-specific individual single-site models. In the case, where sites do not fall into these two extremes, the use of single-site models for precipitation generation at multiple sites is inadequate because it disregards the existence of the spatial dependence in the system.

The 3-D plot in Figure 1.1 shows that it is not uncommon to observe precipitation amounts at multiple sites that are neither completely dependent nor completely independent. For example, on Julian day 246 in the year 1971, ten climatic stations in Manitoba have been selected for illustration and additional detail on these stations is presented in Chapter 4. As shown in the figure, the change in precipitation amounts between sites close together (e.g., coordinates (96.6, 49.2) and (98.9, 49.4)) do not have to be gradual because of spatial intermittency [*Bardossy*, 1992]. In Figure 1.2, a traditional two-stage single-site model using Markov chain and Gamma distribution has been adopted for illustration. It is apparent from the figure that the probabilities of simultaneous occurrences of wet-wet cases at the two selected stations have not been properly preserved in precipitation simulations of the two independent single-site models. Clearly, the use of independent single-site models does not preserve the spatial dependence/intermittency among any two stations.

Over the last two decades, an increasing amount of literature related to weather generation at multiple sites has emerged [*Srikanthan and McMahon*, 2001], however,

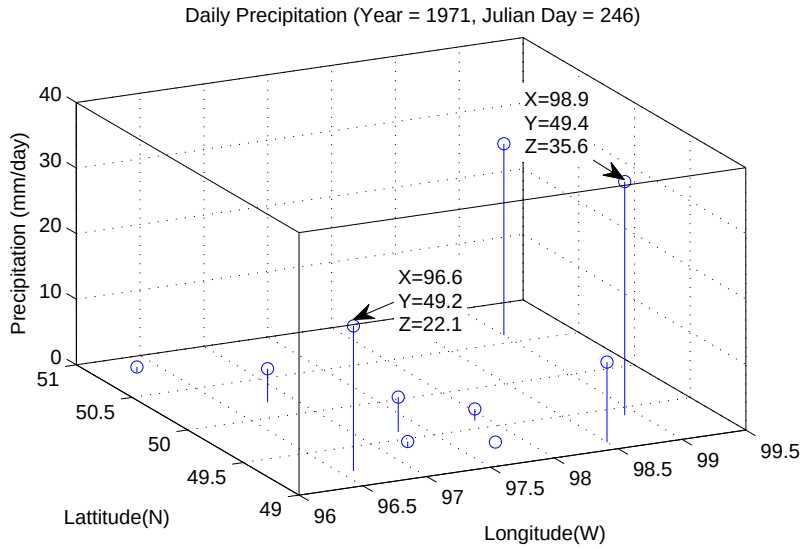


Fig. 1.1: Illustration of spatial variability.

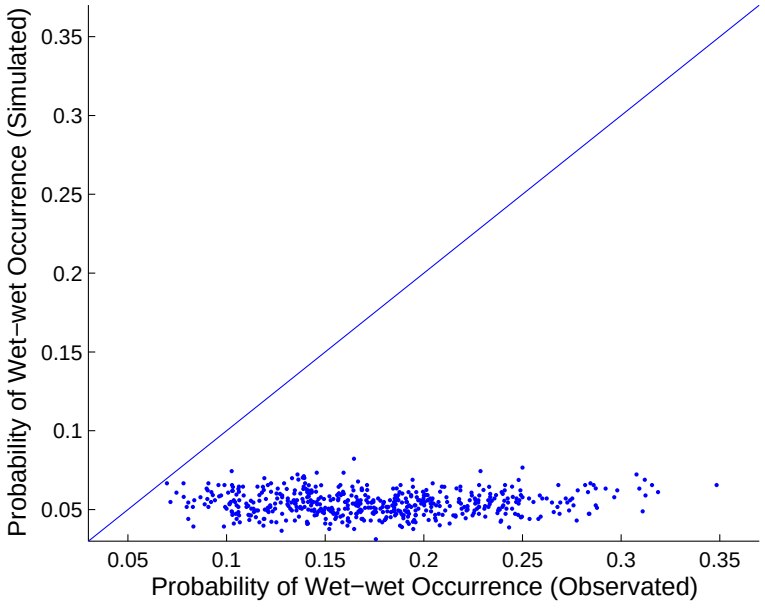


Fig. 1.2: Probability of wet-wet simultaneously occurrences of two stations between observations of historical precipitation and simulations of single-site model.

many weather generation models do not take into account the spatial correlation of the weather elements [*Baigorria and Jones*, 2010]. Research in this area is still continuing.

1.1.2 Precipitation Generation at Multiple Sites

Precipitation modeling at multiple sites is a challenging task due to difficulties in modeling the spatial dependence of precipitation amounts that have mixed distributions (i.e., consisting of discrete zero and distributed as continuous function for above-zero records). It is inappropriate to directly take the mixed distributed data for the estimation of cross-correlations among multiple-site. Therefore, many multiple-site models focus on developing alternative or indirect ways to preserve the spatial dependence in estimated and/or generated precipitation records. These models are typically classified as parametric, non-parametric, and hybrids, depending on the type of techniques involved.

For parametric modeling, many models [*Brissette et al.*, 2007; *Srikanthan and Pegram*, 2009] have completely or partially been developed using the concept *Wilks* [1998], where a single-site model based on the two-stage approach is extended to a multiple-site model. The Wilks model preserves temporal dependence through the assumption of a Markovian dependence structure. The spatial dependence is simulated based on the conditional probability distribution of uniform random variates for the generation of precipitation occurrences. Such a model has been found to have an overall advantage in offering a mechanism for modeling with varied orders of temporal dependence at each site and is capable of preserving and/or reproducing many important statistical properties of precipitation at multiple-site [*Mehrotra et al.*, 2006]. A

drawback of the model is that it does not effectively deal with the spatial variability in the generated precipitation fields (i.e., spatial intermittency) [*Breinl et al.*, 2013]. Also, the formation of empirically derived curves for each pair of sites relating correlations of random numbers to the observed correlated precipitation events is difficult to automate [*Baigorria and Jones*, 2010].

For the non-parametric modeling, many models have been developed based on the k-nearest-neighbor (k-NN) sampling method (e.g., [*Apipattanavis et al.*, 2007]) that simulates spatial dependence by simultaneously generating the precipitation occurrences at all sites. Temporal dependence in this type of modeling is still simulated through a Markov model. The advantage of non-parametric modeling is that it does not require the rather restrictive assumptions regarding the probability distributions and the correlation structure. Also, such a model can easily be extended to multiple-site for precipitation simulations [*Buishand and Brandsma*, 2001]. A drawback of the model is that it cannot reproduce precipitation amounts beyond those present in the data sequences used as base material [*Leander and Buishand*, 2009].

Despite the recent progress of multiple-site modeling, in general, the traditional two-stage approach still has several shortcomings. First, the Markov model tends to underestimate the spatial dependence of seasonal precipitation totals [*Zheng et al.*, 2010]. Second, the two-stage approach separates the modeling process into two components that theoretically increases the number of parameters involved and the complexity in modeling. Third, since the spatial dependence of precipitation is commonly established by alternative or indirect methods without directly estimating the correlation and normalizing the data, such models may be restricted to specific applications.

On the other hand, a model consisting of a generic structure (e.g., normally distributed data sets and correlation) is always well-adapted and preferable in statistical analyses because it facilitates new applications of the model.

Transformations of precipitation data into a multivariate normal distribution have been considered for the generation of precipitation at multiple sites [*Bardossy*, 1992]. The model is well-suited for the generation of spatio-temporally dependent precipitation using a multivariate autoregressive model that is constructed based on the multivariate normal distribution. The multivariate autoregressive model is relatively parsimonious, concise in its analytical form, and simple in operation. With such an integrative structure that avoids separating the simulation task into two stages; simulation can be conducted simultaneously at multiple sites and the aspects of spatio-temporal dependence can be jointly addressed. Such a model, with a generic structure, is also well-suited for subsequent statistical analyses and further applications. The model was found to be sophisticated and performed well; however, the procedures for parameter estimation were relatively cumbersome and complicated, which tends to discourage usage of the model. Simplifying the procedures for parameter estimation is an appealing idea for desirable improvements in the model.

1.1.3 Applications of Precipitation Generation Models

Over the last few decades, stochastic precipitation models have gained wider acceptance and use. Generation models are not restricted to the simulation of a long series of precipitation but are also used in a broad spectrum of applications. This is another important motivation to develop a model with a generic structure that facilitates further implementations. Common uses of stochastic precipitation models include the

simulation of precipitation at ungauged sites, the infilling of missing precipitation records, and the downscaling of regional climate change scenarios [*Wilks and Wilby*, 1999], each of which is briefly described as follows.

Simulation at Ungauged Sites

Precipitation generation models typically simulate data at the gauged sites where historical observations are available. The generated precipitation from the model can be applied to a certain region around the gauged-site where the weather patterns are assumed to be consistently similar. In practice, weather patterns are rarely spatially homogeneous and the generated precipitation may sometimes be inappropriate for applications at sites located even a few kilometers away from the gauged-site. To generate precipitation data at ungauged sites, spatial interpolation of information through weighted regression is commonly adopted. The method transfers attribute information on the desired variate (i.e.. daily precipitation) from the gauged sites with known values to the ungauged sites with unknown values. For spatial interpolations, the weighted regression method is efficient but may suffer from sensitivity problems caused by outliers; the existence of outliers is not uncommon in observed daily precipitation records. A non-parametric method may offer a robust alternative for the interpolation of daily precipitation data, especially in the presence of outliers.

In general, spatial interpolation methods are typically based on the assumption of spatial stationarity in the region of interest and the variability in orography at ungauged sites are rarely addressed. Precipitation generation at ungauged sites based on such an assumption will yield output that ignores spatial variability in data sets.

Development of a method capable of embedding locally varying parameters and uncertainty characterization is crucial to reflect the spatial variability among ungauged sites. Recently, a method based on the assumption of spatial non-stationarity among ungauged sites utilizing the Kriging technique has been developed [*Kleiber et al.*, 2012] for precipitation simulations. An advantage of this method is that the uncertainty at ungauged sites can be immediately quantified and spatially varying precipitation characteristics can be preserved. The method can capture fairly well the statistical properties of precipitation at a wide range of time scales including daily, monthly, and annually. Although, a Kriging method has been applied in conjunction with a two-stage approach, it has not yet been applied to an integrative model such as the multivariate autoregressive model. It would be desirable and advantageous to further explore and evaluate the applicability and performance of the Kriging method with other precipitation generation models.

Infilling of Missing Precipitation Records

The adequacy of generated precipitation relies on the generation model that was established based on the observed precipitation from gauged sites. Precipitation records obtained from gauged sites, however, frequently suffer from missing observations in data sets that increase the difficulty and uncertainty in stochastic modeling. Omitting the missing observations in analyses reduces the size of the data sample and also hinders the full utilization of the available information. Infilling of missing observations is one of the common approaches to handling missing observations [*Little and Rubin*, 1987].

Many infilling methods are parametric because infilling of missing observations

relies on a predetermined assumption on the relationship between the missing and observed data. The relationship can be statistically established through considerations of a probability distribution, spatial dependence, temporal dependence, or other forms of dependence. Missing observations can be infilled in a number of ways such as through sampling from a conditional distribution, prediction using correlated time series at nearby locations, forecasting using auto-correlated time series, and prediction using other variables. There is no single best infilling method that can handle all sorts of problems due to missing data, but methods that can establish a higher correspondence between the missing and observed data are preferable.

Missing data problems are common in many studies related to modeling of daily precipitation. For the infilling of missing precipitation records, methods based on assumptions related to the temporal and/or spatial dependence are normally considered [*Thyer and Kuczera*, 2003b]. The use of temporal dependence for infilling of missing observations in a precipitation series is uncommon because temporal dependence normally suffers from the adverse effect of temporal intermittence that in general, is more apparent than the adverse impact induced by the spatial intermittence. However, completely neglecting the effect of temporal dependence related to the infilling of missing observations may be unjustified. It would be prudent to further explore the possibility of infilling missing observations that preserves comprehensive statistics of historical precipitation while taking into consideration such aspects as probability distribution and spatio-temporal dependence in the historical observations. Methods of infilling missing observations capable of preserving the spatio-temporal dependence of observed precipitation are in need of further investigations.

Based on the multivariate normal distribution, the precipitation model proposed

in this thesis can serve as a conditional distribution from which estimates of missing records can be obtained. Conditional sampling from a normal distribution is a good example of the advantage of a generation model with a generic structure that facilitates use in specific applications.

Downscaling of Regional Climate Change Scenarios

The impact of climate change during the last few decades has drawn a lot of attention at all governmental levels (federal, provincial, and municipal) due to the significant damages caused by the climate change. It has resulted in the establishment of the Intergovernmental Panel on Climate Change (IPCC) in 1988. Thousands of scientists and experts participating in IPCC research provide comprehensive scientific assessments of technical information about the risks of climate change. Many physically-based regional climate models (RCMs) have subsequently been developed and the outputs of these models forecast the future change in climate patterns. The forecast outputs are normally reported in a low spatial resolution (typically in the order of 50 kilometers grid pattern) that may not be practical and useful in hydrological studies at local levels. To resolve the important information into a sub-grid scale, statistical downscaling is required to simulate the local weather conditions in greater detail using the output of RCM.

Statistical downscaling techniques have been categorized as: pattern classification methods, regression methods, and stochastic weather generators [*Wilby et al.*, 2004]. Many of the downscaling approaches require extensive effort in the estimation of parameters and statistical verifications which may limit their use [*Wilks*, 1998]. Alternatively, a relatively simple downscaling method, the Delta change method [*Hay*

et al., 2000], has often been adopted. The main advantage of the Delta change method is its simplicity.

The Delta change method modifies the observed baseline data corresponding to the future changes in statistics, typically monthly mean and standard deviation, as reflected in the output of RCM. Historical monthly precipitation records are normally adopted as the baseline data for modification. The Delta method involves shifting and amplifying the historical data. Taking the modified historical precipitation as a future precipitation scenario implies that the future precipitation has a similar precipitation pattern as the historical precipitation. Instead of taking the historical precipitation as the baseline data, a more sensible approach would be to take the output of a precipitation generation model as the baseline data. This provides the variation of precipitation pattern at future stage but the monthly statistics of historical precipitation are preserved up to a certain degree.

Many downscaling methods are capable of reflecting the future changes of RCMs in terms of their monthly statistics (i.e., monthly mean, monthly standard deviation). However, reflecting the future changes with respect to the daily statistics, such as the probability of wet-day occurrences and probability distribution of daily precipitation amounts, can be challenging. Furthermore, the future changes in daily statistics with respect to the spatio-temporal dependencies multiple sites of interest are seldom investigated in the literature. Conducting a comprehensive study in relation to the future change of daily statistics of precipitation data based on the output of RCMs is required.

1.2 Objectives of the Research

In view of the above challenges in modeling of daily precipitation, the present study is aimed at achieving the following objectives.

- To develop a simplified, integrative stochastic model based on the assumption of a multivariate normal distribution for the generation of daily precipitation that preserves the statistical characteristics related to the probability distribution and spatio-temporal dependence of the historical precipitation at multiple sites.

- To develop and evaluate methods for the simulation of precipitation at ungauged sites based on the assumptions of spatial stationarity and non-stationarity of a domain. The methods are developed utilizing the distribution corresponding to the proposed model as a core structure with the approaches of non-parametric regression and Kriging for the simulation at ungauged sites.

- To develop and evaluate methods for the infilling of missing precipitation records that preserves the spatial or spatio-temporal dependence of historical precipitation records. The methods are developed utilizing the component of the proposed model as a conditional distribution for the sampling of estimates of the missing observations.

- To develop and evaluate methods for downscaling of regional climate scenarios of precipitation that reflects the change of monthly and daily statistics of the RCM data. The methods are developed utilizing the proposed model as a core structure with the approach of Delta change for downscaling of the RCM data.

1.3 Structure of the Thesis

This thesis is divided into eight chapters. In the second chapter, the scope of the research, existing approaches to the problems and focuses of the research will be further clarified. The development of the proposed model and corresponding applications will be briefly presented. In the third and fourth chapters, the development of the proposed models will be discussed in detail. Parameters of the proposed model will be estimated and evaluated for ten climatic stations located in Manitoba, Canada. The evaluation measures used in this study will be presented. The capability of the model in preserving different statistical characteristics will be examined in detail. In the fifth chapter, the focus is on the first application of the proposed model that uses non-parametric and Kriging approaches for the generation of precipitation at multiple ungauged sites. Efficacy of the approaches will be examined with the same data sets. Results are presented at the end of the chapter. In the sixth chapter, the focus is on the second application of the proposed model that uses a sampling method for the infilling of missing observations that preserves the spatio-temporal dependence of the historical observations. Again, the method is further discussed under the "Methodology" subsection. The performance of the method is evaluated based on a comparison of statistical characteristics between the incomplete and complete data set. In the seventh chapter, the focus is on the third application of the proposed model, where it is used as a core structure for a downscaling methodology to assess the future impact on the local precipitation due to climate change. The method is discussed under the "Methodology" subsection. The changes in statistics before and after downscaling will be evaluated and discussed. The last chapter of this thesis summarizes and discusses the overall findings highlighting the particular contributions of the thesis and

making recommendations for future research.

2. A REVIEW OF STOCHASTIC PRECIPITATION GENERATION MODELS

Precipitation form and quantity are influenced by climatic factors such as wind, temperature, and atmospheric pressure, and falls mainly in the form of drizzle, rain, sleet, snow, and hail. Precipitation is measured by rain/snow gauge networks and is the most important input for meteorological and hydrological analyses. In hydrology, daily precipitation denotes the amount of precipitation in depth units measured over a 24-hour period. A day with a precipitation amount less than 0.25 mm is typically regarded as a dry day [*Yang et al.*, 1998]. A succession of wet/dry-days can be viewed as a binary sequence consisting of digits "1" and "0" signifying the wet days and dry days respectively.

Records of daily precipitation are probably the most extensively used data in environmental, climatological, hydrological, and water resources studies. For example, conducting a flood risk analysis for a river requires comprehensive historical river flow records, but availability of records may be limited in length. In such cases, rainfall-runoff models are required to provide the river flow information through the input of precipitation data. Precipitation series at times are too short or contain missing records, thus making reliable and meaningful analyses difficult. In such situations, stochastic precipitation generation models can be a great asset in providing alternate (i.e., simulated or synthetic) precipitation series that are consistent with the observed characteristics of the historical precipitation records.

Stochastic modeling of daily precipitation is more challenging than the modeling of monthly precipitation and of other weather variables because of the inherent and unique statistical properties of daily precipitation records. For instance, daily precipitation has a mixed probability distribution with a high frequency of zeros which makes it quite different compared to the distribution of monthly precipitation, which is often regarded to be close to normal distribution. Furthermore, in northern environments such as that of Canada, monthly precipitation rarely displays zero values and hence makes modeling on a monthly basis a lot easier. Daily precipitation forms a temporally intermittent time series due to the existence of zeros in the series. Because of its crucial role in weather, hydrologic and water resources related studies, daily precipitation is the variable of interest in this thesis.

2.1 Models for Daily Precipitation Generation

Generation of precipitation data has attracted attention among meteorologists and hydrologists alike. As a result, meteorologic and hydrologic literature abounds with numerous simulation or generation models of precipitation data with varying degrees of success. In this chapter, a brief literature review is presented related to the models for the generation of daily precipitation at single and multiple sites.

2.1.1 Generation of Daily Precipitation at a Single Site

Traditionally, the modeling of daily precipitation at a single site separates the task into two stages where precipitation occurrences and amounts are individually generated. A sequence of precipitation occurrences is first generated and precipitation amounts are subsequently generated for the days identified to be wet. A good generation model should reasonably preserve the statistical characteristics of the historical

precipitation records, including the temporal dependence of successive occurrences and the underlying probability distribution of precipitation amounts.

Modeling of Daily Precipitation Occurrences

Methods for modeling of daily precipitation occurrences include alternating renewal processes (ARP) [*Green*, 1964], Poisson models [*Duckstein et al.*, 1972], discrete autoregressive moving average models [*Chang et al.*, 1984], point process models [*Foufoula-Georgiou and Lettenmaier*, 1986], and Markov chain models. Markov models and ARP have been extensively used because of their relative simplicity.

Markov chain dependence structure has been used in the development of a variety of models, e.g., Markov renewal models [*Foufoula-Georgiou and Lettenmaier*, 1987], Hidden Markov models [*Thyer and Kuczera*, 2003b], and non-homogenous Hidden Markov models [*Rasmussen and Akintug*, 2004]. Early applications of Markov models were simple and represented a reasonable mechanism for the generation of daily precipitation occurrences [*Gabriel and Neumann*, 1962]. However, Markov models have been found to be inappropriate in some situation for modeling the clustering tendencies that exist in daily precipitation sequences [*Foufoula-Georgiou and Lettenmaier*, 1987]. Often, these models are reported to underestimate the observed frequency of long dry spells [*Racsko et al.*, 1991]. Some researchers have advocated the inclusion of higher-order memory components for improving the simulation of long-term dependencies [*Semenov et al.*, 1998], but higher order models cannot be considered parsimonious due to the large number of parameters involved.

The generation of daily precipitation occurrences using an ARP is done by alternatingly drawing samples from distributions of wet-spell lengths and dry-spell lengths. A wet/dry-spell length signifies a succession of pure wet-day or dry-day

occurrences that captures the persistence of wet/dry-day occurrences. The distribution of wet/dry-spell can be empirically represented or through parametric probability functions (e.g., truncated negative binomial or geometric distribution [*Rolda'n and Woolhiser*, 1982]). The best distribution model can be selected based on the goodness-of-fit of the observed dry/wet-spell lengths [*Semenov and Barrow*, 1997].

The performance of ARP models, in general, is comparable to the Markov models [*Semenov et al.*, 1998]. Both model types have been used as components in many existing weather generators (e.g., the ARP used in the LARS-WG [*Semenov et al.*, 1998] and the Markov models used in the WGEN [*Richardson and Wright*, 1984], SIMMETEO [*Geng et al.*, 1986], WeatherMan [*Pickering et al.*, 1994], and WXGEN [*Hayhoe and Stewart*, 1996]). Comparisons of the WGEN and LARS-WG models indicate that the LARS-WG model tends to match the wet/dry spells in observed data more closely than the WGEN model, but the WGEN model was found to be simple and has a smoothing effect on the observed precipitation [*Semenov et al.*, 1998]. In general, because of their simplicity, Markov models appear to be more commonly used for the generation of daily precipitation occurrences than ARPs.

Modeling of Daily Precipitation Amounts

For the generation of precipitation amounts, a probability distribution is selected based on its fit to the observed data. Once a suitable probability distribution has been identified, precipitation amounts on wet days can be generated by drawing samples from the chosen probability distribution of wet-spell. The probability distribution of daily precipitation amounts for a single site tends to be highly skewed and commonly modeled by the exponential distribution, the gamma distribution, or the three-parameter mixed exponential distribution. These distribution functions have

been compared and were found satisfactory for the generation of daily precipitation amounts [*Woolhiser and Rolda'n*, 1982].

2.1.2 Generation of Daily Precipitation at Multiple Sites

Many of the models proposed for precipitation simulation at multiple sites [*Qian et al.*, 2002; *Brissette et al.*, 2007; *Srikanthan and Pegram*, 2009] are based on principles presented by *Wilks* [1998]. The Markov model has been used for the generation of precipitation at multiple-site by employing spatially correlated random variables derived from correlation curves of pairwise stations. Performance of this model has been compared with the Hidden Markov and k-nearest neighbor models, and it was found that many important statistical properties of daily precipitation at multiple sites can be reproduced [*Mehrotra et al.*, 2006]. However, the model is not considered effective in dealing with the spatial variability of the generated precipitation fields including the spatial intermittency [*Bardossy*, 1992; *Breinl et al.*, 2013], and the estimation of parameters is difficult to automate [*Baigorria and Jones*, 2010]. Although the model produces some time dependence in consecutive non-zero precipitation amounts, the model was not specifically designed to address the temporal dependence among daily precipitation occurrences.

Some precipitation generation models represent physical aspects to some degree in the precipitation generation mechanism. One example of this is the so-called Hidden Markov models (HMM) [*Akintug and Rasmussen*, 2005]. The HMM generates the precipitation distribution conditional on a discrete weather state representing certain identified spatial precipitation patterns. The spatial dependence is maintained via the assumption of a common weather state in a domain while the temporal dependence is simulated based on the assumption that the weather state is Markovian in nature.

The HMM model is conceptually appealing because of its reliance on atmospheric conditions in order to capture some physical processes. However, the estimation of parameters is intricate and may limit its practical use [*Mehrotra et al.*, 2006].

Another kind of model that captures some of the physical aspects of precipitation is the Neyman-Scott Rectangular Pulses (NSRP) model [*Burton et al.*, 2008]. The NSRP model uses clustered point processes to describe the origin of precipitation events, the number of rain cells in events, and the origin of the cells. A concern related to the model is that the estimation of parameters is sensitive to the selection of the timescale for an analysis.

An alternative approach to dealing with physical aspects of the atmosphere into the generation model is by conditioning on exogenous atmospheric predictors. This approach is common in downscaling studies that are discussed later in this chapter.

Recently, copula-based models have been proposed for simulation of precipitation at multiple-site [*Bardossy and Pegram*, 2009; *Serinaldi*, 2009]. An advantage of these models is that correlations of precipitation amounts of pair-wise stations can be directly described by the spatial structure of the copula without the need for normalizing the precipitation data. The components responsible for the generation of temporal dependence series may still rely on the Markov model. The model has been found to provide a good representation of the covariance structure, but is not yet good enough to fully capture the spatial structure of precipitation records at multiple sites. The ability of the model to preserve the temporal dependence in precipitation series, especially among the distributions of wet/dry-day spells, requires further investigations.

Non-parametric models, including combinations of k-nearest-neighbor (k-NN) models and Markov models [*Beersma and Buishand*, 2003; *Apipattanavis et al.*, 2007], simulation of weather types [*Wilby et al.*, 2003], and reshuffling algorithms [*Clark*, 2004] have been studied. Many models have been developed based on the k-nearest neighbor (k-NN) approach that resamples observed precipitation, thereby preserving the statistical characteristics of observations. The main advantage of non-parametric models is that such models do not require restrictive assumptions regarding the probability distributions and the correlation structure. However, the k-NN model requires long sequences of high-quality data to initialize, and further, this model usually resamples values from the historical database. Another drawback of the k-NN model is that it cannot produce precipitation amounts beyond those present in the historical records [*Leander and Buishand*, 2009].

The semi-parametric modeling approach combines parametric and non-parametric model components for the generation of daily precipitation. Semi-parametric models include methods related to re-simulation of atmospheric circulation patterns [*Brandsma and Buishand*, 1997], the re-simulation of weather types with Markov models, [*Palutikof et al.*, 2002], the use of Bernoulli-Gamma density networks [*Cannon*, 2008], and the k-NN with perturbation of the highest re-sampled values [*Leander and Buishand*, 2009].

In general, Markov models are most commonly used for the generation of daily precipitation occurrences at multiple sites, but no single model has been identified as the most suitable for the generation of daily precipitation amounts. Many multiple-site generation models have been developed based on the two-stage approach that generates precipitation occurrences and amounts individually. Some multiple-site models

(e.g., NSRP, Wilks' model, HMM, or k-NN) are essentially developed by extending single-site models ([*Cowpertwait et al.*, 2002; *Wilks*, 1998; *Thyer and Kuczera*, 2003a; *Buishand and Brandsma*, 2001], and [*Cowpertwait*, 1995; *Richardson and Wright*, 1984; *Hughes et al.*, 1999; *Rajagopalan and Lall*, 1999]).

A spatially dependent precipitation generation model must provide a means of generating daily precipitation time series that reproduce the spatial statistical characteristics of the observed data at multiple sites. Generally, spatial dependence is modeled by using a multivariate normal distribution. However, such a probability distribution cannot directly be used for generating daily precipitation because precipitation amounts at each site are non-normally distributed and follow a mixed distribution with a discrete probability at zero. Alternatives to the multivariate normal distribution exist (e.g., Copulas), but are less developed and may not necessarily be better suited for modeling of daily precipitation.

At each site, if precipitation data were normally distributed, it would be relatively easy to develop a model for generating daily precipitation at neighboring sites. One such possibility is to view the mixed distribution of daily precipitation as a censored normal distribution, with a left-censoring point at zero. This can be achieved by assigning the probability associated with zero values to the portion of the distribution imaginably assumed to spread in the negative domain. The parameter estimation for a univariate censored normal distribution based on the maximum likelihood method is well developed [*Cohen*, 1959]. Transformation of a non-normally distributed daily precipitation data at each site into a multivariate normally distributed data corresponding to multisite has been considered for modeling of daily precipitation by *Bardossy* [1992]; *Hutchinson* [1995]; *Stehlik and Bardossy* [2002], and *Sanso and Guenni*

[2004]. For this purpose, at times, a power transformation [*Sanso and Guenni*, 2000; *Yang et al.*, 2005] is incorporated into the parameter estimation process at each site to better represent the distribution of positive precipitation amounts.

General Concluding Remarks

The advantage of the two-stage approach is that the modeling tasks can be separated to facilitate statistical modeling. However, several disadvantages of the two-stage approach can be identified: [1] different attributes (e.g., precipitation occurrences and amounts, or spatial and temporal dependencies) may not be well addressed. [2] more parameters may be required for the modeling of the two separate components. [3] the analytical form of the model can be intricate. [4] model use may be cumbersome because two-generation mechanisms are involved. Hence, an integrated approach that combines different aspects into a single model would be an attractive alternative for the simulation of daily precipitation at multiple sites.

Describing the spatial dependence of daily precipitation series at two stations using correlation coefficient may not be appropriate because marginal distributions of the daily precipitations at each such station are non-normally distributed. Although methods for describing the dependence of non-normal distribution (e.g., truncated and censored distribution) exist (e.g., [*Singh*, 1960]), the methods tend to be intricate and may not be easily to understood by practitioners. Instead, *Bardossy* [1992] used the quantile data corresponding the censored normal distribution of each site to estimate the dependence parameter between time series at two sites. In this thesis, an alternative approach utilizing the concept of a multivariate censored Normal distribution for the estimation of spatial dependence of daily precipitation is proposed.

The temporal dependence in precipitation records can be modeled using the alternating renewal process or Markov process. However, many of the existing multiple-site models using the two-stage approach may be inadequate or lack the necessary statistical structure capable of integration of the temporal dependence as well as spatial dependence. On the other hand, one advantage of the acceptance and use of the multivariate normal distribution for the development of a multi-site generation model is that it can easily be used to formulate a multivariate autoregressive model for precipitation generation [*Bras and Rodriguez-Iturbe*, 1985; *Bardossy*, 1992] and is capable of integrating both spatial as well as temporal dependencies into one unified single model. With such a unified model, precipitation amounts and occurrences inclusive of spatial and temporal dependencies among multiple sites can be modeled and simulated. In this thesis, the parameters of a multivariate normal distribution are estimated in the form of a multivariate censored distribution.

2.2 Applications of Daily Precipitation Generation Models

In view of the foregoing discussion, an advantage of using the multivariate normal distribution to form a multivariate autoregressive model is that it is amenable to various statistical analyses and easily applies to other scenarios. The multivariate normal distribution allows direct applications into many statistical tests and analyses requiring the assumption that datasets be normally distributed. As stated earlier, an extension of a multivariate autoregressive model capable of preserving the temporal dependence in precipitation sequences into a single model is relatively straight forward. This thesis examines the potential of such a model for three types of problems related to the generation of daily precipitation.

2.2.1 Generation of Daily Precipitation at Ungauged Sites

Modern rain gauges can provide precipitation rates in real time and at a fine temporal resolution. The selection of rain gauge location is affected by factors such as: accessibility, ease of maintenance, topographical features, etc. The lack of rain gauge stations, especially in remote areas, is a common problem. The spatial variation of precipitation is difficult to characterize without a rain gauge network of sufficient density in space. Recent advances in satellite remote sensing seem to have potential to provide spatial coverage in the form of pixel precipitation estimates. However, estimation of precipitation through this technique may already have encouraged meteorological agencies to reduce the number of rain gauge networks. This is an unfortunate outcome since the accuracy of satellite precipitation does not match the accuracy of rain gauge measurements [*Ali et al.*, 2005]. Therefore, a methodology for the augmentation of data sets of an existing rain gauge network has been considered.

A parameter interpolation method, based on the Parameter-elevation Regressions on Independent Slopes Model (PRISM) statistical interpolator [*Johnson et al.*, 2000], was found to be useful for the simulation of precipitation at ungauged sites. The PRISM software relies on a coordinated set of rules, decisions, and calculations designed to accommodate the decision-making process that an expert climatologist normally would invoke. The PRISM software, however, is not publicly available and requires a certain degree of specialized knowledge and judgment for use. In this regard, *Wilks* [2008] has suggested the use of elevation as a covariate to spatially interpolate the parameters of a selected weather generator. This method constructs local weighted regressions for each parameter to be interpolated over a space and has been found to be flexible in capturing non-linear parametric variations in space and to

perform well in a region of the northeastern U.S. However, the use of elevation as the sole predictor may overlook the possible relationships of the interpolated parameters to landform characteristics *Wilks* [2008].

Spatial generation of precipitation at ungauged sites is commonly conducted through geostatistical interpolation using a type of weighted regression method (e.g., Kriging) [*Johnson et al.*, 2000; *Wilks*, 2008]. At ungauged sites, the approach assumes spatial stationarity in model parameters, and accordingly, the estimates of precipitation at ungauged sites are obtained based on interpolated parameters from gauged sites [*Cressie*, 1993]. In this regard, the weighted regression method is an efficient parametric method for spatial interpolation of precipitation based on a predetermined form of the data structure. The regression method, however, is sensitive to outliers that are not uncommon in daily precipitation records. On the other hand, a non-parametric approach may offer a robust alternative for the interpolation of spatial precipitation.

Advances in computing technology over the past few decades have prompted the use of machine learning approaches (e.g., artificial neural network (ANN)) in hydrological analyses [*Panu et al.*, 2000; *Bowden et al.*, 2005]. For instance, ANN techniques establish a non-linear relationship between independent and dependent variables through an architecture consisting of input, hidden, and output layers representing a combination of non-linear functions. The ANN techniques, in some cases, can yield more accurate results under the same set of constraints when compared to the weighted regression methods [*Chowdhury et al.*, 2010]. However, there are drawbacks such as the black-box nature of the ANN technique, greater demand in computation, proneness to over fitting, and subjective judgment in the design of the

architecture.

A comparable alternative to the ANN techniques is a generalized regression neural network (GRNN), which provides a distinct advantage in handling regression-type problems because of its simplicity in structural configuration. The GRNN is a non-parametric technique that has been used in hydrological analysis and its prediction capabilities have been demonstrated [*Ng et al.*, 2009]. Moreover, due to its non-parametric nature, GRNN is a robust predictor that is not sensitive to outliers.

Based on the assumption of spatial stationarity at ungauged sites, many spatial interpolation methods have been developed. These methods, because of their simplicity, are widely adopted but do not adequately address the issue of variable orography. Recently, a precipitation generator based on the Kriging technique has been proposed in an attempt to capture the uncertainty related to variable orography. In the sense of quadratic loss considerations, the Kriging technique is considered to be the best linear unbiased predictor, which also coincides with conditional expectations for normally distributed variables [*Kleiber et al.*, 2012]. Such a technique has been adopted along with spatial Gaussian processes for the interpolation of statistical parameters from gauged to ungauged sites.

Based on the foregoing discussion of concepts and ideas, methods have been developed either utilizing GRNN or Kriging in this thesis for the simulation of precipitation at ungauged sites.

2.2.2 Missing Observations in Daily Precipitation Records

The existence of missing observations is a common problem in hydrological data sets. Improper treatment of missing observations can affect the fitting of statistical models [*Little and Rubin*, 1987]. Missing observations in a data set can be caused by loss

of records, malfunctioning of instruments, non-responses of measuring instruments, and interruption in data collection (due to lack of interest, finances, human resources or technical support). Assessment of incomplete data sets invariably raises the level of complexity in statistical analyses. The varying nature of missing observations under different circumstances requires practical and rational treatments and estimation techniques to adequately handle the presence of missing observations.

For the estimation of missing daily precipitation at multi-site, the selection of an appropriate method should consider the type of neighboring observations, the duration of the time series, the length of the data set with missing observations, the distance between sites, and the impact of temporal/spatial intermittence. Basically, missing observations can be estimated based on assumptions related to probability distributions, spatial dependence, temporal dependence, or other forms of dependence. For instance, missing observations can be estimated using the neighboring observations within a time series itself or using observations from nearby sites.

