Statistical Downscaling of Climate Data by Nearest Neighbor Resampling

by

Mark Lee

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfilment of the requirements of the degree of

MASTER OF SCIENCE

Department of Civil Engineering

University of Manitoba

Winnipeg

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES

*****

COPYRIGHT PERMISSION

Statistical Downscaling of Climate Data by Nearest Neighbor Resampling

By

Mark Lee

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

Of

Master of Science

Abstract

The potential of climate change to affect hydrological regimes has increased the need for simulation of future trends in hydrological variables. Global Climate Models (GCMs) are commonly used to provide possible future climate scenarios. The coarse resolution of GCMs makes it difficult to use the data directly for hydrological modelling. Post-processing of the GCM data is necessary to provide data of appropriate scale. Application of statistic methods to downscale the data is a common solution to the disparity of scales.

A k-nearest neighbor resampling model is presented to downscale hydrological variables from large-scale atmospheric data. Although the nonparametric nature of the resampling algorithm avoids the extensive parameterization required by other statistical downscaling methods, it was necessary to develop an optimization routine to maximize the performance of the model. The algorithm was able to adequately reproduce historical weather data, and was applied to generate data for hydrological modelling under climate change scenarios for multiple sites in the Nelson River and Winnipeg River drainage basins.

Acknowledgements

I would like to thank all those who have provided assistance during my graduate studies. I would like to especially like to thank my supervisor Dr. Peter Rasmussen, and also my fellow colleagues in the Hydrological Processes Laboratory at the University of Manitoba.

Financial assistance was provided by Natural Sciences and Engineering Research Council, Manitoba Hydro, and Prairie Adaptation Research Collaborative.

NCEP/NCAR Reanalysis 1 and North American Regional Reanalysis data were provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA. Global climate model data were provided by the Canadian Centre for Climate Modelling and Analysis, Victoria, BC. Canadian daily climate data were provided by Environment Canada.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem

Global climate change is becoming an issue of growing concern. Legitimate concern over how a changing climate will affect many aspects of human life have made adaptation strategies a common engineering requirement. Industries based on the use of water resources may be particularly vulnerable as changes in temperature, evaporation, and precipitation cause changes in available streamflow. Local changes in hydro-climatic variables and the effects on water availability are not well understood but are imperative to long-term water resource management.

Currently global climate models (GCMs) are the best available tool for simulating climate change. These complex models generate climate data over long periods of time for different scenarios of atmospheric forcing conditions based on global population growth and resource development. However, the resolution of GCMs makes it difficult to model hydrological regimes using GCM data directly as input. For example, the GCM data provided by the Canadian Centre for Climate Modelling and Analysis (CCCma) from their CGCM3.1/T47 model has a spatial resolution of approximately 3.75° by 3.75° latitude and longitude. Hydrological modelling requires data on a much smaller scale. A solution to the disparity of scales is to statistically downscale GCM data to points where weather stations exist. Weather data downscaled from the GCM data can then be used for effective hydrological modelling of the climate trends present in GCM scenarios.

Many statistical methods have been adapted to downscale GCM data (Wilby and Wigley, 1997). Common statistical downscaling methods include transfer functions, weather typing, and weather generators. The difficulty of using these models is the large number of parameters that are required to adequately capture the relationships between large-scale atmospheric variables and local weather.

An alternative approach is a k-nearest neighbor statistical downscaling approach. Nearest neighbor resampling is a nonparametric approach that has the primary advantage of avoiding the complex parameterization process. Local weather data is produced by strategically resampling from a historical record based on similarity of the daily large-scale atmospheric patterns of the GCM. Reanalysis data from the National Centre for Environmental Prediction (NCEP) provides the historical record of atmospheric data. The k days from the historical record that are most similar to the simulated day are extracted and referred to as nearest neighbors. One of these nearest neighbors is selected by random sampling. Since the resampled day has similar large-scale weather conditions, which are correlated to local weather conditions, this day provides the desired local weather variables for the simulated atmospheric conditions. In this report, a k-nearest neighbor downscaling model will be developed to generate data for hydrological modelling of future climate change scenarios.

1.2 Objectives

The objectives of this report are to:

1. **Review climate change principles**

   Review climate change in the context of the recent changes in climate as well as future projections by GCMs. Changes in climate will be reviewed in a global context, and then in greater depth for the Canadian Prairie region.

2. **Review downscaling techniques**

   A multitude of downscaling techniques are available in literature. A cursory review of downscaling techniques will be presented. The importance of developing a k-nn downscaling model as an addition to the realm of existing downscaling methods will be apparent after this objective is met.

3. **Explore GCM data**

   An important step in the project will be to explore the availability of GCM data within Canada and to evaluate the GCMs' ability to reproduce current climatological patterns and statistics in the prairie region of Canada. Biases in the GCM data, if they are present, will be identified in this portion of the project..

4. **Explore relationships in large-scale and local weather variables**

   The basis of downscaling methods is the link between large-scale and local weather variables. Statistical techniques will be used to identify the relationships between the different scales of data.

5. **Develop and apply a k-nn downscaling model**

   The essence of the report and its contribution to science is to develop, apply, and evaluate a k-nearest neighbor downscaling model to downscale GCM output. A successful development and application of a k-nn downscaling would contribute to the advancement of statistical downscaling and climate change impact assessment technology.

1.3 Context of Work

The work presented in this report is part of a larger ongoing project in the Department of Civil Engineering at the University of Manitoba. The project is entitled "Effect of Climate Change on Water Supply for Manitoba Hydro Systems" and will be completed in mid 2009. The project is using statistical methods to downscale GCM data. The downscaled GCM data is being used for hydrological modelling to quantify the effect of climate change on water supply for Manitoba Hydro's hydropower generating system. The k-nearest neighbor downscaling model developed in this report will be one of the downscaling methods used to supply a hydrological model with weather variables for climate change scenarios.

1.4 Organization of Report

Chapter 2 provides a review of climate change, including a description of GCMs, a review of GCM downscaling methods, and a review of climate change data and trends. Chapter 3 is a technical description of the methodology, including background information on the numerical methods used throughout the project as well as a literature review of the k-nn methodology and applications. The different data sets employed in the project are described in Chapter 4. Results from GCM validation and the exploration of large-scale and local-scale climate relationships are provided in Chapter 5. Chapter 6 provides the description and results of four applications of the k-nn downscaling model developed for this project. Chapter 7 evaluates the model's ability to improve upon raw GCM output and simulate plausible future climate scenarios, and also summarizes the model development with recommendations for developing future k-nn downscaling models. Chapter 8 summarizes the overall conclusions from the work completed.

Chapter 2

Background

2.1 Climate Change

2.1.1 Global Climate Change

Climate is the long term average weather that an area experiences. Climate of a specific location may be defined by the average temperature, precipitation, wind patterns, days of sunshine, frequency of severe events, etc.

The Earth's climate has always been changing. Even in the past 100,000 years the Earth has seen extreme variation including periods of glaciation and periods of warming. The Earth's climate is a highly complex, chaotic, non-linear, dynamic system. Many factors can create changes in the Earth's climate. For example, changes in climate have can be attributed to volcanism, plate tectonics and changes in the sun's strength over time called solar variation. Over the last few centuries, human activities have played a role in the Earth's climate. Human activities affecting climate include land use change, livestock, release of aerosols, and the burning of fossil fuels.

The Earth has experienced surface temperature increases of 0.74 ± 0.18°C during the hundred years ending in 2005 (IPCC, 2007). The Intergovernmental Panel on Climate Change (IPCC), a scientific body representing the work of 2500 scientists on climate change, concluded that most of the observed increase in globally averaged temperatures since the mid-twentieth century is very likely due to the observed increase in anthropogenic greenhouse gas concentrations. Global climate models are able to reproduce the mean global temperature over the last one hundred years by incorporating various climate change drivers such as variations in greenhouse gas concentration, solar energy, ozone, volcanic activity, and concentrations of sulfates (Meehl, 2004). The work by Meehl (2004) demonstrates that the rise in greenhouse gases is the leading contributing factor to climate change over the last one hundred years.

As humans continue to burn fossil fuels at high rates, the changes this will have on the climate of the next one hundred years is an issue of growing concern. As a result, groups of atmospheric scientists have developed computer algorithms to model the Earth's climate. These large, complex models will be discussed in greater depth in Section 2.2. The possible changes in climate over the next one hundred years simulated by these models will be discussed in the next section.

2.1.2 Future Global Climate

Estimates of future climate trends are primarily based upon the simulations of GCMs. GCMs are currently the best available tool for simulating future climate change. GCMs are forecasting that large changes will occur in Earth's climate in the next one hundred years.

Figure 2.1 is a summary of the historical global temperature, as well as the future global temperatures simulated by a variety of GCMs for different emission scenarios. The figure shows mean global temperatures compared to the year 1990. Proxy data were used to extend the data back 1000 years using information from air trapped in glacial ice and tree ring analysis. The grey shadow shows the margin of error in these data may be quite high. From the late 1800's global temperatures are estimated from measurements taken around the world. The figure shows that since the early 20th century, temperature has been gradually increasing. The envelope created by the multitude of GCM and scenario combinations show that temperatures will continue to rise at an increasing rate. By the end of the 21st century, temperatures are forecasted to rise between 1.4°C and 5.6°C. The temperature simulated for the 21st century is significantly warmer than temperatures experienced in the last 1000 years.

Figure 2.1 shows the range of results for each of the scenario families from the IPCC (2000) special report on emission scenarios (SRES). The A1 family has a wide range of warming, with the A1Fl showing the highest warming out of all emission scenarios, the A1B showing a medium degree of warming, and the A1T showing a medium to low degree of warming. The A2 scenario shows a high degree of warming, the B2 shows a medium to low degree of warming, and the B1 scenario shows a low degree of warming.

The rate of warming around the globe is forecasted to occur in a heteroge-

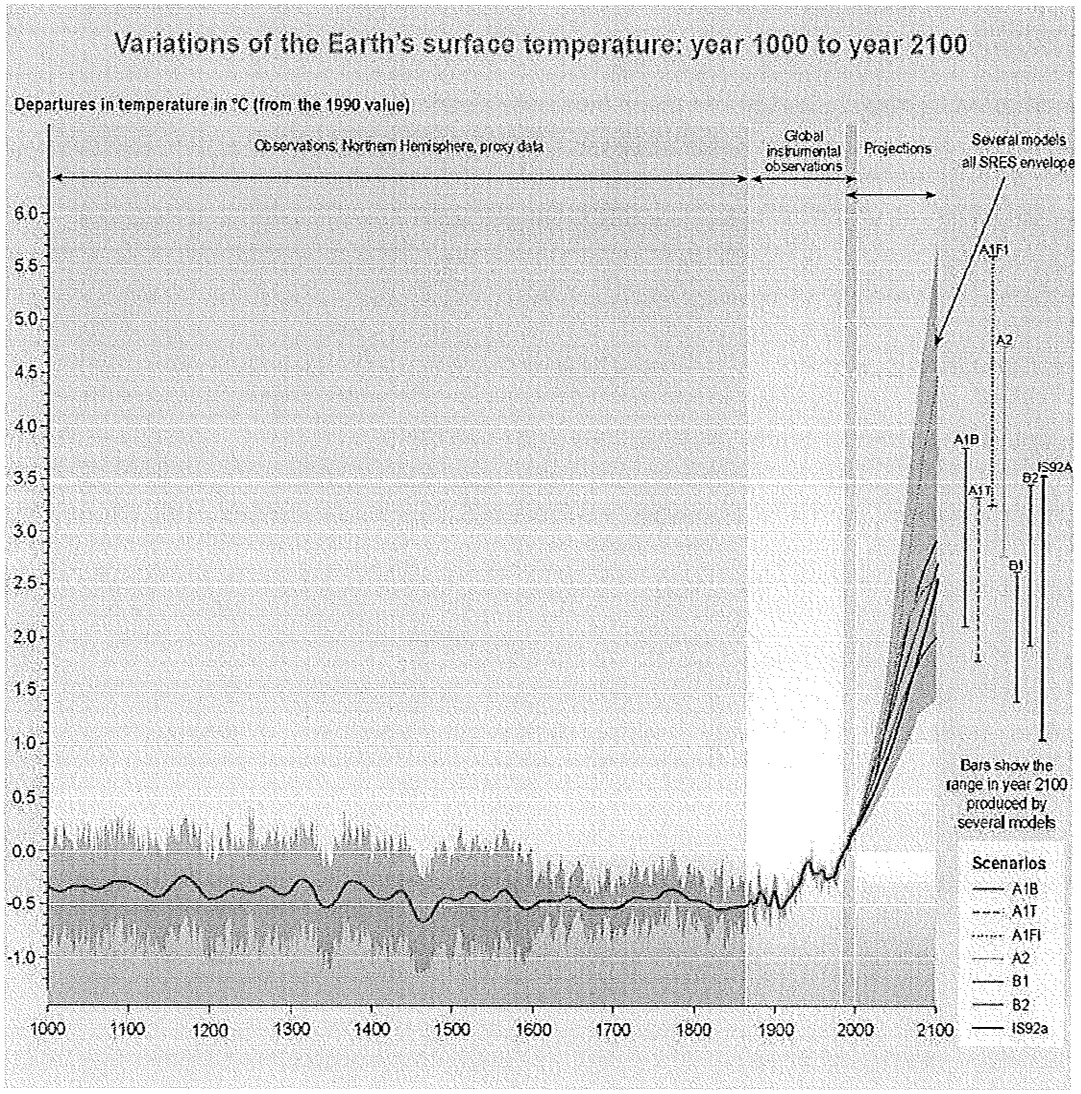


Figure 2.1: Past estimates and future simulation of mean global temperature (Adapted from Climate Change 2001: Synthesis Report, IPCC, 2001, pg. 34).

neous manner (IPCC, 2007). Although the mean global temperature will likely rise as shown in Figure 2.1, some areas may only warm slightly while others will see much higher rates of warming. Generally air above oceans and land near the oceans will have temperatures moderated by the ocean water. Areas in the middle of continents, farthest away from the moderating effects of oceans, will see the highest amounts of warming. Since most of the Earth's land mass is in the Northern hemisphere, the Northern hemisphere, and especially the Arctic, will see an amplification of global warming trends. Higher latitudes will see relatively higher increases than low latitudes.