For daily precipitation data, the probability distributions and spatial dependence among nearby sites are normally more significant than the temporal dependence in the precipitation series. Therefore, the missing observations at a given site should be estimated through the correlation structure among nearby sites expressed in parameters of the multivariate normal distribution model. Missing observations can be infilled or estimated through a conditional sampling obtained from the underlying distribution.

Methods based on Markov Chain Monte Carlo (MCMC) in a Bayesian-type framework have been proposed for dealing with missing data. Among many different

MCMC techniques, Gibbs sampling has proven to be an efficient technique for infilling of missing hydrological observations [*Thyer and Kuczera*, 2000]. Gibbs sampling can be used to obtain the candidate observations (i.e., missing observations) where the form of the full conditional distribution of parameters is available.

The Gibbs sampling approach is typically applied to normally distributed data. For precipitation data, the method must be modified to address the need of simulating negative values representing the dry-day occurrences. The data sets consisting of infilled data should be comparable to data sets with missing observations because the statistical characteristics of historical precipitation records are preserved after the infilling of missing observations. The statistical characteristics refer to the mean and covariance of the underlying distribution that describes certain aspects of the probability distribution and spatial dependence of precipitation at multiple sites.

The temporal dependence is rarely considered due to difficulties encountered in dealing with temporal intermittence. However, completely neglecting the existence of temporal dependence in the precipitation series may be unjustified. Therefore, in the second part of the application related to infilling of missing precipitation data, an attempt has been made to adopt a method for infilling of missing observations that preserves comprehensive statistics of historical precipitation records while taking into consideration certain aspects including probability distribution, spatial dependence, and temporal dependence of observations.

In this study, a Gibbs sampling approach has been utilized for the infilling of missing precipitation observations.

2.2.3 Climate Change Implications on Daily Precipitations Records

The proposed precipitation generator, thus far, has been regarded as a model under a stationary process. However, the change in future precipitation patterns cannot be ignored, particularly at the present juncture when rising global temperatures are causing alarms and are impacting atmospheric processes [*IPCC*, 2007]. Many researchers have been focusing on studies related to the impact assessment of climate change because of its global socio-environmental impacts and damages over the last few decades. The influence of climate change on future precipitation patterns can be quantitatively estimated from the simulated outputs of global climate models (GCMs), also known as general circulation models. These models are based on mathematical algorithms representing the physics of ocean-atmospheric processes and land surface interactions. Atmospheric circulation is the major driver of regional climates and a good correspondence exists between local precipitation patterns and atmospheric processes [*Stehlik and Bardossy*, 2002]. However, the GCMs have low spatial resolutions (e.g., 250 km $\times$ 250 km) that are inadequate for use in studies related to local hydrologic and water resource systems, and further, they do not even provide realistic precipitation series for the present climate scenarios [*IPCC*, 1996]. Although regional climate models offer fine resolutions (e.g., 45 km $\times$ 45 km), they are still considered to be deficient for generating precipitation time series in smaller regions or at sites. Downscaling approaches, therefore, are required to bridge the gap between the coarse resolution of climate models and the finer-resolutions at which information is required by end-users at the local scale.

Downscaling approaches can usually be divided into two categories: dynamic and statistical. Regional climate models (RCM), also called limited-area models (LAM),

are examples of dynamic downscaling in which the climatic data are modeled into a fine grid scale. The output of a GCM can be downscaled using a dynamic approach that adopts the boundary conditions from a GCM as input to a RCM set up for a limited area of interest. Although dynamic downscaling has been reported to be successful in several studies [*Giorgi and Bates*, 1989], it is computationally intensive. On the other hand, statistical downscaling has proven to be a reasonable alternative to regional climate models. Statistical downscaling techniques have been categorized as pattern classification methods, regression methods, and stochastic weather generators. Comparisons of these three schemes have been provided by [*Wilby et al.*, 2004].

Precipitation generation models can be developed based either on the observed precipitation records corresponding to local monitoring sites or the observed precipitation records in combination with exogenous inputs (e.g., assumed/projected monthly changes depicted in Figure 2.1). Many statistical downscaling approaches are similar to the traditional weather generating models, except that their parameters are modified based on external information obtained from climate models [*Wilks*, 1992].

The downscaling of weather data, particularly of daily precipitation, adds further complexity to modeling. Many of the downscaling approaches require extensive effort in the estimation of parameters and statistical verification, which may limit their practical use [*Wilks*, 1998]. In addition to the three categories of statistical downscaling techniques, the impact of climate change can also be assessed by the widely-used Delta change downscaling method [*Hay et al.*, 2000].

In addition, the Delta change method has the advantage of being simple and transparent [*Andreasson and Rosberg*, 2006]. The method produces a future time series of a local meteorological variable (e.g., precipitation) by perturbing the observed

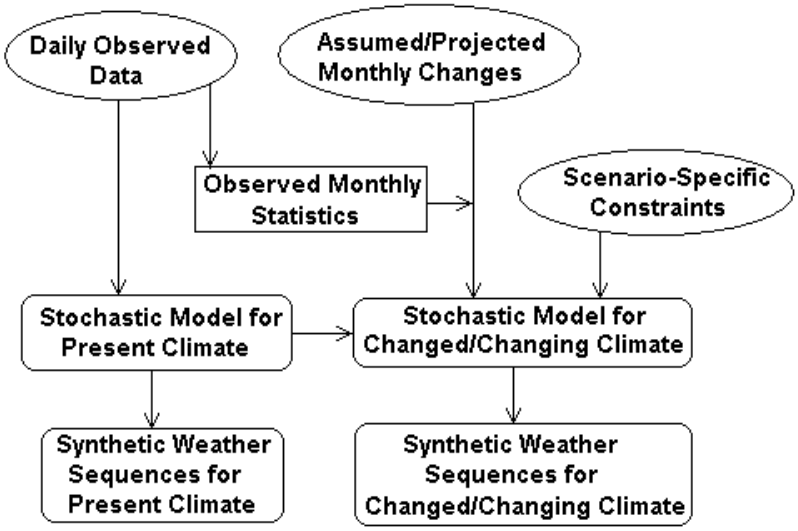


Fig. 2.1: Schematic conceptual framework of the weather generation. (After: [*Wilby et al.*, 1998])

baseline series. The magnitude of perturbation is based on a factor determined by the monthly-expected change of precipitation obtained from a climate model. In other words, the Delta change method modifies the baseline data to capture the future changes in statistics (i.e., monthly mean and standard deviation) as reflected in the output of climate model. The baseline series at the local scale typically refers to the observed historical precipitation.

Many downscaling methods are capable of reflecting the future changes of climate in terms of their monthly statistics (i.e., monthly mean and monthly standard deviation). However, attempts to reflect future changes in the daily statistics (e.g., probability of wet-day occurrences, probability distribution of daily precipitation amounts) can be challenging. Furthermore, future changes in daily statistics with respect to the spatial and temporal dependencies at multiple sites of interest have seldom been reported in the literature.

To assess the impact of climate change, a parametric Delta change method is proposed and the parameters of the proposed precipitation generation model are modified to capture the future change in statistics as reflected in the RCM data. Details of impact assessment are further discussed in chapter 7. The future changes in statistics are described by the change in parameters of the proposed generation model using the RCM data from current and future time periods.

2.3 Scope and Focus of the Research - a Summary

In summary, this thesis intends to develop a stochastic daily precipitation generation model capable of preserving the spatio-temporal characteristics at multiple sites. It is anticipated that this model, with a relevant generic structure, can be applied to studies as follows:

- Data generation at gauged and ungauged sites: To generate daily precipitation data at multisite using the notions of the multivariate autoregressive censored precipitation model, while invoking the assumptions of stationarity and non-stationarity in both space and time;

- Infilling of missing observations: To develop suitable data substitution procedures for the estimation of missing observations using the notions of the multivariate censored Normal distribution and multivariate autoregressive censored process in combination with Gibbs sampling. These two models intend to preserve the spatial and spatial-temporal dependencies of the historical precipitation;

- Statistical downscaling: To downscale the precipitation outputs of a Regional

Climate Model (RCM) in conjunction with proposed precipitation generation models using the Delta change method for addressing impacts of climate change on a local scale.

3. DEVELOPMENT AND EVALUATION OF A WEATHER GENERATION MODEL BASED ON MULTIVARIATE CENSORED DISTRIBUTION (WG-MCD)

A stochastic model is needed to generate the weather that should recognize the spatial and temporal variation of daily precipitation. Daily precipitation is characterized by zero and positive values and therefore its modeling requires a distribution that is tractable using easily accessible probability functions. One simple distribution is the censored normal pdf, whose properties are well documented in statistical literature, which is used to describe the probabilistic features of such data sequences. Furthermore, since we are dealing with the precipitation at several sites, such a scenario represents a multivariate case. In succinct terms, the formulation of the weather generation model is based on the concepts of a multivariate censored distribution (MCD), and is named hereafter as the WG-MCD (weather generator- multivariate censored distribution) model.

3.1 Formulation of the Multivariate Censored Normal Distribution

Daily precipitation data are complex and are not easy to model through a single probability function. The challenge arises due to the nature of the data, i.e., occurrences of zero values (dry days) and positive values (wet days). The probability of occurrence of dry days is significant in almost all regions across the globe, Canada being no exception. In the Canadian Prairies, the probability of dry days could hovers

around 0.70 and that of wet days could be around 0.30. The precipitation magnitude and occurrence corresponding to above zero records, or the precipitation on wet days, can be easily modeled through some known probability function, preferably using a normal probability by suitably transforming the data. Occurrences of zero values comprise the significant probability component that need to be accounted for modeling applications, particularly for simulation purposes. One way of accounting for the occurrences of zero values is to extend the normal probability curve into the negative domain (0 to $-\infty$) such that the area under the curve is equal to the probability of dry days. In other words, occurrences of zero values are considered as the censored portion of the normal distribution as depicted in Figure 3.1. The daily precipitation data, therefore, can be fitted into a normal distribution, whose left limb is imaginary but represents a powerful way to generate precipitation components of the weather. The next complex step is the estimation of parameters (mean and variance) of the hypothetical normal distribution, which is described below.

A stepwise procedure for its development is shown schematically in Figure 3.2. The parameter estimation of the MCD is accomplished through three steps: 1. Estimation of parameters of marginal/univariate censored normal distributions through normalization of the data set, 2. Pair-wise estimation of correlation parameters of bivariate distributions, and 3. Estimation of parameters of the multivariate censored normal distribution, including modification of non-positive definite matrices. The second stage of the model development involving temporal featues is described in the next chapter.

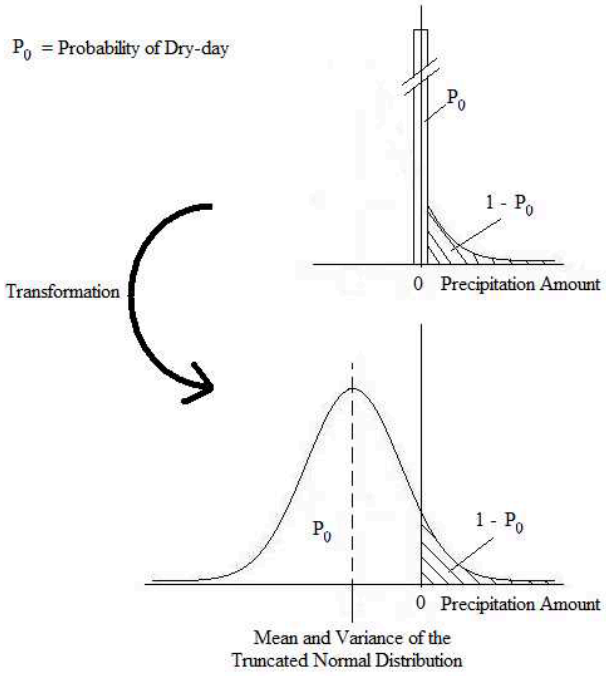


Fig. 3.1: Transformation of a mixed distribution to normal distribution.

3.1.1 Step 1: Estimation of Parameters of Marginal/Univariate Distributions and Data Normalization

The parameters of a censored distribution from precipitation data sets for each Julian day at each site can be estimated by the method of maximum likelihood as described *Cohen* [1950]. Since the distribution is left censored and the number of zeros in the precipitation data set is known, the relevant equations in the maximum likelihood methodology are modified accordingly. The probability density function of a univariate censored distribution may be considered as a mixture of a binary and a normal variable as follows.

$$f(x,\tau) = [\Phi(\xi)]^{1-\tau}\left[\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}(\xi + x/\sigma)^2\right)\right]^{\tau} \tag{3.1.1}$$

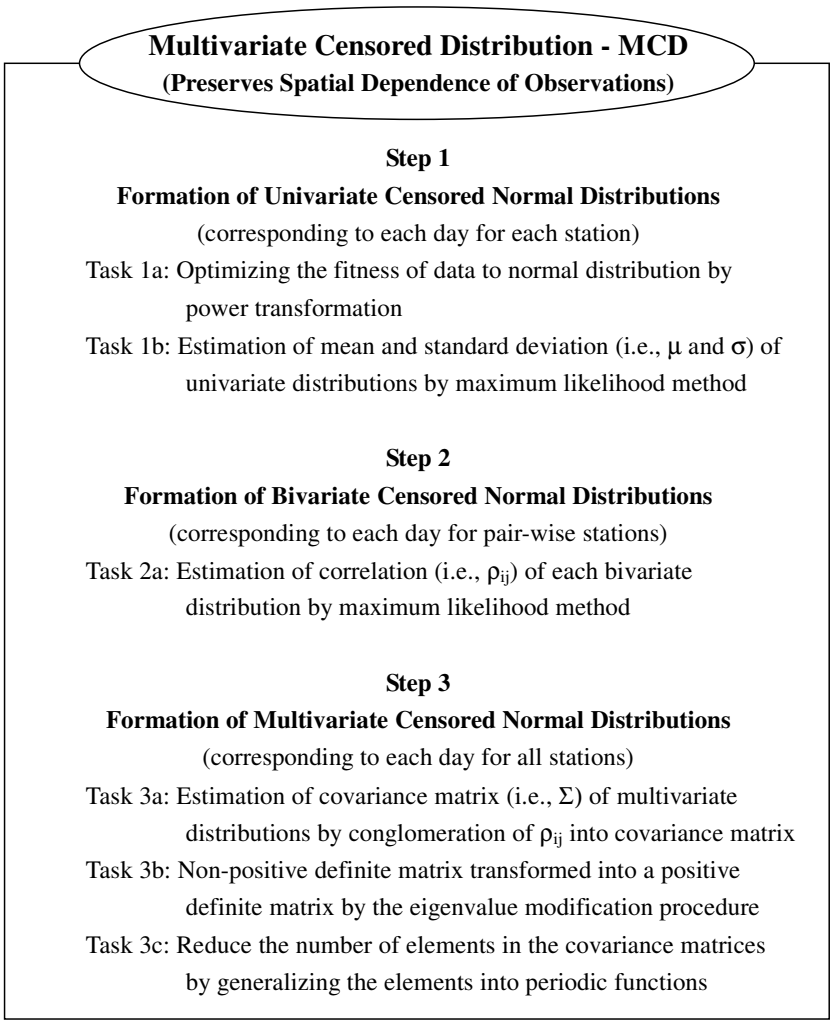


Fig. 3.2: Schematic structure of the MCD formulation.

where x denotes the daily precipitation observations. $\tau = 0$ and $\tau = 1$ respectively denote the dry-day and wet-day occurrence, and ξ denotes the truncation point in the form of a standard censored normal variate (Cohen refers to it as standardized unit), that is,

$$\xi = \frac{x_0 - \mu}{\sigma} \tag{3.1.2}$$

where x_0 is the censoring point. In this research, $x_0 = 0$ because this point separates the observations of wet-day and dry-day under a censored distribution. Furthermore,

$$\Phi(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\xi} \exp\left(-t^2/2\right) dt \tag{3.1.3}$$

where $\Phi(\xi)$ is the cumulative distribution function of the standard normal distribution.

For the observed precipitation data on a Julian day, the likelihood function corresponding to Equation 3.1.1 becomes:

$$L(\xi, \sigma | x_1, x_2, ..., x_{n_1}, x_{n_1+1}, ..., x_{n_1+n_0}) = (\Phi(\xi))^{n_0} (\frac{1}{\sigma\sqrt{2\pi}})^{n_1} \exp[-\frac{1}{2} \sum_{i=1}^{n_1} (\xi + x_i/\sigma)^2] \tag{3.1.4}$$

where n_0 and n_1 respectively denote the number of observed dry days (i.e., zero records) and the number of observed wet days (i.e., above-zero records).

Taking natural logarithms of Equation 3.1.4 and setting partial derivatives equal to zero yields the following maximum likelihood equations:

$$\frac{\partial \ln L}{\partial \xi} = \frac{n_0 \ \phi(\xi)}{\Phi(\xi)} - \sum_{i=1}^{n_1} (\xi + \frac{x_i}{\sigma}) = 0 \tag{3.1.5}$$

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n_1}{\sigma} + \frac{1}{\sigma^2}\sum_{i=1}^{n_1}\{x_i(\xi + \frac{x_i}{\sigma})\} = 0 \tag{3.1.6}$$

where $\phi(\xi) = \frac{1}{\sqrt{2\pi}}\exp(-\xi^2/2)$ is the ordinate of the standard normal probability curve,

Letting

$$Y(\xi) = \frac{n_0}{n_1}\frac{\phi(\xi)}{\Phi(\xi)} \tag{3.1.7}$$

Equations 3.1.5 and 3.1.6 can also be written as:

$$\sigma[Y(\xi) - \xi] - v_1 = 0 \tag{3.1.8}$$

$$\sigma^2[1 - \xi(Y(\xi) - \xi)] - v_2 = 0 \tag{3.1.9}$$

where v_1 and v_2 are the first and second sample moments of the non-zero records, i.e., $v_k = \sum_{i=1}^{n_1} x_i^k/n_1$.

Equations 3.1.8 and 3.1.9 can be solved simultaneously for the two unknowns, σ and ξ. The mean value μ can subsequently be estimated from Equation 3.1.2.

An approximation proposed by *Cohen* [1950] was also applied as a preliminary estimator. Since the number of observed wet and dry days (n_1 and n_0) are known, it is possible to obtain an estimate of ξ directly from the standard normal curve as:

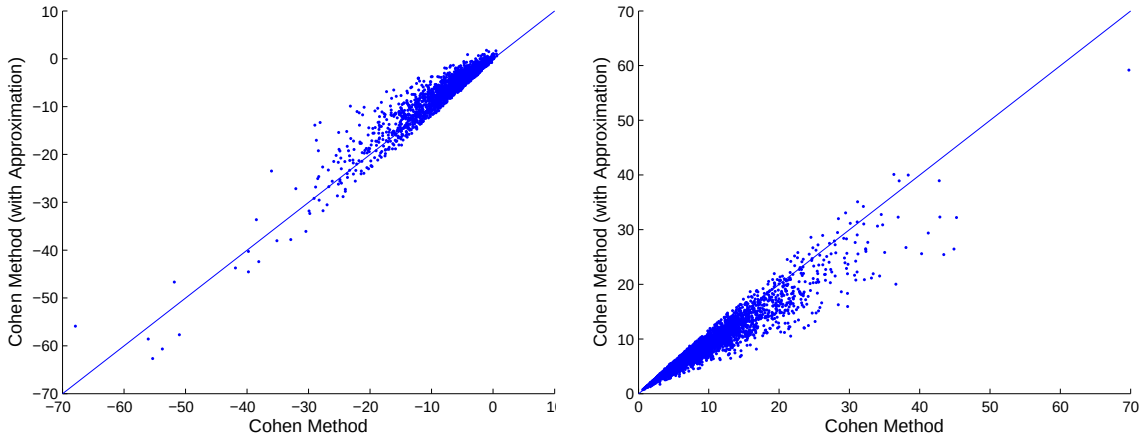
$$\Phi(\xi) = \frac{n_0}{n_1 + n_0} \tag{3.1.10}$$

from which ξ can be calculated directly as:

$$\xi = \Phi^{-1}(\frac{n_0}{n_1 + n_0}) \tag{3.1.11}$$

A simple example will illustrate the above calculation. If in a precipitation sequence $n_0 = 70$ and $n_1 = 30$ then $n_0/(n_0 + n_1) = 0.70$. The cumulative probability in the standard normal curve is therefore equals 0.70 for which the corresponding standard normal variate is 0.53. Therefore, the value of ξ can be taken as 0.53. $\hat{\sigma}$ can then be calculated using either Equation 3.1.8 or 3.1.9, and μ can be calculated using Equation 3.1.2. From a numerical example presented in Cohen's paper, errors of approximately 1% and 8% in estimates of $\hat{\mu}$ and $\hat{\sigma}$ (relative to the unbiased exact maximum likelihood estimates) were obtained using the first approximation method. To further evaluate the loss in accuracy by using the approximation method, the historical observations of this study were employed for the analysis. In Figure 3.3, values of the coefficient of determination (R^2) were calculated between the estimated parameters obtained by the Cohen's method and the Cohen's approximation method. The Cohen method refers to the exact maximum likelihood method corresponding to the solution of Equations 3.1.8 and 3.1.9. The R^2 for both $\hat{\mu}$ and $\hat{\sigma}$ were found to be 0.92. Figure 3.3 also shows that the discrepancies between the parameters estimated from these two methods tends to overestimate the mean and underestimate the standard deviation.

Obtaining the parameters using Cohen's method without approximation provides the most unbiased estimates of a censored normal distribution. However, data on precipitation amounts are unlikely to be represented by a normal distribution. In view of the suggestion by [*Bardossy*, 1992], a power transformation was applied to precipitation data to bring them closer to a normal distribution. The power transformation



(a) Mean Value *(b)* Standard Deviation

Fig. 3.3: Parameters of univariate censored distribution estimated by the exact maximum likelihood method and by Cohen's approximate method for 10 stations in Manitoba.

is expressed as follows:

$$w_i = x_i^{1/\beta} \tag{3.1.12}$$

where w_i is the transformed variable.

Normalization of transformations is not restricted to the power transformation and other transformation functions could be considered. The portion of the probability distribution corresponding to the above-zero records can be modeled using a one-parameter exponential distribution, a two-parameter gamma distribution, or any other similar distribution function [*Woolhiser and Rolda'n*, 1982]. The best transformation is selected based on the maximum value of either the Akaike or the Bayesian information criterion:

$$AIC = -2\ln(Likelihood\ estimate) + 2Q \tag{3.1.13}$$

$$BIC = -2\ln(Likelihood\ estimate) + Q\ln(n_1) \tag{3.1.14}$$

where Q is the number of independently adjusted parameters, and n_1 is the number of observed wet days.

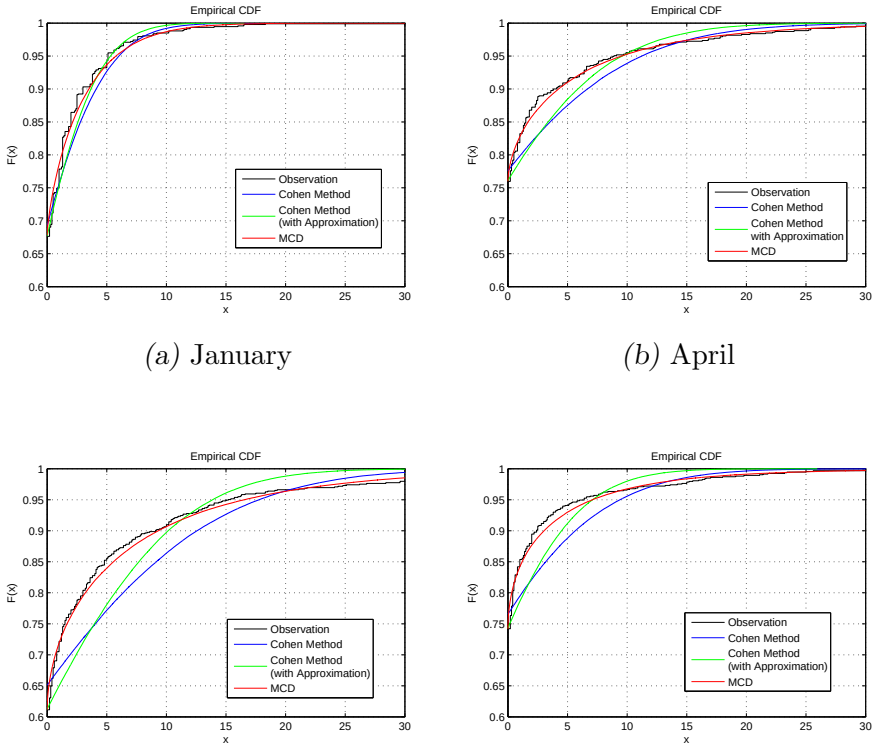
Although other transformation methods are available, for the sake of simplicity in this thesis, the power transformation is considered. With the additional parameter β, the estimation of the parameters of a censored distribution becomes more intricate. Clearly Equations 3.1.8 and 3.1.9 alone are insufficient to estimate three parameters. One way to handle this problem is to employ an approximation method that provides an additional Equation 3.1.11 to determine the three unknowns. To estimate the parameters, the σ in Equation 3.1.8 is substituted into Equation 3.1.9 and ξ is calculated using Equation 3.1.11. As a result, the combined equation only has β as an unknown and can then be determined using numerical root-finding. The parameters of σ and μ can also be calculated using Equations 3.1.8 and 3.1.2.

In Figure 3.4, various modeling options are compared using the data set of station 1. The options are:

1. Cohen method: Exact maximum likelihood estimation of censored normal distribution

2. Cohen method with approximation: Approximate maximum likelihood estimation of censored normal distribution

3. Method involving MCD: Power-transformation of relevant data, combined with approximate maximum likelihood method.



(a) January *(b)* April

(c) July *(d)* October

Fig. 3.4: CDFs of observation and simulation of models at station 1.

Above options do not include "Power-transformation of relevant data, combined with exact maximum likelihood method" due to the challenge of solving three unknowns (i.e., μ, σ, and β) with only two maximized likelihood equations available.

Figure 3.4 clearly shows that a power transformation of precipitation data shown for four months is suitable to obtain a good fit with censored normal distributions.

It is to be noted that the Cohen method does not ensure that the CDFs at truncation point $F(0)$ match well with the observed wet-day (dry-day) probability as the method is not designed to preserve this property and therefore:

$$F(0) \neq \frac{n_0}{n_1 + n_0} \tag{3.1.15}$$

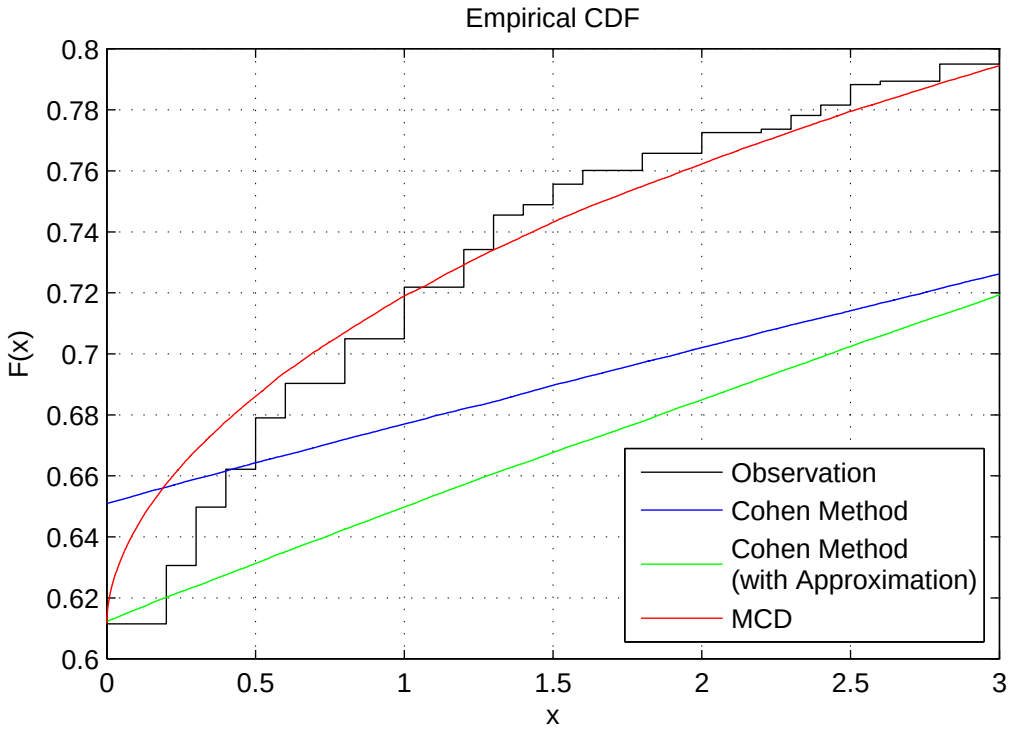


Fig. 3.5: CDFs of observation and simulation of models at station 1 (an extended view of Figure 3.4c.

As shown in the example in Figure 3.5, when $x = 0$, the $F(x)$ of Cohen method does not match the observed probability of dry-day, which is 0.61. On the other hand, the remaining two methods that are based on the approximation do match well at $x = 0$, just as one would expect.

There are several advantages of using a censored normal distribution. First, the method provides a generally good fit to the observations. Second, the method is relatively simple and easy to implement as it does not require the solution of two equations simultaneously. Simplicity and efficiency are important for engineering practitioners, especially when large data sets are involved. The use of a censored normal distribution can match well the probability of observed wet-day. The drawback

is an additional parameter β has to be estimated.

It is worthy to note that the parameter ξ in a censored normal distribution can be estimated based on the occurrences of wet-day and dry-day in the relevant historical data. The parameters (mean and variance) of a censored normal distribution can be estimated at each site for a specific time period (i.e., a Julian day, a week, or a month). In this study, the Julian day has been used.

3.1.2 Step 2: Estimation of Correlation in a Bivariate Censored Normal Distribution

A bivariate situation arises when two precipitation gauging sites are considered and, in such cases, the probability distribution of precipitation data at each site is represented by a censored normal distributions. There is some association or dependence between precipitation amounts at any of two sites and such an association is customarily expressed through the coefficient of cross correlation (or simply as correlation) designated as ρ. The distribution of precipitation amounts at any of two sites can be expressed through the joint censored normal probability density function containing ρ. The estimation of parameter ρ is not simple, as the historical precipitation sequences contain a significant number of zero's (in addition to positive values) and routine methods (e.g., Pearson correlation) are inappropriate. This problem has long been recognized and various methods have been proposed for the estimation of ρ. For a censored normal distribution *Singh* [1960] suggested a method in which maximum likelihood based equations are iteratively solved. In hydrologic literature, *Bardossy* [1992] proposed a method for the estimation of correlation in a bivariate distribution, which is somewhat cumbersome. The estimation procedure becomes unwieldy in higher dimensions (i.e., multivariate cases) because of the existence of non-positive

definite matrices, which may arise due to missing observations and/or numerical inaccuracies in the formation of multivariate correlation matrices. In this section, an optimization approach based on the method of maximum likelihood is proposed.

For the estimation of parameter ρ, the proposed method requires information on estimates of mean and variance of each of univariate censored normal distribution corresponding to any two gauged sites, and probabilities of four joint events (wet-wet, dry-wet, dry-dry, and wet-dry). To estimate ρ, the maximum likelihood function as given below is used.

$$L(\rho|\boldsymbol{x},\boldsymbol{y},\boldsymbol{\mu},\boldsymbol{\sigma}) \propto \prod_{i=1}^{N_1} f(x_i,y_i|\rho) \times \prod_{i=1}^{N_2} \int_{-\infty}^{0} f(x,y_i|\rho)dx \times [F(0,0|\rho)]^{N_3} \times \prod_{i=1}^{N_4} \int_{-\infty}^{0} f(x_i,y|\rho)dy \tag{3.1.16}$$

where N_j refers to the number of joint events (wet-wet, dry-wet, dry-dry, and wet-dry) that fall into quadrant j.

The bivariate censored normal distribution of joint events is divided into four quadrants: (1) wet-wet events in the first quadrant, (2) dry-wet events in the second quadrant, (3) dry-dry events in third quadrant, and (4) wet-dry events in the fourth quadrant. Using historical data at each site, parameters (mean and variance) of a censored normal distribution are obtained followed by the computation of ρ corresponding to any combination of two sites by maximizing the log likelihood function as follows:

$$\hat{\rho} = \arg\max_{\rho} \log L(\rho) \tag{3.1.17}$$

In principle, the maximization is straightforward as only one unknown ρ is involved. The challenge is to compute the integral terms in Equation 3.1.16. In this

research, however, the parameter ρ is estimated numerically using Matlab.

It is worthy to note that, Equation 3.1.16 does not produce the joint maximum likelihood estimation for (μ, σ, and β) but is used for its convenience. The μ and σ are known and assumed to be $\hat{\mu}$ and $\hat{\sigma}$.

3.1.3 Step 3: Estimation of Parameters of a Multivariate Censored Normal Distribution

The parameters (mean vector, covariance matrix) of a multivariate censored normal distribution corresponding to more than any two precipitation sites can be obtained by combining the mean of individual sites and covariances for all possible pairs of any combination of two sites. Each covariance is calculated from the variances and correlation of two sites. The possible pairs of covariance are designated as σ_{12} (sites 1 and 2), σ_{13} (sites 1 and 3), σ_{23} (sites 2 and 3), and so on. The estimated values of the covariance corresponding to pair-wise sites constitute elements of the covariance matrix (i.e., a multivariate parameter containing values of spatial covariance among all sites), that will preserve the spatial dependence of precipitation at multiple sites. Because the covariances are estimated in a pair-wise manner, the covariance matrix may not be positive definite. Some matrix modification approaches are available to transform a non-positive definite covariance matrix in to a positive definite [*Rasmussen et al.*, 1996] and some of them are appropriately utilized in this thesis.

The problem of non-positive definite matrix commonly occurs when covariances are estimated from data series of uneven lengths or with missing observations. In order to transform a non-positive definite matrix into a positive definite matrix, one can use either of two methods described as follows.

The first method involves increasing the diagonal elements of the covariance matrix

until the matrix becomes positive definite. The rationale behind this method is that some of the off-diagonal elements in the covariance matrix may be too large. The resulting matrix must then be re-scaled so that the diagonal elements in the matrix will return to the original values.

The second method is based on the adjustment of eigenvalues of the covariance matrix. The reasoning behind this method is that when a covariance matrix is non-positive definite then some of the associated eigenvalues of the matrix are negative. Hence, the negative eigenvalues are set equal to zero to recalculate elements of the covariance matrix based on the modified eigenvalues. Again, some re-scaling will be required to ensure that the original variances are preserved.

Let i be a selected time period of interest (e.g., a Julian day or a month). In this approach, i refers to a specific Julian day meaning that there are 365 sets of parameters corresponding to the 365 Julian days. The modification of the covariance matrix at a specific day i through eigen decomposition is as follows. We write

$$\Sigma_i = Q_i^T \Lambda_i Q_i \tag{3.1.18}$$

where Σ_i is the covariance matrix with eigenvectors, $q_{i,k}$ $(k = 1, ..., K)$; K is the number of sites involved in the study. Q_i is the square $(K \times K)$ matrix whose k^{th} column is the eigenvector $q_{i,k}$ of Σ_i. Λ_i is a diagonal matrix whose diagonal elements are the corresponding eigenvalues, i.e., $\Lambda_{i,kk} = \lambda_{i,k}$. If negative eigenvalues are present, these are set to zero as follows.

$$\lambda_{i,k}^* = \begin{cases} \lambda_{i,k} & : \lambda_{i,k} \geq 0 \\ 0 & : \lambda_{i,k} < 0 \end{cases}$$

where $\lambda_{i,k}^*$ is the modified $\lambda_{i,k}$.

With the modified Λ_i^* matrix, Σ_i is recalculated using (3.1.18) and a modified Σ_i^* is obtained. Additional re-scaling is required to ensure that the original variances are preserved. The re-scaled Σ_i^* becomes:

$$\Sigma_i^{**} = P_i^{1/2} \Sigma_i^* P_i^{1/2} \tag{3.1.19}$$

where Pi is the square $(K \times K)$ diagonal matrix whose k^{th} element is $\sigma_{k,k}/\sigma_{k,k}^*$.

Re-scaling Σ_i^* preserves the diagonal elements of the Σ_i and the variance of each site can be preserved. However, re-scaling Σ_i^* will also increase the off-diagonal elements of Σ_i that sacrifices the accuracy of the correlation estimates in Σ_i^*. In this thesis, Σ_i^* is used without re-scaling to ensure that the correlations of the matrix are reasonably preserved; however, the variances may be inaccurate. For simplicity, the modified matrix Σ_i^* hereafter is called Σ_i.

The estimated mean vector and covariance matrix of a multivariate censored normal distribution hereafter for brevity is referred to as simply the multivariate censored distribution (MCD). In brief, the above steps can be listed as follows.

- Estimation of univariate parameters (μ and σ) of each site by the Cohen's approximation method with power transformation,
- Estimation of bivariate parameters (ρ) of pairwise sites,
- Formulation of multivariate parameters ($\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$) with covariance matrix modification.

3.2 Model Development of the Weather Generation Using Multivariate Censored Distribution

The model developed in this section involves three components: 1. simulation of multivariate normally distributed data using parameters of MCD, 2. introduction of a periodic function to reduce the number of parameters involved, and 3. transformation of normally distributed data to synthetic precipitation amounts. In this thesis, hereafter, the data simulation model based on MCD is known as the Weather Generation-Multivariate Censored Distribution (WG-MCD) Model.

The first step in simulating precipitation at K sites involves simulation from a multivariate censored normal distribution with parameters $\boldsymbol{\mu_i}$ and $\boldsymbol{\Sigma_i}$, $i = 1, ..., 365$. In practice, the simulation is accomplished using the following model structure.

$$\mathbf{u_t} = \mathbf{H_i z} + \boldsymbol{\mu_i} \tag{3.2.1}$$

where t denotes the day that the precipitation is to be simulated, i refers to a specific Julian day, where i and t are related as follows:

$$i = 365(t/365 - \lfloor t/365 \rfloor) \tag{3.2.2}$$

where $\lfloor t/365 \rfloor$ is the largest integer not greater than $t/365$.

The term $\boldsymbol{\mu_i}$ is the estimated mean vector of MCD at a selected time period i, $\mathbf{z}$ is a vector of K variables drawn from independent standard normal distributions, and $\mathbf{H_i}$ is the Cholesky factorization of covariance matrix $\mathbf{\Sigma_{0i}}$. The use of subscribe 0 is to differentiate this term from the lag-1 covariance matrix $\mathbf{\Sigma_{1i}}$ that is to be discussed in the next chapter. In this thesis, this term is obtained from the modified covariance matrix stated in the step 3 of subsection 3.1.3. The model corresponding to a selected

time period i for the analysis therefore depends on the day t to be simulated where leap years are ignored.

The Cholesky factorization is given by:

$$\mathbf{\Sigma_{0i}} = \mathbf{H_i}\mathbf{H_i'} \tag{3.2.3}$$

where $\mathbf{H_i'}$ is the transpose of $\mathbf{H_i}$.

The Cholesky factorization decomposes $\mathbf{\Sigma_{0i}}$ into $\mathbf{H_i}$ and thus $\mathbf{H_i}$ is similar to the square root of the covariance matrix. $\mathbf{\Sigma_{0i}}$ must be a positive semi-definite matrix, and therefore suitable matrix modifications as discussed in Section 3.1.3 are invoked. It is noted that $\mathbf{\Sigma_{0i}}$ is the lag-0 covariance matrix which is different from the lag-1 covariance matrix as discussed in the next chapter.

A relevant parameter matrix $\mathbf{H_i}$ must be obtained for each Julian day. Each $\mathbf{H_i}$ matrix consists of $K \times K$ elements. To reduce the number of parameters involved, the 365 values of each entry (m, n) in the $\mathbf{H_i}$ matrices will be represented by a periodic function.

A series of elements from the matrices $\mathbf{H_1}$ to $\mathbf{H_{365}}$ corresponding to row m and column n is represented by the vector $\mathbf{h}_{mn} = \{h_{1,mn}, h_{2,mn}, ..., h_{365,mn}\}$. m and n are the selected site numbers in the range 1 to K. The value of h on day i can be approximated by the generalized periodic function in Equation 3.2.4 if one harmonic is considered.

$$\hat{h}_{i,mn} = [1 \quad \cos(2\pi i/365) \quad \sin(2\pi i/365)] \begin{bmatrix} \alpha_{0,mn} \\ \alpha_{1,mn} \\ \alpha_{2,mn} \end{bmatrix} \tag{3.2.4}$$

The parameter vector $\boldsymbol{\alpha}_{mn} = [\alpha_{0,mn}\ \alpha_{1,mn}\ \alpha_{2,mn}]'$ can be estimated by ordinary

least-square as follows:

$$\boldsymbol{\alpha}_{mn} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{3.2.5}$$

where $\mathbf{y}$ is $\mathbf{h}'_{mn}$ and $\mathbf{X}$ is

$$\mathbf{X} = \begin{bmatrix} 1 & \cos(2\pi \cdot 1/365) & \sin(2\pi \cdot 1/365) \\ 1 & \cos(2\pi \cdot 2/365) & \sin(2\pi \cdot 2/365) \\ \vdots & \vdots & \vdots \\ 1 & \cos(2\pi \cdot 365/365) & \sin(2\pi \cdot 365/365) \end{bmatrix} \tag{3.2.6}$$

Using Equation 3.2.5, the number of parameters required to represent $\mathbf{h}_{mn}$ in Equation 3.2.4 is reduced from 365 to 3 when only one harmonic is involved.

Equation 3.2.6 can be generalized into an equation for any number of harmonics (R) by taking

$$\hat{h}_{i,mn} = \alpha_{0,mn} + \sum_{r=1}^{R} [\alpha_{2r-1,mn} \cos(2r\pi t/365) + \alpha_{2r,mn} \sin(2r\pi t/365)] \tag{3.2.7}$$

where R defines the number of harmonics. In this thesis, a periodic function with four harmonics is adopted to depict the change of weather patterns in four seasons. The use of four harmonics is a sensible choice because four seasons are marked by changes in weather that captures the physical characteristics of seasonal weather patterns into the stochastic weather generation model. Although adopting a larger number of harmonics provides a better fit to the series of elements from the parameter matrices, using four harmonics is considerably parsimonious and sufficient to cover up the periodic function with less harmonics.

For simulated positive values, an inverse power transformation is applied to obtain

actual precipitation amounts and simulated negative values are replaced by zeros to represent the dry days as expressed below.

$$s_t = \begin{cases} u_t^{\beta} & : u_t \geq 0 \\ 0 & : u_t < 0 \end{cases}$$

where u is the simulated data obtained from Equation 3.2.1.

3.3 Evaluation of the WG-MCD model

The evaluation of the proposed WG-MCD model consists of two steps: model verification and model validation. The aim of model verification is to ensure that computer coding of mathematical structure and logistics of the model are correctly implemented. Idealised/control data from a normal distribution with known parameters are used to ensure that the model is properly coded. Model validation examines the adequacy of model in preserving the statistical characteristics of observed precipitation data.

3.3.1 Verification of the Computer Coding

For verification purposes, a control/idealised precipitation data set containing 10,000 values was synthesized from a bivariate normal distribution with arbitrarily assigned parameters ($X_1 = N$(-2.5, 3.5), $X_2 = N$(3.3, 4.7), with correlation parameter ρ_{12}= -0.74) as shown in Figure 3.6. Synthesized negative values of the control data set were replaced by zero values to mimic the dry-day precipitation events. The probability associated with zero values and positive values in the control/idealised precipitation data are used as input into a computer program for providing numerical values (-∞ to +∞) for a normal probability curve. This computer program, in essence, represents a computational algorithm of the multivariate censored normal distribution. The

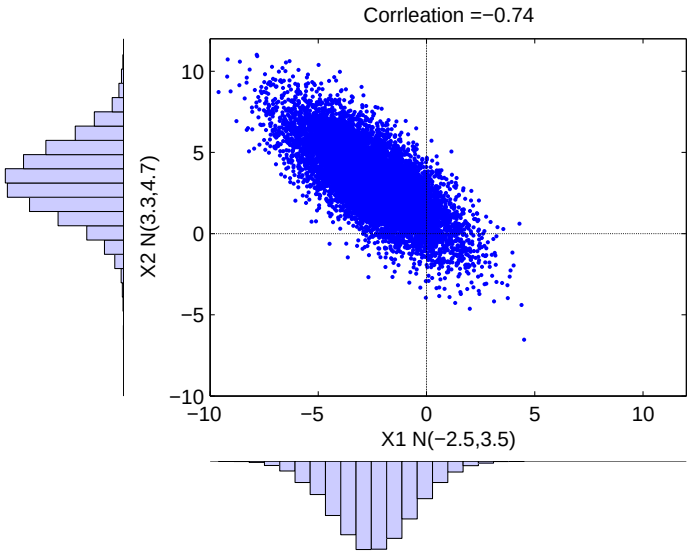


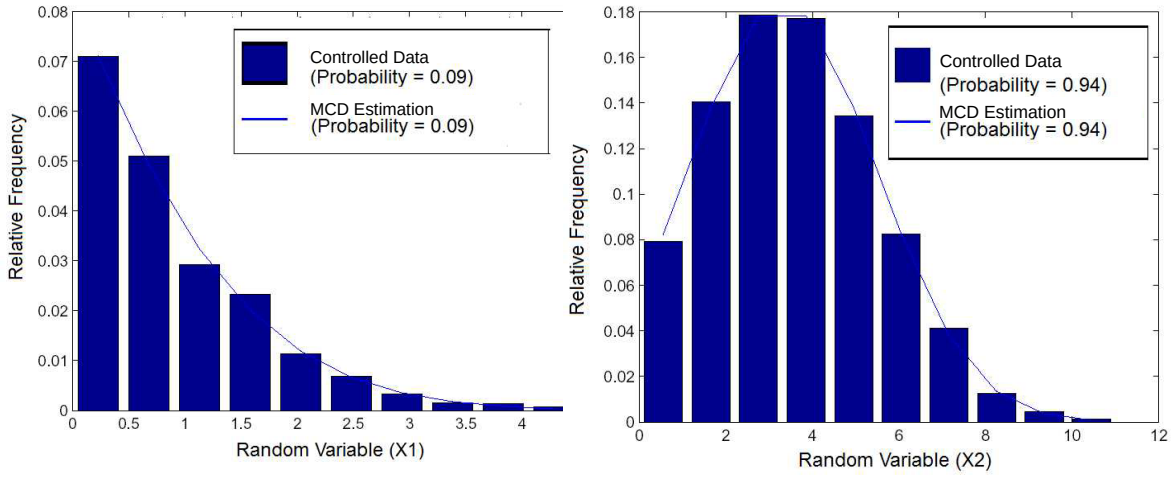
Fig. 3.6: Bivariate normal distribution.

estimated parameters from the MCD were then compared with the control parameters for verifying the logistics and coding in the computer program of the WG-MCD model.

For illustrative purposes, the verification of the WG-MCD model's computer coding will be examined first for two marginal/univariate censored normal distribution each for X_1 and X_2 variables, and then for a bivariate censored normal distribution corresponding X_1 and X_2 variables.

Step 1: Verification of Marginal/Univariate Censored Normal Distribution

In Figure 3.7, the standardized histogram corresponds to normally distributed data (labeled as "Controlled Data") which was generated using "known (i.e., control) parameters". Likewise, the probability density function obtained from the WG-MCD model (labeled "MCD Estimation") are compared visually while using the positive (i.e., above zero) values. For the variables X_1, and X_2, the control and estimated



(a) Variable X1. *(b)* Variable X2.

Fig. 3.7: Normalized histogram and distribution of synthetic variables.

parameters respectively are: N(-2.5, 3.5; -2.6, 3.8) and N(3.3, 4.7; 3.3, 4.7) that refer to N(controlled μ_1, controlled σ_1^2; estimated μ_1, estimated σ_1^2) and N(controlled μ_2, controlled σ_2^2; estimated μ_2, estimated σ_2^2). Table 3.1 summarizes the differences between the control and estimated parameters. For both variables, the probabilities of wet-day in particular are well estimated by the WG-MCD model, as one would expect from the use of the approximation method in Equation 3.1.10. The relatively higher error observed for the variable $X1$ is due to the limitation of data points for analysis, as only 9% of the data correspond to positive values (i.e., wet days).

Steps 2: Verification of a Bivariate Censored Normal Distribution

A plot of the bivariate distribution corresponding to the control and the WG-MCD model based parameters are shown in Figure 3.8. The contour lines in the figure refer to the distributions corresponding to the WG-MCD model and control parameters based simulated data. The very outer contour lines in the figure are not very smooth due to sensitivity of the plotting algorithm in Matlab to extreme values in the data.

Tab. 3.1: Summary of results of the estimation of parameters of bivariate censored normal distribution.

Parameters	Controlled	MCD Estimated	Error (%)
μ_1	-2.5	-2.6	4.0
σ_1^2	3.5	3.8	7.8
$P(X1 > 0)$	0.1	0.1	0
μ_2	3.3	3.3	0.1
σ_2^2	4.7	4.7	0.2
$P(X2 > 0)$	0.94	0.94	0
ρ_{12}	-0.74	-0.75	1.8

The value of correlation (ρ) between variables (X_1 and X_2) for the control and simulated data based on the WG-MCD model were respectively found to be - 0.74 and -0.75.

Quadrants $P1, P2, P3$, and $P4$ in Figure 3.8 refer to the percentage of data lying in a specific quadrant of the bivariate normal distribution. For example, $P1 = 5.5\%$ means only 5.5% of the bivariate data are observed as positive for both variables. As shown in the figure, the probability differences of control and estimated data at the four quadrants are less than 1% which may be deemed insignificant. The above illustrations based on the hypothetical data amply demonstrate that the computer coding is correctly formulated and can be used for running the WG-MCD model using the historically observed data of daily precipitation.

3.3.2 Validation of the WG-MCD Model

A model must be validated using external data, i.e. the data which have not been used in the exercise related to verification of the computer coding. Therefore, the historically recorded daily precipitation data were used. The details of the the relevant gauging sites for monitoring of the precipitation data are described in the following

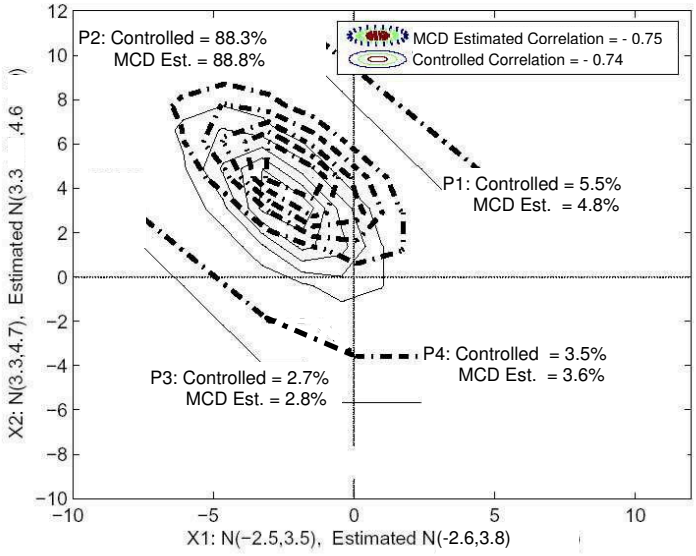


Fig. 3.8: Contour plot between estimated and theoretical joint distributions.

Tab. 3.2: Geographical information on selected stations.

Station	Name	Station Number	Latitude	Longitude	Elevation (m)
1	Deer Wood	5020720	49.4N	98.9W	338
2	Emerson	5020880	49.4N	97.2W	238
3	Morden	5021848	49.1N	98.5W	298
4	Plum Coulee	5022245	49.3N	97.8W	265
5	Steinbach	5022780	49.2N	96.6W	254
6	Altona	5020040	49.6N	97.3W	248
7	Morris 2	5021965	49.6N	97.9W	238
8	Minnedosa	5011760	50.6N	99.0W	521
9	Beausejour	5030160	50.2N	96.8W	251
10	Arborg	5030080	50.8N	96.3W	302

section.

Historical Precipitation

Daily precipitation records from year 1961 to 1990 from 10 stations located in southern Manitoba, Canada, were utilized in this thesis. Availability of sufficient and reliable daily precipitation data is the reason for the selection of this region for the analysis. The 10 stations are described in Table 3.2 and their locations are shown in Figure 3.9. Stations 1 to 7 are located inside the Red River Basin, station 8 is located in the Assiniboine River Basin, and stations 9 and 10 are located inside the Winnipeg River Basin.

Precipitation records were obtained from the Canadian Daily Climate Data 2002 West CD-ROM [*Environment Canada*, 2007]. Daily precipitation records obtained from the period between 1961 and 1990 is common in precipitation and climate change studies, as it is neither too short, nor too recent to include a strong global change signal [*Wilby et al.*, 2004]. With 30 years of daily precipitation records, a sample size of 30 is available to estimate the parameters for each Julian day at each station.

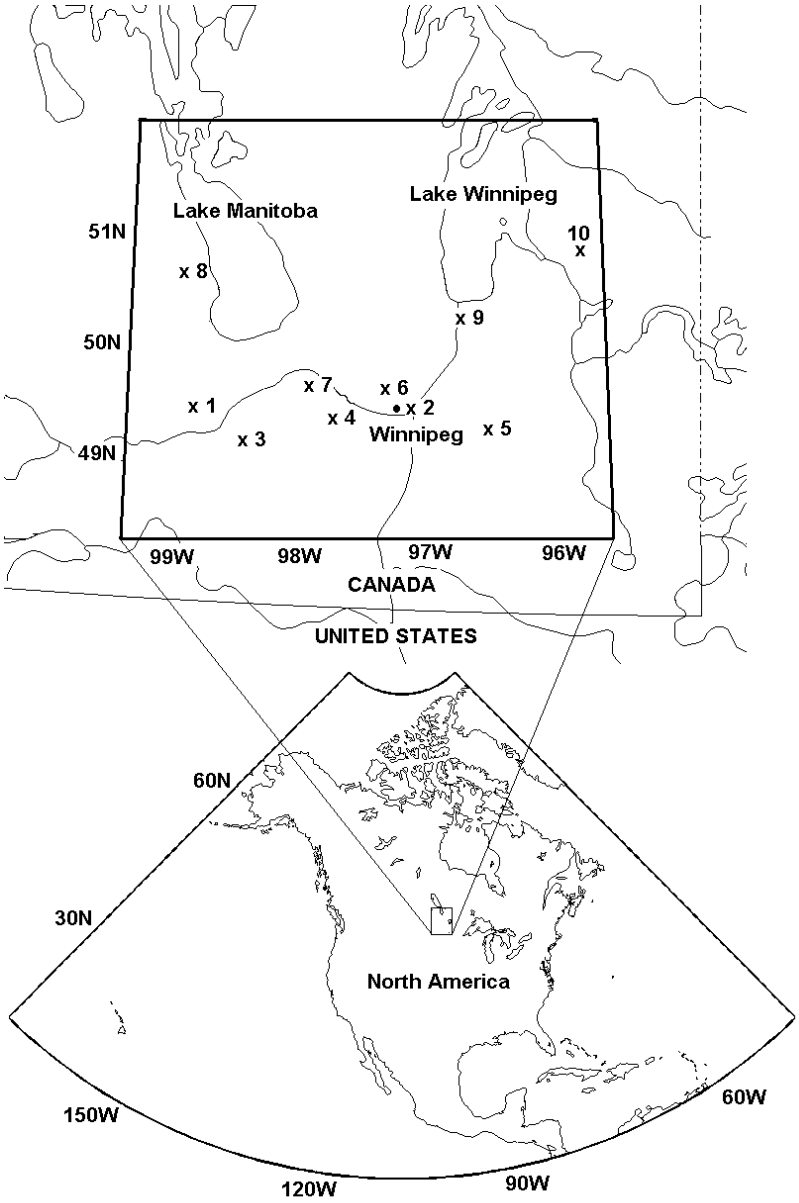


Fig. 3.9: Location of precipitation stations in the Winnipeg region, Manitoba, Canada.

At 10 stations with 30-years records, a total of 109,575 daily precipitation records were available for the analysis. Of these, 29,069 (27%) records are above zero; 75,986 (69%) records are zeros, and 4,520 (4%) records are missing. The number of above-zero records was used for the estimation of parameters (mean vectors and correlation matrices) of the the WG-MCD model and its variant forms.

Visual Comparison of Frequency Distributions

Since the WG-MCD model essentially deals with normal probability distributions, it is imperative that the observed data must be fitted by a censored normal pdf. In order to do that the observed precipitation of the wet days must be normalized. In general, it is well known that daily precipitation data are skewed and a power transformation has been found to be a successful method for normalizing the skewed wet day precipitations. Therefore, a power transformation was used in the present case and the majority of data from all sites showed good results with the power transformation. A simple measure of the usefulness of the transformation is to compare the graphical normal probability plot of the non-transformed and transformed data as shown in Figure 3.10. A straight line after transformation of data set on the normal probability graph indicates a good approximation to a normal distribution in the graph of the transformed data. In other words, a near straight line on a normal probability paper, affirmed the need of the power transformation for normalization the precipitation data.

When such a procedure is carried out for all sites, the transformed data would satisfy the characteristics of a multivariate censored normal distribution (MCD). The extrapolated normal pdf will have parameters mean and variance which will be different from the raw precipitation data of wet days. In this context, MCD(m) denotes the

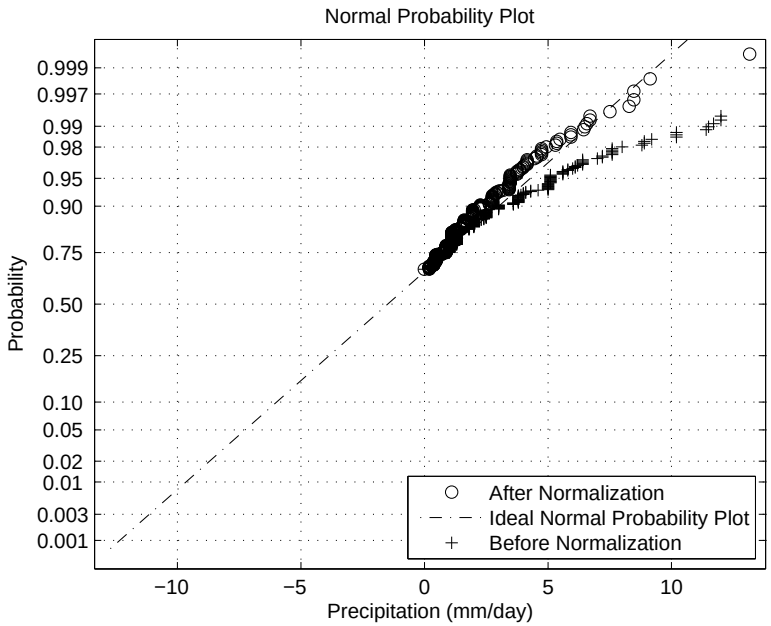


Fig. 3.10: Normal probability plot of MCD(m,1) for the month of January at station 1. (Annotation of MCD(.) refers to Table 3.3.)

multivariate censored normal distribution, in which parameters have been obtained using the historical data as input. The distribution consists of negative and positive values that may result in negative values of mean.

After the parameters of MCD have been estimated, they can now be substituted in Equation 3.2.1 and a string of say 1,000 normally distributed numbers can be generated. The negative values can be discarded and positive values are retained. The positive values can be back transformed (through the inverse of power transformation, Equation 3.1.12). The distribution obtained from the above simulation route is designated as MCD(s), where s denotes that parameters are based on simulated values. The mean and variance of this simulated data can be estimated from the back transformed positive values. Table 3.3 gives the notational conventions for MCD(.).

For a good simulation, the mean and variance of the observed and simulated data

Tab. 3.3: Annotation of types of data used for the calculation of attributes.

Notation	Description
MCD(m)	Refers to the multivariate censored normal distribution generated by the WG-MCD model, in which attributes (e.g., μ and σ) have been estimated using the historical data as input. The distribution consists of negative and positive values. Such attributes can also be calculated using a sufficient size of simulated data generated by the model.
MCD(s)	Refers to the right portion of a truncated point under a multivariate censored normal distribution generated by the WG-MCD model, in which s denotes that attributes have been calculated using the simulated data generated by the model. The distribution only consists of above-zero values (wet day) that has been back transformed by the inverse of power transformation, Equation 3.1.12.
MCD(m,1) or MCD(s,1)	Equivalent to the MCD(m) or MCD(s) except the model does not involve the use of the periodic function discussed in the Section 3.2.4.
Notes:	Annotation is not only restricted to WG-MCD model, but also applicable to the model to be discussed in Chapter 4.

should show 1:1 correspondence. To elucidate the point, for the month of January at station 1, the mean and variance of observed wet daily data are respectively 2.92 and 11.63 (non-normal pdf). The corresponding estimates from MCD(m) model respectively turned out to be -1.66 and 11.19. The final simulated parameters of MCD(s) were 3.14 and 9.86 respectively (after ignoring negative values and back transformation). The model seems to provide satisfactory results as the observed and simulated parameters are in close proximity of each other. This comparison has been done in chapter 4 for all the stations and all scenarios. Another example is illustrated in Figure 3.11 using the precipitation data for all stations for the daily data of the month of January. The distribution of actual precipitation is displayed through histograms, whereas the simulated counterparts are shown by the dashed curve generated by the WG-MCD model. The close correspondence between observed histograms and simulated curve is a good evidence of the capability of the model to simulate the frequency distribution of the daily precipitation data.

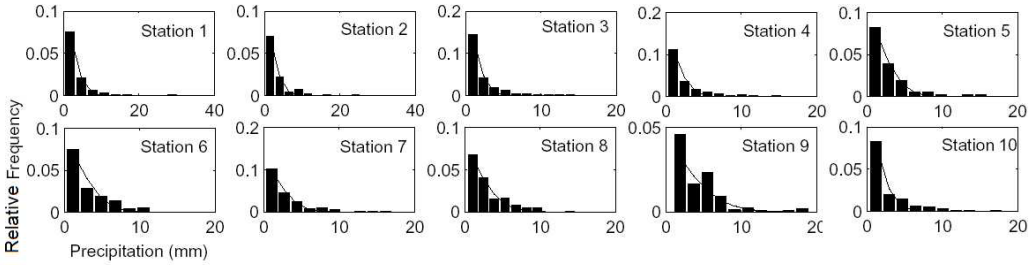


Fig. 3.11: Observed and estimated probability distributions of MCD(m,1) for the month of January.

Description of Non-Parametric Tests For Frequency Distributions

Since preservation of the frequency distribution of the observed (historical) data is an important feature, the capability of the WG-MCD(s) model was further evaluated using the statistical measures. Since the observed data follow some skewed distribution, one recourse is to follow the nonparametric methods of testing the frequency distribution. For testing the equivalence of frequency distributions of simulated and historical data, two commonly used test are: (a) Kolomogirov-Smirnov (K-S) test (b) Mann-Whiteney(M-W) test. The K-S and M-W tests are nonparametric and are better suited for the comparison of two mixed distributions. The K-S test is sensitive to shape, spread, and median in two distributions and will result in a small P value when there is a substantial departure from the hypothesized distribution. In contrast, the M-W test is sensitive only to the change in median value. A short description of the aforesaid statistical measures follows.

Kolomogorov-Smirnov (K-S) Test

The null hypothesis in the K-S test is that the true underlying CDFs (cumulative distribution functions) for the simulated and historical data are identical, i.e.,

$$H_0 : \ F_s = F_o \tag{3.3.1}$$

where H_0 is the null hypothesis. F_s and F_o are the empirical distribution functions of simulated and historical data. The alternative hypothesis is that the distribution of simulated data is not the same as the distribution of the historical data.

$$H_1 : \ F_s \neq F_o \tag{3.3.2}$$

where H_1 is the alternative hypothesis.

The K-S test statistic is based on the maximum separation between the two empirical distribution functions:

$$D = \sup |F_s - F_o| \tag{3.3.3}$$

The K-S P-values can be obtained and interpreted using the table found in DeGroot [*DeGroot*, 1975] in which values are based on the modified D_{mn} values. A rejection or non-rejection of the null hypothesis would depend on the selected acceptable level of significance (critical value) chosen by the practitioner.

Mann-Whitney(M-W) Test

The Mann-Whitney U-test first ranks all the values of two data sets from low to high, and then computes a P-value depending on level of the discrepancy between ranks of simulated and historical data sets. In this study, the M-W test is used to test whether the underlying probability distributions of simulated and historical data are identical and therefore, the null hypothesis is that the distributions of simulated data and historical data are identical. The details of the aforesaid non-parametric tests are

well documented in standard hydrological texts, however, are also reproduced below for posteriority:

$$H_0 : \ F_s = F_o \tag{3.3.4}$$

where H_0 is the null hypothesis. F_s and F_o are the empirical distribution functions of simulated and historical data sets. The alternative hypothesis is that the distribution of the simulated data is not the same as the distribution of the historical data:

$$H_1 : \ F_s \neq F_o \tag{3.3.5}$$

where H_1 is the alternative hypothesis.

The M-W test statistic is denoted U and is the smaller of U_1 and U_2, as defined below:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \tag{3.3.6}$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 \tag{3.3.7}$$

where n_1 and n_2 are the sizes of the two data sets. R_1 and R_2 denote the respective sum of ranks in the two data sets. The P-values of the M-W can be obtained using the critical values in table for the M-W test [*Hollander and Wolfe*, 1999].

Interpretation of Frequency Distributions Based on K-S and M-W Tests

For the K-S and M-W based null hypothesis tests, when the P-value is less than 5%, the null hypothesis is rejected in this case study. The tests were performed on the daily data within a specific month and at a specific site, corresponding to model

Tab. 3.4: Results summary of two-sample Kolmogorov-Smirnov test.

Level of Significance (P-values)

	Station									
Month	1	2	3	4	5	6	7	8	9	10
Jan	0.37	0.23	0.34	0.18	0.32	0.27	0.37	0.48	0.15	0.17
Feb	0.24	0.32	0.22	0.52	0.53	0.20	0.15	0.37	0.12	0.28
Mar	0.23	0.49	0.36	0.31	0.52	0.30	0.41	0.31	0.25	0.40
Apr	0.35	0.70	0.46	0.58	0.65	0.54	0.64	0.60	0.23	0.47
May	0.19	0.52	0.21	0.40	0.67	0.54	0.50	0.44	0.42	0.52
Jun	0.51	0.60	0.59	0.52	0.55	0.58	0.49	0.44	0.51	0.35
Jul	0.44	0.49	0.57	0.46	0.50	0.40	0.48	0.42	0.34	0.46
Aug	0.42	0.11	0.39	0.23	0.41	0.50	0.21	0.43	0.35	0.53
Sep	0.47	0.60	0.46	0.40	0.22	0.32	0.35	0.55	0.18	0.06
Oct	0.39	0.42	0.36	0.31	0.30	0.39	0.50	0.42	0.28	0.29
Nov	0.23	0.36	0.31	0.35	0.40	0.22	0.31	0.43	0.26	0.30
Dec	0.24	0.42	0.13	0.34	0.44	0.08	0.37	0.34	0.11	0.33

simulations and historical observations. The calculated P-values of the K-S and M-W tests are respectively reported in Tables 3.4 and 3.5. Since all P-values were found to be outside the rejection region (i.e., larger than 0.05), the null hypothesis is not rejected in any case. In other words, an inference can be drawn that simulated and historical data belong to the same population.

Finally, a bivariate K-S test was conducted using a 5% significance level. The test compares observed and simulated daily data within a specific month and for a specific pair of sites. The calculated P-values of the K-S test are reported in Table 3.6. Since all P-values are outside the rejection region (i.e., larger than 0.05 or 5%), the null hypothesis is not rejected. One can therefore draw the inference that the bivariate frequency distributions involving any two stations also fulfill the premise that these precipitation sequences hail from the same parental population.

Tab. 3.5: Results summary of Mann-Whitney U-test.

Level of Significance (P-values)

	Station									
Month	1	2	3	4	5	6	7	8	9	10
Jan	0.53	0.54	0.58	0.48	0.44	0.50	0.54	0.42	0.54	0.34
Feb	0.13	0.48	0.15	0.51	0.63	0.56	0.13	0.54	0.15	0.27
Mar	0.33	0.49	0.48	0.59	0.59	0.52	0.42	0.39	0.40	0.56
Apr	0.51	0.60	0.56	0.50	0.56	0.47	0.60	0.59	0.15	0.57
May	0.32	0.55	0.40	0.39	0.58	0.58	0.51	0.53	0.34	0.59
Jun	0.55	0.55	0.60	0.55	0.58	0.55	0.57	0.51	0.58	0.44
Jul	0.53	0.46	0.49	0.41	0.54	0.53	0.40	0.51	0.50	0.57
Aug	0.50	0.30	0.55	0.41	0.46	0.47	0.45	0.55	0.46	0.51
Sep	0.58	0.58	0.57	0.33	0.30	0.20	0.24	0.52	0.24	0.21
Oct	0.44	0.57	0.54	0.23	0.42	0.31	0.55	0.33	0.39	0.42
Nov	0.54	0.30	0.45	0.55	0.52	0.55	0.54	0.56	0.34	0.37
Dec	0.25	0.48	0.09	0.56	0.49	0.09	0.51	0.46	0.15	0.57

From a visual examination of the relevant plots and the non-parametric statistical tests, it can be concluded that the WG-MCD model preserves the frequency distribution of the historical data of daily precipitation. The other capabilities of the model such as the simulation of the mean, variance at single and at multiple sites along with cross correlations (spatial dependency) are discussed at length in chapter 4, where its extended version named as WG-MACP is discussed.

3.4 Remarks on Parameters and their variations in the WG-MCD Model

The following describes the manner in which parameters in the WG-MCD model tend to vary and the nature of such variations. The observations are made on the nature of variations in parameters under the sub-heading as temporal and spatial variations.

Tab. 3.6: Results summary of bivariate two-sample Kolmogorov-Smirnov test for January.

Level of Significance (P-values)

	Station								
Station	2	3	4	5	6	7	8	9	10
1	0.34	0.34	0.33	0.35	0.34	0.34	0.34	0.33	0.34
2	-	0.35	0.34	0.34	0.33	0.33	0.35	0.34	0.34
3	-	-	0.32	0.34	0.33	0.33	0.33	0.33	0.33
4	-	-	-	0.33	0.34	0.33	0.33	0.34	0.34
5	-	-	-	-	0.34	0.34	0.33	0.33	0.34
6	-	-	-	-	-	0.35	0.32	0.34	0.33
7	-	-	-	-	-	-	0.33	0.35	0.34
8	-	-	-	-	-	-	-	0.33	0.33
9	-	-	-	-	-	-	-	-	0.33

3.4.1 Temporal Changes of Mean, Maximum Precipitation and Wet Day Probabilities

The daily precipitation data was organized to facilitate the presentation and comparison of the attributes in twelve groups, each group corresponding to a month. For example, if one considers the month of January, in a sample of 30 years, there are 930 days. One can count the number of wet days over these 30 years and calculate the observed mean and standard deviation of daily precipitation for the month of January and likewise for the rest of the months.

The means and variances of MCD(m) estimated from observations at 10 stations are presented as box-plots in Figure 3.12. Each month consists of 10 values of a chosen attribute (10 stations) estimated from the observations. Each attribute is calculated as the average of the MCD(m) parameters estimated based on the Julian-day period within a specific month over the 30 years of records. This figure illustrates regional variation of the estimated attributes over the 12-month period. The lowest and

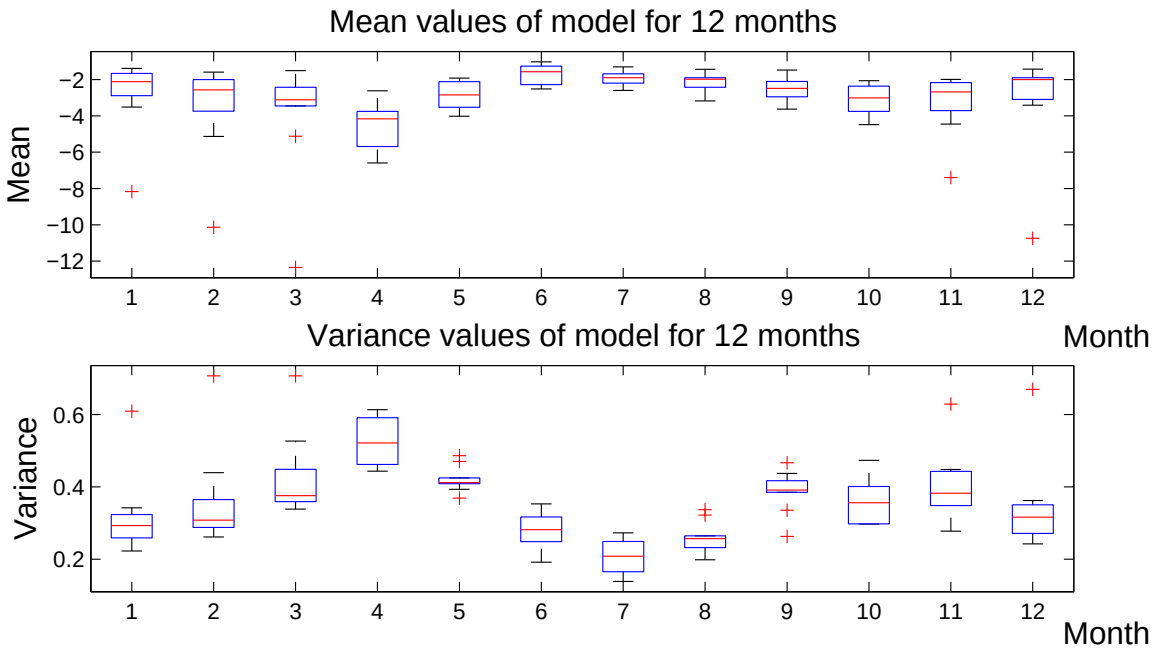


Fig. 3.12: Estimated mean and variance of MCD(m,1) model.

highest mean values observed in April and June suggest that the least and highest average precipitation occurs at the beginning of spring and summer respectively. Outliers observed from November to March suggest a relatively higher local variation in weather during the winter time.