2.1.3 Climate Change on the Canadian Prairies

Sauchyn and Kulshreshtha (2008) summarized the possible effects of climate change on the Canadian Prairie Provinces. Their report provided a discussion on climate change in the Canadian Prairie Provinces, including the specific effects on water resources, ecosystems, soil landscapes, agriculture, forestry, transportation, communities, health, energy, and tourism and recreation.

For water resources, closed basin lakes are a good indicator of trends in climate. Closed basin lakes, lakes without natural outlets, are indicators of long term water balances in prairie watersheds. Their water levels provide a *memory* of water balance conditions over a number of years, even decades. Water levels are generally constant but will change as they gradually rise or fall over time as changes in climate affect the water balance of the watershed. A summary of water levels of a number of closed basin prairie lakes (Whitewater Lake, Big Quill Lake, Manito Lake, Redberry Lake, Upper Mann Lake, Spring Lake, Little Fish Lake) all show decreasing water levels since the mid 1950's (Sauchyn and Kulshreshtha, 2008). This trend shows that, overall, runoff over the last sixty years has generally decreased on the prairies compared to the short term conditions prior to 1950.

For future changes in the climate of the Canadian Prairie Provinces, Sauchyn and Kulshreshtha (2008) provide a summary of temperature and precipitation projections from a variety of GCMs and emission scenarios for both grassland and forest areas of the prairies. Table 2.1 summarizes the range of trends in mean annual temperature for the Prairie Provinces. Table 2.2 summarizes the range of trends in annual precipitation. These tables show the range of trends and the ensemble mean simulated by a variety of GCMs.

Table 2.1: Simulated °C change in temperature for the Prairie Provinces.

	Grasslands			Forest		
	Minimum	Mean	Maximum	Minimum	Mean	Maximum
2020's	0.5	1.8	2.8	0.9	1.7	2.9
2050's	1.7	3.1	5.6	1.9	2.4	6.8
2080's	2.2	5.0	8.9	2.3	4.4	10.8

Table 2.2: Simulated % change in precipitation for the Prairie Provinces.

	Grasslands			Forest		
	Minimum	Mean	Maximum	Minimum	Mean	Maximum
2020's	-10	2	15	-3	5	6
2050's	-4	5	18	2	9	16
2080's	-6	9	29	2	12	25

Mean annual temperatures are shown to continually increase throughout the next century. By the year 2100, temperatures are likely to have increased 5.0°C in grasslands and 4.4°C in forested areas. The warmest models simulate that temperature could rise as much as 8.9°C in grasslands and 10.8°C in forested areas.

For the climate of the next one hundred years, the GCMs simulations suggest a wide range of possible precipitation trends. Some models show slight decreases, while others show increases of up to 29%.

The Canadian Center for Climate Modelling and Analysis (CCCma) CGCM2 model was included in the summary information. Among the GCMs compared, the CCCma CGCM2 model, the precursor to CCCma's most recent model, the CGCM3, was near the average for temperature changes, and one of the drier models in terms of precipitation.

While higher CO_2 levels and warmer temperatures may provide some benefits to the prairie region, the large warming trends that are projected for the prairies bring many disadvantages. For example, the cold winters in the prairies help limit pest and diseases, and also facilitate access to northern communities and resources via winter ice roads. Warmer winters may allow new pest and diseases to move into the prairies, similar to how the mountain pine beetle has seen increasing populations in British Columbia and severely threatens forest ecosystems. A possible benefit to higher CO_2 levels and higher temperatures could be increased forest, grassland, and crop productivity. However, these benefits may be limited by the availability of water.

Sauchyn and Kulshreshtha (2008) cited that the most serious climate risk to the prairies is increases in water scarcity. Their report stated that recent trends and future projections for water resources include lower summer streamflows, falling lake levels, retreating glaciers, and increasing soil moisture deficits. The frequency of dry years is also likely to increase. Although the number of dry years are expected to increase, risk due to flooding may also increase due to more severe rainfall events.

In 2004, the Canadian Government published a summary report regarding projections of climate change and impacts of climate change in Canada (Lemmen and Warren, 2004). This report cites that due to its inland location, the prairie region could see a greater increase in temperature than the rest of the country. Although precipitation on the prairies is forecasted to slightly increase, there may be significant negative changes in the annual hydrologic cycle due to these changes in climate. The rise in temperature reduces the frost season significantly and increases the rate of evapotranspiration.

The 2004 government report (Lemmen and Warren, 2004) cites some potential changes in water resources in the Prairie Provinces as:

- Changes in annual flow regime, reduced summer flows,
- Increasing likelihood of severe drought, and
- Increases or decreases of irrigation demand and water availability.

The threat climate change poses to water resources in the Canadian Prairies and elsewhere around the world makes climate change impact studies an important component in long-term water resource management plans.

2.2 Global Climate Models

2.2.1 Introduction

A global climate model (GCM), also referred to as a general circulation model, is a large computer model used for simulating long periods of weather over the entire globe. GCMs are very complex and constructing a GCM is a massive undertaking. Therefore they are usually developed by government organizations or universities. A list of current GCMs and the organizations who developed them is given in Table 2.3.

GCMs are combinations of other large computer models. The two major building blocks of a GCM are an atmospheric general circulation model (AGCM) and an oceanic general circulation model (OGCM). The combination of an AGCM and an OGCM is referred to as a coupled global climate model (CGCM). Additional models are added to complete the description of the Earth's climate driving forces. These supplementary models include ice models, river routing models, evapotranspiration models, and chemical transport models. Each GCM employs similar but slightly different types and combinations of these components, and as a result GCMs provide different results when forced with the same atmospheric conditions. The variation in the models provide a spectrum of possible projections of the Earth's climate.

GCMs typically complete their computations using either a finite difference method or a spectral method. The model is discretized into a three dimensional grid. Since the models are complex, to be computationally feasible the model

Table 2.3: List of current GCMs.

Model Name	Source
CM1	Beijing Climate Centre (BCC), China
BCM2.0	Bjerknes Centre for Climate Research (BCCR)
CGCM3	Canadian Centre for Climate Modelling and Analysis (CCCma), Canada
CM3	Centre National de Recharshes Meteorogiques (CNRM)
Mk3.0	Australia's Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia
ECHAM5-OM	Max-Plank-Institute for Meteorology (MPI) and Deutsches Klimarechenzentrum (DKRZ), Germany
ECHO-G	Meteorological Institute, University of Bonn (MIUB) Meteorological Research Institute of KMA (METRI) Model and Data Groupe at MPI-M (M& D)
FGOALS-g1.0 CM2.0 CM2.1	Institute of Atmospheric Physics (LASG) Geophysical Fluid Dynamics Laboratory GFDL, USA
AOM E-H E-R	Goddard Institute for Space Studies (GISS) Atmosphere Ocean Model
CM3.0	Institute for Numerical Mathematics (INM), Germany
CM4	Institut Pierre Simon Laplace (IPSL), France
MICROC3.2	Center for Climate Research Studies & National Institute for Environmental Studies, Japan
MRI	Meteorological Research Institute (MRI), Japan
PCM CCSM3	National Centre for Atmospheric Research, USA
HadCM3 HadGEM1	Hadley Centre for climate Prediction and Research HCCPR, Uk Meteorological Office (UKMO)
SXG 2005	National Institute of Geophysics and Volcanology (INGV) Italy

grids are coarse. Typical grid resolutions are on the order of between one and five degrees in latitude and longitude. The Hadley HadCM3 model uses a grid of 2.5° in latitude and 3.75° in longitude, giving a global grid of 73 by 96 points. The T47 version of the CCCma CGCM3 has a grid resolution of approximately 3.75° in latitude and longitude, giving the global grid of 48 by 96 points shown on Figure 2.2. Such resolutions result in grid cells with side lengths in the order of

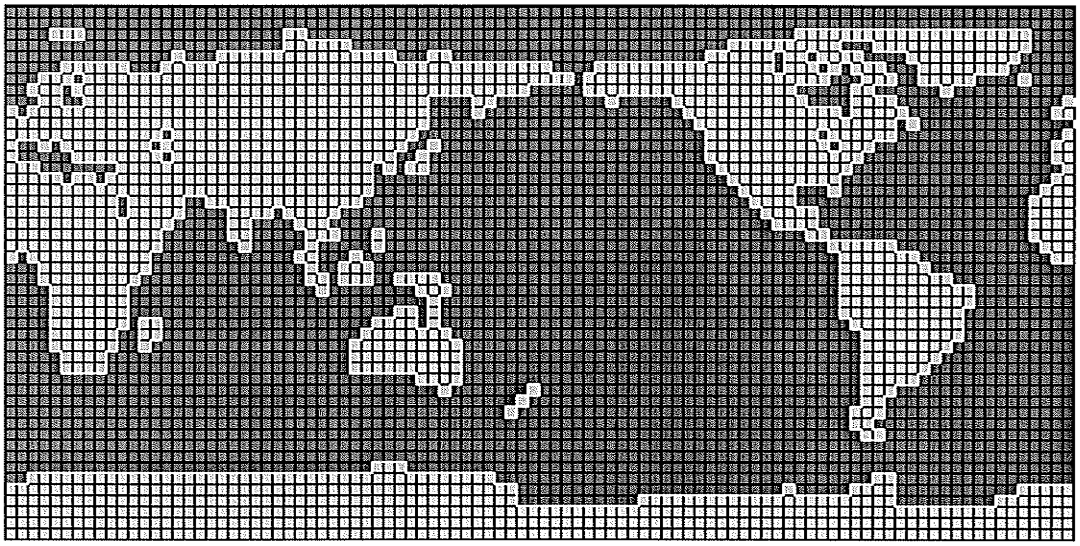


Figure 2.2: Discretization grid of the CCCma CGCM3/T47 model.

300 to 400 kilometers. Discretization in the vertical is also necessary. The Hadley HadCM3 model uses 19 levels in the vertical while the CCCma CGCM3 uses 31 levels in the vertical.

The chemical transport models are important to the modelling of climate change. The changes in future climate can be modeled by adding one or more chemical transport models for the atmospheric chemicals important to climate. For example, a chemical transport model is developed to describe the carbon cycle. The carbon cycle is then modified by adding greenhouse gas emissions according to a plausible future anthropogenic emission scenario. The changes in the Earth's climate, such as changes in temperature or precipitation patterns, in response to change in the carbon cycle can then be studied. For consistency between modelling agencies, the IPCC has defined sets of emission scenarios for future anthropogenic releases of greenhouse gases. The IPCC emission scenarios will be discussed in detail in the following section.

2.2.2 Emission Scenarios

The IPCC has been the leading organization for developing future emission scenarios to be used in climate change studies. In 1992, the IPCC released the first set of emission scenarios named the IS92 scenarios. These were the first future emission scenarios used by GCMs and the first global scenarios to provide estimates for the full suite of greenhouse gases.

In the years following the development of the IS92 scenarios many scientific advancements were made in the field. In 1996, the IPCC decided to develop a new set of emission scenarios to be used in the IPCC Third Assessment Report. As a result, in the year 2000 the IPCC released a report on updated possible future emission scenarios entitled Special Report on Emissions Scenarios (SRES) (IPCC, 2000). In this report, the IPCC developed *families* of emission scenarios that explore alternative global development pathways, covering a wide range of demographic, economic and technological driving forces and resulting greenhouse gas emissions. The scenarios are grouped into four scenario families: A1, A2, B1, and B2.

The following is the description of the different families of emission scenarios according to the IPCC Special Report on Emission Scenarios (IPCC, 2000):

- **SRES A1 Scenario Family**

  The A1 storyline and scenario family describes a future world of very rapid economic growth, global population that peaks in mid-century and declines thereafter, and the rapid introduction of new and more efficient technologies.

Major underlying themes are convergence among regions, capacity building, and increased cultural and social interactions, with a substantial reduction in regional differences in per capita income. The A1 scenario family is divided into three groups that describe alternative directions of technological change in the energy system. The three A1 groups are distinguished by their technological emphasis: fossil intensive (A1FI), non-fossil energy sources (A1T), or a balance across all sources (A1B).

- **SRES A2 Scenario Family**

  The A2 storyline and scenario family describes a very heterogeneous world. The underlying theme is self-reliance and preservation of local identities. Fertility patterns across regions converge very slowly, which results in continuously increasing global population. Economic development is primarily regionally oriented and per capita economic growth and technological change are more fragmented and slower than in other storylines.

- **SRES B1 Scenario Family**

  The B1 storyline and scenario family describes a convergent world with the same global population that peaks in mid-century and declines thereafter, as in the A1 storyline, but with rapid changes in economic structures toward a service and information economy, with reductions in material intensity, and the introduction of clean and resource-efficient technologies. The emphasis is on global solutions to economic, social, and environmental sustainability, including improved equity, but without additional climate initiatives.

- **SRES B2 Scenario Family**

  The B2 storyline and scenario family describes a world in which the emphasis is on local solutions to economic, social, and environmental sustainability. It is a world with continuously increasing global population at a rate lower than A2, intermediate levels of economic development, and less rapid and more diverse technological change than in the B1 and A1 storylines. While the scenario is also oriented toward environmental protection and social equity, it focuses on local and regional levels.

When these emission scenarios are used with different GCMs, they produce a spectrum of possible future climates. Although the futures simulated by different GCMs will vary for the same scenario, general conclusions can be made. In general, the A2 scenario family shows the highest degree of global warming, the A1B and B1 scenario families show a medium level of warming, and the B2 family shows the lowest level of warming.

2.3 Downscaling of Global Climate Models

Global climate models are currently the best tools available for climate change impact assessments. However, one of the primary difficulties with utilizing GCM data is that the coarse resolution of the data grids makes it difficult to directly apply in a meaningful way, particularly for water resources applications. Even large watersheds may only have a few GCM grid cells covering the watershed. Because of the disparity in scale between hydrologic processes and GCM data, these few cells poorly represent the variability in hydrological weather variables, especially the distributed nature of precipitation. In addition, GCM cells are too large to simulate some important features of the local weather and hydrological cycle such as cloud cover. GCMs are designed to reproduce the fluid dynamics of the hydrological cycle at the continental scale and therefore precipitation at individual weather station locations are generally not well reproduced.

For the above reasons, GCM output requires postprocessing before it is acceptable to use in climate change impact assessments. The GCM output must be downscaled to a finer spatial, and possibly temporal, resolution. The goal of downscaling GCM output is to produce new output that is on the scale of subcatchment hydrology.

Downscaling models can be divided into two large groups, dynamic downscaling and statistical downscaling. In dynamic downscaling a regional climate model uses GCM output as boundary conditions and runs a higher resolution (10 to 50 km) climate model over the area of interest. In statistical downscaling models, a variety of statistical methodologies are used to parameterize the relationships between large and small scale climate variables. The chart on Figure 2.3 shows the general classes of downscaling models.

2.3.1 Dynamic Downscaling

Running GCMs at a high enough resolution to be acceptable for assessing local climate change is not computationally possible. Dynamic downscaling involves nesting a regional climate model (RCM), also known as a limited area model

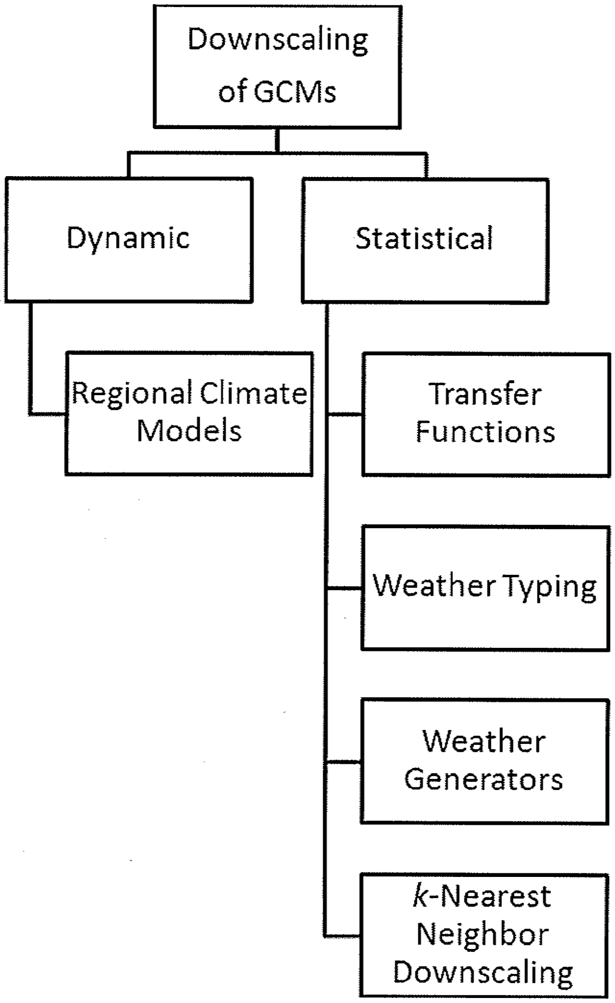


Figure 2.3: Organization chart of downscaling methodologies.

(LAM), within a GCM. The embedded model uses the GCM data as boundary conditions to produce weather on a much finer resolution, typically on grids with 10 to 50 km horizontal resolution and 100 to 1000 m vertical resolution. Running the model over only a portion of area of the GCM's global grid is computationally much more affordable than running a GCM at a higher resolution.

There are some limitations of downscaling GCM data with a RCM (Wilby and Wigley, 1997). RCMs still require considerable computing resources, much more than current personal computers can provide. RCMs are somewhat inflexible in the sense that the computational demands apply each time that the model is transferred to a different region and for each emission scenario. Another potential drawback of RCMs is the fact that they are completely dependent upon the veracity of the GCM grid-point data that are used to drive the boundary conditions of the region. If biases or errors are present in the GCM for the area the RCM is nested in, those inaccuracies are transferred into the downscaled data through the RCM.

Many RCMs have been developed around the world. A RCM has been developed in Canada by the Canadian Regional Climate Modelling and Diagnostics (CRCMD) Network. The CRCMD Network is made up of researchers from various institutions; Université du Québec à Montréal (UQAM - Centre ESCER), University of Victoria (BC), Ouranos Consortium (QC), Environment Canada (Canadian Centre for Climate Modelling and Analysis) and Recherche Prévision Numérique (RPN) (QC). The Canadian Regional Climate Model (CRCM) was first developed in the early 1990's (Caya and Laprise, 1999) and is now in its fourth generation, CRCM4.2.

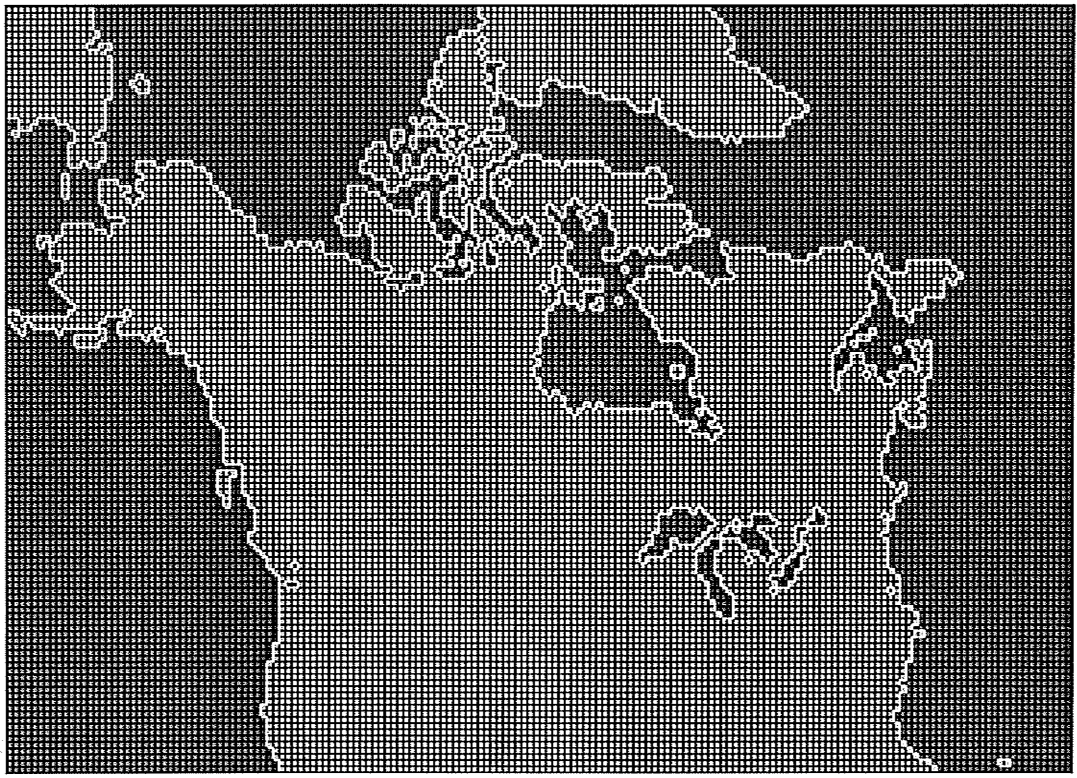


Figure 2.4: CRCM discretization of Pan-Canadian region.

The 172 x 124 grid discretization over the Pan-Canadian region used by the CRCM is shown on Figure 2.4. The CRCM has a 45 km horizontal grid-size mesh and 18 vertical levels. The time step between the CRCM calculations is fifteen minutes, which is slightly more frequent than the twenty minute time step used by the CCCma CGCM3.

2.3.2 Statistical Downscaling

Statistical downscaling methods use the statistical relationships that exist between large-scale and local climate variables to downscale GCMs. While not physically based, statistical downscaling methods have the primary advantage that they require significantly less computational resources than dynamic downscaling. Statistical downscaling can be carried out in a short period of time on inexpensive personal computers. The models can be run for different areas for different emission scenarios in short periods of time.

The following is a brief description of some of the most common categories of statistical downscaling models.

Transfer Functions

Downscaling by transfer functions usually refers to the application of linear or nonlinear regression methods. Regression methods were among the earliest downscaling approaches (Wibly and Wigley, 1997). Regression methods generally involve relating the weather variables to the coarse resolution GCM predictor variables.

Other transfer functions such as artificial neural networks or canonical correlation analysis may also be used to derive the empirical relationships to downscale the large-scale climate variables.

A software application based on regression methods was developed by Wilby et al. (2002) as a methodology that could be applied in a wide variety of downscaling applications. According to the developers, the software package, named Statistical DownScaling Model (SDSM), facilitates the rapid development of multiple, low-cost, single-site scenarios of daily surface weather variables under current and future regional climate forcing. The software performs ancillary tasks of predictor variable pre-screening, model calibration, basic diagnostic testing, statistical analysis and graphing of climate data. The general process SDSM uses to produce downscaled simulations is as follows (Wilby et al., 2002):

1. Screening of predictor variables;
2. Model calibration;
3. Synthesis of observed data;
4. Generation of climate change scenarios;
5. Diagnostic testing and statistical analysis.

Although SDSM performs the downscaling and some other ancillary tasks as described above, the application of SDSM is still a time consuming process as NCEP reanalysis and GCM data must be downloaded and manipulated to conform to SDSM's protocols. The data gathering and preprocessing tasks are generally the most time consuming and difficult steps for other statistically downscaling methodologies as well.

The utility of the SDSM software package to assess the hydrological impacts of climate change on the Canadian Prairies is being investigated by others at the University of Manitoba concurrently with the development of the k-nn statistical downscaling model in this project.

Weather Typing

Downscaling by weather typing methodologies involves statistically relating common weather patterns to observed station variables. If a strong relationship exists between the circulation patterns and the local weather, the local weather can be simulated conditional on the circulation pattern classes.

Days are divided into groups based on their large-scale circulation variables using a classification scheme. Yarnal (1993) describes a variety of classification schemes, including manual classification, correlation-based map pattern classification, eigenvector-based classification and compositing, indexing and specification.

Once a classification scheme has been utilized to divide the historical days into groups with similar circulation patterns, the distributions of local surface variables such as temperature and precipitation conditional on the occurrence of each weather pattern are calculated. For example, the conditional probability of a wet day following a wet day, or the mean wet-day amount associated with a given atmospheric circulation pattern can be derived. The conditional probabilities may also be calculated on a season or monthly basis to improve correlation between the weather patterns and local variables.