The mean values, maximum values, and wet-day probabilities are presented in Figures 3.13 to 3.15. For clarity, the attributes from Julian-day 1 to 180 at station 1 are presented. In these figures, notation (s) means periodic function is included in the simulation, (s,1) means that periodic function has not been included. The "MCD(s,1) using 2 stations" case in these three figures shows a good fit in the output of the bivariate WG-MCD model, which does not employ periodic functions. As more stations are involved in the analysis (e.g., "MCD(s,1) using 10 stations" in the figure), a slight loss of accuracy in parameter estimation is apparent. Involving

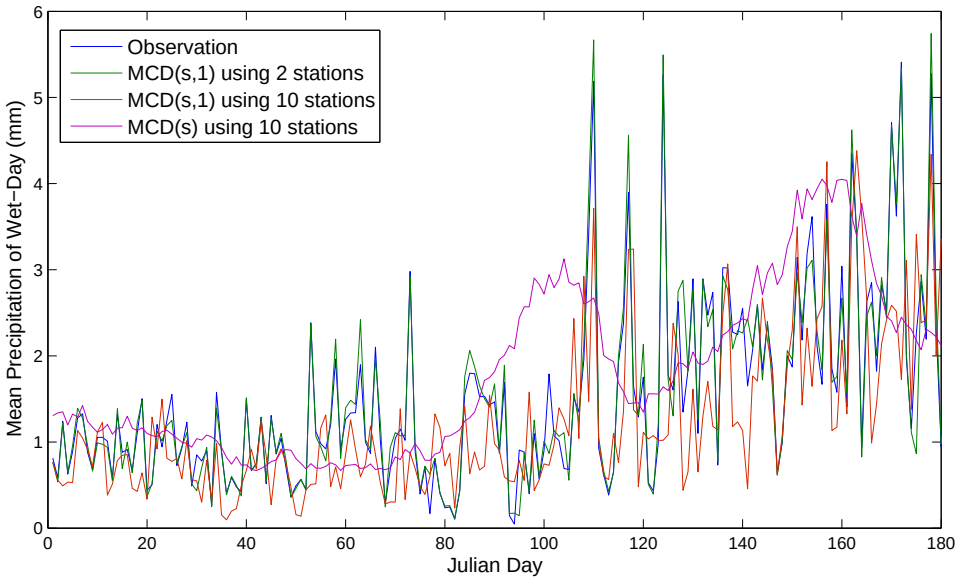


Fig. 3.13: Mean precipitation series of wet-day at station 1.
(Annotation of MCD(.) refers to Table 3.3.)

more stations in the analysis increases the size of the covariance matrix which also increase the chance of elemental inconsistency observed in the matrix, i.e., that the covariance matrix is not positive definite. The inconsistency can be handled by matrix modification, as described in Section 3.1.3. The modification, however, will reduce the model's ability to preserve attributes of the historical observations. Further, it is found that increasing the number of stations and applying the periodic function reduce the accuracy of the WG-MCD model.

3.4.2 Assessment of Characteristics of Spatial Dependency Among Stations

In this section, the results of model validation related to the spatial dependence are presented. To illustrate the correlation of pairwise combinations of the 10 stations, Figure 3.16 shows the variation of the estimated model correlation of the

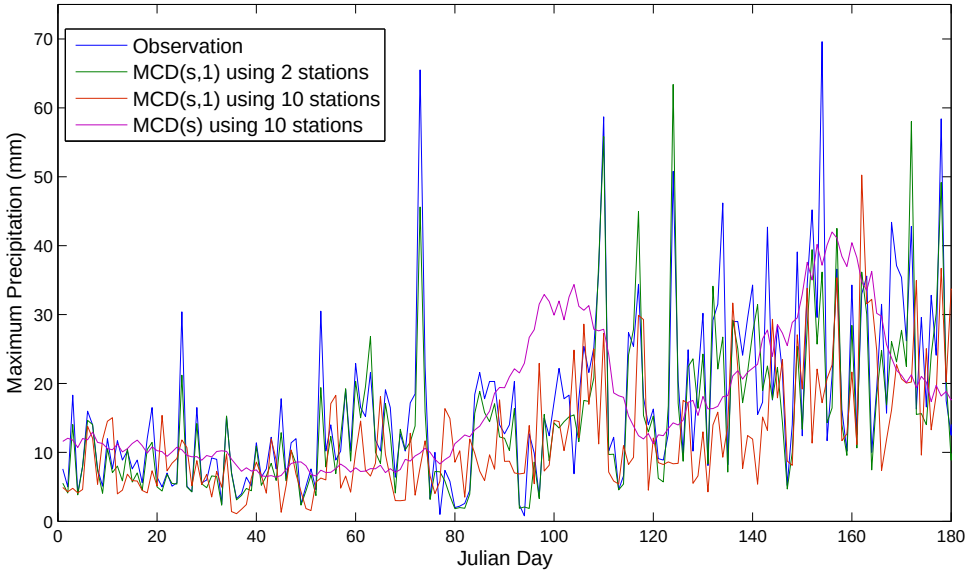


Fig. 3.14: Maximum precipitation series of wet-day at station 1.
(Annotation of MCD(.) refers to Table 3.3.)

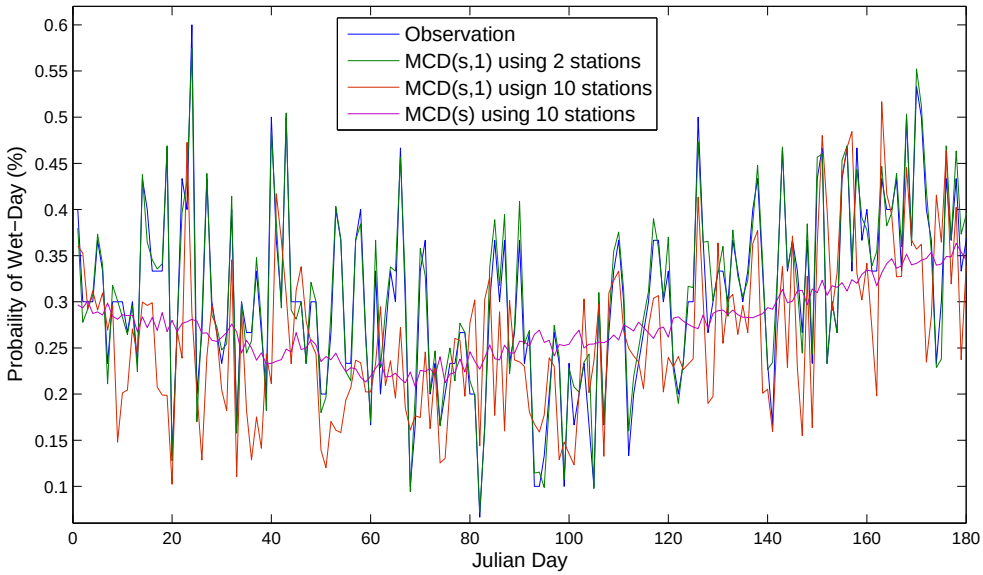


Fig. 3.15: Probability of wet-day series at station 1.
(Annotation of MCD(.) refers to Table 3.3.)

12 months. For a specific month, each box consists of 45 pairwise correlations of the 10 selected stations. In terms of spatial dependence, correlations between pairs of stations (930 values of precipitation data, for example, for the month of January) were computed. Since there are 10 stations, the number of correlations ($\rho_{1,2}, \rho_{1,3}, ...\rho_{1,10}, \rho_{2,3}, \rho_{2,4}, ...\rho_{9,10}$) for a month (say January) will amount to 45. The total number of correlations involving all 12 months will sum to 540 (45×12).

Figure 3.16 shows the range of correlation in this study as this model may not be required in situations where station values are either independent or highly dependent. A highly correlated data set is more likely to lead to problems with covariance matrices that are not positive definite. Highly correlated stations, therefore, should be amalgamated into a single station. On the other hand, in cases where observations are relatively uncorrelated, accounting for spatial dependence may not be desirable. Therefore, these stations can be considered as independent stations for the analysis.

The correlations in Figure 3.16 are relevant for further analysis of the spatial dependence. The variation of the estimated model correlations seems to be consistent over the 12 months and the dependence between stations are smaller during the summer period, where small convective precipitation events dominate.

To gain a better understanding of how data are distributed at two spatially dependent stations, bivariate distributions are presented in Figure 3.17. Stations 1 and 2 for the month of January were selected for illustration. The figure shows that contours of the estimated bivariate distribution reasonably describe the distribution of above-zero observed data. The estimated bivariate distribution depicts: (1) a reasonable match between the observed data and the theoretical contours at the right upper corner for a joint distribution, (2) a good fit between observed histograms and

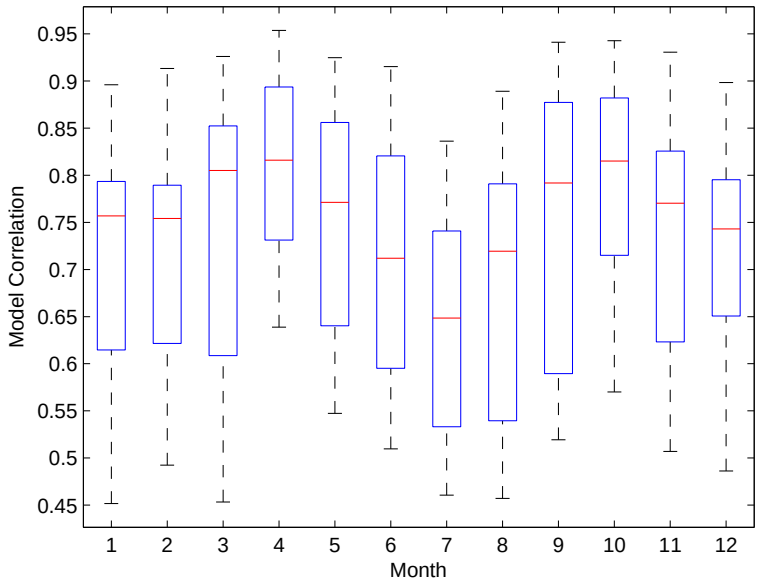


Fig. 3.16: MCD(m,1) Model correlation of the pairwise combinations of the 10 stations.

theoretical pdfs at stations 1 and 2, (3) and a good agreement for the correlation between observed and simulated precipitation.

Additional investigation was conducted to quantify the average correlation change before and after matrix modification. Considering 10 stations involved in the analysis, the percentage change in elements of correlation matrices to transform them into positive definite matrices is summarized in Table 3.7. Each value in the table is calculated based on the average changes to elements of the correlation matrix in a specific month. The minimum and maximum changes are found to be 0.6% and 13.5% for the months of April and November, respectively. The average change is 4.8% and therefore is modest. This means modeling of multivariate scenario is less accurate than the univariate or bivariate scenarios.

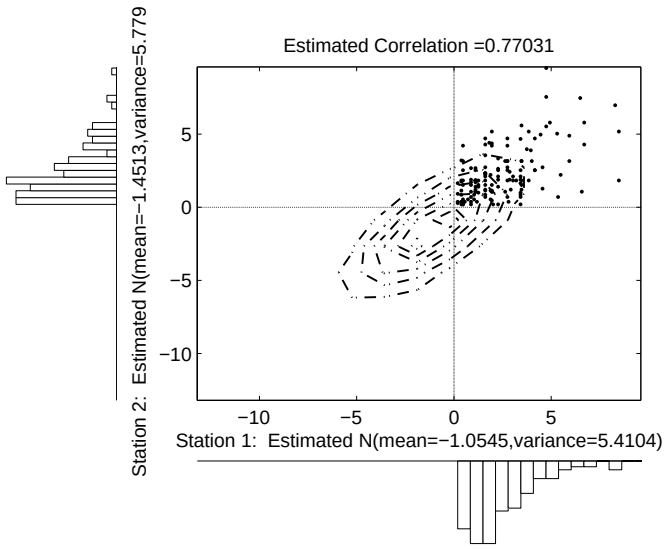


Fig. 3.17: Bivariate distribution of observation and MCD(m,1) between station 1 and 2 for the month of January. (Annotation of MCD(.) refers to Table 3.3.)

Tab. 3.7: Summary of average correlation change (%) in monthly correlation-matrices between before and after modification of positive-definite matrices applied to the parameters of WG-MCD without periodic function.

Month	1	2	3	4	5	6
Change (%)	5.1	4.0	6.5	0.6	4.0	2.9
Month	7	8	9	10	11	12
Change (%)	2.3	4.0	4.6	3.8	13.5	5.7

4. DEVELOPMENT AND EVALUATION OF A WEATHER GENERATION MODEL BASED ON MULTIVARIATE AUTOREGRESSIVE CENSORED PROCESS (WG-MACP)

In the last chapter, focus was directed to the modelling of daily precipitation at multiple sites using the multivariate censored normal probability distribution. The spatial dependence was modeled using the correlation parameter between the precipitation sequences at any two sites. However, it must be recognized that there is a temporal dependence in a precipitation sequence, which was not addressed in the last chapter. Such temporal dependence has been investigated in meteorological and hydrological literature using the notions of an autoregressive process (Markovian process). This chapter deals with the modelling of precipitation sequences with an explicit consideration for temporal dependence while incorporating the role of spatial dependence among multiple sites. The developed spatio-temporal model is referred to as weather generation based on the multivariate autoregressive censored process (WG-MACP).

4.1 Formulation of Multivariate Autoregressive Censored Process (MACP)

The development of the MACP is an extension of the MCD with additional consideration of temporal dependence in daily precipitation sequences (Figure 4.1). This section will discuss step 4 for the development of the MACP, since relevant details

pertaining to steps 1 to 3 have already been presented in context of the MCD development in chapter 3. In succinct terms, the MACP involves, at each time step (i.e., each Julian day), the lag-1 covariance matrix of daily precipitation sequences, while incorporating the mean vector and lag-0 covariance matrix derived from the MCD.

4.2 Model Development for the Weather Generation Using Multivariate Autoregressive Censored Process

In the formulation of a weather generation model based on MACP, the spatially dependent WG-MCD model is fitted into a multivariate autoregressive structure. A lag-1 autoregressive multivariate model is considered for the analysis in view of its promise for modeling the temporal dependence of precipitation sequences [*Bardossy*, 1992]. This formulation involves both spatial as well temporal considerations. Hereafter, such a formulation is referred to as a weather generation model based on multivariate autoregressive censored process (WG-MACP). The mathematical structure of the lag-1 WG-MACP model can be expressed as follows:

$$\mathbf{s_t} = \mathbf{B_i}(\mathbf{s_{t-1}} - \boldsymbol{\mu_i}) + \mathbf{C_i z} + \boldsymbol{\mu_i} \tag{4.2.1}$$

where

$$\mathbf{B_i} = \boldsymbol{\Sigma_{1i}}\boldsymbol{\Sigma_{0i}^{-1}} \tag{4.2.2}$$

and

$$\mathbf{C_i C_i'} = \boldsymbol{\Sigma_{0i}} - \boldsymbol{\Sigma_{1i}}\boldsymbol{\Sigma_{0i}^{-1}}\boldsymbol{\Sigma_{1i}'} \tag{4.2.3}$$

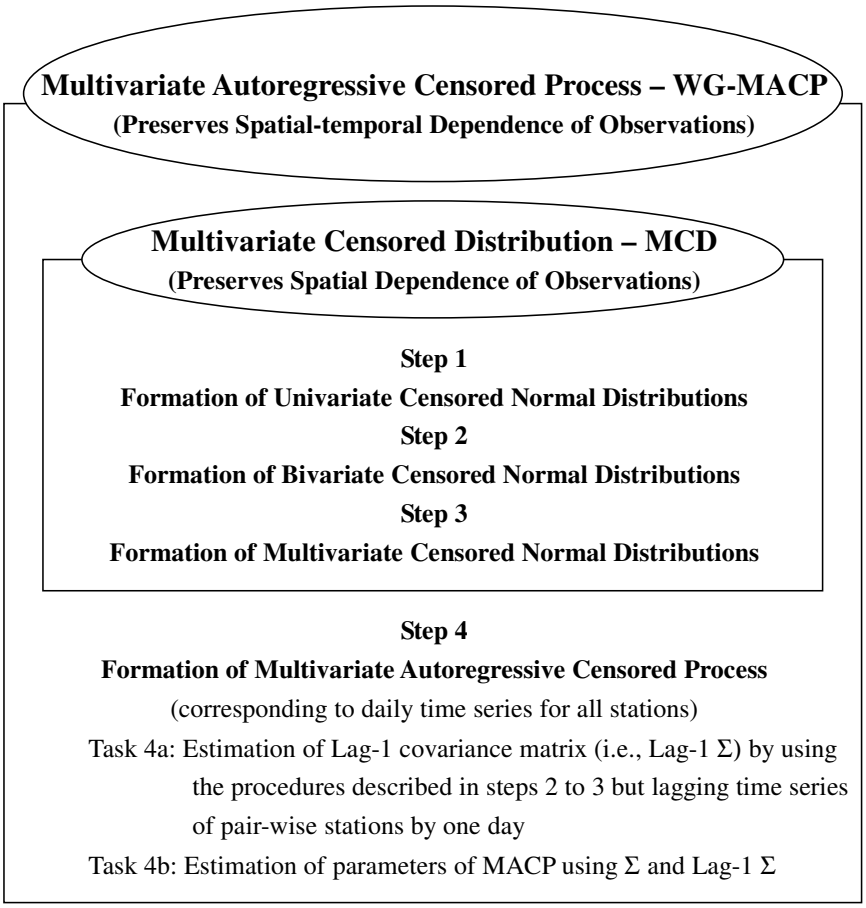


Fig. 4.1: Schematic structure of the WG-MACP formulation.

in which, $\mathbf{s_t}$ is the simulated precipitation on day t at any specified number of sites (in this thesis, the number of sites is equal to 10). $\mathbf{z}$ is a vector of K independent observations drawn from the standard normal distribution.

The term $\mathbf{B_i}$ (Equation 4.2.2) describes the temporal dependence in a daily precipitation sequence through the conditional information obtained from the previously generated vector, $\mathbf{s_{t-1}}$. The second term describes the randomness and spatial dependence in a daily precipitation sequence through the randomly generated vector, $\mathbf{z}$.

The matrix $\mathbf{C_i}$ is the Cholesky factorization of the matrix obtained from the right-hand side matrices of Equation 4.2.3. The terms on the right-hand side of Equations 4.2.2 and 4.2.3 are:

$$\mathbf{\Sigma_{0i}} = E[(\mathbf{x_i} - \mu_\mathbf{i})(\mathbf{x_i}^{'} - \mu_\mathbf{i})] \tag{4.2.4}$$

$$\mathbf{\Sigma_{1i}}^{'} = E[(\mathbf{x_{i-1}} - \mu_\mathbf{i})(\mathbf{x_i}^{'} - \mu_\mathbf{i})] \tag{4.2.5}$$

The term $\mathbf{\Sigma_{0i}}$ of the MCD, defined in Equation 3.2.3 was previously estimated through steps 1 to 3. The term $\mathbf{\Sigma_{1i}}$ is estimated in the same manner as the estimation of term $\mathbf{\Sigma_{0i}}$ but considers the lag-1 covariance of the pairwise precipitation series among sites. For instance, for a specific Julian day i,

$$\mathbf{\Sigma_{1i}}^{'} = \begin{bmatrix} E(\mathbf{x}_{i-1,1} - \mu_{i-1,1})(\mathbf{x'}_{i,1} - \mu_{i,1}) & E(\mathbf{x}_{i-1,1} - \mu_{i-1,1})(\mathbf{x'}_{i,2} - \mu_{i,2}) & \cdots & E(\mathbf{x}_{i-1,1} - \mu_{i-1,1})(\mathbf{x'}_{i,K} - \mu_{i,K}) \\ E(\mathbf{x}_{i-1,2} - \mu_{i-1,2})(\mathbf{x'}_{i,1} - \mu_{i,1}) & E(\mathbf{x}_{i-1,2} - \mu_{i-1,2})(\mathbf{x'}_{i,2} - \mu_{i,2}) & \cdots & E(\mathbf{x}_{i-1,2} - \mu_{i-1,2})(\mathbf{x'}_{i,K} - \mu_{i,K}) \\ \vdots & \vdots & \ddots & \vdots \\ E(\mathbf{x}_{i-1,K} - \mu_{i-1,K})(\mathbf{x'}_{i,1} - \mu_{i,1}) & E(\mathbf{x}_{i-1,K} - \mu_{i-1,K})(\mathbf{x'}_{i,2} - \mu_{i,2}) & \cdots & E(\mathbf{x}_{i-1,K} - \mu_{i-1,K})(\mathbf{x'}_{i,K} - \mu_{i,K}) \end{bmatrix} \tag{4.2.6}$$

where K refers to the total number of stations, and $k = 1, 2, ..., K$.

Tab. 4.1: Summary of the relevant information used for the model evaluation.

Figures	Number of Site(s)	Types of Generated Data used for Parameter Estimation	Attribute(s) Evaluated	Remarks
4.2	10	$MCD(s,1), MCD(s), MACP(s)$	μ, σ	ND
4.3	10	$MCD(s,1), MCD(s), MACP(s)$	$[P_{11}, P_{01}, P_{00}, P_{10}]$ of 4 Quadrants	S
4.4	10	$MCD(s,1), MCD(s), MACP(s)$	$[P_{00}, P_{11}]$ of Markov Model	T
4.5	1	$MACP(m), MACP(s), MACP(m,1)$	Σ_{Lag-1}	T

Notes: ND, S or T refer to the purpose of the evaluation is related to univariate normal probability distribution (ND), spatial dependence (S), and temporal dependence (T).
(Annotation of MCD/MACP(.) refers to Table 3.3.)

$\mathbf{x}_{i,k} = [x_{i,k,1},\ x_{i,k,2}, ..., x_{i,k,n}]'$ where n is the number of years. When Julian day $i = 1$, the matrix will be modified accordingly and takes Julian day 365 as $i - 1$. For the parameter estimation, steps 1 to 3 are still adopted in the same manner but consider a pairwise precipitation of $\mathbf{x}_{i-1,k}$ and $\mathbf{x}_{i,k}$.

To reduce the number of parameters, coefficients of $\mathbf{B_i}$ and $\mathbf{C_i}$ in the Equation 4.2.1 can be obtained using periodic functions (Equation 3.2.7).

Thus, steps 1 to 4 complete the development of weather generation based on a multivariate autoregressive censored process (WG-MACP) model at multiple sites capable of preserving spatio-temporal dependence in simulated data sequences (i.e., precipitation data sequences).

4.3 Evaluation of the WG-MACP Model

The efficacy of the WG-MACP model is evaluated in terms of such attributes as mean, variance, spatial correlations between any two sites, and lag-1 temporal correlation in historical precipitation series at multiple sites. Relevant information pertaining to various attributes of historical and simulated data is summarized in Table 4.1. The measures such as R^2, E and $RMSE$ for evaluating the efficacy of the model were used and are briefly described as follows.

R^2 is given by:

$$R^2 = \frac{[\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})]^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{4.3.1}$$

where x_i and y_i respectively are the historical observations and model simulations. n is the number of pairwise data.

The efficiency coefficient (Equation 4.3.2) is expressed as follows:

$$E = 1 - \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{4.3.2}$$

This measure ranges from $-\infty$ to 1 and the value of $E = 1$ denotes a perfect match of model simulations and historical observations. When the value of E is less than or equal to 0, the applicability of the model is questionable.

A value RMSE (Equation 4.3.3) is computed is follows.

$$RMSE = \left[\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2\right]^{1/2} \tag{4.3.3}$$

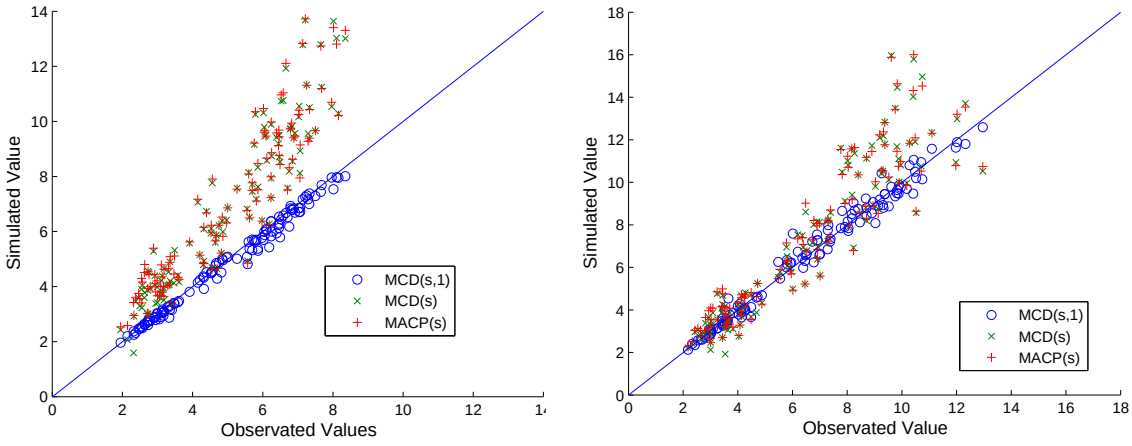
It is noted that though R^2 and $RMSE$ values are always positive, the E values may turn out to be negative for a weak model. Essentially, E is based on the regression between observed and simulated values on a 1:1 line, whereas R^2 is based on the line of best fit based on the linear least-squares, which may not be 1:1. The $RMSE$ is tantamount to the standard deviation of errors (deviations) between observed and simulated data and thus, a smaller value of $RMSE$ is an indicator of a better quality fit and vice-versa.

4.3.1 Validation of the WG-MACP Model

The adequacy of the WG-MACP model is validated by comparing the statistical characteristics of observed and simulated daily precipitation data sets. Evaluation of the model is based on daily values clumped over a month so as to facilitate the interpretation of results and enhance the clarity of pictorial representations. Therefore, each attribute/parameter estimate (e.g., μ stated in Table 4.1) is calculated for a specific site based on the observed and simulated daily precipitation data. Results of the evaluation are presented using the features (1) probability distribution, (2) spatial dependence, and (3) temporal dependence as follows.

Features of a Censored Normal Probability Distribution

At each site, the historical data and the simulated data were fitted to their respective censored Normal distributions, and their respective parameters (mean and variance) were computed. For the estimation of parameters (mean and standard deviations) in case of historical records, only above-zero values were considered. For example, the observed μ value for the month of March at station 5 is calculated by averaging the above-zero values of daily precipitation spread over the wet days in the month over a record period of 30 years. This means that although there are 930 days (31 days in March $\times$ 30 years of recoded data) in the sample, only wet days were used for computing the mean daily precipitation for the month. The corresponding simulated μ value is calculated in the same manner but there are 27,900 days (31 days in March 30 years of simulated data $\times$ 30 realizations of simulated data sequence) taken into consideration. Similarly, standard deviations of the observed and simulated data were computed. The observed and simulated values of μ constitute a coordinate (point)



(a) Mean value *(b)* Standard deviation

Fig. 4.2: Comparison of parameters of normal distribution from historical data with simulated sequences: averaged out values for each of 12 months and 10 stations. (Annotation of MCD/MACP(.) refers to Table 3.3.)

shown in Figure 4.2 from the above described data sets for the month of March. In a similar manner, parameters for all sites and all months were computed.

It is noted that this μ is not the mean of the MCD and to clarity their differences refer to Table 3.3.

The estimated parameters (mean and standard deviations) of the historical and simulated data were plotted as shown in Figure 4.2. It is noted that in these scatter plots corresponding to means and standard deviations, each point represents a value (as computed in the above) corresponding to each of the twelve months and each of the ten sites and thus there are a total of 120 points in each scatter plot.

A visual inspection of these scatter plots reveals that the best fit is provided by the WG-MCD(s,1) model, where simulated daily means or standard deviations were used without using periodic function. When the smoothing is adopted, the performance of WG-MCD(s,1) model turns out to be poorer than the WG-MCD(s). Even though

Tab. 4.2: Values of R^2, E, and $RMSE$ for attributes of the normal distributions obtained from observed and simulated data using 3 forms of models.

Evaluation Measures	Attributes	$MCD(s,1)$	$MCD(s)$	$MACP(s)$
R^2	mean	**0.99**	0.87	0.86
	standard deviation	**0.98**	0.86	0.86
E	mean	**0.98**	-0.77	-0.84
	standard deviation	**0.98**	0.66	0.65
$RMSE$	mean	**0.23**	2.36	2.41
	standard deviation	**0.40**	1.64	1.65

the temporal dependence has been explicitly considered using AR(1) dependence in the WG-MACP model, the performance was not any better.

To further affirm the aforesaid observations from the graphical plots, the evaluation yardsticks viz. R^2 (coefficient of determination), E (coefficient of efficiency), and $RMSE$ (root mean square error) were computed and are shown in Table 4.2. The bold numbers in this table indicate the best model under a specific evaluation measure. It is evident from the table that the highest values of R^2 and E, as well as the lowest values of $RMSE$ were yielded by the WG-MCD(s,1) model. In other words, the features of the simulated censored normal distribution are best preserved by this model, followed by the WG-MCD(s) model. The consideration of temporal association appears to play an insignificant role as far as preservation of the features of the simulated censored normal distribution is concerned.

Features of the Spatial Dependence

In this section, the spatial dependence of observed and simulated data were investigated. Conventionally, the correlation coefficient is used as an attribute to express

the degree of spatial dependence between precipitation data at any two sites. However, it is well documented that when data are non-normally distributed (e.g., daily precipitation records riddled with zero values), a test of significance using the Pearson correlation may inflate Type I error and reduce power of the test [*Bishara and Hittner*, 2012]. Therefore, the use of correlation coefficient to evaluate a bivariate mixed distribution of the precipitation data may not be appropriate and thus was not used in this evaluation. Alternatively, the probability of joint events in each of four quadrants of a bivariate distribution is considered as a reasonable attribute to express the degree of spatial dependence between data sets at any two sites. The four joint events refer to the simultaneous occurrences of wet-wet, dry-wet, dry-dry, and wet-dry day observed at any two sites. By using a counting method, the joint relative frequencies of the aforesaid events were evaluated from the observed and simulated data sets. That is, for the month of March, as an example, there are 930 observed daily precipitation events at each site with discrete values as 0 (dry-day) and 1 (wet-day). These 930 values were used in computing observed joint probabilities P_{11} (wet-wet), P_{01} (dry-wet), P_{00} (dry-dry) and P_{10} (wet-dry). Similarly, the simulated values of the aforesaid probabilities were computed from 27,900 simulated data points as mentioned in the aforesaid section. The values of joint probabilities (P_{11}, P_{01}, P_{00} and P_{10}) corresponding to historical and simulated data were graphically compared in Figure 4.3. The 45 combinations are the joint events observed at any two out of 10 sites. Each point on the graph indicates probabilities corresponding to the observed and simulated data in any of the four quadrants for a specific time between any two sites. A total of 2,160 (i.e., 45 combinations $\times$ 12 months $\times$ four events) joint probability values were computed for each version of the model.

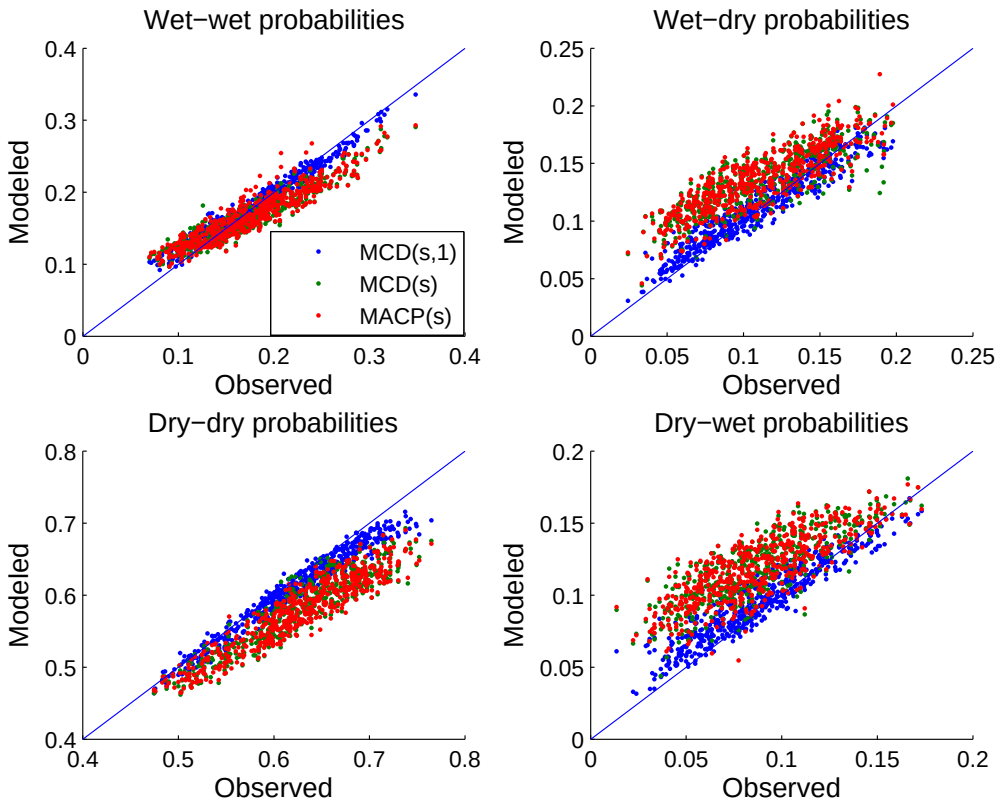


Fig. 4.3: Probabilities of simultaneously occurrences of wet-wet, wet-dry, dry-dry, and dry-wet stages at any two stations estimated by daily precipitation of 12 months at 10 stations.
(Annotation of MCD/MACP(.) refers to Table 3.3.)

In general, the WG-MCD(s) and WG-MACP(s) models show a good fit (top left figure) between observed and simulated values of P_{11} (i.e. wet-wet joint probabilities). In the same top-left, the WG-MCD(s,1) model tends to be slightly off the ideal diagonal (1:1) line and thus the model tends to underestimate. For simulations of the wet-dry and dry-wet probabilities, the WG-MCD(s,1) model seems to perform satisfactorily while the WG-MCD(s) and WG-MACP(s) models show a tendency for overestimation. For the simulation of dry-dry probabilities, all models are slightly off the diagonal line. However, a closer examination reveals that all models exhibit a tendency for underestimation. Visual observations of both models can be interpreted as simulation of dry-dry probabilities was weak and simulation of wet-wet probabilities appeared satisfactory.

In the graphical plots, all three forms of the model appeared to perform closely while either slightly under- or over- simulating the joint-probabilities. However, to discern which model is superior for the preservation of joint-probabilities, the evaluation measures R^2, E and $RMSE$ were computed as shown in Table 4.3. It can be inferred from this table that the best simulation is offered by the WG-MCD(s,1) model. A negative value of the statistic E is a striking indicator that the WG-MCD(s) and WG-MCDP(s) models are inferior to the WG-MCD(s,1) model, in which E values are always positive (Table 4.3).

Features of the Temporal Dependence

Attributes of temporal dependence are commonly expressed as lag-1 autocorrelation or transitional probabilities. It is known that the value of lag-1 autocorrelation in the case of mixed data such as daily precipitation data is not a true representative of temporal dependence in a strict sense [*Bishara and Hittner*, 2012]. The recourse

Tab. 4.3: Values of R^2, E, and $RMSE$ for attributes of the probabilities of simultaneous occurrences obtained from observed and simulated data using 3 forms of models.

Evaluation Measures	Attributes	$MCD(s,1)$	$MCD(s)$	$MACP(s)$
R^2	P_{11} of Quadrant 1	**0.97**	0.90	0.86
	P_{01} of Quadrant 2	**0.88**	0.67	0.65
	P_{00} of Quadrant 3	**0.96**	0.87	0.87
	P_{10} of Quadrant 4	**0.91**	0.69	0.73
E	P_{11} of Quadrant 1	**0.94**	0.79	0.79
	P_{01} of Quadrant 2	**0.81**	-0.35	-0.36
	P_{00} of Quadrant 3	**0.91**	0.32	0.25
	P_{10} of Quadrant 4	**0.90**	0.27	0.28
$RMSE$	P_{11} of Quadrant 1	**0.013**	0.024	0.024
	P_{01} of Quadrant 2	**0.013**	0.036	0.036
	P_{00} of Quadrant 3	**0.019**	0.052	0.054
	P_{10} of Quadrant 4	**0.012**	0.032	0.032

is to evaluate transitional probabilities rather than lag-1 autocorrelations. In view of the foregoing reasoning, transitional probabilities (P11, P10, P01, and P00) from historical and simulated daily precipitation time series (or sequences) were computed and plotted in Figure 4.4. The sequential occurrences of wet-wet, wet-dry, dry-wet, and dry-dry are essentially the same as the P11, P10, P01, and P00 of a Markov Chain Model.

A visual examination of the figure indicates that, in all models, simulation of wet-wet transition probabilities is weak and the simulation of dry-dry transition probabilities is satisfactory. In addition, the transitional probabilities of wet-dry or dry-wet are satisfactorily simulated by the WG-MACP(s) model, while other models were found to overestimate these transitional probabilities. Although the plots in figures for the wet-dry and dry-wet cases appear to be a bit off from the diagonal line, it is worthy to note that all these plots are clustering within a narrow range (0.1 to 0.25). In this

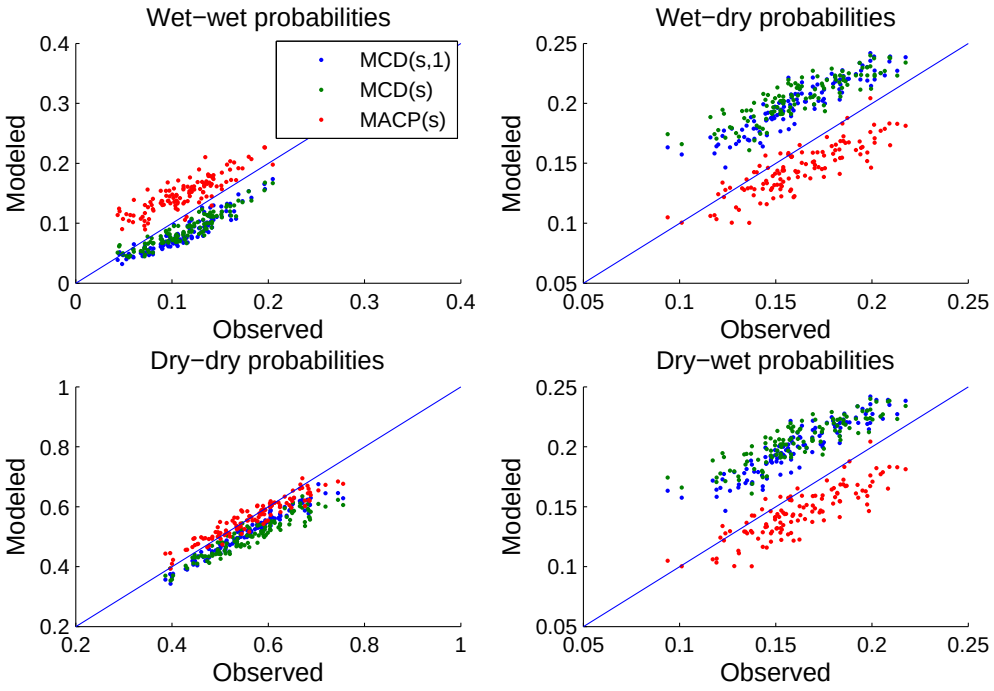


Fig. 4.4: Probabilities of sequential occurrences of wet-wet, wet-dry, dry-dry, and dry-wet stages estimated by daily precipitation of 12 months at 10 stations. (Annotation of MCD/MACP(.) refers to Table 3.3.)

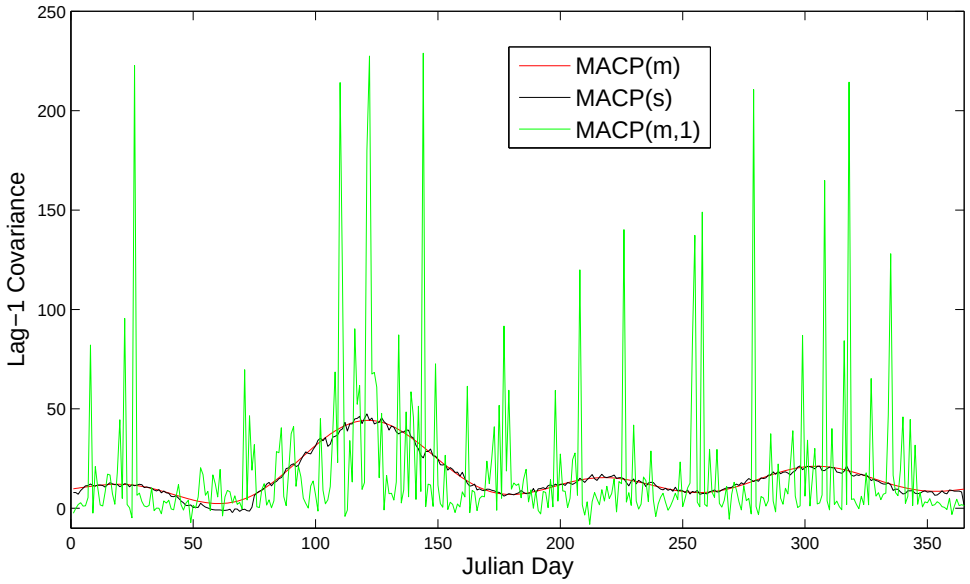


Fig. 4.5: Lag-1 covariance series at station 10 using 1 station in analysis. (Annotation of MCD/MACP(.) refers to Table 3.3.)

figure, the red coloured points refer to the simulation generated by the WG-MACP(s) model, when the daily values of lag-1 auto-covariance were smoothed using a periodic function. A typical graph showing the daily values of lag-1 auto-covariance in green colour is depicted in Figure 4.4. The temporal variation of lag-1 covariance can be smoothed by a periodic function involving 4 harmonics and the smoothed graph is shown by the curve in red colour. The smoothed values were found to lie in close proximity of the historical values (Figure 4.5) and therefore 4 harmonics were considered sufficient. In other words, the smoothing by a periodic function improved the model performance while reducing the number of parameters in the model structure.

To further assess the relative effectiveness of the models in simulating transition probabilities, the evaluation measures R^2, E and $RMSE$ were computed and are shown in Table 4.4. It can be seen from the values of E and RMSE in the table that

Tab. 4.4: Values of R^2, E, and $RMSE$ for attributes of the transition probabilities of sequential occurrences obtained from observed and simulated data using 3 forms of models.

Evaluation Measures	Attributes (Transition Probabilities)	$MCD(s,1)$	$MCD(s)$	$MACP(s)$
R^2	P_{11}	**0.87**	0.87	0.72
	P_{10}	**0.85**	0.74	0.72
	P_{01}	**0.85**	0.74	0.72
	P_{00}	**0.96**	0.89	0.90
E	P_{11}	0.17	**0.33**	-0.38
	P_{10}	-1.40	-2.13	**0.28**
	P_{01}	-1.44	-2.18	**0.27**
	P_{00}	0.62	0.30	**0.88**
$RMSE$	P_{11}	0.032	**0.029**	0.041
	P_{10}	0.039	0.045	**0.021**
	P_{01}	0.039	0.045	**0.021**
	P_{00}	0.050	0.067	**0.028**

the WG-MACP(s) model offers a relatively better preservation of transition probabilities (P11, P10, P01, and P00) based on the evaluation measures E and $RMSE$. The WG-MCD(s,1) model offers a better simulation of transition probabilities based on the evaluation measure of R^2. Although the WG-MCD(s,1) model is not specifically designed for the preservation of the temporal dependence of historical precipitation series, the performance of this model is comparable to the WG-MACP(s). It is due to the fact that parameters of the WG-MCD(s,1) model is estimated based on 365 Julian days, and therefore the corresponding temporal dependence are indirectly preserved through the sequential change of these parameters. As a result, the temporal dependence of historical series is preserved.

In summary, the WG-MCD(s,1) model tended to be satisfactory for simulating the parameters of the multivariate censored normal distribution. The same model

did the satisfactory job in preserving the spatial features, meaning that the smoothing of the cross co-variances among stations offered insignificant benefits. As far as temporal features are concerned, the WG-MACP(s) model can be regarded as acceptable since it reasonably well preserved the lag-1 dependency in the daily precipitation sequences. In this case study, it was found that the WG-MCD(s,1) model, in general, outperform the WG-MACP(s) model, especially with regarded to preserving precipitation amounts and spatial dependence. Since the temporal dependence in the observed precipitation is not very strong, attributes of spatial dependence become very critical to the formation of the simulation model in this study. However, in some cases it may be difficult to determine whether spatial or temporal dependence is more critical. For example, precipitation stations located far apart have weak spatial dependence; therefore, formation of the generation model relying on spatial dependence is questionable. For another example, when hourly time interval is considered for the analysis, formation of the generation model relying on temporal dependence is reasonable. In general, an integrative model consisting of two components that correspond to spatial and temporal dependencies is desirable. This model, while incorporating the features of smoothing of the lag-1 auto-covariances, provided a parsimonious and robust in terms of efficiency.

4.4 Preview of Model Applications

The structure of the proposed weather generation model and corresponding applications is shown in Figure 4.6. First, the distributional characteristics of precipitation amounts and the corresponding spatial dependencies are modeled as a multivariate

normal distribution. The parameters of the distribution are estimated using the multivariate censored normal distribution. The temporal dependencies of precipitation sequences are preserved using a multivariate autoregressive model. The spatial and temporal dependencies are then integrated into a single model. Second, local precipitation, at gauged and ungauged sites, can then be generated using the proposed model in association with the technique of regression. Third, the missing observations at a specific site can be estimated using the obtained parameters of the proposed model with Gibbs sampling. Last, the future change in precipitation patterns, due to the issue of global climate change, are addressed by a downscaling approach using the Delta change method in association with the proposed multivariate regressive model. For this purpose, the Canadian regional climate model (CRCM) has been selected for the analysis. Details of the three implementations are presented in the following chapters.

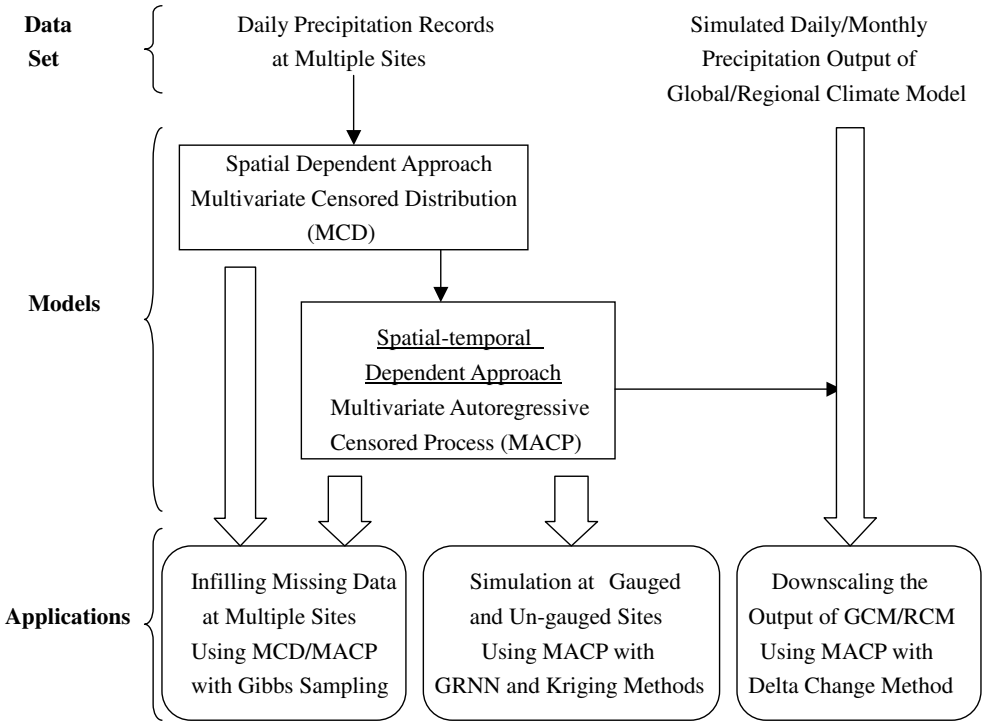


Fig. 4.6: Schematic structure of the thesis.

5. APPLICATION I: SIMULATIONS AT GAUGED AND UNGAUGED SITES

A versatile stochastic model for the generation of daily precipitation not only requires the capability of preserving the spatio-temporal dependence of precipitation but also requires expansion and modification to its structure to accommodate and facilitate implementations of the model to different applications. For evaluating the applicability of the proposed models, three applications are presented to illustrate: (1) Simulation of daily precipitation at gauged and ungauged sites, (2) infilling of missing records at multiple sites, and (3) assessment of climate change sensitivity.

This chapter illustrates how the WG-MCD model can be applied to problems related to simulation of daily precipitation at gauged and ungauged sites. The WG-MCD model was presented in the chapter 4 where it was demonstrated that the model satisfactorily preserve the spatial dependence among sites. Such a satisfactory capability of the model represents the key element, which is required for the simulation of daily precipitation at ungauged sites. For this purpose two procedures, namely the generalized regression neural networks (GRNN) and Kriging are considered. Kriging (or Gaussian process regression) is a method of interpolation for which the interpolated values are modeled by a Gaussian process governed by prior covariances compared to the piecewise-polynomial spline chosen to optimize smoothness of the fitted values. Under suitable assumptions on the priors, Kriging tends to provide the best linear unbiased prediction of the intermediate values. The WG-MCD model

in conjunction with GRNN and Kriging is chosen in this chapter to demonstrate as to how this model should be utilized in resolving a specific problem. In addition to the GRNN, such as multiple linear regressions, artificial neural network, and nearest neighbor methods are also be considered.

5.1 Methodology

The core structure of the WG-MCD model is designed for the simulation of occurrences and amounts of daily precipitation at multiple gauged sites. To extend its applicability to ungauged sites, the procedures such as the GRNN and the Kriging as described below were used. These two procedures, respectively, represent scenarios involving spatial stationarity and non-stationarity in orographic conditions. Figure 5.1 briefly summaries the manner in which these two procedures were integrated into the structure of the WG-MCD model to develop variant forms such as WG-MCD-GRNN and WG-MCD-Kriging.

5.1.1 Simulation of Data at Gauged Sites

The parameter estimation from data sets is followed the same procedure as previously described in Chapter 3 or 4 depending on which model is adopted for the simulation. For instance, when simulation is conducted using the WG-MACP model, Equation 4.2.1 was used for simulation of precipitation at 10 gauge sites and the periodic functions in Equation 3.2.7 were applied for generalizing the parameters B_i and C_i of this equation. To adequately represent seasonal variability, four harmonics are used in the historical precipitation data sets at 10 sites.

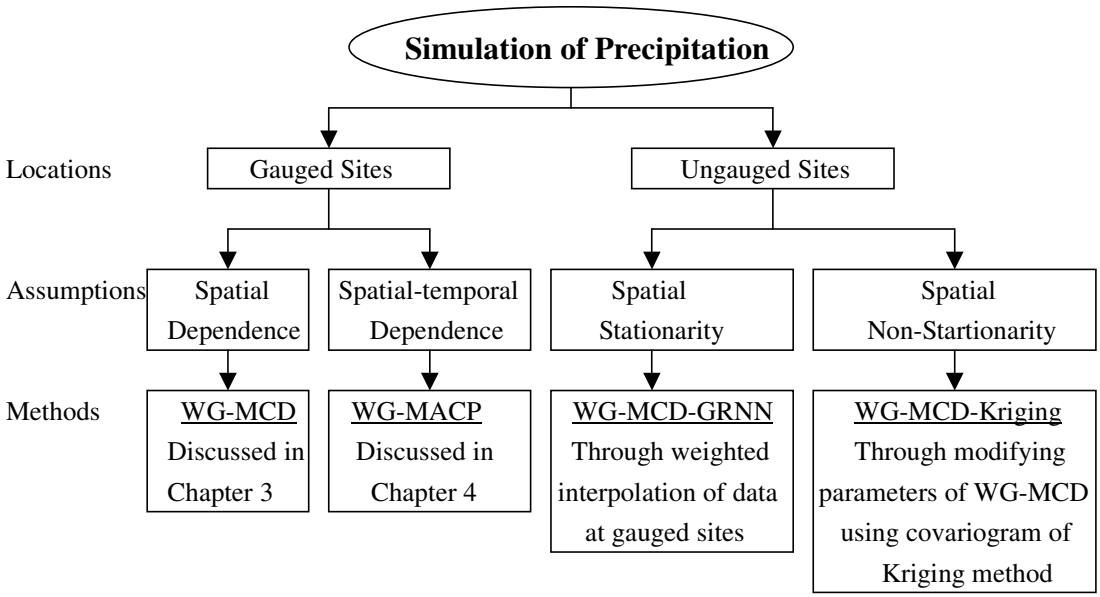


Fig. 5.1: Schematic diagram depicting assumptions and models used for infilling of missing data at gauged and ungauged sites.

5.1.2 Simulation of Data at Ungauged Sites

For the simulation of daily precipitation at ungauged sites, the procedures of GRNN and Kriging are integrated to formulate two variant forms of the WG-MCD model as the WG-MCD-GRNN and WG-MCD-Kriging. A brief description of the manner in which each procedure is integrated into the WG-MCD model for the generation of daily precipitation at ungauged sites is described below.

WG-MCD-GRNN Procedure

The WG-MCD-GRNN procedure initially generates precipitation at gauged sites using the WG-MCD model and in turn interpolates this information to the ungauged sites involving the GRNN procedure. Conceptually, the simulated data (i.e., predictand) at ungauged sites is predicted based on the simulated data at gauged sites (i.e., predictors). Therefore, the closer a gauged site is located to an ungauged site, the higher weight will be assigned to it. The procedure is simple and can be used to interpolate the information from gauged sites to ungauged sites.

The GRNN procedure is relatively robust to infrequent extreme observations (i.e., outliers) [*Specht*, 1991]. This aspect of the GRNN procedure is especially attractive in the prediction of daily hydrological data [*Ng et al.*, 2009]. For the generation of precipitation at ungauged sites, the procedural steps involved in the integration of GRNN into the WG-MCD model and its implementation are described as follows.

General structure of the GRNN is shown in Figure 5.2, where the first layer of the GRNN (i.e., input neurons) is responsible for the reception of a given input vector, x_p, which is used to estimate the scalar output y_p. The second layer (i.e., pattern neurons) measures the impact contributed by each of the x_j using a radial basis

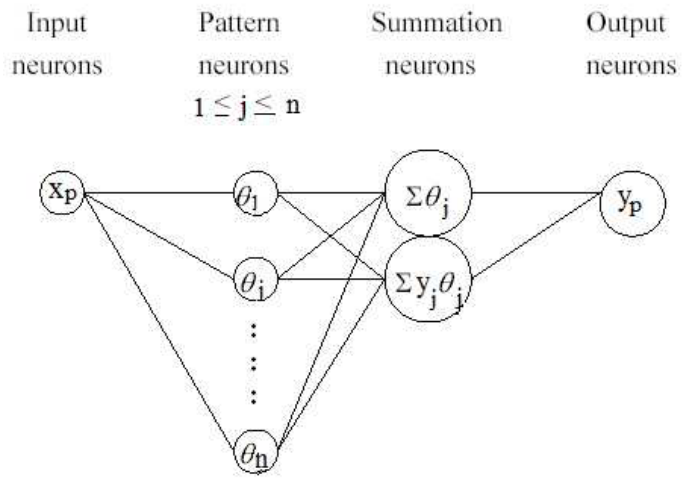


Fig. 5.2: Generalized regression neural network.

function.

Equation 5.1.1 is the multivariate Gaussian function that was adopted [*Specht*, 1991] in the design of the GRNN procedure:

$$\theta_j = \exp[-\Delta_j^2/(2\sigma^2)] \tag{5.1.1}$$

where $\Delta_j = \|x_p - x_j\|$ is a scalar distance (e.g., a Euclidean distance) between x_p and x_j. The x_p and x_j are not necessarily scalars. When multiple predictors variables are used for the prediction, the x_p and x_j are in vector form but the calculated distance Δ_j will still be a scalar. σ is the smoothing parameter.

The relationship between Δ_j and θ_j is illustrated in Figure 5.3. If x_p is equal to x_j, input Δ is equal to 0 on the x-axis and θ_j is equal to 1. If x_p is located at a distance from x_j, input Δ will be away from 0 on the x-axis and θ_j will be less than 1. If σ is small, the radial basis function is very steep. Consequently, the network tends to respond drastically when x_p is located close to x_j.

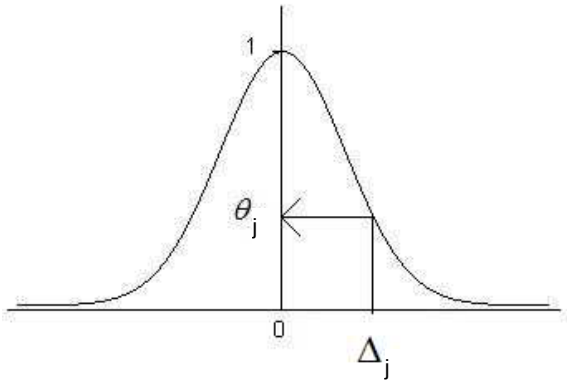


Fig. 5.3: Radial basis function.

The outputs of pattern neurons are then forwarded to the third layer (summation neurons in Figure 5.2), where these outputs are summed. The output neurons then perform the operation in Equation 5.1.2 to obtain the output y_p of the GRNN procedure. The term, y_j represents the simulated output at gauged sites using the WG-MCD model.

$$y_p = \sum_{j}^{n} (y_j \theta_j) / \sum_{j}^{n} \theta_j \tag{5.1.2}$$

The optimum smoothing parameter can be estimated through cross-validation. Instead of using a complete training set for the evaluation(i.e., verification and validation) of an optimum setting, cross validation evaluates the optimum setting based on results obtained from subsets of the training set. For example, a training set is first divided into n subsets. Then, an optimum GRNN setting is estimated using all but excluding one sample subset and one records the error of prediction in the withheld sample (i.e., RMSE). The procedure is repeated n times by withholding different sample subsets. Finally, the optimum GRNN setting based on the average errors obtained from the n times of evaluation [*Chaudill*, 1993] is selected.

WG-MCD-Kriging Procedure

Regression approaches have been used for the simulation of daily precipitation data at ungauged sites, but such procedures may overlook the variability of orography. To address the inherent variability of such systems, the Kriging procedure is embedded into the structure of the WG-MCD model. The following illustrate as to how the WG-MCD model can be modified to integrate the variability of orography through the consideration of the Kriging procedure for the simulation of daily precipitation data at ungauged sites.

The use of WG-MCD-Kriging procedure for the simulation of daily precipitation is similar to the use WG-MCD model for data simulation, except that of redefining values of parameters (i.e., mean, covariance, and lag-1 covariance) based on the gridded sites. The redefined mean and covariance at gridded sites can be calculated and considered as the conditional mean and covariance corresponding to a multivariate distribution. The procedural detail on the use of Kriging concept in data generation of redefined mean and covariance at gridded sites is explained below.

A covariogram of the Kriging concept and its normalized form are by far the most intuitive techniques for summarizing the structure of spatial dependencies in a covariance-stationary process. First step in the Kriging concept is to construct a variogram from interpolations of a set of scatter points. The use of a variogram is based on the assumption that a relationship exists between the distance and covariance. The spatial dependence in precipitation amounts can then be expressed as a function of the distance between two sites. The variogram, initially takes into consideration the gauged sites and establishes a relationship function between the elements of the covariance and the distance of two points for each time period (i.e.,

Julian day). The elements of the covariance matrix covering all gridded locations can then be established using the variogram function based on the distance of a pair of ungauged sites.

Once the experimental variogram has been computed, the next step is to define a model variogram. A variogram can only take certain special forms such as spherical, exponential, or Gaussian covariance matrices, satisfying the requirement of positive-definite.

To facilitate the analysis, the covariogram is utilized for summarizing the structure of spatial dependencies. The covariogram function $C(h)$ can be calculated based on the variogram $\gamma(h)$ as follows:

$$C(h) = \sigma^2 - \gamma(h) \tag{5.1.3}$$

where $\sigma^2 = C(0)$ is estimated by averaging the variances at sites.

In this thesis, the exponential function in Equation 5.1.4 has been selected. Maximum likelihood and the least squares are the most popular methods for fitting the variogram model [*Cressie*, 1993].

$$\gamma(h) = C(0) + C(1 - \exp(-h/r)) \tag{5.1.4}$$

where $h > 0$ refers to the distance between two gridded sites. $C(0)$ is the nugget that is similar to the y-intercept for the case of variogram and is considered to be zero.

The parameters r is selected by fitting the variogram to observed data. The best fit in the least squares sense minimizes the sum of squared residuals that are differences between observed values (i.e., covariances between two sites) and the fitted values of the model (i.e., $C(h)$).

Once, the model covariogram is constructed, it is used to reformulate the covariance matrices of the WG-MCD model. For example, in the generation of daily precipitation at K ungauged sites on a specific Julian day, the corresponding covariance matrix and mean vector of the WG-MCD model are required. Elements of the covariance matrix can be calculated based on the distance, h, between two sites using Equation 5.1.3. On a given day, the mean vector is assumed to be constant and is estimated by averaging the mean values of the WG-MCD model from the gauged sites. The formulated covariance matrix and mean vector are then used for the generation/simulation of precipitation at ungauged sites.

In this case study, the study area is divided into a 10 by 18 grid that covers up the 10 stations/sites of interest. Each divided grid represents an ungauged site with 0.2 degree of interval in the direction of longitude and latitude.

5.2 Results and Discussion

The results using the WG-MCD model for the simulation of precipitation at gauged sites was provided in Chapter 4. Figure 5.4 illustrates a 3-D plot showing the results of simulation for the Julian day 246 using the WG-MCD model at gauged sites. For illustration purposes, the Julian day 246 was arbitrary selected.

A number of different values for the smoothing parameter are first selected for the simulation of daily precipitation data to illustrate as to how various smoothing parameters affect the generation of precipitation. Second, the most suitable smoothing parameter is then estimated by the cross-validation procedure.

Figure 5.5 illustrates the results of simulation of precipitation amount on Julian day 246 using the WG-MCD-GRNN procedure with smoothing parameter equal to 1.

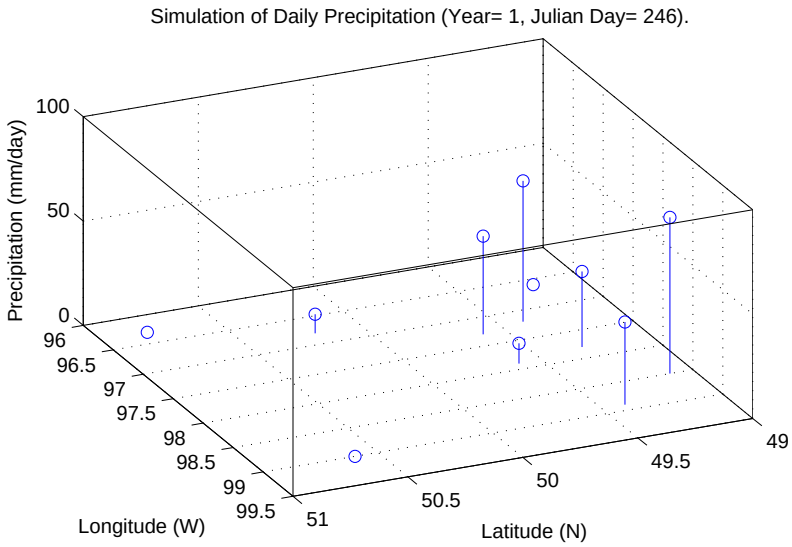


Fig. 5.4: Simulation of the WG-MCD at Julian day 246.

This figure shows a very smooth pattern of simulated precipitation to the area. The figure is a frame of an animation generated by the Matlab software using the simulated data. The 3-D representation provides a better visual presentation to observe the change in weather pattern in the study area through weight interpolations of data at gauged sites.

Figure 5.6 illustrates the results of simulation of precipitation amount on a given day using the WG-MCD-GRNN procedure with the smoothing parameter equal to 0.1 for comparative purposes only. The GRNN is used to estimate data at each ungauged site by using the data at gauged sites on the same given day. For the purpose, a total of 81 gauged sites ([9*9] grid points) are located from latitude(N) [49 to 51] and longitude(W) [96 to 100] for the estimation. For instance, the precipitation data at grid point [latitude 96.5 and longitude 50] is estimated using a weighted interpolation from data at gauged sites. The closer a gauged site is located to an ungauged site; the higher the weight given to the site. The data at gauged site can be based on

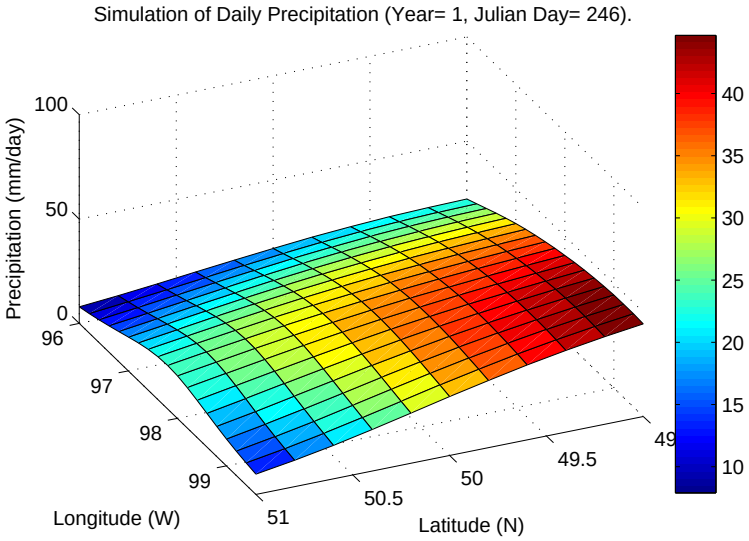


Fig. 5.5: Simulation of the WG-MCD-GRNN with smoothing parameter equal to 1 at Julian day 246. The 3-D plot covers the ten sites within the study area.

historical record if they are available, else a simulation at gauged site using the WG-MCD model can be utilized. The results in this figure do not exhibit gradual change in precipitation intensities. Such a presentation clearly reveals the existence of individual storms of interest in the region through the use of a smaller smoothing parameter.

Figure 5.7 depicts simulated precipitation on Julian day 246 using the WG-MCD-GRNN procedure with smoothing parameter equal to 0.5. The selection of smoothing parameter equal to 0.5 seems to be a reasonable choice for visual examination, as the precipitation pattern seems to be more physically and rationally representative of the variation of storm cells within the study region.

The 3-D presentation in conjunction with the WG-MCD-GRNN procedure is helpful in revealing the scope, magnitude, location, direction, center of rainstorm and rainstorm pattern. Such a visual presentation facilitates the study and analysis related to the simulated precipitation in the study area. The smoothing parameter

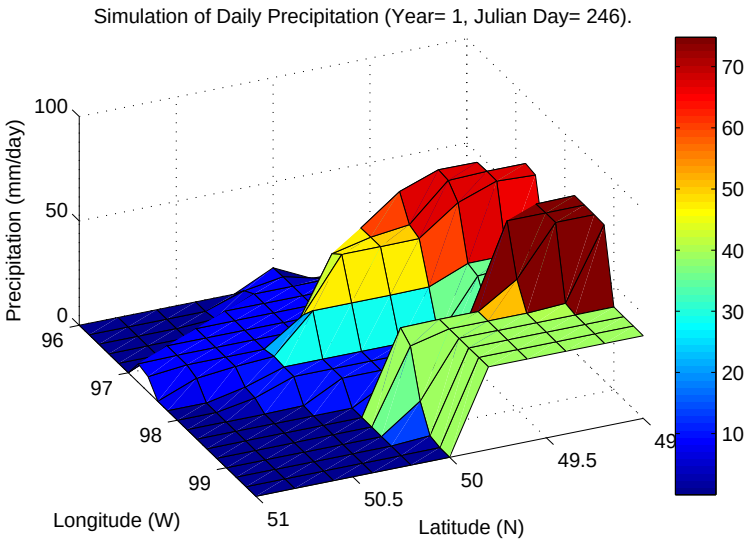


Fig. 5.6: Simulation of the WG-MCD-GRNN with smoothing parameter equal to 0.1 at Julian day 246. The 3-D plot covers the ten sites within the study area.

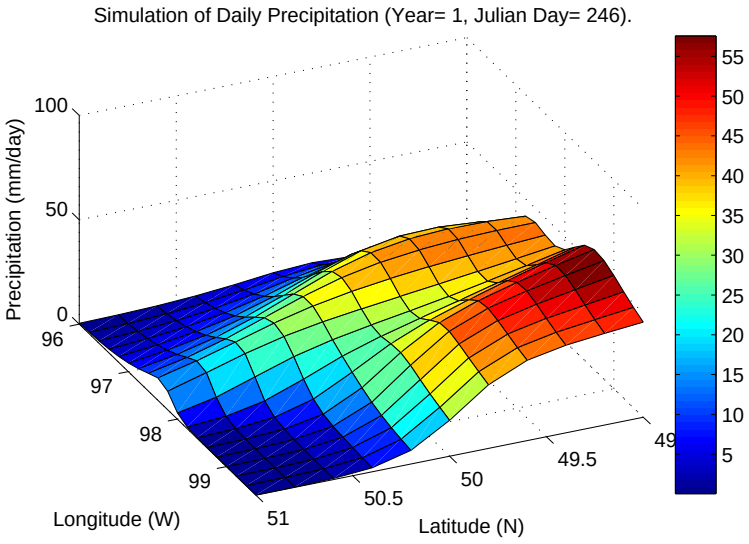


Fig. 5.7: Simulation of the WG-MCD-GRNN with smoothing parameter equal to 0.5 at Julian day 246. The 3-D plot covers the ten sites within the study area.

provides flexibility to adjust the smoothness of storm patterns, which in turn aids in identifying individual storm. For a small study area, it is expected that a large smoothing parameter will be more appropriate because data sets from gauged sites within a study area are expected to be highly correlated and thus getting a very smooth precipitation pattern as shown in Figure 5.5. On the other hand, if the study area is relatively large, the use of a smaller smoothing parameter may be more reasonable to capture individual storm patterns, such as those illustrated in Figure 5.6.

Figure 5.8 shows that an optimum value of the smoothing parameter is approximately 150, thus suggesting the existence of a smooth precipitation distribution. A very large value of the optimum smoothing parameter signifies that assignment of equal weight to observations of gauged stations can provide the best estimate of precipitation at ungauged stations.

The WG-MCD-GRNN procedure may predict the mean value and may also provide the best prediction of the dependent variable, but it is not a realistic representative of real precipitation variability. For such a purpose, the WG-MCD-Kriging procedure is considered.

For an analysis with the WG-MCD-Kriging procedure, Figure 5.9 shows covariograms on Julian Day 1 using covariance and lag-1 covariance respectively. Each dot in the plot represents a covariance/lag-1 covariance between a specific pair of stations selected from the 10 stations in the data set.