To downscale GCM data, the same large-scale circulation variables used in the classification scheme are used to classify each day of the GCM output into the weather types determined from the classification of historical data. The local weather is then simulated based on the relationships that exist between the historical weather patterns and historical local weather. The change in frequency of the different weather patterns in the GCM data will reveal trends in the future local-scale climate.

Weather classification can be used to form Markov models where the probability of precipitation occurrence is conditional on the occurrence of precipitation on the previous day as well as the previous and/or current circulation pattern class. The most sophisticated of the weather typing schemes is the nonhomogeneous hidden

Markov model (NHMM) as defined by Hughes and Guttorp (1994) and Hughes et al. (1999). The basis of the model is the existence of an unobservable discrete-value stochastic process which links the large-scale atmosphere to the local-scale precipitation. This unobserved process is referred to as a hidden weather state. The hidden weather state is assumed to follow a Markov chain conditioned on the current day large-scale climate to capture the persistence of wet and dry conditions.

Weather Generators

Weather generators are computer models that generate long synthetic time series of weather data based on parameters derived from historical data.

Weather generators first simulate whether or not rainfall occurs. This is commonly based on a Markov renewal process conditioned on the occurrence of precipitation on the previous day. The Markov process can be first-order and be based only on the single previous day, or be multiple-order and be based on the occurrence of precipitation on multiple previous days. Other weather parameters are then generated conditional on the occurrence of precipitation.

To downscale GCM data, the parameters of the weather generator are adjusted using data from the GCM. The weather generator is then run with the new parameters to generate time series of climate change data. One of the difficulties in applying stochastic weather generators to future climate scenarios has been the method of adjusting the parameters in a physically realistic and internally consistent way (Wilby and Wigley, 1997).

A popular weather generator model is WGEN (Richardson, 1981). The WGEN model is able to produce a daily time series of precipitation amount, maximum and minimum temperature, and solar radiation. WGEN is the most common weather generator used for climate impact studies. While WGEN is the most popular weather generator, a multitude of other models exist: WXGEN (Sharpley and Williams, 1990), CLIGEN (Nicks and Gander, 1993), USCLIMATE (Johnson et al., 1996), ClimGen (Semenov et al., 1999) and LARS-WG (Semenov et al., 1998). The LARS-WG and ClimGen models have been previously applied in Canada.

The utility of the LARS-WG software package to assess the hydrological impacts of climate change on the Canadian Prairies is being investigated by others at the University of Manitoba concurrently with the development of the k-nn statistical downscaling model in this project.

k-Nearest Neighbor Resampling

Nearest neighbor resampling is a nonparametric statistical downscaling method that has the primary advantage of avoiding the complex parameterization process of other statistical downscaling models. Local weather data is produced by strategically resampling from a historical record based on similarity of the daily large-scale atmospheric patterns of the GCM. A data set such as the NCEP/NCAR Reanalysis 1 data set provides the historical record of large-scale atmospheric data while data from weather station measurements typically provide the historical local weather. The nearest neighbors, or most statistically similar days, to the simulation day in the historical record are determined. One of the nearest neighbors is selected by random sampling. Since the resampled day has similar large-scale

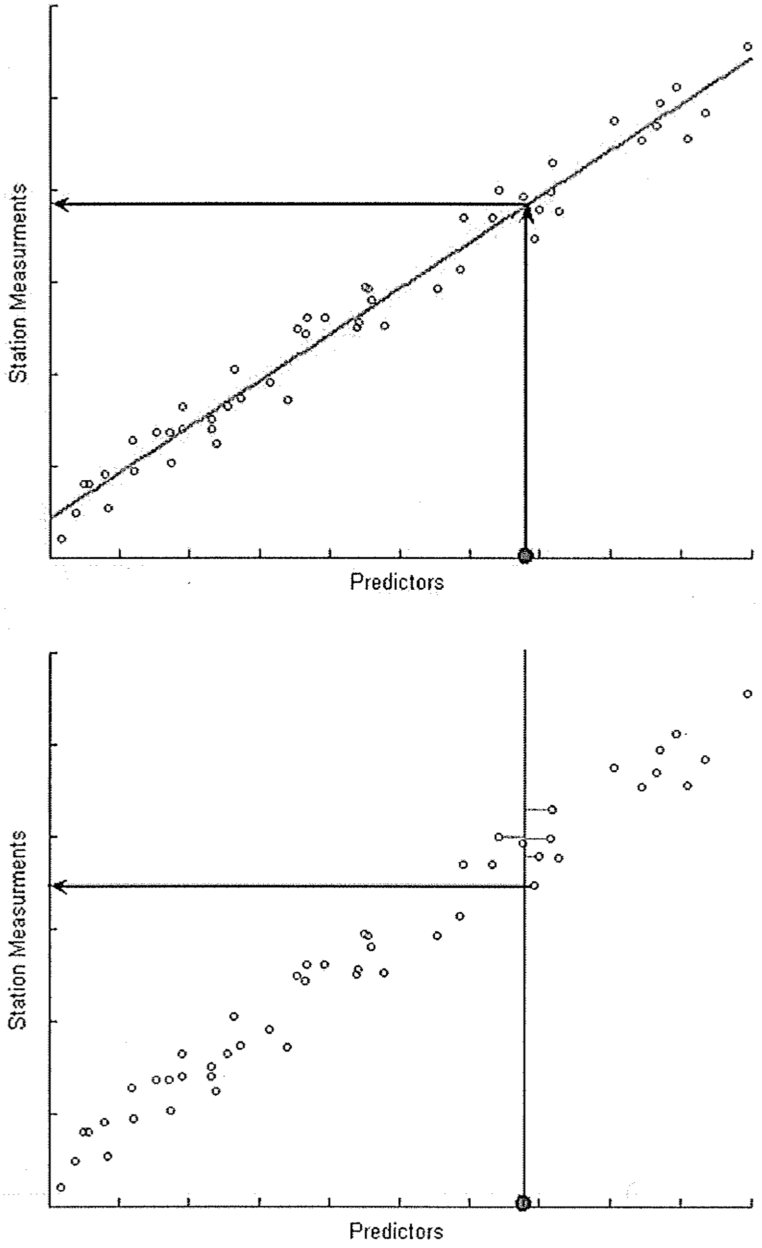


Figure 2.5: Schematics of regression (top) and k-nn (bottom) downscaling approaches.

weather conditions, which are correlated to local weather conditions, this day provides the desired local weather variables for the simulated atmospheric conditions. The process is repeated for each simulated GCM day to generate a time series of local weather.

Figure 2.5 shows a basic comparison of a regression based methodology to the methodology of nearest neighbor resampling. The top figure (Figure 2.5) shows the basic concept of parameterizing the relationships between predictors and local weather and then using the relationship to estimate the local weather. The bottom figure (Figure 2.5) shows the basic concept of resampling from the nearest neighbors of the set of predictor variables. Nearest neighbor resampling is explained in full detail in the methodology section.

2.3.3 Statistical vs. Dynamic Downscaling

In the above description of statistical and dynamic downscaling methods, some of the advantages and disadvantages of the different methodologies were noted. Both statistical and dynamic downscaling have unique advantages and disadvantages. Which methodology is preferred should be determined for each climate change assessment. Wilby et al. (2002) provided a summary of the general strengths and weaknesses of statistical and dynamic downscaling methods as shown in Table 2.4. While not an exhaustive list, the information in Table 2.4 provides basic information to assist in the choice of model.

Table 2.4: Comparison of statistical and dynamic downscaling (Wilby et al., 2002).

	Statistical downscaling	**Dynamical downscaling**
Strengths	- Station-scale climate information from GCM-scale output	- 10 to 50 km resolution climate information from GCM-scale output
	- Cheap, computationally undemanding and readily available	- Respond in physically consistent ways to different external transferable forcings
	- Ensembles of climate scenarios permit risk/uncertainty	- Resolve atmospheric processes such as orographic precipitation
	- Flexibility	- Consistency with GCM
Weaknesses	-Dependent on the realism of GCM boundary forcing	- Dependent on the realism of GCM boundary forcing
	- Choice of domain size and location affects results	- Choice of domain size and location affects results
	- Requires high quality data for model calibration	- Requires significant computing resources
	- Predictor/predictand relationships are often non-stationary	- Ensembles of climate scenarios seldom produced
	- Choice of predictor variables affects results	- Initial boundary conditions affect results
	- Choice of empirical transfer scheme affects results	- Choice of cloud/convection scheme affects (precipitation) results
	- Low-frequency climate variability problematic	- Not readily transferred to new regions

Chapter 3

Methodology

3.1 Numerical Methods

3.1.1 Principal Component Analysis

Often in multivariate data sets there can be a high degree of correlation among variables. This leads to a redundancy in the information contained in the variables. If it were possible to remove the redundancy among variables, the information contained in the data could possibly be represented in only a few variables. Principal component analysis (PCA) is a method to reduce the number of variables required to explain the variation within a multivariate data set. For an in-depth derivation of PCA, Wilks (1995) is an excellent resource.

Climate data often contains significant spatial correlations making principal components a useful tool for data analysis. PCA will be applied later in Section 3.1.1 to geopotential height data. Before the PCA is shown for climate data, the simplified case of a bivariate data set will be discussed.

Bivariate Example of Principal Component Analysis

Consider a bivariate data set with a high degree of correlation (Figure 3.1). The data has a 2 x 2 covariance matrix, Σ, which has eigenvalues of λ_1 and λ_2. A shift in the coordinate system can be made where the first axis, Z_1, is aligned in the direction of maximum variation. This axis is in the direction of the eigenvector corresponding to λ_1. The second axis is aligned in the direction of second greatest variation. The second axis, Z_2, is in the direction of the second eigenvector, which is perpendicular to the first eigenvector. The Z_1 axis contains $\lambda_1/(\lambda_1+\lambda_2)\times 100\%$ of the variance contained in the original data set (approximately 95% in the example in Figure 3.1), and the Z_2 axis contains $\lambda_2/(\lambda_1 + \lambda_2) \times 100\%$ of the variation (approximately 5% in the example in Figure 3.1).

The values of the first principal component (PC) scores are the values of the data points along axis Z_1, and the second PC scores are the values of the data points on axis Z_2. Figure 3.2 shows the two sets of PC scores plotted against each other. It can be seen that PC_1 has much more variation than PC_2, and that the two variables have no correlation. Therefore, much of the variation in the original scatter plot in Figure 3.1 can be described using only one variable, PC_1.

Principal Component Analysis with Multivariate Data

The above process can be applied to data sets with any number of variables. Consider a multi-variable data set X with n variables. The $n \times n$ covariance matrix Σ will have n eigenvalues and n eigenvectors. A shift in coordinates using PCA will produce a new data set with n principal components. The sum of all

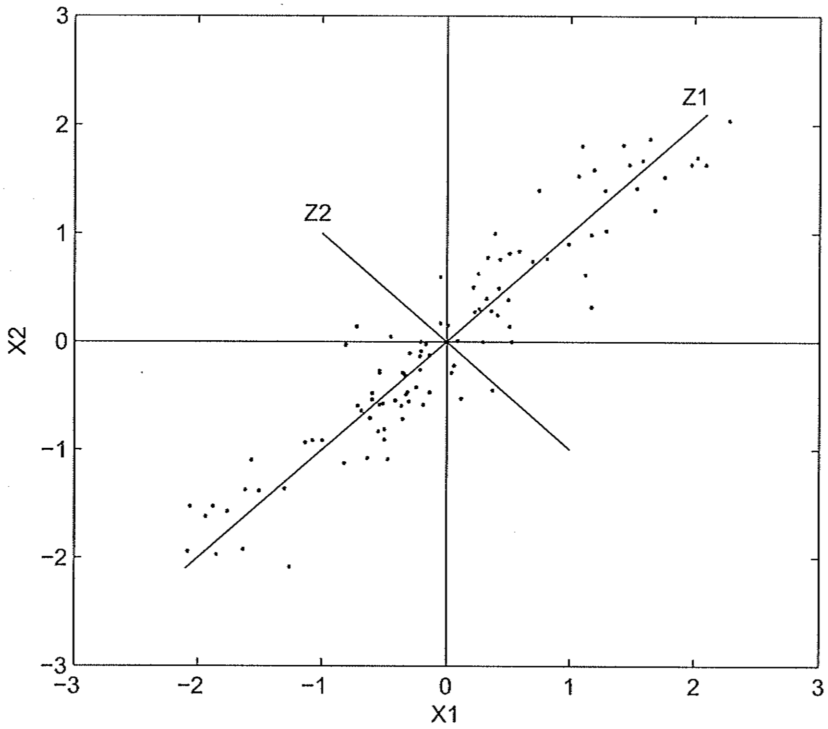


Figure 3.1: Scatter plot of correlated bivariate data.

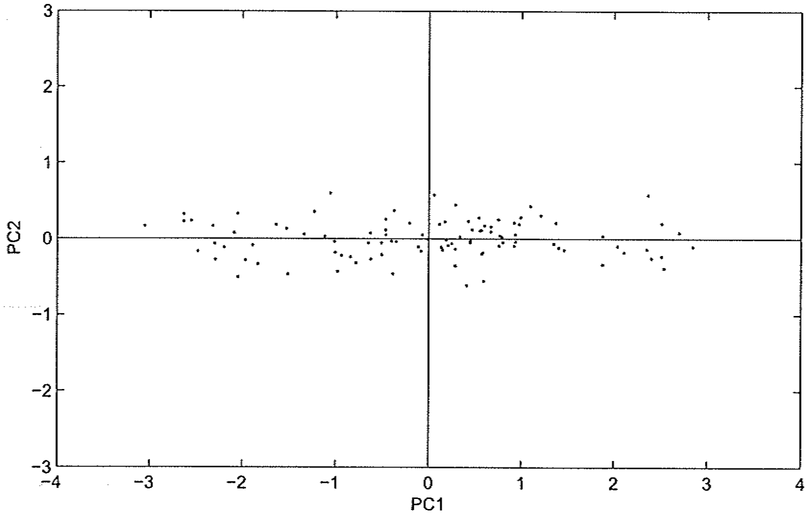


Figure 3.2: Plot of PC_2 versus PC_1.

eigenvalues equals the sum of the variance of each variable. Each PC describes a unique portion of variance in the original data set, which can be computed as

$$\text{Percent variation explained by } \mathrm{PC}_i = \frac{\lambda_i}{\Sigma_{j=1}^{n}\lambda_j}. \tag{3.1}$$

Each PC is a linear combination of the original n variables. The coefficients for the i^{th} PC are given by the i^{th} eigenvector. A set of PCs are generated for each time step,

$$PC_i = e_i X = e_{i1}X_1 + e_{i2}X_2 + \ldots + e_{in}X_n. \tag{3.2}$$

Application of Principal Component Analysis

To illustrate the usefulness of PCA when analyzing spatial climate data, a sample application of PCA will be demonstrated with a 500 mb geopotential height data set. The data is comprised of thirty years of data from the NCEP/NCAR Reanalysis 1 data set. The data is on a 9 × 15 grid with spacing of 2.5° in latitude and longitude. Geopotential heights vary gradually across the large distances between grid points and therefore nearby grid points experience high correlations with each other. For this reason, geopotential height is an excellent example of data sets that can be easily reduced to a few variables using PCA.

The values of the first few eigenvalues and the percentage of the total variance (sum of the variances of individual data points) explained by a given number of PCs are shown on Figure 3.3. As expected with the high degree of correlation in the data set, the first few eigenvalues are large and explain most of the variance in the large data set. Only five PCs are needed to explain over 90% of the variance

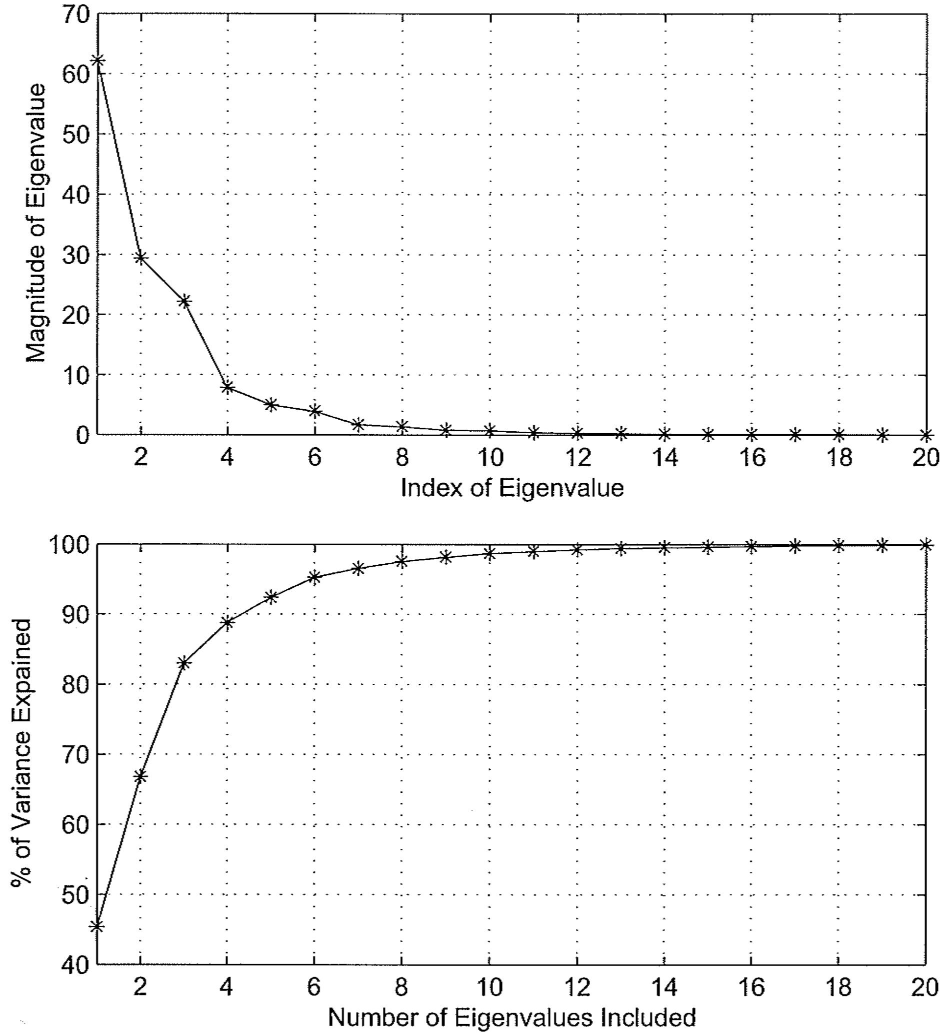


Figure 3.3: Eigenvalues and the % variation explained.

contained in the original data set of 135 variables.

The weights associated with the eigenvectors of each PC can be plotted in space. Figure 3.4 shows the weights of the first eight eigenvectors plotted in space. The most weight of the first PC is in the center of the grid. The first PC describes if the geopotential grid is above or below average for that day of the year. The second PC has weights distributed in a west to east direction. If there is a strong pressure gradient in this direction, the second PC will have a large value. The third PC is similar to the second but describes the north to south gradient. Each PC there after describes a characteristic pressure pattern until after a handful of PCs the eigenvectors dissolve and contain only noise. The stronger a particular pattern is, the larger the corresponding PC will be.

The ability of PCA to reduce large gridded climate data sets to only a handful of variables made it a very useful numerical method for this project. PCA was used extensively to reduce the variables in the large-scale NCEP/NCAR Reanalysis 1 and GCM data sets.

3.1.2 Canonical Correlation Analysis

Canonical correlation analysis (CCA) is a statistical method used to explore the connections between two multivariate data sets. As principal component analysis explores a single multivariate data set by projecting the data onto a new set of variables that describe maximum amounts of variation, CCA projects two multivariate data sets onto two new projections with maximum correlation between them.

Another description of CCA may be to envision it as multivariate regression with two sets of predictor variables. Instead of having one set of weights as regression coefficients to simulate a single variable from a multivariate set, CCA finds pairs of sets of regression coefficients that define new variables with maximized correlation.

CCA transforms two multivariate data sets, x and y, into new variables v_m and

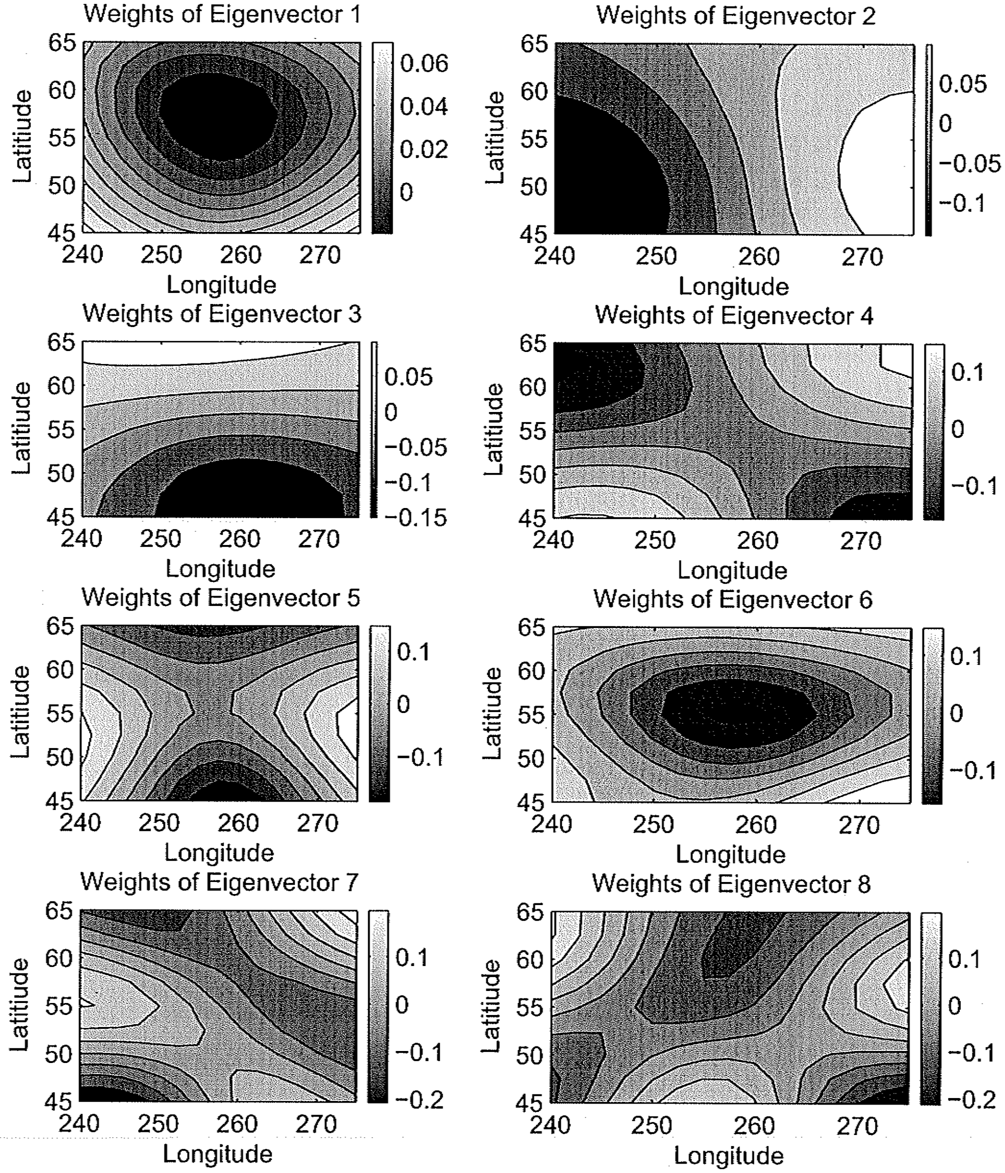


Figure 3.4: Weights of eigenvectors plotted in space.

w_m called canonical variates. The new variables are defined by

$$v_m = a_m^T x = \Sigma_{i=1}^{I} a_{m,i} x_i, \quad m = 1, \ldots, \min(I, J) \tag{3.3}$$

and

$$w_m = b_m^T y = \Sigma_{j=1}^{J} b_{m,j} y_j, \quad m = 1, \ldots, \min(I, J), \tag{3.4}$$

where I is the number of elements in x and a_m, and J is the number of elements in y and b_m. I and J do not need to be the same. However, the number of canonical pairs, M, that can be produced is equal to the smaller of the two.

The selection of the canonical vectors a_m and b_m are done so that the following are satisfied:

$$\mathrm{Corr}[v_1, w_1] \geq \mathrm{Corr}[v_2, w_2] \geq \ldots \geq \mathrm{Corr}[v_M, w_M], \tag{3.5}$$

$$\mathrm{Corr}[v_k, w_m] = \begin{cases} r_{C_m}, & k = m \\ 0, & k \neq m \end{cases}, \tag{3.6}$$

and

$$\mathrm{Var}[v_m] = \mathrm{Var}[w_m] = 1, \quad m = 1, \ldots, M. \tag{3.7}$$

Equation 3.5 states that each of the M successive pairs of canonical variates has a weaker correlation than the previous pair. The correlations between pairs of canonical variates are referred to as canonical correlations. Equation 3.6 states that a canonical variate has no correlation with all other variates, except for its counterpart in the m^{th} pair. Equation 3.7 states that each variate has unit variance.