In terms of seasonal pattern that was observed through the animation of the simulation results, the 3-D plot obtained based on results of the WG-MCD-Kriging is

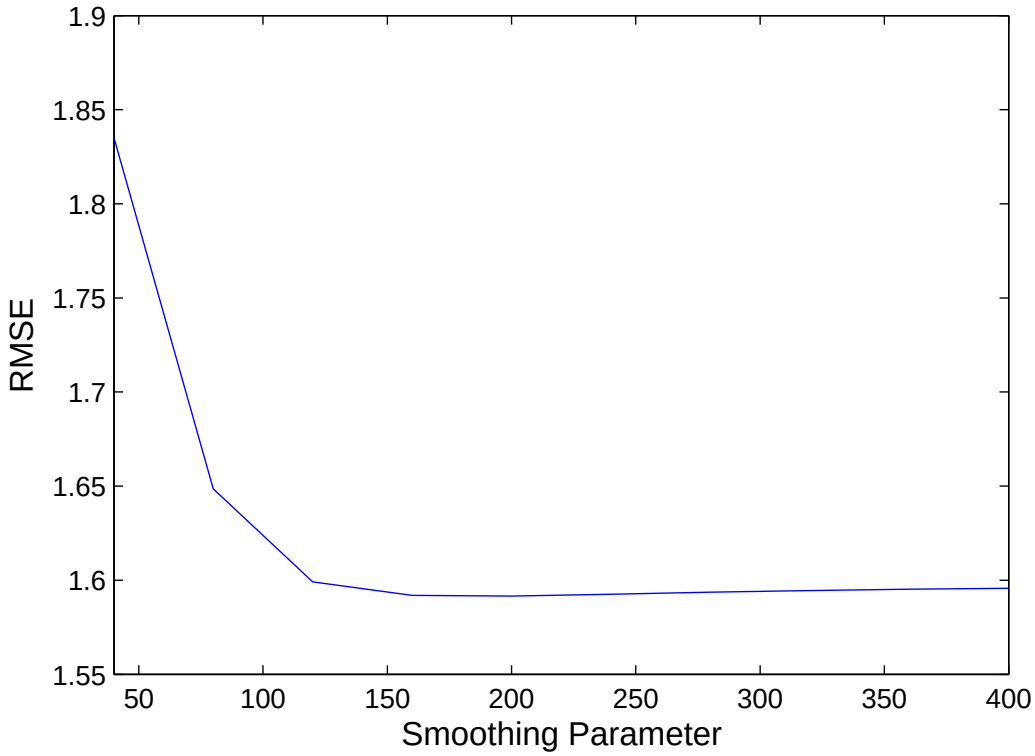


Fig. 5.8: Optimum smoothing parameter.

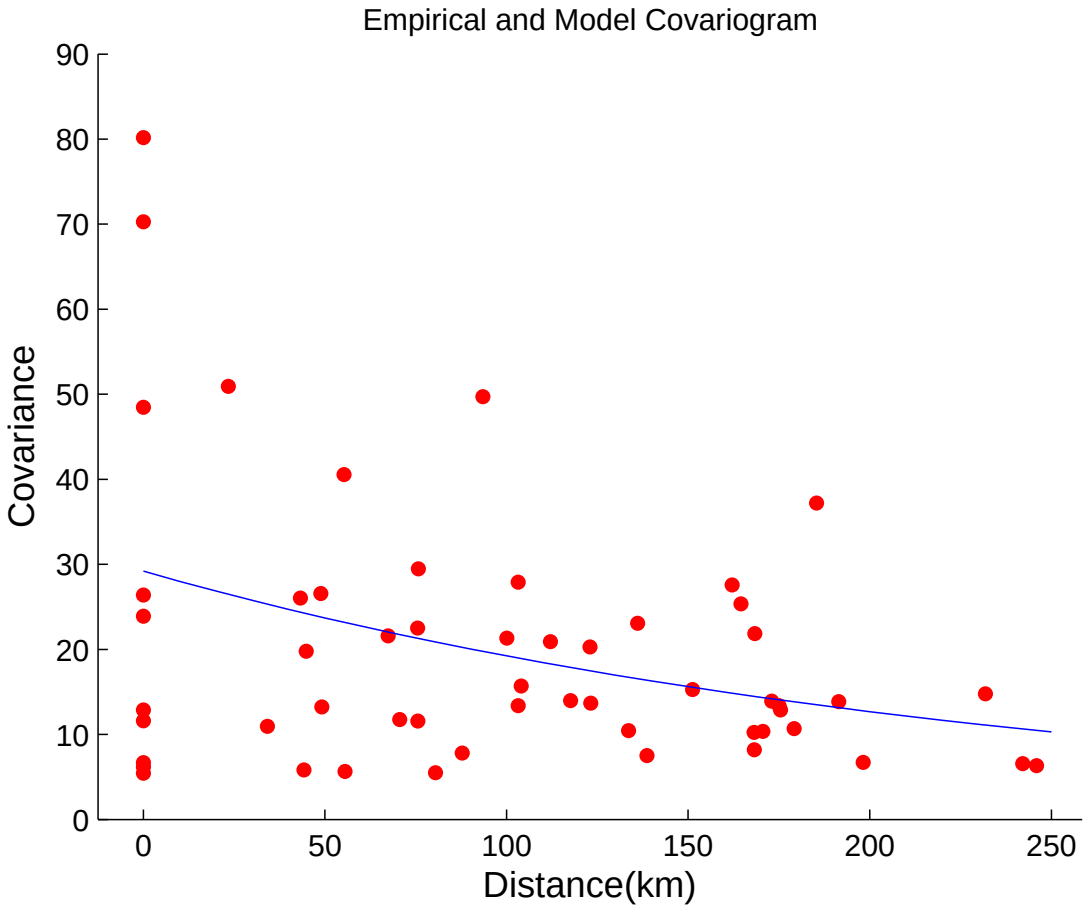


Fig. 5.9: Covariogram plots based on Julian day 1 of 10 stations.

Tab. 5.1: R^2, E, and $RMSE$ calculated based on WG-MCD and WG-MCD-GRNN/MCD-Kriging.

Evaluation Measures	Attributes	$WG-MCD-GRNN$	$WG-MCD-Kriging$
R^2	mean	0.82	**0.84**
	standard deviation	0.59	**0.69**
E	mean	**-0.97**	-5.04
	standard deviation	**-0.42**	-2.31
$RMSE$	mean	**2.46**	4.31
	standard deviation	**3.37**	5.14

similar to the 3-D plot based on results obtained from the WG-MCD-GRNN procedure. However, it was observed that the variability is quite obvious and is not as smooth as the results obtained by the WG-MCD-GRNN procedure. Essentially for this very reason that the use of the WG-MCD-Kringing procedure offers an advantage in embedding and preserving the variability due to orography.

To further assess the relative effectiveness of models in simulating precipitation at ungauged sites, the evaluation yardsticks viz. R^2 (coefficient of determination), E (coefficient of efficiency), and $RMSE$ (root mean square error) were computed and are shown in Table 5.1. The bold numbers in this table indicates the best model under a specific evaluation measure.

For the evaluation of these two procedures, a comparison can be conducted by cross validation based on information obtained from the 10 gauged stations. Specifically, precipitation is estimated at each of the gauged stations based on these two procedures, namely the WG-MCD-GRNN and WG-MCD-Kriging. For evaluation purposes, a procedure similar to cross-validation is adopted in which each station was deemed as ungauged station for which precipitation is estimated from the remaining

nine stations using one of the two procedures (i.e., WG-MCD-GRNN or WG-MCD-Kriging). This step was repeated for all the remaining nine stations to obtain a simulated data set for each procedure. Evaluation measures, such as R^2 from the table, is then calculated based on the attributes of observed and simulated data. For instance, the R^2 of 0.82 in the table is calculated based on the μ of the observations and simulations using the WG-MCD-GRNN procedure. Each μ is calculated only from the wet-days (excluding zero records) and are then calculated corresponding to each calendar month and station. A total of 120 μ values (12 months $\times$ 10 stations) are then obtained from historical observations and simulated data set. These pairwise 120 values can then be used for the calculation of R^2.

It is evident from the table, that highest value of R^2 was yielded by the WG-MCD model with Kriging procedure; however, the highest values of E and $RMSE$ were yielded by the WG-MCD model with GRNN procedure. In other words, the results are inconclusive depending on types of the evaluation measure being considered.

6. APPLICATION II: INFILLING OF MISSING OBSERVATIONS

Missing observations are common occurrences in hydrologic data sets and precipitation records are no exception. The existence of missing observations may arise due to human errors or equipment malfunctioning. Therefore, suitable procedures for infilling of missing observations in data sets are usually required. In this chapter, the utility of two data substitution procedures, namely the WG-MCD with Gibbs sampling and the WG-MACP with Gibbs sampling for data infilling is demonstrated. Firstly, in the Methodology section, each method is discussed along with necessary details related to data substitution in both of these procedures. Secondly, historical daily precipitation observations at 10 stations are then transformed into a normally distributed data set. In the last section of this chapter, the results of data substitution/infilling are discussed with a view to show their applicability for infilling of missing observations in daily precipitation.

The parameters estimation for the WG-MCD and WG-MACP models can also be extended to incomplete data sets. The completed data sets (also include infilled observations and are also known as treated data), therefore should exhibit the same statistical characteristics as that of the incomplete data sets (i.e., data set with missing observations and are also known as untreated data).

For infilling purposes, daily precipitation records for the years 1961 to 1990 corresponding to 10 stations located in the Southern Manitoba, Canada were utilized. This

data set has also been used for model validation in Chapters 3 and 4. The data set corresponding to 10 stations and involving 30 years constitutes a total of 109,575 daily records in which 29,069 (26.5%) days show precipitation above zero; 75,986 (69.4%) days show precipitation to be zero; and 4,520 (4.1%) days lack any information on the status of precipitation (i.e., commonly referred to as missing observations). That means to complete the data set, there is a need to infill the precipitation data for the days with missing records.

To validate the adequacy of the proposed substitution/infilling procedures (i.e., the WG-MCD and WG-MACP models (both with Gibbs Sampling), the parameters (i.e., mean, covariance, or lag-1 covariance) of the before-infilled (i.e., untreated case) and the after-infilled (i.e., treated case) data sets will be compared. Such results can provide a reasonable basis for discerning the relative effectiveness of the proposed procedures in infilling/substitution of missing observations in data sets.

6.1 Methodology

The missing observations in a data set can be replaced by appropriate estimated values using a data transformation procedure that consists of a component of data simulation. Two procedures corresponding the use of WG-MCD and WG-MACP models are formulated based on the Gibbs sampling for the simulation of missing observations. In this context and to serve the purpose of data transformation, a daily precipitation observation (above-zero or zero) can be classified as missing and can be infilled depending on its class (i.e. whether it is 0, above zero, or missing) based on a specific data transformation procedure. The simulation of observations by the WG-MCD and WG-MACP models respectively relies on Equations 3.2.1 and 4.2.1.

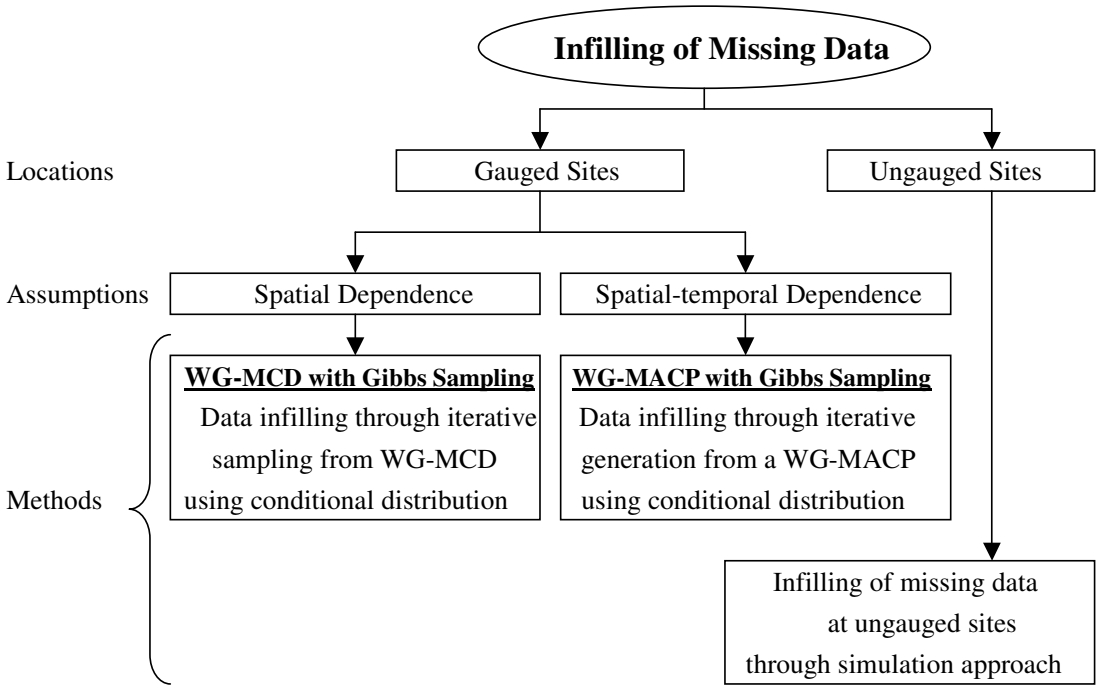


Fig. 6.1: Schematic structure of infilling the missing observations by procedures of the WG-MCD-Gibbs sampling and the WG-MACP-Gibbs sampling.

The use of the WG-MCD model with Gibbs sampling is intended to generate data preserving the statistical characteristics, such as mean and covariance. On the other hand, the use of the WG-MACP model with Gibbs sampling takes into account lag-1 covariance as an additional statistical characteristic for infilling of missing observations. Figure 6.1 briefly summaries the flow chart related to data infilling procedure utilized in the thesis.

6.1.1 Data Infilling Procedure - Preserving Spatial Dependence

Missing observations have been an area of statistical research for many years and multiple imputation (data augmentation) procedures have been extensively studied for the infilling of missing observations [*Little and Rubin*, 1987]. To preserve spatial dependence of the precipitation at multiple sites, the WG-MCD model with Gibbs sampling is proposed to handle the problem of missing observations. The use of multiple imputations procedures requires that a known probability distribution function be fitted to the existing data to produce estimates of the overall mean and variance. In this thesis, the distribution function utilized is the MCD (multivariate censored normal distribution) and the procedural details of parameter estimation are the same as presented in Chapter 3 through steps 1 to 3.

Multiple imputations are carried out by using the conditional distribution of the missing values given the observed values. Let $\mathbf{X} = (\mathbf{X}_{miss}, \mathbf{X}_{obs})$ be the daily precipitations at multiple sites on a specific day. The terms $\mathbf{X}_{miss}$ and $\mathbf{X}_{obs}$ respectively refer to the vectors of missing and observed precipitation records. The $\mathbf{X}_{obs}$ consists of above-zero values $\mathbf{X}_{pos}$ and zero values $\mathbf{X}_{zero}$. Similarly, the mean and covariance of the MCD function can be partitioned. For the MCD function, the distribution of $\mathbf{X}_{miss}$ given $\mathbf{X}_{obs}$ is:

$$\mathbf{X}_{miss}|\mathbf{X}_{obs} \ \sim \ MCD(\boldsymbol{\theta_t}) \tag{6.1.1}$$

where $\boldsymbol{\theta_t}$ is the conditional mean and variance of MCD function on day t. The X_{obs} typically refers to a normal distributed data set; however, X_{obs} in this study refers to a mixed distribution of precipitation amounts. Therefore, the Gibbs sampling method can not directly be applied. To facilitate the use of Gibbs sampling, the X_{obs}

requires to be transformed into a normally distributed data set. The transformation procedure is described in the following section.

The Gibbs sampling procedure is conducted through a Bayesian analysis and is a kind of Markov chain algorithm that has proven to be effective in managing high dimensional problems [*Tanner and Wong*, 1987]. The steps for implementing the Gibbs sampling procedure with an iterative step for the MCD function is as follows, where the terms MCD means MCD function, i.e. multivariate censored normal distribution function.

Let s_k^i and $\boldsymbol{\theta_k}$ respectively denote the simulated precipitation and the conditional parameters of a prior MCD. The i and k refer to the iteration from 1 to n and the gauged station from 1 to K. The $\boldsymbol{\theta_k}$ can be estimated based on the observed precipitation obtained for a specific month. Additional detail and relevant discussion on parameter estimation of $\boldsymbol{\theta_k}$ are presented through steps 1 to 3 in Chapter 3. For day t, the sampling procedure is as follows:

$$\begin{aligned}
s_1^{i+1} &\sim MCD(\boldsymbol{\theta_1}|s_2^i, s_3^i, s_4^i, ..., s_K^i) \\
s_2^{i+1} &\sim MCD(\boldsymbol{\theta_2}|s_1^{i+1}, s_3^i, s_4^i, ..., s_K^i) \\
s_3^{i+1} &\sim MCD(\boldsymbol{\theta_3}|s_1^{i+1}, s_2^{i+1}, s_4^i, ..., s_K^i) \\
&\vdots \\
s_K^{i+1} &\sim MCD(\boldsymbol{\theta_K}|s_1^{i+1}, s_2^{i+1}, s_3^{i+1}, ..., s_{K-1}^{i+1})
\end{aligned} \tag{6.1.2}$$

where the drawn samples on the left-hand side of the equations denote the substituted/infilled data at sites with missing information. The right-hand side of the equations refers to the sites at which the data is available. The procedure is explained as follows.

The implementation of the MCD with Gibbs sampling is conducted through a repetitive data simulation procedure. The simulation of above-zero, missing, or zero

values are processed differently. Therefore, each data point to be simulated is first categorized and then treated according to the class (i.e. whether it is 0, above zero, or missing) of the data. The treated data essentially is expected to preserve statistical characteristics of the observed precipitation data and thus a normally distributed set is obtained. An observation recorded as zero has to be replaced by a negative value $[-\infty, 0]$ to facilitate the formation of a normally distributed data set. For the infilling of missing observations, both negative and positive values are acceptable. Negative values will subsequently be censored to zero.

For case studies in this thesis, the data sets consisting of a total of 109,575 daily precipitation observations (10 stations $\times$ 30 years $\times$ 365 days + days of a leap year) were utilized. Nearly 4% of the records are missing that spread across the Julian days and stations. To facilitate the infilling of missing data using Gibbs-sampling procedure, the precipitation records require to be transformed into a normally distributed data set. The transformation may have each data modified and re-substituted/infilled. The substitution/infilling allows the missing data in the set to be replaced by an infilled value.

The data transformation requires each historical record be classified as above-zero, zero, or missing observation. Each record will then be evaluated one at a time, and treated in a particular way depending on their class. For example, for a specific day and station, when the evaluated data is classified as an above-zero record, the record will stay to be the same without modification. When the evaluated data is classified as a missing record, the data will be substituted by a sample drawn from the conditional distribution of the MCD with that the given data obtained from the remaining of nine stations for the same day as described in the Equation 6.1.2. When the evaluated

data is classified as a zero record, a procedure similar to one applied to the missing data is performed. However, the simulated data is retained as a negative value to facilitate the transformation of precipitation records into a normally distributed set that is required when the Gibbs-sampling is applied. Through this procedure, the missing data can be estimated. The evaluation and data treatment operations will be iteratively continued several times until stable values of the output are obtained.

A detailed algorithm pertaining to the data substitution procedure is presented as follows. Let s_{tk}^i and x_{tk} respectively denote a simulated and observed daily precipitation; $i = 1 : n$, where n refers to the number of iterations; $t = 1 : T$ day and $k = 1 : K$ station. $\mathbf{s_t}$ and $\mathbf{x_t}$ are the vector of s and x at day t.

for Day $t = 1$ *to* T **do**

- $\mathbf{s_t^1} = \mathbf{x_t}$.

for Iteration $i = 2$ *to* n **do**

for Station $k = 1$ *to* K **do**

if $x_{tk} > 0$ **then**

- $s_{tk}^i = x_{tk}$.

end if

if x_{tk} = missing observations **then**

- $s_{tk}^i \sim N(E(\mathbf{X_1}|\mathbf{X_2}), Cov(\mathbf{X_1}|\mathbf{X_2}))$, where details of conditional parameters are defined in Equations 6.1.3 and 6.1.4. $\mathbf{X_1}$ is the observation that is to be simulated s_{tk}^i. $\mathbf{X_2}$ is the vector $\mathbf{s_t^{i-1}}$ excluding station k.

end if

if $x_{tk} <= 0$ **then**

- $s_{tk}^{i} \leq 0$, and
- $s_{tk}^{i} \sim N(E(\mathbf{X_1}|\mathbf{X_2}), Cov(\mathbf{X_1}|\mathbf{X_2}))$.

end if

end for

end for

end for

As shown in the above algorithm, the data (i.e., above-zero, missing observations, or zero records) are first classified based on the observed precipitation x and each simulated precipitation is then individually evaluated and treated. In cases, where the observed precipitation constitutes an above-zero value, then the same value is retained. In cases where the precipitation is a missing observation, then the data is simulated based on the conditional parameters and conditional observation for the same day. In cases, where the observed precipitation shows a zero value, an additional requirement is invoked for the simulation such that the simulated value is less than or equal to zero to mimic a physical dry day.

The conditional covariance is essentially the variance of $\mathbf{X_1}$ in Equations 6.1.3 and 6.1.4, which is a scalar quantity equivalent to s_{tk}^{i}.

$$E(\mathbf{X_1}|\mathbf{X_2}) = \boldsymbol{\mu_1} + \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}^{-1}}(\mathbf{X_2} - \boldsymbol{\mu_2}) \tag{6.1.3}$$

$$Cov(\mathbf{X_1}|\mathbf{X_2}) = \boldsymbol{\Sigma_{11}} - \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}^{-1}}\boldsymbol{\Sigma_{21}} \tag{6.1.4}$$

where $\mathbf{X_1}$ and $\mathbf{X_2}$ respectively refer to the observations that are intended to be estimated and the given conditional observations. $\mathbf{X}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ respectively refer to partitioned multivariate observations for a specific day, mean vector, and covariance

matrix.

$$\mathbf{X} = [\frac{\mathbf{X_1}}{\mathbf{X_2}}]; \quad \boldsymbol{\mu} = [\frac{\boldsymbol{\mu_1}}{\boldsymbol{\mu_2}}]; \quad \boldsymbol{\Sigma} = [\frac{\boldsymbol{\Sigma_{11}}}{\boldsymbol{\Sigma_{21}}}|\frac{\boldsymbol{\Sigma_{12}}}{\boldsymbol{\Sigma_{22}}}] \tag{6.1.5}$$

The term $\mathbf{X_2}$ is the updated simulated precipitation of the nine stations excluding the data s_{tk}^i that is to be determined. A new estimated s_{tk}^{i+1} can then be calculated through sampling from the calculated conditional distribution $s_{tk}^i \sim N(E(\mathbf{X_1}|\mathbf{X_2}), Cov(\mathbf{X_1}|\mathbf{X_2}))$ or expressed analytically as follows:

$$s_{tk}^i = Cov(\mathbf{X_1}|\mathbf{X_2})z + E(\mathbf{X_1}|\mathbf{X_2}) \tag{6.1.6}$$

where the $z \sim N(0,1)$.

This sampling procedure is applicable to missing observations and to records displaying zero values. For the case with missing observations, any data drawn from the conditional distribution is adopted without restriction. For the case showing zero values, however, the requirement of $s_{tk}^i \leq 0$ is applied. The use of Equation 6.1.7 ensures that inequality constraint must be satisfied by generating a value to be less than or equal to zero:

$$s_{tk}^i = F^{-1}(u|E(\mathbf{X_1} \mid \mathbf{X_2}), Cov(\mathbf{X_1}|\mathbf{X_2})) \tag{6.1.7}$$

where u is a uniform random number drawn between $[0, p_{dry}]$, and p_{dry} refers to the probability of dry day:

$$p_{dry} = F(0|\mathbf{X_1} \mid E(\mathbf{X_2}), Cov(\mathbf{X_1}|\mathbf{X_2})) \tag{6.1.8}$$

The conditional parameters will be constant throughout a specific month but the conditional observations will be continuously modified and will result in a sample

from the conditional distribution.

For obtaining a data set with infilled values that preserves the temporal dependence, the WG-MACP-Gibbs sampling procedure as proposed below is followed.

6.1.2 Data Infilling Procedure - Preserving Spatio-temporal Dependence

The approach of the WG-MCD and WG-MACP models with Gibbs sampling are very similar except that samples are obtained from the MACP instead of drawn from the MCD. Thus, the temporal dependence of the system is preserved in the infilled data.

Similar to Equation 6.1.2, the sample of MACP will be drawn via Gibbs sampling. For day t,

$$\begin{aligned} s_1^{i+1} &\sim MACP(\boldsymbol{\theta_1}|s_2^i, s_3^i, s_4^i, ..., s_K^i) \\ s_2^{i+1} &\sim MACP(\boldsymbol{\theta_2}|s_1^{i+1}, s_3^i, s_4^i, ..., s_K^i) \\ s_3^{i+1} &\sim MACP(\boldsymbol{\theta_3}|s_1^{i+1}, s_2^{i+1}, s_4^i, ..., s_K^i) \\ &\vdots \\ s_K^{i+1} &\sim MACP(\boldsymbol{\theta_K}|s_1^{i+1}, s_2^{i+1}, s_3^{i+1}, ..., s_{K-1}^{i+1}) \end{aligned} \tag{6.1.9}$$

where notations are the same as in Equation 6.1.2.

Equation 4.2.1 is adopted for sample generation from the WG-MACP model. As discussed in Chapter 4, the parameters of the WG-MACP (i.e., **B** and **C**) model are obtained from Equations 4.2.2 and 4.2.3. Thus, the mean, covariance, and lag-1 covariance are the required parameters for the data generation using the WG-MACP model. Similar to the procedure with the WG-MCD with Gibbs sampling, the parameters of the WG-MACP model are simulated conditional to the data at the remaining nine stations. The calculation of the conditional mean and covariance has been achieved using Equations 6.1.3 and 6.1.4. The calculation of conditional lag-1 covariance is similar to the calculation used for the conditional covariance. Parameters

of the WG-MACP model for the full set of data have been estimated through steps 1 to 4 as described in Chapter 4.

The evaluation procedure for the WG-MACP model with Gibbs sampling is similar to the WG-MCD model with Gibbs sampling. The parameters of the WG-MACP model are also estimated on a monthly basis, as was achieved in Chapter 4, to facilitate the comparison of results before and after data substitution/infilling. To enhance accuracy in the infilling/substitution procedure in these case studies, the periodic function based calculations (Equation 3.2.7) for the WG-MACP model will be omitted.

6.2 Results and Discussion

It is noted that all stations were involved for infilling of missing data, simply because each station suffers from the need to infill/substitute missing data (Table 6.1). In such a case, it was considered prudent not to set any specific experiments but rather evaluate the efficacy of the WG-MCD and WG-MACP models with regard to infilling of missing data involving all stations in no particular order or consideration.

For this case study, parameters of the priori distribution will be based on the parameters of the WG-MCD or the WG-MACP model corresponding to 365 Julian days. However, the evaluation of procedures is based on daily values corresponding to a month such that there are 31 $\times$ 30 $\times$ 10 days of records in the month of January (i.e., number of days in January x number of years of precipitation records x number of stations in the study area). Likewise, there are 28 (29 in a Leap-year) $\times$ 30 $\times$ 10 days of records in the month of February; 31 $\times$ 30 $\times$ 10 days of records in the month of March; 30 $\times$ 30 $\times$ 10 days of records in the month of April and so on.

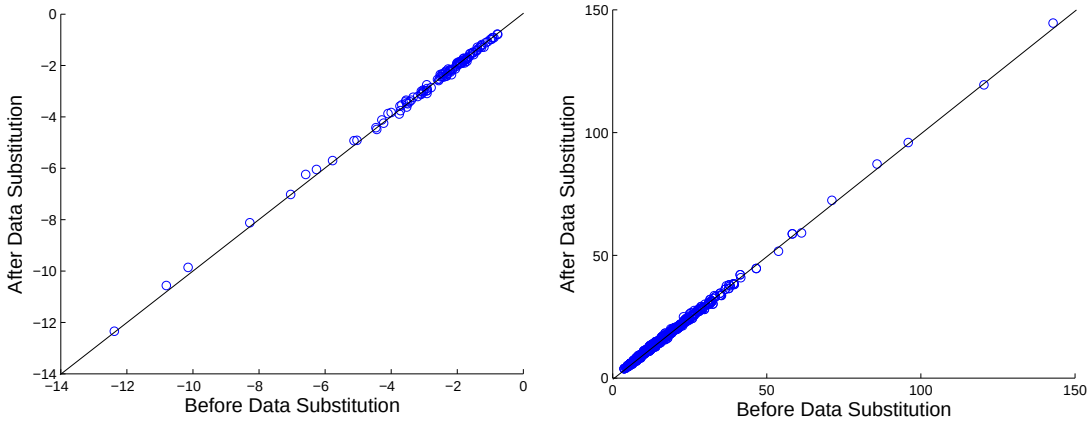
Such a pooling together of data into groups facilitates the interpretation of results and enhances the clarity in graphical presentations.

6.2.1 Performance Evaluation of the WG-MCD model with Gibbs Sampling

The WG-MCD model with Gibbs sampling is first to be evaluated by comparing the differences in attributes (e.g., means and covariances) that were obtained using parameters of the WG-MCD model and thus infilling of missing data. These two data sets are named as after substituted (treated/filled) data and before substituted (untreated/unfilled) data in the Figures 6.2a and 6.2b. A total of 120 mean values (10 stations $\times$ 12 months) and 540 covariance values (45 combinations of pairwise stations $\times$ 12 months as discussed in Chapter 4) were obtained from the WG-MCD model or calculated from the data after successful infilling/substitution.

Figures 6.2a and 6.2b illustrate the differences in parameters (i.e., mean, covariance) that were estimated from the infilled data set (before and after substitution/infilling) and by using the WG-MCD model. Each plot in Figure 6.2a represents a mean precipitation for a specific station and month and is calculated from infilled as well as untreated data. Likewise, each plot in Figure 6.2b represents elements in a covariance matrix between two stations for a specific month that were calculated using infilled and untreated data. As shown in these figures, the use of Gibbs sampling for infilling/substitution of the missing data can preserve the statistical characteristics of the historical observations that existed in the untreated data.

The proposed infilling procedure using the WG-MCD model with Gibbs sampling appears to be quite effective in preserving the statistical characteristics (i.e., mean and covariance) of a priori distribution (i.e., MCD) through the structure of covariance matrix. The WG-MCD model with Gibbs sampling thus can be regarded a promising



(a) Mean values

(b) Elements in covariances matrices

Fig. 6.2: Comparison of parameters obtained from the WG-MCD model and infilled data set of the first month.

procedure for infilling of missing observations in view of aforesaid encouraging results.

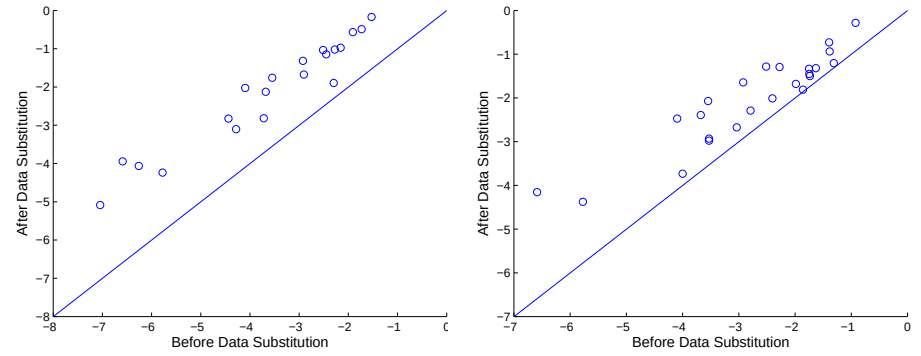
6.2.2 Performance Evaluation of the WG-MACP Model with Gibbs Sampling

The WG-MACP model with Gibbs sampling is then evaluated by comparing the differences in attributes (e.g., means, covariances, and lag-1 covariances) that were obtained using parameters of both untreated as well as treated data. Firstly, all stations were involved in the analysis as was considered for the evaluation of the WG-MCD model with Gibbs sampling and the corresponding results are presented in Figures 6.3a, 6.4a, and 6.5a. It can be apparent from the graphs that the correspondence between these parameters is less satisfactory as they tend to plot on one side of 1:1 line. One sided scatter of points in Figure 6.3 through 6.5 is an indication that the WG-MACP model with Gibbs method is less reliable for a scenario involving a large number of stations. Therefore, the worth of this procedure was tested by forming the groups with lesser number of stations. Consequently, five groups consisting of

stations [1 to 8], [1 to 6], [1 to 4], [1 to 3], and [1, 2] were formed and the aforesaid analysis repeated. It may be noted that the station numbers are labeled in the same order as shown in Table 3.2 (Chapter 3). These results are displayed in Figures 6.3 to 6.5 (b to f).

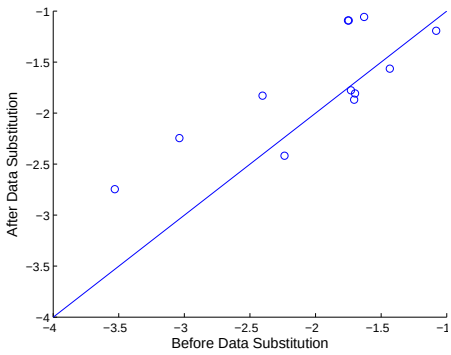
It is evident from the scatter of points in the graphs that the correspondence between statistics of treated and untreated data increases as the number of stations used in joint analysis decreases. In fact when only two stations are used (which lie in close proximity of each other), the results of data infilling show the best correspondence (Figures 6.3f through 6.5f). The reasons of such results can be attributed to the eigenvalue modification, which tend to decrease accuracy of the model while modifying covariance matrices to turn them into positive definite matrices. Also, to address an additional criteria of temporal dependence, the parameters responsible for the preservation of spatial dependence will need to be modified and subsequently the accuracy of the model is compromised. It is an interesting finding, which corroborates the common observation that precipitation events over a region display a large spatial variability (i.e. on a given day, it may not be raining at all sites), thus the model may prove less reliable for data infilling at multiple stations. However, there is a good chance that two or a few more adjoining sites may be receiving precipitation on that day. That means, a few stations probably with high spatial correlation should be considered for the purpose of data infilling.

Although, the correspondence in terms of means and covariance appears satisfactory when two adjacent stations are involved but the correspondence in terms of lag-1 covariance is still weak (Figure 6.5f). This behavior further affirms that due to spatial-temporal intermittency of precipitation events, the infilling capability of the

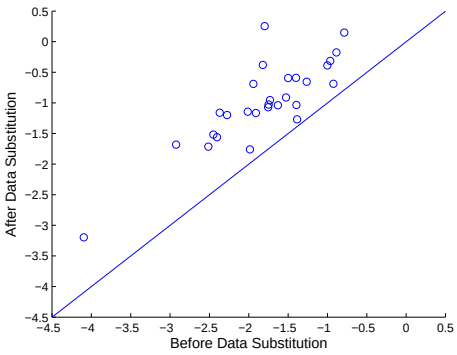


(a) 10 stations involved

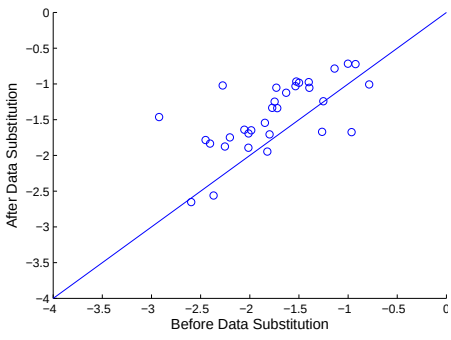
(b) 8 stations involved



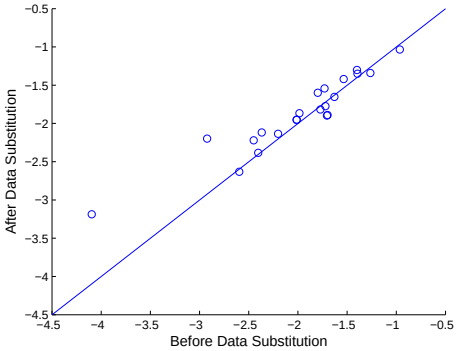
(c) 6 stations involved



(d) 4 stations involved

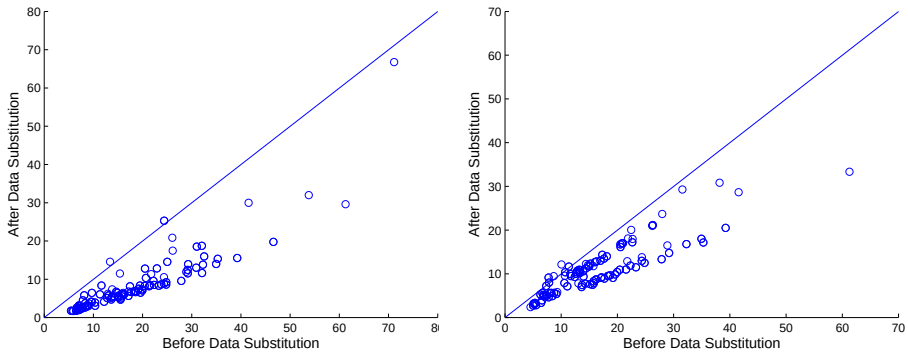


(e) 3 stations involved



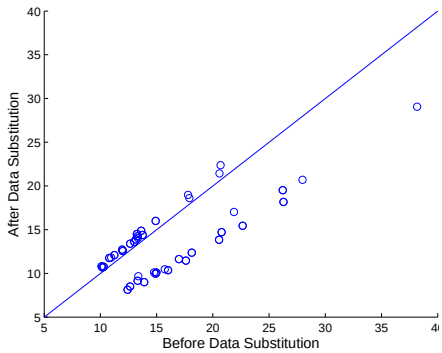
(f) 2 stations involved

Fig. 6.3: Comparison of mean values obtained from the WG-MACP model and substituted data set of 12 months at the 10 stations.

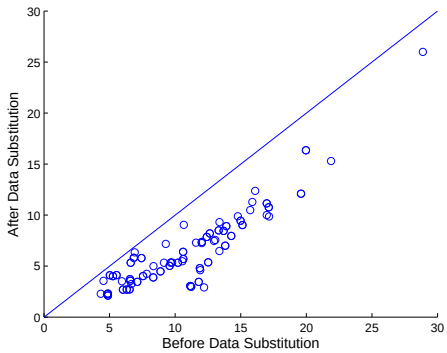


(a) 10 stations involved

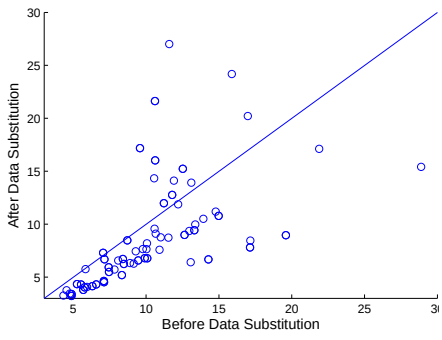
(b) 8 stations involved



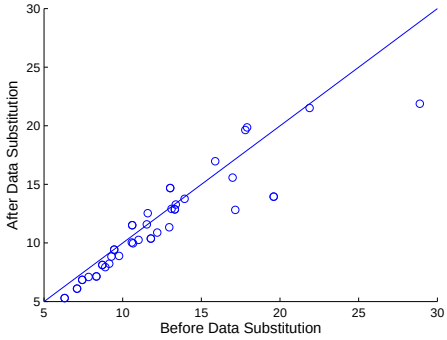
(c) 6 stations involved



(d) 4 stations involved

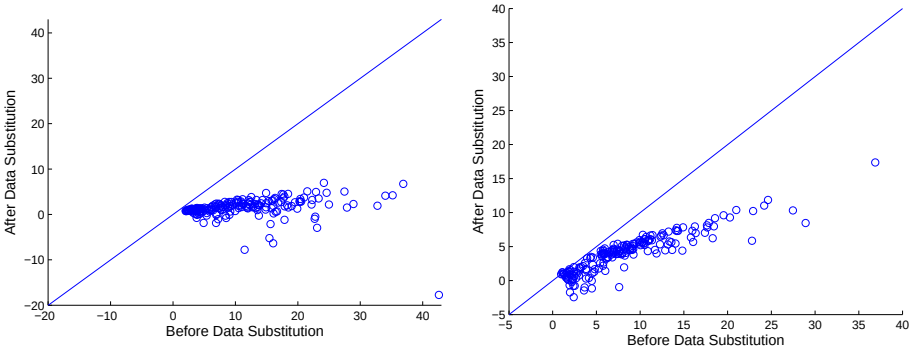


(e) 3 stations involved



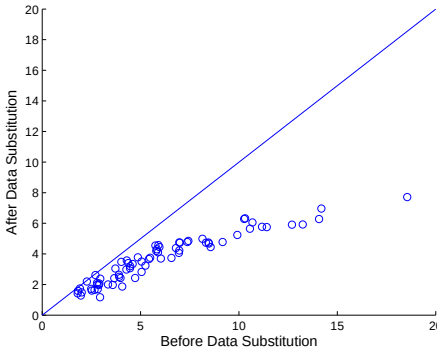
(f) 2 stations involved

Fig. 6.4: Comparison of covariance obtained from the WG-MACP model and substituted data set of 12 months at the 10 stations.

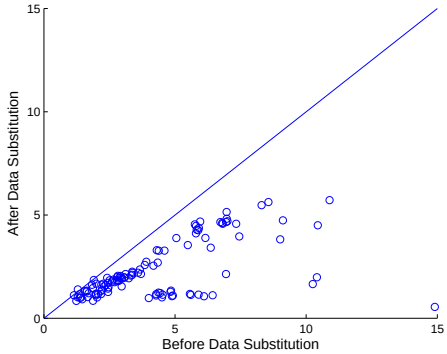


(a) 10 stations involved

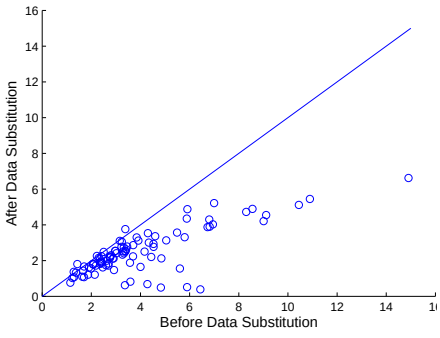
(b) 8 stations involved



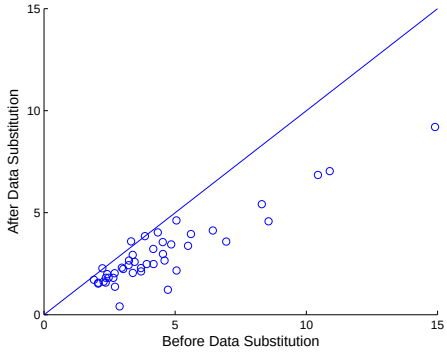
(c) 6 stations involved



(d) 4 stations involved



(e) 3 stations involved



(f) 2 stations involved

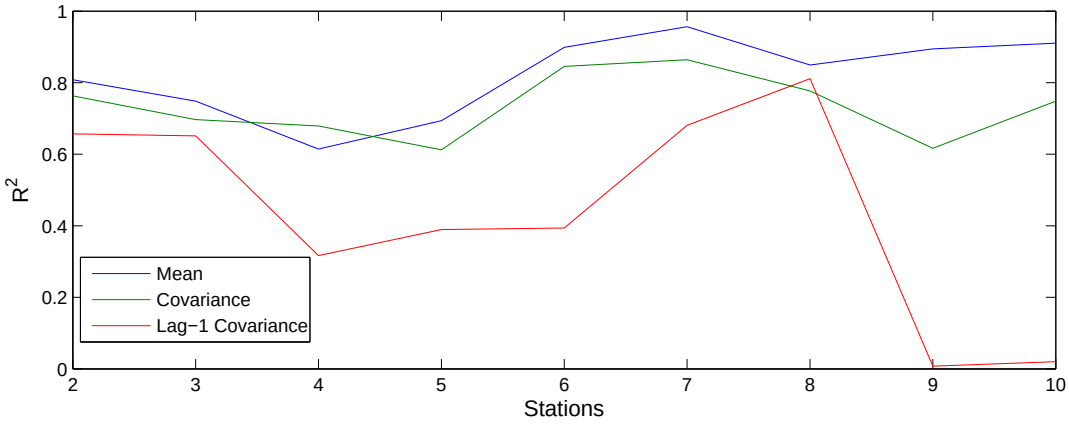
Fig. 6.5: Comparison of covariance at lag-1 obtained from the WG-MACP model and substituted data set of 12 months at the 10 stations.

method is modest.

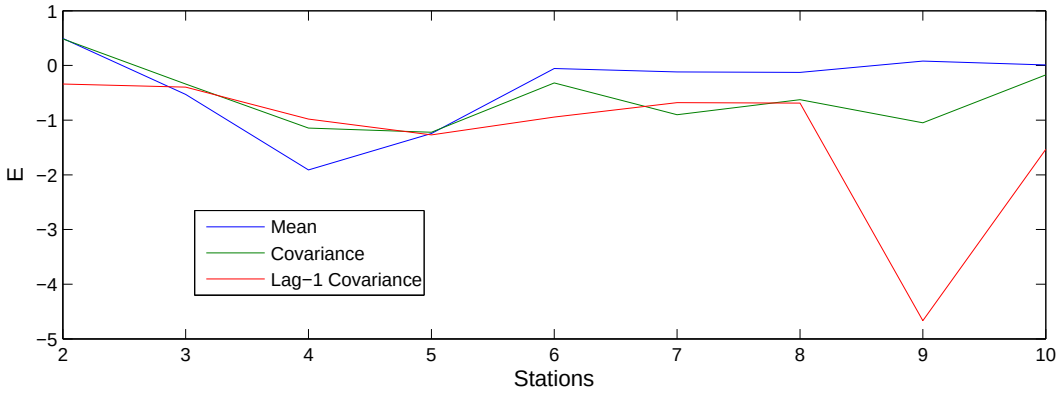
Although, the scatter of points in Figures 6.3 through 6.5 provides the first hand information on the capability of model in data infilling, the efficacy of model should be quantified through suitable statistical measures. Therefore, statistical measures R^2 (coefficient of determination), E (efficiency coefficient) and $RMSE$ (root mean square error) as utilized in chapter 3 were adopted. The values of the aforesaid statistical measures were computed for the attributes viz. mean, covariance and lag-1 covariance and are also displayed in Figure 6.6. The plots in Figure 6.6 generally suggest that no particular advantage can be achieved when more than two stations are involved in data infilling. In fact most striking inference can be drawn from the graph of E, in which the values of efficiency turn negative when more than two stations are used, meaning that the quality of infilled data is unlikely to improve with the involvement of several stations. The values of $RMSE$ also increase with increasing number of stations that further affirm the above inference. Similar performance is portrayed by the graph of R^2. Further, in general the graphs allude to the weak preservation of lag-1 covariance (values shown by red lines are lower than shown by green and blue lines), compared to the other attributes viz. mean and covariance.

6.2.3 Comparison of models (WG-MCD and WG-MACP) with Gibbs Sampling

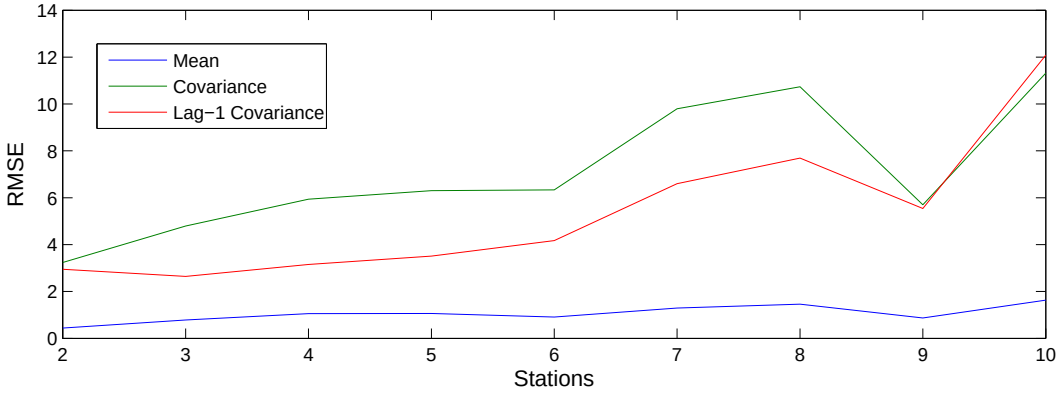
The mean and covariance are the common parameters that WG-MCD model with Gibbs sampling and WG-MACP model with Gibbs sampling, should preserve in the treated and untreated data. The plots of comparison of these parameters obtained from treated and untreated data (involving all 10 stations) are shown in Figure 6.2a and 6.3a (for means) and Figure 6.2b and 6.4a (for covariances). In general, it is apparent from graphical plots that the WG-MCD model with Gibbs sampling performs



(a) Result of R^2



(b) Result of E



(c) Result of RMSE

Fig. 6.6: Graphical displays of relationships between evaluation measures and precipitation stations in the study area.

better compared to the WG-MACP model with Gibbs sampling for which graphical plot displays a significant scatter. However, the statistical measure viz. R^2, E, and $RMSE$ were computed using the relevant data and are summarized in Table 6.2. It is clear from the values in Table 6.2 that the performance of the WG-MCD model with Gibbs sampling is much superior to the WG-MACP model with Gibbs sampling.

In summary, the use of the WG-MCD or WG-MACP models with Gibbs sampling is a satisfactory approach for infilling of missing observations that can reasonably preserve the statistical characteristics of the model. Through simulations based on the prior covariance structure, the spatial dependence among multiple sites can be preserved. The use of the WG-MCD model with Gibbs sampling for the infilling of missing data appears to be promising because the infilled data can still preserve the mean and covariance of a-priori distribution. However, if temporal dependence is to be considered, the WG-MACP model with Gibbs sampling would be a feasible option for infilling for cases involving fewer records.

Tab. 6.1: Comparison of differences in attributes obtained from before and after data substitution using MACP(m,1).

R^2 summary

	Station								
Month	2	3	4	5	6	7	8	9	10
Mean	0.81	0.75	0.61	0.69	0.90	0.96	0.85	0.89	0.91
Covariance	0.76	0.70	0.68	0.61	0.85	0.86	0.78	0.62	0.75
Lag-1 Covariance	0.66	0.65	0.32	0.39	0.39	0.68	0.81	0.01	0.02

Efficiency coefficient summary

	Station								
Month	2	3	4	5	6	7	8	9	10
Mean	0.49	-0.53	-1.91	-1.24	-0.06	-0.12	-0.13	0.08	0.01
Covariance	0.49	-0.34	-1.15	-1.22	-0.32	-0.90	-0.63	-1.05	-0.17
Lag-1 Covariance	-0.34	-0.40	-0.98	-1.27	-0.94	-0.68	-0.69	-4.67	-1.53

RMSE summary

	Station								
Month	2	3	4	5	6	7	8	9	10
Mean	0.44	0.78	1.05	1.06	0.91	1.29	1.46	0.87	1.63
Covariance	3.24	4.79	5.94	6.30	6.33	9.80	10.73	5.70	11.32
Lag-1 Covariance	2.95	2.65	3.15	3.51	4.17	6.60	7.69	5.54	12.08

Tab. 6.2: Values of R^2, E, and $RMSE$ obtained from attributes of normal distributions between observation and simulation of three models.

Evaluation Measures	Attributes	$WG-MCD$	$WG-MACP$
R^2	mean	**0.99**	0.44
	covariance	**0.99**	0.73
E	mean	**0.99**	0.04
	covariance	**0.99**	0.67
$RMSE$	mean	**0.18**	1.74
	covariance	**0.34**	1.62

7. APPLICATION III: IMPACT ASSESSMENT OF CLIMATE CHANGE

The WG-MACP model in Chapter four was demonstrated to be capable of generating synthetic realizations at multiple sites that preserve the statistical characteristics of the historical observations. At a precipitation station, the generated precipitation have been found to reasonably describe the current state of precipitation patterns; however, generation of precipitation patterns into the future may not be representative because of the impacts of global climate change. To ameliorate the proposed model for the purposes of addressing the impact of climate change scenarios in the future, additional statistical descriptors, corresponding to future precipitation patterns, are needed.

Conspicuous changes in regional precipitation patterns over the past few decades due to climatic change has been the focus of the intense attention of many researchers. In this regard, many physically-based numerical climatic models have been developed to generate the change in atmospheric patterns. Data from global numerical models have low spatial resolution that is usually inadequate for use in studies related to local hydrologic and water resource systems. Dynamic downscaling is capable of resolving the scale issues, but has the drawback of high computational costs. Therefore, there is a need to consider appropriate statistical downscaling approaches. Studies related to downscaling, along with appropriate output of the numerical climatic models, have been presented among others by *Wilks* [1992] and *Brandsma and Buishand* [1997].

Statistical downscaling methods are designed for use in the transformation of output from low to high spatial resolution in an effort to capture the impacts of global climate change at a local scale. Employing the output of the numerical global climatic model to generate local precipitation data not only includes the information on the physical process of the atmospheric circulation, but also aids in the projection of future change scenarios in precipitation patterns.

Many of the existing downscaling approaches require extensive estimation of parameters and statistical verification, which tend to limit the implementation of these methods [*Wilks*, 1998]. In addition, many of the existing downscaling approaches consistently underestimate the variance of the generated monthly and annual precipitation totals [*Goodess et al.*, 2003]. Therefore, approaches with simple structure and the capability of adequately describing the change of variance under the impact of climatic change, are preferable.

Although, a variety of methods are available to estimate values of weather variables at forthcoming times that are appropriate for local climate change impact assessment, a statistical downscaling approach often referred to as the "Delta change method" or the "change factor method" is commonly used [*Drogue et al.*, 2005]. Comparison between Delta change and other downscaling methods has been described by *Hay et al.* [2000] and *Diaz-Nieto and Wilby* [2005]. The Delta change method has been reported in the literature to be simple, flexible, and general enough to be applied in different climatic scenarios. As there are a number of ways by which the Delta change method can be used to estimate future climate scenarios, *Anandhi et al.* [2011] presented guidelines to help decide the best possible way for a specific implementation.

This chapter describes the details of precipitation generation with considerations

of global climatic change scenarios. Various aspects related to future statistical characteristics arising due to climate change have been investigated. The applicability of the Delta change method in capturing the impact of future climate change is specifically investigated and illustrated through a case study. The output of the Canadian regional climate model (CRCM) was selected for the demonstration of the Delta change method. The results and discussion are presented at the end of this chapter.

7.1 Methodology

The CRCM was developed through collaboration between the Canadian Regional Climate Modeling and Diagnostics Network, the ESCER Center of Université du Quebéc à Montréal, Environment Canada (CCCma), and the Ouranos Consortium. The CRCM simulations were done by the Ouranos Consortium and provided by the Ouranos Climate Simulation Team. For historical simulations, CRCM is driven either by NCEP/NCAR or by ERA40 global atmospheric re-analysis data, or by a Global Climate Model (GCM) data, and follows observed greenhouse gas and aerosol (GHG+A) concentrations. Re-analysis data sets were used at a 2.5 x 2.5 global latitude/longitude resolution. In future climate projections, the CRCM has followed various GHG+A scenarios from the IPCC (e.g., IS92a, A2), according to the driving GCM. The selected CRCM version was a CRCM 3.7.1 time-slice simulation for 2041-2070 driven by CGCM2, following IPCC scenario "A2" over the North-American domain (AMNO) with a 45-km horizontal grid-size mesh, 29 vertical levels and spectral nudging of large-scale winds. A total of 360 monthly precipitation data obtained from the CRCM data set has been considered for analysis. A grid area (Latitude from 50 to 49 and Longitude from 98 to 99) corresponding to the location of station

1 (Table 3.2) in the Manitoba region has been selected for illustration.

Daily and monthly precipitations represent two broad categories of data that are normally obtained from the output of climate models. These two categories of data possess different statistical characteristics that the downscaled data intends to capture. Due to inherent uncertainty in the generated daily data by climate models for future stages, many studies pertaining to precipitation generation and climate change are based on analyses of monthly data. Therefore, the discussion presented in this chapter also focuses on the monthly data.

Climate model output of daily data can provide additional information to describe future climate patterns. For instance, the use of monthly data is insufficient to describe changes in precipitation occurrences (wet-day frequency) at future stages. On the other hand, the downscaling of daily data can provide additional information. Therefore, how the WG-MACP model can be modified in association with the daily precipitation obtained from the climate model is briefly discussed in the Methodology section. Figure 7.1 shows a layout of a procedural plan that was followed to achieve the objectives of this chapter.

7.1.1 Downscaling the Monthly Data from Canadian Regional Climate Model

The basic idea of the Delta change method is to quantify the degree of change in a meteorological variable to be estimated by a climate model. Subsequently, a change is imposed on the variable of observed/historical series. In other words, the values of Delta are measured from the relative change between current and future climates that can be estimated from the output of CRCM. The changes are then applied to the simulation of the WG-MCD/WG-MACP model to ensure the simulated data possess

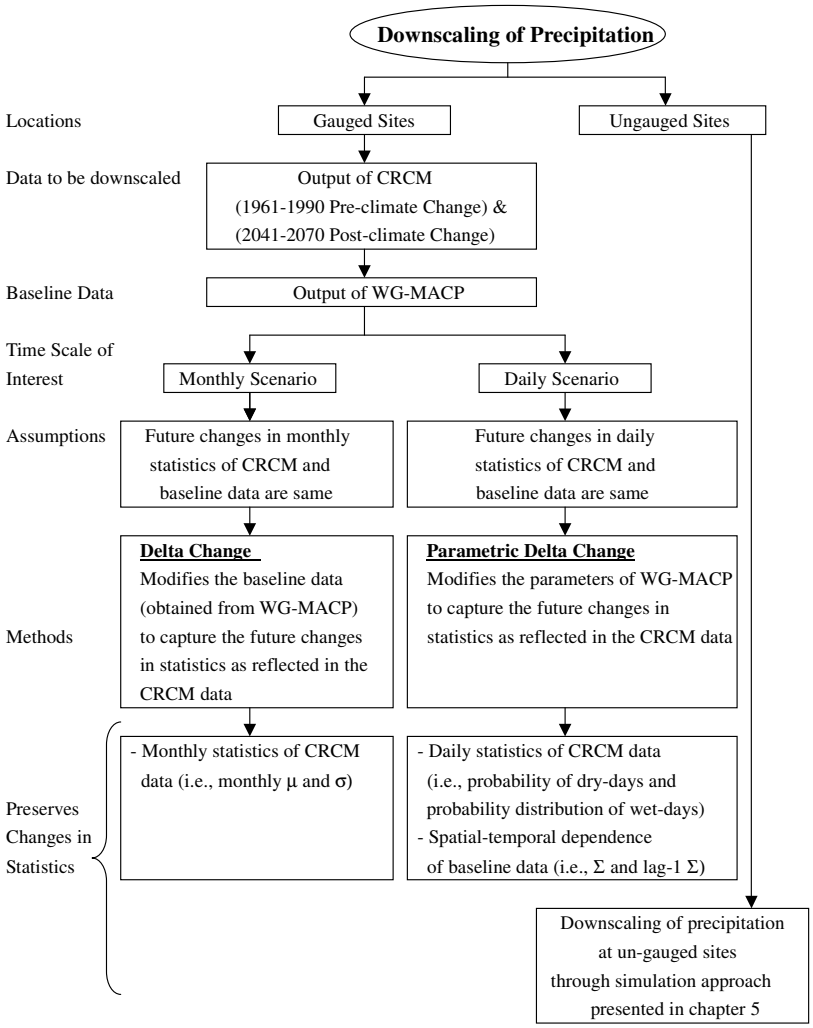


Fig. 7.1: Schematic structure of implementation III: downscaling monthly and daily output of CRCM using methods of Delta change and Delta adjustment of the WG-MACP parameters.

the similar effect of climate change as that of CRCM. A new scenario is created by perturbing the observed baseline series according to the monthly change factors [*Prudhomme et al.*, 2003].

The future change in atmospheric synoptic patterns described by the output of CRCM will serve as a reference for the change of the baseline time series at a local scale. In this thesis, the baseline time series is the simulated precipitation from the WG-MACP model. The baseline data obtained from the output of the WG-MACP model will be modified using the Delta change method to capture future changes in statistics (mean and standard deviation) as reflected in the CRCM data. Using the procedure discussed in Chapter 3, a baseline time series can be generated. Adopting the the WG-MACP model with Kriging or with-GRNN procedure, the downscaling can also be conducted at ungauged sites. For brevity, only the case of downscaling at the gauged sites will be considered. The necessary steps required for the implementation of the Delta change method are outlined below:

1. Select output (e.g., monthly precipitation amounts) of a numerical climatic model (e.g., CRCM).

2. Spatially interpolate [*Semenov and Brooks*, 1999] a monthly precipitation amount at a specific local site I from monthly precipitation amount of the region that was earlier obtained from a numerical climatic model.

3. For each month and location

$$x_{Baseline}^{future} = \overline{x}_{Baseline}^{present} + (x_{Baseline}^{present} - \overline{x}_{Baseline}^{present}) \times \Delta_{\sigma} + \Delta_{\mu} \tag{7.1.1}$$

where $x_{Baseline}$ and $\overline{x}_{Baseline}$ respectively denote the values and mean monthly

precipitation, as a baseline of observations, obtained from the local stations. In this case study, the baseline is described by the simulated output of the WG-MACP model, instead of historical observations. Δ_σ and Δ_μ denote the degree of change in monthly standard deviation and mean values.

$$\Delta_\sigma = \sigma_{CRCM}^{future} / \sigma_{CRCM}^{present} \tag{7.1.2}$$

where σ_{CRCM}^{future} and $\sigma_{CRCM}^{present}$ denote respectively the standard deviation of monthly precipitation obtained from the CRCM at the future and present stages.

$$\Delta_\mu = \overline{x}_{CRCM}^{future} - \overline{x}_{CRCM}^{present} \tag{7.1.3}$$

where $\overline{x}_{CRCM}^{future}$ and $\overline{x}_{CRCM}^{present}$ denote respectively the mean monthly precipitation obtained from the CRCM at the future and present stages. Equation 7.1.1 essentially modifies the current baseline time data using the Delta change obtained from CRCM output.

4. The $x_{Baseline\ Modified}$ is the expected monthly precipitation amounts obtained based on the local output of the WG-MACP model and using the change information from the CRCM. The monthly expected mean value at the future stage, $x_{Baseline\ Modified}$, is then to become the reference values that the simulated daily precipitation of the WG-MACP model must match. To match with the $x_{Baseline\ Modified}$, multiplying a factor to the simulated output of the WG-MACP model (i.e., daily precipitation of a specific month) allows the output of the WG-MACP model to match with the $x_{Baseline\ Modified}$. Another way of adjustment to match the daily precipitation output of the WG-MACP model

to statistical moments of the monthly precipitation output from the CRCM is to shift the mean values of the censored distribution. The shifting not only increases the overall monthly precipitation, but also changes the probability of wet-day occurrences.

7.1.2 Downscaling the Daily Data from Canadian Regional Climate Model

It is generally accepted that global climate change will produce differences in occurrences and amounts of daily precipitation such as more frequent precipitation occurrences and more intense rainstorm (i.e., precipitation amount). Statistically, this means that the probability of wet-day is likely to increase and the distribution of precipitation is likely to become more fat-tailed to the right. The change in these statistical characteristics can be captured by a change in the censored distribution. For example, a shift in the mean of the censored distribution will increase the probability of wet-days. Therefore, the Delta change method can be considered to modify the parameters of the underlying distribution. The following describes the necessary procedural detail related to the Delta Change method.

1. Use estimation procedures of the WG-MACP model (Steps 1 to 4 in Chapter 4) to calculate the parameter set (i.e., $\boldsymbol{\Upsilon}_{\boldsymbol{CRCM}}^{\boldsymbol{Present}}$) for the daily precipitation output obtained from the CRCM at the present stage.

2. Apply the above parameter estimation procedure but to the daily output of the CRCM at the future stage to obtain (i.e., $\boldsymbol{\Upsilon}_{\boldsymbol{CRCM}}^{\boldsymbol{Future}}$).

3. Calculate the changes (i.e., Delta) of these parameters through:

$$\Delta_{\Upsilon} = \Upsilon_{CRCM}^{Future} \; / \; \Upsilon_{CRCM}^{Present} \tag{7.1.4}$$

4. Use the estimation procedures of the WG-MACP model to calculate parameters of the daily observations at present stage (i.e., $\boldsymbol{\Upsilon_{Baseline}}$) that also serves as the baseline for the method.

5. Modify the parameters (i.e., $\boldsymbol{\Upsilon_{Baseline}}$) of the baseline model by multiplying Δ_{Υ} with the corresponding parameters: $\Upsilon_{Baseline\ Modified} = \Upsilon_{Baseline} \times \Delta_{\Upsilon}$

6. Simulate the WG-MACP model using the parameters $\boldsymbol{\Upsilon_{Baseline\ Modified}}$).

The above described method provides a simple and straightforward method for the modification of the WG-MACP model that addresses the change in the climate patterns of the local daily precipitations. The proposed method primarily involves daily precipitation data.

7.2 Results and Discussion

This section examines the capability of the Delta change method to preserve the change in statistical characteristics (i.e., mean and standard deviation) of the monthly precipitation output of the CRCM.

Simulated output of the CRCM provides a quantitative forecast of the change in weather variables (e.g., precipitation) due to global climate change. The statistical change in the output of CRCM suggests the required changes in the generated precipitation of the proposed model at a local scale. The generated precipitation is modified using the Delta change method.

For the baseline time series, a total of 360 monthly precipitation values obtained from the simulated daily precipitations of the WG-MACP were adopted for the analysis. The WG-MACP output at station 1 in the Manitoba region was selected for evaluation of the method. The parameters of the WG-MACP model were evaluated based on the observed/historical data that were obtained from the Canadian Daily Climate Data (CDCD), Environment Canada. The procedural details of the parameter estimation have been presented in Chapter 3 and the historical observations for the period from January 1961 to December 1990 were used for parameter estimation. For a comparative analysis, the monthly precipitations of the baseline time series are modified using one of the two procedures, namely, the regression analysis and the Delta change method. The goal of the modification is to create a baseline time series with similar standard deviation as that of the output of the CRCM.

As a first modification, the regression analysis was adopted to establish a relationship between the baseline time series of the WG-MACP model (i.e., dependent variable) and the output of the CRCM (i.e., independent variable). For this purpose, the baseline time series was modified using a two-parameter linear regression model. The two parameters were estimated by the ordinary least square method in Equation 7.2.1. The use of parameters in the regression model allows the output of the WG-MACP model to match in terms of statistical characteristics with output of the CRCM.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y} \tag{7.2.1}$$

where $\hat{\boldsymbol{\beta}}$ denotes two regression parameters to be estimated with size of $[2 \times 1]$. $\mathbf{X}$ refers to the monthly precipitation of the CRCM with size of $[30 \times 2]$ in which the

first and second columns of the matrix refer respectively to the 30 years of monthly precipitation of a specific month and values of 1, which is corresponding to the constant parameter to be estimated. The $\mathbf{y}$ refers to the baseline monthly time series obtained from the WG-MACP model with size of [30 $\times$ 1], which refers to the 30 years of monthly precipitation in a specific calendar month.

For the second modification, Equation 7.1.1 was utilized. Theoretically, the use of Equations 7.1.2 and 7.1.3 ensures the degree of change in parameter will be consistent between the baseline time series and output of the CRCM. However, the implementation of the delta change to the baseline time series does not have to result in the same monthly mean and standard deviation as parameters of the CRCM at a future stage. It is noted that only the degree of change is comparable in the two cases. To facilitate a comparison of parameters, the $\sigma_{CRCM}^{present}$ and $\overline{x}_{CRCM}^{present}$ of Equations 7.1.2 and 7.1.3 will be replaced by $\sigma_{Baseline}$ and $\overline{x}_{Baseline}$ for the evaluation. Such a change allows a direct comparison in terms of magnitude of the mean and standard deviation obtained from the modified baseline time series and the output of CRCM at a future stage.

For evaluating the adequacy of the modified Delta change method, a preliminary comparative study was conducted. Methods of the regression analysis and the Delta change are first applied to modify the baseline time series for estimating the parameters corresponding to a specific month. The monthly mean and standard deviation of the modified observed daily precipitation data are then calculated and compared with the expected precipitation output of the CRCM to evaluate the methods.

To evaluate the performance of modification procedures, the mean and standard deviation of the baseline time series, output of the CRCM, the baseline time series with modification 1, and the baseline time series with modification 2 are calculated

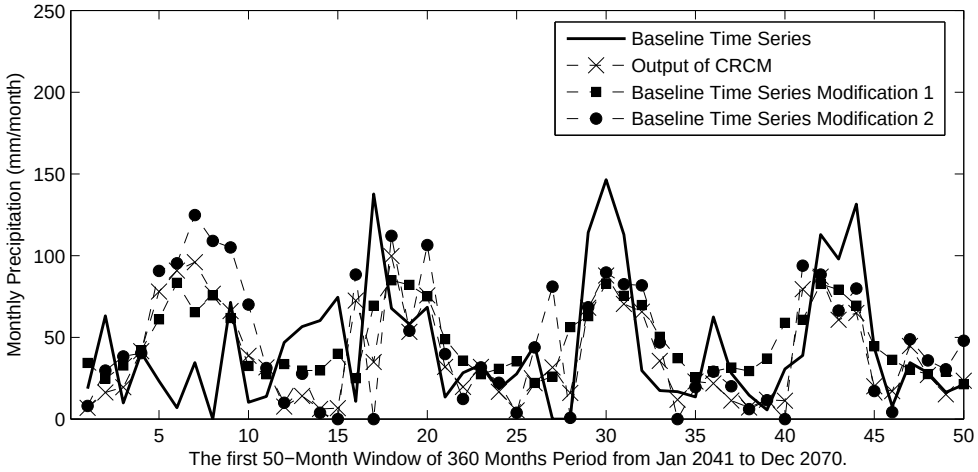


Fig. 7.2: Comparison of monthly precipitations from Jan 2041 to Feb 2045.

and compared. Figures 7.2 to 7.4 respectively illustrate the differences with respect to the monthly precipitation amounts, their means, and standard deviations.

In Figure 7.2, the outputs of CRCM underestimate the variability in monthly precipitation during the summer period. Both modification methods appear to be functioning well in modifying values of monthly precipitations closer to output of the CRCM; however, the figure is not easy to interpret and it is not clear to discern as to which of the modification procedure is superior.

The mean and standard deviation of monthly precipitation were calculated based on the 360 monthly precipitation data and are shown in Figures 7.3 and 7.4. In Figure 7.3, the difference in mean values between the baseline time series and output of the CRCM is modest. According to the output of CRCM, the figure shows a reduction in precipitation during the winter period (December to March) as impacted by climate change, but the precipitation during the spring period (April to June) seems to remain unchanged with respect to the current stage.

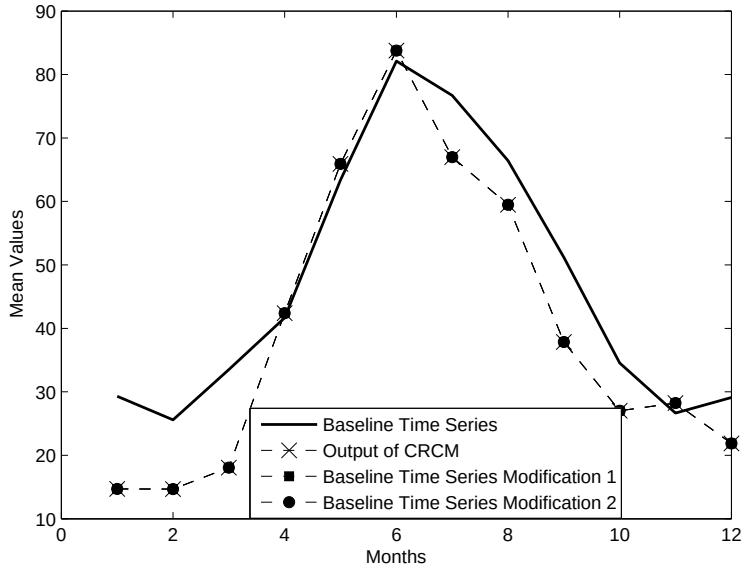


Fig. 7.3: Comparison of the mean values of monthly precipitations.