The goal in CCA, as in PCA, is to select the weights such that the new projection produces variables that contain useful information. The selection of canonical coefficients is based on the variance-covariance matrix of x and y. The joint variance-covariance matrix, S_c, of the variables combined into one variable set, $c^T = [x^T, y^T]$, is

$$S_c = \frac{1}{1-n}[c']^T[c'] = \begin{bmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{bmatrix}, \tag{3.8}$$

where the prime in c' denotes that the variables are centered on the sample means.

The canonical coefficients and canonical correlations are related to the eigenvectors and eigenvalues of the matrices, M_x and M_y, where

$$[M_x] = [S_{xx}]^{-1}[S_{xy}][S_{yy}]^{-1}[S_{yx}] \tag{3.9}$$

and

$$[M_y] = [S_{yy}]^{-1}[S_{yx}][S_{xx}]^{-1}[S_{xy}]. \tag{3.10}$$

The canonical vectors, a_m and b_m, are the eigenvectors of M_x and M_y, satisfying

$$[M_x]a_m = r_{C_m}^2 a_m, \quad m = 1, \ldots, M \tag{3.11}$$

and

$$[M_y]b_m = r^2_{C_m} b_m, \quad m = 1, \ldots, M. \tag{3.12}$$

The canonical correlations, r_C, are the square root of the eigenvalues. High canonical correlations are the result of strong relationships existing between the two data sets.

3.1.3 Fourier Series Analysis

In most cases where a signal varies with time, the signal is analyzed as a time series with the signal as a function of time. However, in some cases, particularly when a signal is periodic, problems can be solved more easily if the signal is transferred into frequency domain. In the frequency domain, a signal is separated into sine and cosine functions with varying frequencies, amplitudes and phase shifts. A continuous periodic function can be expressed as a linear combination of these sinusoids known as the Fourier series.

For a signal with a length of time T, composed of N observations at intervals of δt, the longest sinusoid component has a period of $T = N\delta t$, and the shortest has a period of $2\delta t$.

Given a function $f(t)$ that varies with time, the Fourier series is expressed as

$$F(t) = \sum_{u=-\infty}^{\infty} \left[a_u \cdot \cos\left(u\frac{2\pi}{T}t\right) + b_u \cdot \sin\left(u\frac{2\pi}{T}t\right)\right]. \tag{3.13}$$

The Fourier series can be determined for discrete functions as well as continuous functions. This is important for engineering applications as often the measurements are recorded at discrete time increments. For a discrete function, the discrete Fourier series is expressed as

$$x_i = \sum_{u=0}^{N-1} \left[a_u \cdot \cos\left(u \frac{2\pi}{N} i \right) + b_u \cdot \sin\left(u \frac{2\pi}{N} i \right) \right]. \tag{3.14}$$

The vectors of coefficients a_u and b_u can by found as

$$a_u = \frac{1}{N} \sum_{i=0}^{N-1} x_i \cdot \cos\left(u \frac{2\pi}{T} i \right) \tag{3.15}$$

and

$$b_u = \frac{1}{N} \sum_{i=0}^{N-1} x_i \cdot \sin\left(u \frac{2\pi}{T} i \right). \tag{3.16}$$

Using the coefficients a_u and b_u, the signal can be described by the amplitude and phase shift, φ, of each of the individual frequencies from $u = [0, 1, \ldots, N-1]$. The frequency index of $u = 0$ is known as the static term. It represents the mean of the signal and has no phase shift. The signal can be expressed as

$$X_u = [(Amp_0, 0), \ldots, (Amp_u, \varphi_u), \ldots, (Amp_{N-1}, \varphi_{N-1})] \tag{3.17}$$

where

$$Amp_u = \sqrt{a_u^2 + b_u^2} \tag{3.18}$$

and

$$\varphi_u = \tan^{-1} \frac{b_u}{a_u}. \tag{3.19}$$

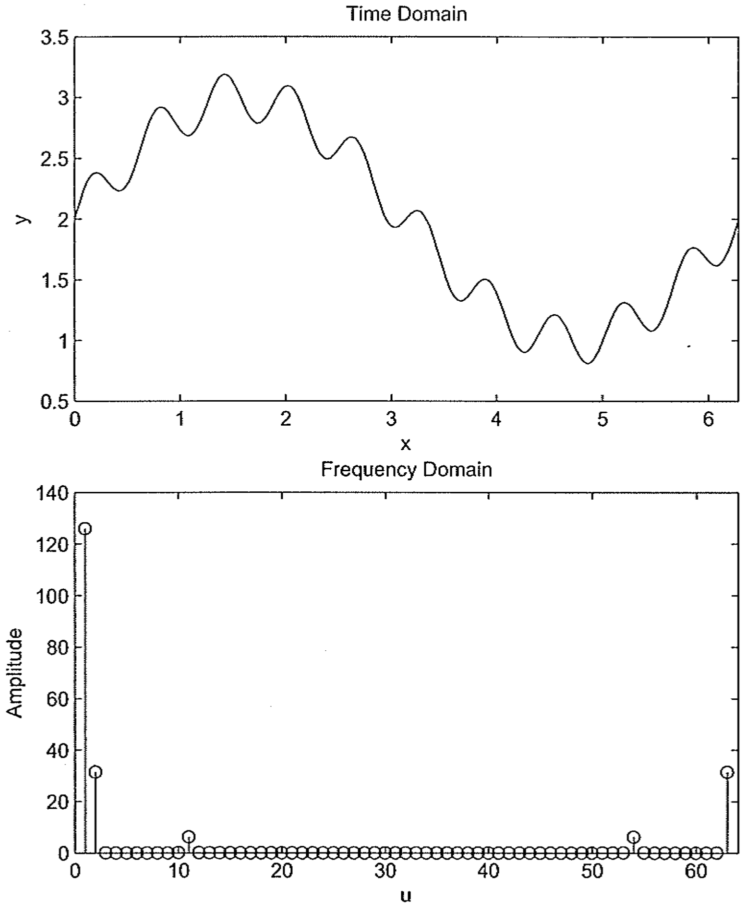


Figure 3.5: Fourier series example of function $y = 2 + \sin(x) + 0.2\sin(10x)$ in time and frequency domain.

The function $y = 2 + \sin(x) + 2\sin(10x)$ is shown in both the time domain and frequency domain on Figure 3.5. The two sine terms of different frequency are shown in the frequency domain as two amplitudes with specific frequencies. Since the first term in the function offsets the signal from having a zero mean, the static term at $u = 0$ is nonzero. This short example shows how easily Fourier series can simplify a problem that is more complex in the time domain.

Fourier series transforms have many useful applications in engineering. One important application of Fourier series is reduction of noise in signals. By separating a signal into its frequency components, the high or low frequency components can be easily identified. If a signal contains noise of a higher frequency than the desired signal, these high frequencies can be removed by using a filter which eliminated high frequencies.

A demonstration of the ability to reduce or remove noise from data is shown on Figure 3.6. In this example the daily mean temperature was calculated for the Thompson weather station from 1967 to 2000. The result should be a gradual increase to a maximum in the summer months and a decrease to the minimum in winter. The signal is noisy due to the natural variations in the day-to-day temperatures. By applying Fourier series analysis and filtering high and low frequencies, the noise portion of the signal is easily identified and removed.

Fourier series analysis and the filtering technique shown above was used extensively throughout this project to standardize station data and climate model grid data to remove seasonality of the variables. It was critical to remove noise from the daily mean and standard deviations before standardization.

3.1.4 Circulation Pattern Classification

Correlation-Based Map-Pattern Classification

Correlation-based map-pattern classification is a circulation-to-environment approach to classification. The categories are developed independent of surface conditions. The goal is to divide days based on their similarity to common pressure patterns. A day is categorized based on the strongest correlation toward one of the common map patterns. The focal point of the method is to find the common map patterns, referred to as *key days*. The classification scheme requires the user to

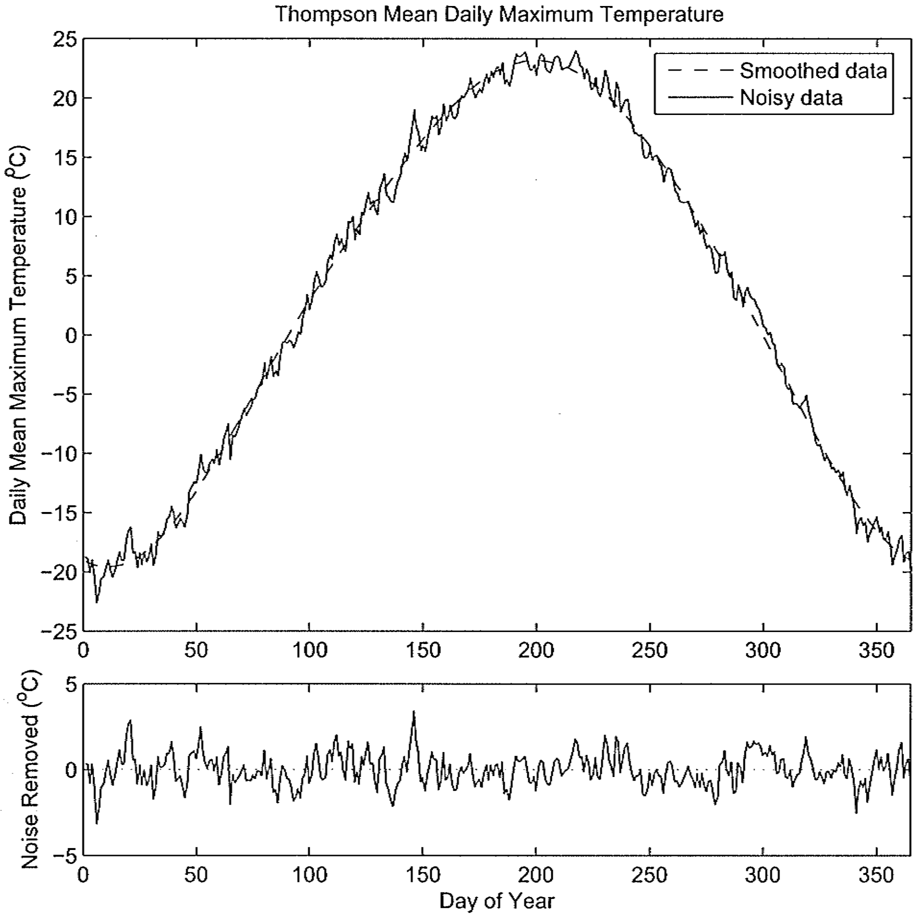


Figure 3.6: Example of noise reduction using Fourier series analysis.

decide features such as how many key days are necessary, correlation thresholds, minimum group sizes, etc.

The first step is to determine the key days which all other days will be compared to and classified. Each key days is an actual historical pressure field selected as representative of a common circulation pattern. These are selected by a comparison of circulation from each historical day to every other day on record.

Usually the data is first standardized to remove the seasonality of the data. For instance, geopotential heights will be greater in summer and have more gradual gradients than in winter. Standardization using monthly or daily means and standard deviations remove these seasonal differences.

The comparison between grids is made using the Pearson product-moment correlation, r_{xy}, defined by

$$r_{xy} = \frac{\sum_{i=1}^{N} \left[(x_i - \bar{X})(y_i - \bar{Y})\right]}{\left[\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{Y})^2\right]^{1/2}}, \tag{3.20}$$

where x represents the grid points of one grid, and y of another. $\bar{X}$ and $\bar{Y}$ represent the mean of the N grid points. The degree of correlation is a measurement of similarity between grids. A threshold value of correlation is used to discern if two grids are significantly similar. Threshold values generally range from 0.5 to 0.7, but could range higher or lower (Yarnal, 1993). The degree of similarity of one day to the rest is now recorded as either 1 or 0. For a historical record of z days, a correlation matrix of size $z \times z$ will be constructed.

The first key day is selected as the day that is significantly correlated to the most days. This is the most typical grid pattern. The first key day is removed from the record along with all days considered significantly similar. The process is repeated to identify the second key day and remove it and the days similar to it from the record. The process is repeated until all days have been accounted for, or until a user defined minimum group size has been reached. The key days represent the circulation patterns to categorize the data.

The days are then reclassified. Reclassification is necessary as a day may have been significantly correlated with more than one key day. A day could possibly be more correlated with key day 2 than key day 1, but first classified into those in key day 1. Correlation between each day and the key days are calculated and days are placed into the key day categories they are most correlated to. Decisions such as minimum group sizes and the number of categories needed to effectively categorize the circulation patterns are made. The last step is to catalog the classification which now becomes another variable to describe the large-scale climate.

Eigenvector-Based Map-Pattern Classification

Eigenvector-based map-pattern classification is another circulation-to-environment approach to classification. There are many different forms of models based on the use of eigenvectors. There are a multitude of data selection options and decomposition methods. In this methodology, gridded pressure data are imputed into the model and PCA is used as the decomposition method. Days are classified based on clusters of days with similarities in principal component scores.