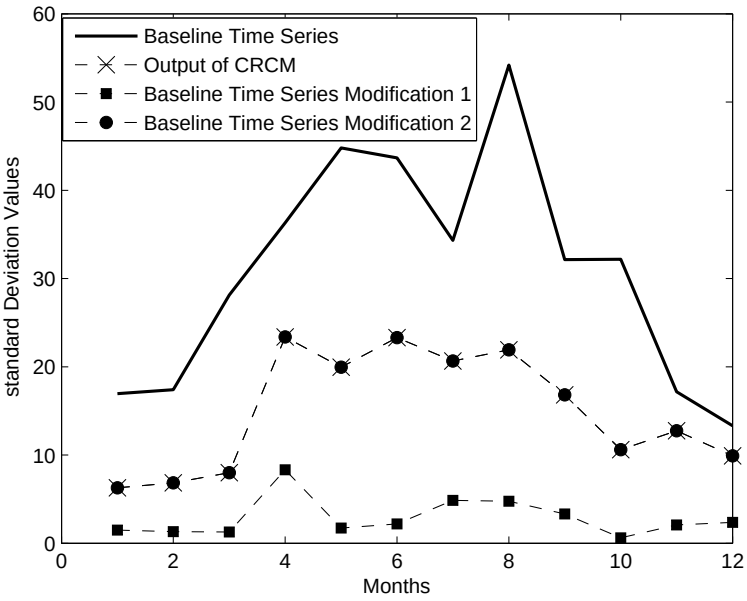


Fig. 7.4: Comparison of the standard deviation values of monthly precipitations.

In Figure 7.4, only the modification 2 improves the fitness of the standard deviation value in each month. The modification 2 can reasonably modify the baseline of daily precipitation data to match the expected precipitation output of the CRCM. Modification 2, which uses the Delta change method seems superior to modification 1, which uses the regression method. From a comparison of the overall standard deviations between the baseline time series and output of the CRCM, suggests that the variance of monthly precipitation is likely to be less than current stage. This finding seems to contradict the general perception, in which the variability in future precipitation events is projected to increase. In recent years, experience with occurrences of extreme precipitation events tend to corroborate the increase in future precipitation events. In other words, there is a need to ameliorate the present modeling techniques to be able to simulate the variance in the generated precipitation sequences.

In conclusion, the Delta change method was found to be preferable over the regression method in modifying the baseline time series to match the statistical characteristics of output of the CRCM. Therefore, the proposed WG-MACP-Delta change method for the generation of spatio-temporally dependent precipitation data is a reasonable method capable of projecting the impact of climatic change at the future stage. Compared to traditional statistical downscaling methods [*Wilby and Wigley*, 1997], the proposed method is simple to implement and yet general enough for wider implementations. Another merit of the proposed method over the traditional methods is that the change in future variance can be described through modification of the generated data (i.e., baseline time series).

It is noted that the aim of this chapter was not to search for the best climate model applicable to the study area, but to illustrate the use of Delta change method with

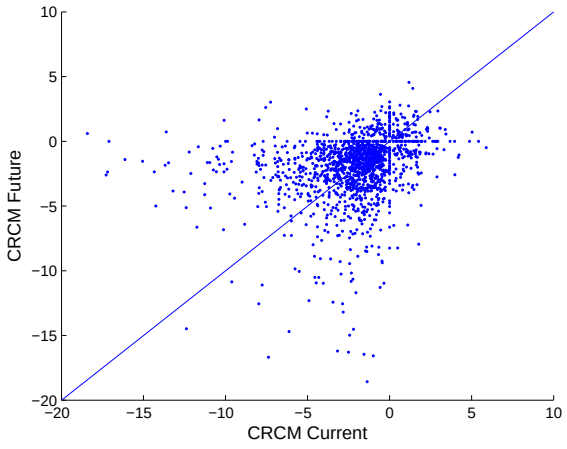
the proposed WG-MACP model in managing the downscaling problem. Therefore, theoretically, the method should be applicable to different kinds of weather variable (e.g., max/min temperature) of climate model and is not restricted to output of the CRCM.

For the analysis of WG-MACP-Parametric Delta change method, the change in daily statistics reflected by the CRCM has been investigated. Figures 7.5 to 7.7 illustrate the estimated change in the mean, standard deviation, lag-1 correlation, correlation, and probability of wet-day occurrences evaluated by the WG-MACP-Parametric Delta Change method.

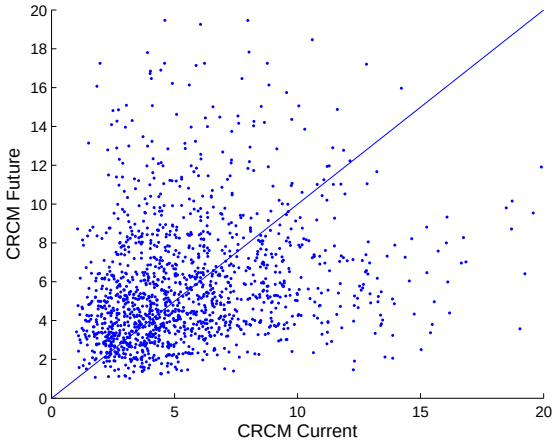
The daily precipitation of CRCM obtained from years 1961-1990 and years 2041-2070 are adopted for the analysis. For a specific Julian day and location, each point shown in these figures represents an attribute (i.e., mean, standard deviation, correlation, or lag-1 correlation) calculated from the parameters of WG-MACP using the current and future data of CRCM. Data of four gridded sites obtained from the CRCM data set were used for the analysis. The four gridded sites encompass the area of interest corresponding ten gauged sites selected in this thesis. A total of 1,460 points (attributes of 365 Julian days $\times$ 4 gridded sites) are presented in each graph of the Figures 7.5 and 7.7. A total of 2,190 points (attributes of 365 Julian days $\times$ 6 pair-wise combinations of gridded sites) are presented in each graph of Figure 7.7. A visual comparison does no yield any visually noticeable change in these figures.

It is noted that the scatter plots in Figure 7.7 are discrete that is because there are only 30 years of realization for the analysis. The results thus are in a grid of 30×30.

To further investigate the degree of change in the parameters, some quantitative

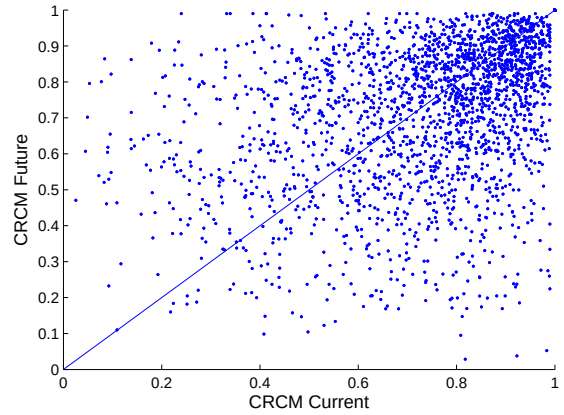


(a) Mean values

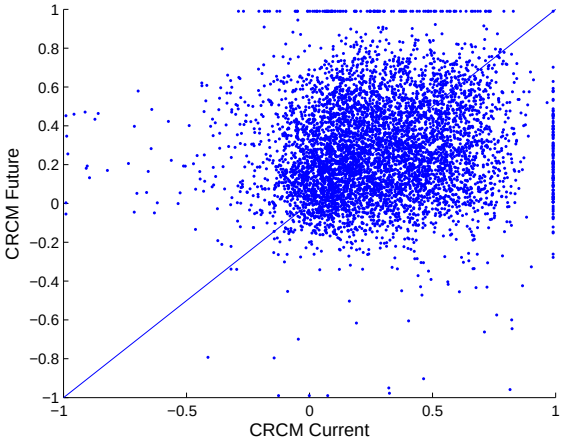


(b) Standard deviations

Fig. 7.5: Comparison of the change on mean values and standard deviations estimated by the WG-MACP using data of CRCM at current and future stages.



(a) Elements in lag-1 correlation matrices



(b) Elements in correlation matrices

Fig. 7.6: Comparison of the change on correlations and lag-1 correlations estimated by the WG-MACP using data of CRCM at current and future stages.

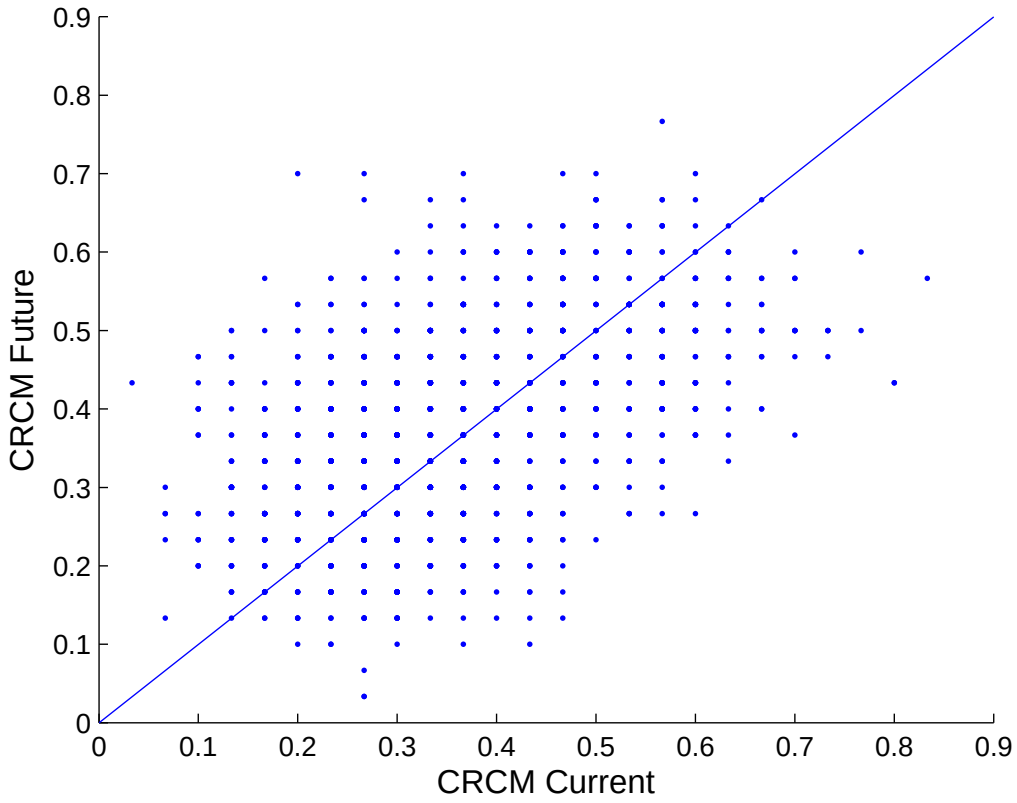


Fig. 7.7: Comparison of the change on probability of wet-day occurrences obtained from CRCM data at current and future stages.

Tab. 7.1: Summary of changes in attributes obtained from CRCM data at current and future stages.

Attributes	Mean	Standard Deviation	Correlation	Lag-1 Correlation	Probability of Wet-day Occurrences
Points above diagonal lines in figures (%)	52	52	46	50	51
Average change (%)	9	31	8	-1	17

measures are given in Table 7.1. The first row in the table refers to the number of points (i.e. a parameter measured for a specific month and station) that indicate an increase. In cases where future stages have no change compared to current stages, one may anticipate all points may fall along the diagonal line, or 50% of points above diagonal lines in these figures. An examination of these figures revealed that more than 50% points were above the diagonal line thus clearly implying that more attributes with respect to different Julian days and gridded sites are likely to increase at future stages. In the second row of the table, the average of means in the case of WG-MACP model has shown a nearly 9% increase, while the standard deviation has shown a noticeable, on the average, an increase of 31%. This suggests that the mean of censored distributions has shifted to the right-hand side and the increase in standard deviation will result in more extreme precipitation observed in the future. The increase of 17% in the probability of wet-day occurrences is the results of the changes in mean and standard deviation. Therefore, the results indicate a high frequency of wet-day occurrences and extreme precipitation is expected in the future.

8. CONCLUSIONS AND RECOMMENDATIONS

The overall objective of this research was to develop a versatile stochastic weather generator, capable of generating daily precipitation amounts that preserve the statistical characteristics of the observed precipitation data at multiple sites. With minor modifications, the weather generator has been found to be applicable for diverse purposes. This thesis is particularly focused on the model development and encompasses studies related to various applications.

The greatest challenge in modeling of daily precipitation at multiple sites is that the spatial dependence at multiple sites cannot be easily described by a simple correlation (Pearson correlation coefficient). This is because daily precipitation events have a mixed distribution, as they are riddled with zero and positive values. Most multiple-site models use indirect ways to preserve the spatial dependence in precipitation and are often based on the two-stage approach in which the occurrence process and the amount process are treated separately.

In this thesis, an alternative method has been proposed for the generation of daily precipitation based on an integrative structure that is different from conventional two-stage approaches because it involves only one-stage approach. A multivariate censored distribution (MCD) and a multivariate autoregressive censored process (MACP) have been developed for the generation of daily precipitation at multiple sites. The

adequacy of MCD and MACP have been validated by comparing synthetic and observed daily precipitation data. Daily precipitation records for the period 1961 to1990 from 10 precipitation stations located in Manitoba, Canada, have been utilized for the analysis. The proposed model has been found to be suitable for reproducing the statistical characteristics of daily historical precipitations at multiple sites. The spatial and temporal dependencies appear to have been reasonably captured by the covariance and lag-1 covariance of the WG-MACP model. Other descriptors, such as probabilities of wet/dry-day, mean values, and variances, show good statistical agreement with similar statistical characteristics of the observed data sets. These models are also found to be capable of capturing the unique characteristics of non-normally distributed precipitation data sets, such as the occurrence of a high frequency of zero records and intermittent precipitation observations.

In the evaluation section, it was found that a better performance of the method could be obtained firstly with a smaller number of stations involved, and secondly with less statistical attributes needed to be preserved. The implementation of modification in the covariance matrix with the periodic function of parameters will reduce the accuracy of the result in preserving the observed precipitation statistics. However, implementing these two measures will ensure that covariance matrices are positive definite and consequently will significantly reduce the number of parameters in analyses. For improved results, researchers or practitioners can consider reducing the number of stations. As well, omission of some of the insignificant attributes may also improve the result. For example, if the temporal dependence in a data set is found to be insignificant, instead of using the WG-MACP model, the use of a WG-MCD model will be considered sufficient.

The potential and procedural details of using the proposed method in handling daily precipitation data have been illustrated. Specifically to evaluate the adequacy of the proposed model, three applications involving precipitation generation were investigated: [1] precipitation simulation at ungauged sites, [2] handling precipitation data set with missing observations, and [3] precipitation generation using a regional downscaling climatic model. The three applications also encompass different scales of analysis. Specifically, the infilling of missing observations represents a study at micro-scale, simulation at ungauged sites represents a study at meso-scale, and downscaling of the output of global climate model is a study at macro-scale.

Some salient findings from the study are as follows:

- In the case on the WG-MACP model for precipitation simulation at ungauged sites, the use of the WG-MACP-GRNN procedure is applied for generating animated simulations (3-D Plots) of precipitation at gridded sites. The use of WG-MACP-GRNN procedure with such a capability of presentation can reveal the scope, magnitude, location, direction, center of rainstorm, and precipitation pattern. Such a visual presentation facilitates the analysis related to the simulated precipitation in the study area. The use of the the WG-MACP-Kriging procedure takes into account the varying orography by the covariance structure of simulated precipitations at gridded sites.

- In the case on the WG-MACP model for handling missing observations, the parameters obtained from the spatially dependent model (i.e., the WG-MCD or WG-MACP) can provide a prior distribution to allow for the estimation of missing observations through conditional drawing of samples from the model via Gibbs Sampling. In cases where weak temporal dependence exists in the

observed data, the WG-MCD model was found to be a simple yet accurate model in preserving the statistical properties of the data set. The WG-MACP model with Gibbs sampling turns out to be less satisfactory. Another stark feature revealed by the aforesaid procedure is that for data infilling less number of stations are more desirable than the inclusion of large number of stations.

- In the case related to the WG-MACP model for purposes of a climate change study, the WG-MACP model can be used for downscaling the output of a regional climate model for projecting the future change in precipitation pattern at a local scale. The generated precipitation was found to agree with the output and the projection of the regional climate model. The Delta change method was found to be a simple method that [1] engages the physical processes into the simulated data, [2] projects future change of weather pattern to the local generated precipitation, and [3] disaggregates data from low to high spatial-temporal resolution. The results of analysis indicate frequent occurrences of wet-days and extreme precipitation events in the future.

The intentions of selecting those techniques with the WG-MACP model is to illustrate the potential of the model as well as the procedural details of how these hybrid procedures can be set up to handle problems pertaining to hydrology/climatology.

Several unique characteristics and advantages of the the WG-MCD and WG-MACP models are identifiable from the thesis:

- The estimation of correlation in the proposed models, via likelihood function estimate based on the four quadrants of a bivariate distribution, is comparatively simple than the traditional estimation method [*Bardossy*, 1992].

- The use of a periodic function with four harmonics to replace the 365 parameters of the WG-MACP model reduces the number of parameters for modeling.

- The estimation of the covariance matrix of the multivariate distribution is conducted by conglomeration of pair-wise covariances that are obtained by maximizing the likelihood function of bivariate distributions corresponding to a pair of any two stations. The proposed method requires less computational resources to handle a large-scale multivariate problem, as the elements of covariance matrix are estimated individually.

- To ensure the consistency among the elements of the covariance matrix, an eigenvalue modification procedure was tested and was found to be effective in ensuring that the covariance matrix become a positive definite.

- Since the missing observations will be omitted from the univariate and bivariate parameters estimation process, all available information is fully utilized and thus modeling is not affected by missing observations.

- In general, models with integrative forms similar to the WG-MCD and WG-MACP models that avoid separating the simulation tasks into two stages are relatively parsimonious, concise in their analytical forms, simple in operation, and can address the aspects of spatial and temporal dependencies equally. Because of the multivariate normal structure, these models are well-adapted to a variety of of potential applications.

In summary, the differences of this research study from the studies in the literature are identified as follows:

- A maximum likelihood method was proposed for the estimation of correlation based on the four quadrants of a bivariate distribution. The method is comparatively simple compared to the traditional estimation method. Periodic functions are applied to reduce the number of parameters and to remove the affects of daily fluctuations in the estimated parameters.

- The non-parametric GRNN method has been integrated in to the structure of the WG-MACP model for the simulation/generation of daily precipitation at ungauged sites. The covariogram of Kriging has been adopted for the expansion of the dimension of the WG-MACP model.

- A substitution method has been implemented in the structure of the WG-MCD model with Gibbs sampling to preserve the spatial dependence in historical observations for the infilling of missing observations. The spatio-temporal dependence exhibited in historical data sets has been preserved through an iterative generation from the WG-MACP model involving conditional distributions for infilling of missing observations.

- The WG-MACP model has been adopted as a baseline series for the downscaling of CRCM scenarios. A parametric Delta change method has been proposed to capture the change in mean, covariance, and lag-1 covariance of the daily precipitation data obtained from the CRCM.

Several possible studies, extensions, and implementation of the WG-MACP model can be considered in a future study. First, one could focus on improving the accuracy of the proposed model in preserving the observed statistics when a large number of stations are involved. It was found in this thesis that there is a trade-off between the

accuracy in preserving the statistics of historical observations and the dimensions of the WG-MACP model. In other words, a better performance is expected with the use of a small number of stations. To reduce the number of stations, techniques including k-mean clustering or supportive vector machine can be considered to group correlated stations together which in turn reduces the dimension of the WG-MACP model. Also, it was found that the temporal dependence in the precipitation sequences was not large, but at the same time cannot be neglected.

Second, for the study on simulations at ungauged sites, other hybrid techniques could be considered with the WG-MACP model to handle a specific problem. For instance, instead of using the GRNN procedure, alternative approaches of artificial neural network or k-NN could also be considered for the simulation of data at ungauged sites.

Third, for the study on infilling missing observations, the performance of WG-MCD model with Gibbs sampling can be compared with other similar methods. For enhancing accuracy in preserving the temporal dependence aspect in estimates of missing observations through the use of the WG-MACP with Gibbs sampling procedure requires further investigations for improvement.

Fourth, for the study on downscaling of CRCM scenarios, the study could be extended to a larger region to further validate the adequacy of the method. Also, a comprehensive sensitivity study could be conducted through downscaling of the output of several regional climate models and then compare the downscaled output of the WG-MACP with Parametric Delta change method. The result related to the changes in parameters can be applied to re-evaluate the adequacy of the existing water resources systems. For example, the existing design of water resources systems

(e.g., reservoir, watershed management systems, or flood mitigation structures etc.) can be re-evaluated based on the generation of future precipitations from the changed parameters of the WG-MACP model.

BIBLIOGRAPHY

Akintug, B., and P. F. Rasmussen, A Markov switching model for annual hydrologic time series, *Water Resources Research*, *41,*(W09424), 2005.

Ali, A., T. Lebel, and A. Amani, Rainfall estimation in the Sahel. Part I: Error function, *Journal of Applied Meteorology*, *44*, 1691–1706, 2005.

Anandhi, A., A. Frei, D. C. Pierson, E. M. Schneiderman, M. S. Zion, D. Lounsbury, and A. H. Matons, Examination of change factor methodologies for climate change impact assessment, *Water Resources Research*, *47*, 3501–3502, 2011.

Andreasson, J., and J. Rosberg, Moving on from delta change towards direct use of RCM output by scaling - A method for transient impact simulations, *Geophysical Research Abstracts*, *8*, 03,820, 2006.

Apipattanavis, S., G. Podesta, B. Rajagopalan, and R. Katz, A semiparametric multivariate and multisite weather generator, *Water Resources Research*, *43*, 1–19, 2007.

Baigorria, G. A., and J. W. Jones, Gist: A stochastic model for generating spatially and temporally correlated daily rainfall data, *Journal of Climate*, *23*, 5990–6008, 2010.

Bardossy, A., Space-time model for daily rainfall using atmospheric circulation patterns, *Water Resources Research*, *28*(5), 1247–1259, 1992.

Bardossy, A., and G. G. S. Pegram, Copula based multisite model for daily precipitation simulation, *Hydrology and Earth System Sciences*, *13*(12), 2299–2314, 2009.

Beersma, J. J., and T. A. Buishand, Multi-site simulation of daily precipitation and temperature conditional on the atmospheric circulation, *Climate Research*, *25*, 121–133, 2003.

Bishara, A. J., and J. B. Hittner, Testing the significance of a correlation with non-normal data: comparison of pearson, spearman, transformation, and resampling approaches, *Psychological Methods*, *3*, 399–417, 2012.

Bowden, G. J., H. R. Maier, and G. C. Dandy, Input determination for neural network models in water resources applications. Part 2. Case study: forecasting salinity in a river, *Journal of Hydrology*, *301*(1-4), 93–107, 2005.

Brandsma, T., and T. A. Buishand, Statistical linkage of daily precipitation in Switzerland to atmospheric circulation and temperature, *Journal of Hydrology*, *198*, 98123, 1997.

Bras, R. L., and I. Rodriguez-Iturbe, *Random Functions and Hydrology*, Addison-Wesley, Reading, Massachusetts, 1985.

Breinl, K., T. Turkington, and M. Stowasser, Stochastic generation of multi-site daily precipitation for applications in risk management, *Journal of Hydrology*, *498*, 23–35, 2013.

Brissette, F. P., M. Khalili, and R. Leconte, Efficient stochastic generation of multi-site synthetic precipitation data, *Journal of Hydrology*, *345*, 121–133, 2007.

Buishand, T. A., and T. Brandsma, Multisite simulation of daily precipitation and temperature in the rhine basin by nearest-neighbor resampling, *Water Resources Research*, *37*, 2761–2776, 2001.

Burton, A., C. G. Kilsby, H. J. Fowler, P. S. P. Cowpertwait, and P. E. O'Connell, RainSim: A spatial-temporal stochastic rainfall modelling system, *Environmental Modelling and Software*, *23*(12), 1356–1369, 2008.

Cannon, A. J., Probabilistic multisite precipitation downscaling by an expanded Bernoulli gamma density network, *Journal of Hydrometeorology*, *9*(6), 1284–1300, 2008.

Chang, T. J., M. L. Kavvas, and J. W. Delleur, Daily precipitation modeling by discrete autoregressive moving average processes, *Water Resources Research*, *20*(5), 565–580, 1984.

Chaudill, M., GRNN and bear it, *AI Expert*, *8*(5), 28–33, 1993.

Chowdhury, M., A. Alouani, and F. Hossain, Comparison of ordinary kriging and artificial neural network for spatial mapping of arsenic contamination of groundwater, *Stochastic Environmental Research and Risk Assessment*, *24*, 2010.

Clark, M. P., A resampling procedure for generating conditioned daily weather sequences, *Water Resources Research*, *40*, 4, 2004.

Cohen, A. C. J., Estimating the mean and variance of normal populations from singly truncated and doubly truncated samples, *The Annals of Mathematical Statistics*, *21*(4), 557–569, 1950.

Cohen, A. C. J., Simplified estimators for the normal distribution when samples are singly censored or truncated, *Technometrics*, *1*(3), 217–237, 1959.

Cowpertwait, P. S. P., A generalized spatial-temporal model of rainfall based on a clustered point process, *Proceedings: Mathematical and Physical Sciences*, *450*(1938), 163–175, 1995.

Cowpertwait, P. S. P., C. G. Kilsby, and P. E. O. Connell, A space-time Neyman-Scott model of rainfall: Empirical analysis of extremes, *Water Resources Research*, *38*(8), 2001WR000,709, 2002.

Cressie, J. P., *Statistics for Spatial Data*, John Wiley, New York, 1993.

DARA, and the Climate Vulnerable Forum, *Climate Vulnerability Monitor Second Edition*, Fundacion DARA International, Madrid, 2012.

DeGroot, M. H., *Probability and statistics*, Addison-Wesley, Massachusetts, 1975.

Diaz-Nieto, J., and R. L. Wilby, A comparison of statistical downscaling and climate change factor methods: impacts on low flows in the River Thames, United Kingdom, *Climatic Change*, *67*, 245–268, 2005.

Drogue, G., P. Matgen, and L. Pfister, Dynamically and statistically downscaled outputs of AO-GCMs for mesoscale hydrological impact assessment, *Geophysical Research Abstracts*, *7*, 02,435, 2005.

Duckstein, L., M. Fogel, and C. C. Kisiel, A stochastic model of runoff-producing rainfall for summer type storms, *Water Resources Research*, *8*(2), 410–421, 1972.

Environment Canada, The climate cds. http://www.weatheroffice.ec.gc.ca, 2007.

Foufoula-Georgiou, E., and D. P. Lettenmaier, Continuous-time versus discrete-time point process models for rainfall occurrence series, *Water Resources Research*, *22*(4), 531–42, 1986.

Foufoula-Georgiou, E., and D. P. Lettenmaier, A Markov renewal model for rainfall occurrences, *Water Resources Research*, *23*, 875884, 1987.

Gabriel, K. R., and J. Neumann, A Markov chain model for daily rainfall occurrence at Tel Aviv, *88*, 1962.

Geng, S., F. P. de Vries, and I. Supit, A simple method for generating daily rainfall data, *Agric For Meteorol*, *36*, 363–76, 1986.

Giorgi, F., and G. T. Bates, The climatological skill of a regional model over complex terrain, *Monthly Weather Review*, *117*, 23252347, 1989.

Goodess, C. M., M. Hulme, and T. J. Osborn, The identification and evaluation of suitable scenario development methods for the estimation of future probabilities of extreme weather events, *Tyndall Centre for Climate Change Research, Technical Report 4*, 2003.

Green, J. R., A model for rainfall occurrence, *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*(2), 345–353, 1964.

Hay, L. E., R. L. Wilby, and G. H. Leavesley, A comparison of Delta change and downscaled GCM scenarios for three Mountainous basins in the United States, *Journal of the American Water Resources Association*, *36*(2), 387–397, 2000.

Hayhoe, H. N., and D. W. Stewart, Evaluation of CLIGEN and WXGEN weather data generators under Canadian conditions, *Canadian Water Resources Journal*, *21*, 53–57, 1996.

Hollander, M., and D. A. Wolfe, *Nonparametric statistical methods*, John Wiley and Sons, New York, 1999.

Hughes, J. P., P. Guttorp, and S. P. Charles, A non-homogeneous hidden Markov model for precipitation occurrence, *Applied Statistics*, *48*(1), 15–30, 1999.

Hutchinson, M., Stochastic space-time weather models from ground-based data, *Agricultural and Forest Meteorology*, *73*, 237–264, 1995.

IPCC, *Climate change 1995. Impacts, Adaptations, and Mitigation of Climate Change: Scientific-Technical Analyses, Cambridge Contribution of Working Group I to the Second Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, Cambridge, UK, 1996.

IPCC, *Climate Change 2007: The Scientific Basis*, In J.T. Houghton and Y. Ding and D. J. Griggs and M. Noguer and P. J. Van Der Linden and X. Dai and K.

Maskell and C.A. Johnson (Eds), Contribution of Working Group 1 to the Fourth assessement report of the Integovernmental Panel on Climate Change, Cambridge University Press, Cambridge, UK, 2007.

Johnson, G. L., C. Daly, G. H. Taylor, and C. L. Hanson, Spatial variability and interpolation of stochastic weather simulation model parameters, *Journal of Applied Meteorology*, *39*, 778–795, 2000.

Kleiber, W., R. W. Katz, and B. Rajagopalan, Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes, *Water Resources Research*, *48*(W01523), 2012.

Leander, R., and T. A. Buishand, A daily weather generator based on a two-stage resampling algorithm, *Journal of Hydrology*, *374*, 185–195, 2009.

Little, R. J. A., and D. B. Rubin, *Statistical Analysis with Missing Data*, John Wiley and Sons, New York, 1987.

Mehrotra, R., R. Srikanthan, and A. Sharma, A comparison of three stochastic multi-site precipitation occurrence generation, *Journal of Hydrology*, *331*, 280–292, 2006.

Ng, W. W., U. S. Panu, and W. C. Lennox, Comparative studies in problems of missing extreme daily streamflow records, *Journal of Hydrologic Engineering*, *14*(1), 91–100, 2009.

Palutikof, J. P., C. M. Goodess, S. J. Watkins, and T. Holt, Generating rainfall and temperature scenarios at multiple sites: examples from the mediterranean, *Journal of Climate*, *15*, 3529–3548, 2002.

Panu, U. S., M. Khalil, and A. Elshorbagy, *Streamflow Data Infilling Techniques Based on Concepts of Groups and Neural Networks*, 235-258 pp., Kluwer Academic Publishers, 2000.

Pickering, N. B., J. W. Hansen, J. W. Jones, C. M. Wells, V. K. Chan, and D. C. Godwin, Weatherman: a utility for managing and generating daily weather data, *Agron J*, *86*, 332 –337, 1994.

Prudhomme, C., D. Jakob, and C. Svensson, Uncertainty and climate change impact on the flood regime of small UK catchments, *Journal of Hydrology*, *277*(1-2), 1–23, 2003.

Qian, B., J. Corte-Real, and H. Xu, Multisite stochastic weather models for impact studies, *International Journal of Climatology*, *22*, 1377–1397, 2002.

Racsko, P., L. Szeidl, and M. Semenov, A serial approach to local stochastic weather models, *Ecological Modelling*, *57*, 27–41, 1991.

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

Rasmussen, P. F., and B. Akintug, Drought frequency analysis with a hidden state Markov model, *Proceedings of the 2004 World Water and Environmental Resources Congress: Critical Transitions in Water and Environmental Resources Management*, pp. 2078–2087, 2004.

Rasmussen, P. F., J. D. Salas, L. Fagherazzi, J. Rassam, and B. Bobee, Estimation and validation of contemporaneous PARMA models for streamflow simulation, *Water Resources Research*, *32*(10), 3151–3460, 1996.

Richardson, C. W., and D. A. Wright, WGEN: a model for generating daily weather variables, *U.S. Department of Agriculture, Agricultural Research Service, ARS-8*, p. 83, 1984.

Rolda'n, J., and D. A. Woolhiser, Stochastic daily precipitation models, 1, A comparison of occurrence processes, *Water Resources Research*, *18*, 1451–1459, 1982.

Sanso, B., and L. Guenni, A nonstationary multisite model for rainfall, *Journal of American Statistics Association*, *95*(452), 1089–1100, 2000.

Sanso, B., and L. Guenni, A Bayesian approach to compare observed rainfall data to deterministic simulations, *EnvironMetrics*, *15*(66), 597–612, 2004.

Semenov, M. A., and E. M. Barrow, Use of a stochastic weather generator in the development of climate change scenarios, *Climatic Change*, *35*, 397–414, 1997.

Semenov, M. A., and R. J. Brooks, Spatial interpolation of the LARS-WG stochastic weather generator in great britain, *Climate Research*, *11*, 137–148, 1999.

Semenov, M. A., R. J. Brooks, E. M. Barrow, and C. W. Richardson, Comparison of the WGEN and LARS-WG stochastic weather generators for diverse climate, *Climate research*, *10*, 95–107, 1998.

Serinaldi, F., A multisite daily rainfall generator driven by bivariate copula based mixed distributions, *Journal of Geophysical Research: Atmosheres*, *114*, 10, 2009.

Singh, N., Estimation of parameters of a multivariate normal population from truncated and censored samples, *Journal of the Royal Statistical Society. Series B (Methodological)*, *22*(2), 307–311, 1960.

Specht, D., A general regression neural network, *IEEE Transactions on Neural Networks*, *2*, 568–576, 1991.

Srikanthan, R., and T. McMahon, Stochastic generation of annual, monthly and daily climate data: A review, *Stochastic Hydrology and Earth System Sciences*, *5*(4), 653–670, 2001.

Srikanthan, R., and G. G. S. Pegram, A nest multisite daily rainfall stochastic generation model, *Journal of Hydrology*, *371*, 142–153, 2009.

Stehlik, J., and A. Bardossy, Multivariate stochastic downscaling model for generating daily precipitation series based on atmospheric circulation, *Journal of Hydrology*, *256*, 120–141, 2002.

Tanner, M., and W. Wong, The calculation of posterior distributions by data augmentation., *Journal of the American Statistical Association*, *82*, 528–540, 1987.

Thyer, M., and G. Kuczera, Modelling long-term persistence in hydro-climatic time series using a hidden state Markov model, *Water Resources Research*, *36*(11), 3301–3310, 2000.

Thyer, M., and G. Kuczera, A hidden Markov model for modelling long-term persistence in multi-site rainfall time series 1. Model calibration using a Bayesian approach, *Journal of Hydrology*, *275*, 12–26, 2003a.

Thyer, M., and G. Kuczera, A hidden Markov model for modelling long-term persistence in multi-site rainfall time series 2. real data analysis, *Journal of Hydrology*, *275*, 27–48, 2003b.

Wilby, R. L., and T. M. L. Wigley, Downscaling general circulation model output: a review of methods and limitation, *Progress in Physical Geography*, *21*, 530548, 1997.

Wilby, R. L., T. M. L. Wigley, D. Conway, P. D. Jones, B. C. Hewitson, J. Main, and D. S. Wilks, Statistical downscaling of general circulation model output. A comparison of methods, *Water Resources Research*, *34*(11), 2995–3008, 1998.

Wilby, R. L., O. J. Tomlinson, and C. W. Dawson, Multi-site simulation of precipitation by conditional resampling, *Climate Research*, *23*, 184–194, 2003.

Wilby, R. L., S. Charles, E. Zorita, B. Timbal, P. Whetton, and L. Mearns, Guidelines for use of climate scenarios developed from statistical downscaling methods,

"supporting material" of the Intergovernmental Panel on Climate Change for consideration by the IPCC at the request of its task group on data and scenario support for impacts and climate analysis (TGICA), 2004.

Wilks, D. S., Adapting stochastic weather generation algorithms for climate change studies, *Climatic Change*, *22*, 67–84, 1992.

Wilks, D. S., Multisite generalization of a daily stochastic precipitation generation model, *Journal of Hydrology*, *210*, 178–191, 1998.

Wilks, D. S., High-resolution spatial interpolation of weather generator parameters using local weighted regressions, *Agricultural and Forest Meteorology*, *148*(1), 111–120, 2008.

Wilks, D. S., and R. L. Wilby, The weather generation game: A review of stochastic weather models, *Progress in Physical Geography*, *23*, 329–357, 1999.

Woolhiser, D. A., and J. Rolda'n, Stochastic daily precipitation models, 2, A comparison of distributions of amounts, *Water Resources Research*, *18*, 1461–1468, 1982.

Yang, C., R. E. Chandler, V. S. Islam, and H. S. Wheater, Spatial-temporal rainfall simulation using generalized linear models, *Water Resources Research*, *41*, 11,415, 2005.

Yang, D., B. E. Goodison, S. Ishida, and C. S. Benson, Adjustment of daily precipitation data at 10 climate stations in Alaska: Application of world meteorological organization intercomparison results, *Water Resoureces Research*, *34*(2), 241–256, 1998.

Zheng, X., J. Renwick, and A. Clark, Simulation of multisite precipitation using an extended chain-dependent process, *Water Resources Research*, *46*, 10,504, 2010.