The data is decomposed into principal components (PCs). The PCs are rotated to produce physically interpretable loadings that appear as pressure patterns. Individual grids will typically not resemble only one of these patterns, but will rather be a combination of the rotated PCs. Therefore, instead of classifying the grids into categories defined by one of the PCs, cluster analysis determines common combinations of the PCs. If one takes the pressure fields within the clusters and finds the average among them, the pressure fields representative of the clusters will be shown. Each cluster should have significantly different representative pressure patterns.

A subset composed of a given number of the first PCs explains the maximum amount of variation of the original data set in a minimum number of variables. In this regard, the orientation of eigenvectors after PCA is optimum. There are no other subset of eigenvectors that can explain more of the variation the original data set. However, these patterns are a result of the statistical process and may not be useful as physical explanations of the patterns which exist in the data. In many applications of PCA, it is useful to rotate the leading principal components to another projection of eigenvectors.

Two primary options for rotation exist: orthogonal or oblique. In an orthogonal rotation, the resulting eigenvectors remain orthogonal and explain unique variance. An oblique rotation results in the PCs sharing a portion of variance. Yarnal (1993) suggests that eigenvector-based map-pattern classification use an orthogonal rotation.

The rotation transforms one set of input eigenvectors, $P = (\vec{p}_{1_1}, \cdots, \vec{p}_K)$, into the output eigenvectors $Q = (\vec{q}_{1_1}, \cdots, \vec{q}_K)$. The rotation is made by the $K \times K$

matrix R such that

$$Q = PR \tag{3.21}$$

and

$$\vec{q_i} = \sum_{j=i}^{K} r_{ij}\vec{p_j}. \tag{3.22}$$

The matrix R determines the type of rotation and is selected such that a constraint $V(Q)$ is optimized. If K is orthogonal, the rotated eigenvectors will also be orthogonal; otherwise they will be oblique. One example of an orthogonal rotation is the 'varimax' method,

$$V(Q) = \sum_{i=i}^{K} fv(\vec{q_j}), \tag{3.23}$$

where fv is defined by

$$fv(\vec{q}) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{q_i}{s_i}\right)^4 - \frac{1}{m^2}\sum_{i=1}^{m}\left(\frac{q_i}{s_i}\right)^2, \tag{3.24}$$

where m is the length of the vectors. The constant s_i is chosen by the user. The raw varimax rotation is obtained when s_i is set to 1 for all i, and the normal varimax rotation is obtained by setting s_i equal to $\sum_{j=1}^{K}(p_{ij})^2$.

The PCs corresponding to the new rotated eigenvectors are the dot products of the data points and rotated eigenvectors,

$$PC_i^r = q_i^T X. \tag{3.25}$$

The pressure grids will be composed of combinations of the rotated eigenvectors. Rather than compare grids to the rotated PCs, groups with similar PCs are identified. Cluster analysis is a statistical tool used to classify multivariate data into previously unknown groups. Cluster analysis is applied to the principal components to divide the days into circulation patterns. The most common methods for clustering data are hierarchical. In the beginning of the analysis, all days belong to their own group or cluster. A distance measurement is made between all of the n groups. The two closest groups are then combined into one group to make $n-1$ groups. The process of clustering the most similar groups could continue until all observations are grouped as one. The process is therefore stopped when a specified distance threshold or minimum number of groups is reached.

Many options exist in clustering algorithms. The distance measurement between two vectors can be made in many ways: Euclidean distance, squared distance, Mahalanobis distance, or Pearson correlation are just a few of many. The most common is the Euclidean distance (Wilks, 1995),

$$d_{ij} = \|x_i - x_j\| = \left[\sum_{i=1}^{K} (x_{i,k} - x_{j,k})^2 \right]^{1/2}, \tag{3.26}$$

where K is the dimension of the vectors.

There are also many methods available to measure the distance between two clusters of vectors, G_1 and G_2. Some of the common methods are:

- Single-linkage clustering (minimum-distance clustering)

  The distance between G_1 and G_2 is the smallest distance between any member of G_1 and any member of G_2,

$$d_{G_1,G_2} = \min\left[d_{ij}\right]. \tag{3.27}$$

- Complete-linkage clustering (maximum-distance clustering)

  The distance between G_1 and G_2 is the maximum distance between any member of G_1 and G_2,

$$d_{G_1,G_2} = \max\left[d_{ij}\right]. \tag{3.28}$$

- Average-linkage clustering

  The distance between G_1 and G_2 is the average Euclidean distance between all possible pairs of points,

$$d_{G_1,G_2} = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=i}^{n_2} d_{ij}, \tag{3.29}$$

  where n_1 is the number of days in G_1, and n_2 is the number of days in G_2.

- Centroid clustering

  The distance between G_1 and G_2 is the distance between the centroids or average vectors,

$$d_{G_1,G_2} = \|\bar{x}_{G_1} - \bar{x}_{G_1}\| . \tag{3.30}$$

- Ward's minimum-variance clustering

  This method makes G groups from $G+1$ groups by minimizing the variance between each day and the centroid of the cluster it belongs to, summed over the G groups. The variance over the G groups is determined as

$$V = \sum_{g=1}^{G} \sum_{i=1}^{n_g} \|x_i - \bar{x}_g\|^2 . \tag{3.31}$$

Yarnal (1993) indicates that no preference in clustering method has developed in the literature and that the selection of clustering method is most likely less of a priority than other user decisions made during eigenvector-based map-pattern classification.

An important user decision which will directly affect the outcome of the classification scheme is the number of clusters to retain. There are no clear-cut rules to follow when selecting the number of clusters. Scree plots of various output statistics can be viewed and subjectively interpreted. The number of clusters retained should be similar to the number of circulation patterns determined from analysis by correlation-based map-pattern classification.

Each cluster represents days with similar circulation patterns that are different from circulation patterns in other groups. The clusters cannot be physically mapped or interpreted. To physically interpret the clusters, the average of the PCs can be found and back transformed to find the gridded values. The resulting patterns should be physically meaningful pressure patterns.

3.1.5 Optimization

In some mathematical problems it may be difficult to find an analytical solution because of the complexity of the problem. Numerical optimization methods can often be used to find solutions to problems that cannot be solved analytically.

Optimization is used to find a set of parameters, $x = [x_0, x_1, \ldots, x_n]$, that can be defined in some way as optimal. Typically, optimization is used to find the minimum of a function, $f(x)$, as described by

$$\min_{x} f(x). \tag{3.32}$$

One of the simplest set of optimization techniques are line search methods. If $f(x)$ is a function, the iterative process can be setup such that

$$x^{(k+1)} = x^{(k)} - \alpha_k d_k \; ; \quad k = 0, 1, 2, \cdots . \tag{3.33}$$

To start, the iteration counter k is set to zero and an initial guess is made for x^0. A search direction, d_k, is chosen or calculated, and the step size α_k which minimizes the function in that direction is found by solving

$$\min_{\alpha} \phi(\alpha) = f(x_k + \alpha d_k). \tag{3.34}$$

Once the optimal α_k is found, the parameters are then updated in Equation 3.33. The process is repeated to find the optimal set of parameters.

In the simplest case, the search could be done one variable at a time. However, this would most likely lead to a long convergence time and many evaluations of the objective function. An improvement is to search in the direction of steepest descent. The direction of steepest descent is the gradient of the objective function at x_k. The gradient can be calculated as the unit vector of partial derivatives of $f(x)$ at x_k.

This section presented a very brief introduction to optimization for the purpose of understanding the methodology applied later in the optimization of the nearest neighbor model. Many textbooks offer in-depth discussions of various optimization techniques (Nocedal and Wright, 1999; Chong and Stanislaw, 2001).

3.2 k-Nearest Neighbor Resampling

Nearest neighbor resampling is a nonparametric method which resamples data from a historical record. The nonparametric aspect of the model makes it appealing to statistical downscaling. Parametric models, such as the regression methods or weather generator models, require extensive parameter estimation.

The basic idea of the model is that if one compares the large-scale variables that a GCM produces for a given day to the same variables of a historical record, a similar day in the historical record can be found. Since there is a direct link between the large-scale and local climates, the simulation day should exhibit a local climate similar to the historical day with a similar large-scale climate. The local climate variables required for hydrologic modelling can be retrieved from the selected historical day and used as the downscaled variables for the GCM simulated day.

The comparison between the simulation day and the historical record is made by using a vector of variables referred to as the feature vector. A distance measurement is made between the feature vector and individual days in the historical record. A group of the k most similar days is retained and one is selected to provide the local climate variables. The process is repeated to produce a time series.

The following section describes the methodology of the k-nearest neighbor resampling model.

3.2.1 Feature Vector

The feature vector, D_t, is used to compare the simulation day to historical days. The feature vector is given as

$$D_t = [v_1, v_2, v_3, \ldots, v_n], \tag{3.35}$$

where n is the number of variables contained in the feature vector. The composition of D_t can be varied from a few climate variables (Buishand and Brandsma, 2001) to many variables (Gangopadhyay et al., 2005). The selection of variables to include in the feature vector is an important step in the development of the nearest neighbor algorithm. Some multivariate statistical methods can be used to explore the relationship of large-scale atmospheric variables and local weather. Canonical correlation analysis or circulation pattern classification and analysis are examples of potential methods to explore relationships between the two scales of variables. An investigation of different combinations of large-scale variables has been done in some studies (Buishand and Brandsma, 2001).

The raw variables require some manipulation before they are used by the algorithm. In the literature, data sets are often standardized to remove seasonality and avoid differences in magnitude between different types of variables. In some instances, principal component analysis is used to reduce the number of variables used in the feature vector (Gangopadhyay et al., 2005; Buishand and Brandsma, 2001; Young, 1994).

3.2.2 Finding the k-Nearest Neighbors

A reduced set of days to resample from is determined by finding the nearest neighbors to the current feature vector in state space. The neighbors are found by calculating the distance between the feature vector of the simulation day and the feature vectors of historical days. The k-nearest neighbors are the k days that are most similar to the simulation day and therefore produce the smallest distances.

To reduce the effect of seasonal variation, Lall et al. (1996) divided the year into four seasons and restricted the selection of neighbors to the season of the simulation day. An alternative to dividing the year into seasons is choosing the neighbors from a moving window around the calendar day of the simulation day. The size of the moving window, W, can be varied. A window of 14 days was used by Gangopadhyay et al. (2005) and by Yates (2003). A larger window of 61 days was used by Buishand and Brandsma (2001) and by Wójcik and Brandsma (2003).

The number of neighbors to retain after the distances are calculated is a feature of the algorithm which the designer can manipulate. However, the number of neighbors retained can have significant effects on the algorithm outcome. The number of nearest neighbors to retain was studied using general cross-validation (GVC) by Rajagopalan and Lall (1999) and by Lall and Sharma (1996). The goal of the GVC studies was to minimize the predictive mean square error of the k-nn algorithm (Rajagopalan and Lall, 1999). In both studies, good results were obtained when the value of k was equal to the square of the sample size,

$$k = \sqrt{n}, \tag{3.36}$$

where the sample size, n, is the number of historical days that could be possible neighbors (number of years of data × W). Buishand and Brandsma (2001) varied the number of neighbors using two, five, twenty, and fifty neighbors in their study. Their study recommended a small k, but larger than two. A k-value equal to five showed the best overall results. Young (1994) also found that a smaller k can produce good results.

The k nearest neighbors are searched for in the historical record by using a distance metric. Two distance metrics are commonly used, a Euclidean distance metric (Gangopadhyay et al., 2005; Buishand and Brandsma, 2001; Rajagopalan and Lall, 1999; Brandsma and Buishand 1998; Lall and Sharma, 1996), or a Mahalanobis distance metric (Yates, 2003; Wójcik and Buishand, 2003).

The weighted Euclidean distance, δ_{tu}, between two feature vectors, D_t and D_u, can be calculated as

$$\delta_{tu} = \sqrt{\sum_{i=1}^{n} w_i \left(v_{ti} - v_{ui}\right)^2}, \tag{3.37}$$

where n is the number of variables in the feature vector, and w_i is the weight given to the variable v_i (Brandsma and Buishand, 1998).

The Mahalanobis distance, δ_{tu}, between two feature vectors, D_t and D_u, can be calculated as

$$\delta_{tu} = \sqrt{(D_t - D_u)' \, B^{-1} \, (D_t - D_u)}, \tag{3.38}$$

where B is the covariance matrix of the feature vector D_t.

Pairs of feature vectors that have smaller distances represent days that have a more similar climate than pairs having larger distance metrics. The historical days within the selection window are ranked according to distance and the k nearest neighbors are retained.

3.2.3 Choosing a Neighbor

Once the k nearest neighbors have been chosen, the next step is to resample one of the neighbors. There are different methodologies to resampling a neighbor. Most applications of the k-nn model have had an objective to generate synthetic time series of weather for short-term forecasting. In these situations, the feature vector is composed of the station variables from the previous time step. If the nearest neighbor is resampled every time, the model would only reproduce the existing time series. To avoid this problem, more than one neighbor are retained.

The selection of which neighbor to use could be done randomly with equal chance given to all neighbors. However, it is more common to use a weighting scheme to favor days that have a smaller distance. There are two common ways to assign weights to the neighbors during resampling. One possibility is to use a decreasing kernel density function (Lall and Sharma, 1996, Brandsma and Buishand 1998, Rajagopalan and Lall, 1999, Buishand and Brandsma, 2001, Wójcik and Buishand, 2003, Yates et al., 2003). The kernel function distributes the probability of the day being selected based on its rank in the set of sorted distances,

$$p_j = \frac{1/j}{\sum_{i=1}^{k} 1/i}, \quad j = 1, \ldots, k \tag{3.39}$$

where p_j is the probability that the day of rank j is resampled. A plot of the probabilities assigned by a kernel density function with $k = 20$ nearest neighbors is shown on Figure 3.7. This is a simple method that has been found to be effective in the literature. Also, using a kernel allows the weights to be calculated once rather than each time the algorithm is used to resample a neighbor because the weights do not depend on the actual distances.

An alternative to pre-calculating the probability weights is to use a weight function that depends on the distances. Gangopadhyay et al. (2005) use the

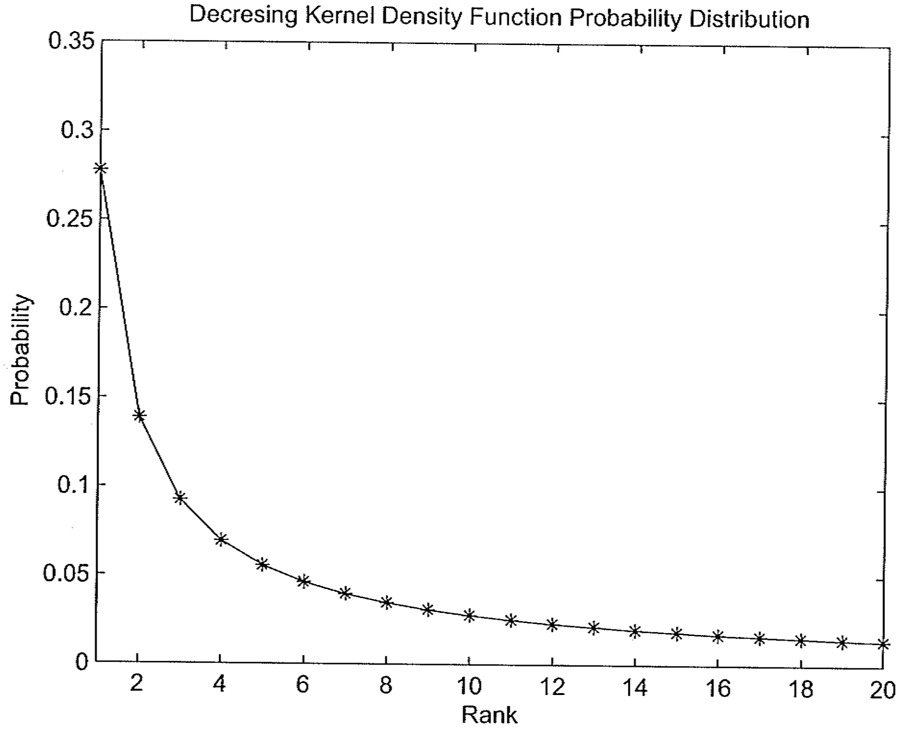


Figure 3.7: Plot of the decreasing kernel density function.

bi-square weight function

$$p_j = \frac{\left[1 - \left(\frac{\delta_j}{\delta_k}\right)^2\right]^2}{\sum_{i=1}^{k}\left[1 - \left(\frac{\delta_i}{\delta_k}\right)^2\right]^2}, \tag{3.40}$$

where δ_k is the distance to the k^{th} neighbor after sorting. The advantage of this form of probability weighting is that the probability is based on the actually similarity of the historical day to the simulation day. The disadvantage is that the weights depend on the distances and need to be calculated each time a day is resampled, increasing computation time.

The application of the resampling algorithm for downscaling GCM data is slightly different than the applications for weather forecasting. In weather forecasting, the station variables for one day are often used to predict the variables for the next day. The simulation day is a day that has actually occurred in the past, and there is a danger that the model will reproduce spans of historical data exactly as they occurred in the past if the nearest neighbor is resampled too frequently. Therefore k must be selected large enough to prevent this from occurring. When downscaling a GCM, the GCM generates simulation days that are separate from the historical record. Therefore, problems created by resampling the closest of the nearest neighbors will not occur when downscaling GCM data. Retaining a small number of neighbors when downscaling a GCM should not adversely affect model performance.

3.2.4 Modelling Climate Change Scenarios

To evaluate a possible climate change scenario, GCM data for a particular scenario will be used as input to the algorithm. Historical days with feature vectors similar to the GCM simulation days will be resampled to generate a time series of downscaled climate data. A variety of GCM generated scenarios are available based on different population growth predictions and consumption models. Another possible concept to simulate a climate change scenario is to strategically resample data. Yates et al. (2003) adapted a k-nn algorithm to generate alternative climate scenarios by using prescribed conditioning data. To apply this concept, years were given an index number. For example, if it was desired to have a climate scenario depicting warm moist springs with cool dry autumns, the weekly means for temperature and precipitation would be used to create an index. The years would then be ordered based on a paired ranking method. The years with above average temperature and precipitation in spring and below average temperature and precipitation in autumn would receive the highest ranking and would be given a higher index. When using the k-nn algorithm, the years with higher indices are biased to favor days from those years in the resampling procedure. The resulting data would have the desired attributes of the annual climate cycle.

In this project, the first procedure of modelling climate change through the use of GCMs will be employed. Data from the Canadian global climate model (CGCM3.1/T47) will be used as input into the k-nn model. Although the indexing methodology of Yates et al. (2003) will not be further explored, the methodology does have merit for developing adaptation strategies, particularly for generating data to test a system's sensitivity against a particular climate trend.

Chapter 4

Data

4.1 Canadian Daily Climate Data

Canadian Daily Climate Data (CDCD) is a set of archived weather station data managed by Environment Canada. The data set includes data from over 10,000 weather stations across Canada. The variables recorded and the length of record vary from station to station, with some records extending as far back as 1830.

Many variables are available in the data set. Temperature and precipitation variables are available at the daily time scale and include:

- Temperature: Temperatures are recorded 1.5 m above the ground in a box called a Stevenson Screen.
  - Maximum Temperature: The highest temperature of a day is recorded as the maximum temperature.
  - Minimum Temperature: The lowest temperature of the day is recorded as the minimum temperature.

  - Mean Temperature: The average between the maximum and minimum temperatures is the mean daily temperature.

- Precipitation: Precipitation as rain, drizzle, freezing rain, freezing drizzle, snow, and hail are all recorded as depth of water. Precipitation is recorded using a standard Canadian rain gauge, a cylindrical container 40 cm high and 11.3 cm in diameter.

- Snowfall: Measured as the depth of newly fallen snow.

- Depth of snow on ground: The depth of accumulated snow on the ground.

The CDCD data is available online through Environment Canada's web site, or available on a CD-ROM (Environment Canada, 2000). Since the data set is large, it has been divided into western and eastern data sets. The western Canada data set contains climate data for all stations in Manitoba, Saskatchewan, Alberta, B.C., Yukon and N.W.T. The western data set is available on a separate CD-ROM. The CD-ROM also includes software to extract the data and data description text files.

One characteristic of the station measurement data is that it is common to have missing entries in a data file. Some stations may have an almost complete data set, while other stations, particularly stations in small towns or northern areas, may have a high rate of missing data. Sometimes the missing entries may be infrequent and only a day long, other times there may be lengthy spans of missing data of months or years as equipment fails or stations are temporarily abandoned. Missing data are marked with an entry of "M" or a numerical marker such as "-9999".

4.2 NCEP/NCAR Reanalysis 1

In a joint effort, the National Center for Environmental Prediction (NCEP) and the National Center for Atmospheric Research (NCAR) generated the NCEP/NCAR Reanalysis 1 data set in the late 1990's (Kalnay et al., 1996). The project first started in 1991 to correct jumps in climate data that occurred over time as a result of changes in equipment and data assimilation methods.

The reanalysis data set is generated by assimilating multiple sources of data by a consistent assimilation method throughout the data period. Some of the sources of data include land surface, ship, rawinsonde, weather balloon, aircraft, and satellite data. By using a consistent assimilation method, biases or jumps in climate caused by changes in the assimilation method are minimized.

The data set is available on a grid with a resolution of 2.5° x 2.5° latitude and longitude over the entire globe. Originally the reanalysis was available for 40 years (1957-1996), but is currently available from 1948 to the present day. Data are available at a temporal resolution of 4-times daily, daily and monthly values.

NCEP/NCAR Reanalysis 1 has a massive array of output variables. Some of the variables are available at multiple pressure levels. There are 17 pressure levels available, including the 1000, 925, 850, 700, 600, 500, 400, 300, 250, 200, 150, 100, 70, 50, 30, 20, 10 mb geopotential heights. A small sample of the variables available include:

- Air temperature (surface and at multiple pressure levels)
- Geopotential height (surface and at multiple pressure levels)

- Relative humidity (surface and at multiple pressure levels)
- Omega (vertical velocity) (surface and at multiple pressure levels)
- U-wind
- V-wind
- Precipitable water (surface)
- Sea level pressure (surface)
- Soil moisture

Variables within the NCEP/NCAR Reanalysis 1 data set are derived by different methodologies. Some of the variables in the data set are assimilated through interpolation directly from observations. Some variables are determined by the model during the data assimilation and do not use any observation data. The reanalysis gridded fields have been divided into four classes, depending on the relative influence of the observational data and the model on the specific variables. Class A indicates that the analysis variable is strongly influenced by observed data and is therefore the most reliable class. Class A variables include upper air temperature and wind. The designation B indicates that, although there are observational data that directly affect the value of the variable, the model also has a very strong influence on the analysis value. Humidity and surface temperature are examples of this category. Class C indicates that there are no observations directly affecting the variable, so that it is derived solely from the model fields forced by the data assimilation to remain close to the atmosphere. Class C variables include clouds, precipitation, and surface fluxes. Finally, the letter D represents a field that is obtained from climatological values and does not depend on the model. Class D variables include plant resistance and land-sea mask.

NCEP/NCAR Reanalysis 1 is retrieved in a data format called netCDF. NetCDF (network Common Data Form) is an interface for array-oriented data access and involves a library that provides an implementation of the interface. The netCDF library also defines a machine-independent format for representing scientific data. Together, the interface, library, and format support the creation, access, and sharing of scientific data. A netCDF toolbox is available for unpacking or creating netCDF files in Matlab. The toolbox simplifies handling the netCDF files.

NCEP/NCAR Reanalysis 1 data is useful for calibrating and validating statistical models for downscaling GCM data. The NCEP/NCAR Reanalysis 1 data are available at a similar resolution as GCM data. Also, many GCM variables are available in the NCEP/NCAR Reanalysis 1 data set. Therefore, NCEP/NCAR Reanalysis 1 data can be easily used as predictors in a statistical downscaling model to generate data at weather stations. The simulated weather can then be compared to the observed record to determine how well a model performs.

A second global reanalysis data set produced by NCEP and the Department of Energy (DOE), called the NCEP-DOE Reanalysis 2, is an improved version of the NCEP/NCAR Reanalysis 1 model that fixes errors and employs updated parameterizations of physical processes. However, this data set is only available for the period of 1979 to 2003. Due to this relatively short temporal span of data, it was not used in this study.

4.3 North American Regional Reanalysis

The NCEP North American Regional Reanalysis (NARR) data set is a very high resolution reanalysis of the North American region (Mesinger et al., 2006). The NARR project is an extension of the NCEP Global Reanalysis over the North American region. The grid resolution is 349 × 277 which is approximately 0.3 degrees (32 km) at the lowest latitude. The higher spatial resolution is better at capturing the regional hydrological cycle. The higher resolution also allows for better data assimilation, including assimilated precipitation rather than model derived precipitation as in the NCEP/NCAR Reanalysis 1 data set. NARR data is downloaded in netCDF files, the same file format as NCEP/CAR Reanalysis 1 data and CCCma CGCM3.1 data.

NARR data is also available at a higher temporal resolution than NCEP/NCAR Reanalysis 1 data. Data are available at time intervals of three hour, daily and monthly means. This improves the model's ability to capture the diurnal cycle in variables.

Although the data has higher spatial and temporal resolution than NCEP/NCAR Reanalysis 1 data, its utility is somewhat limited because it is currently only available for the 25-year period from January 1, 1979, to December 31, 2006.

Despite the NARR data set covering a relatively short time period, it may have many useful purposes in climate change assessment studies. Choi et al. (2007) evaluated the temperature and precipitation data from the NARR data set by comparison with selected weather stations in Manitoba and concluded NARR data have good potential for use as input data for hydrological models. Choi et al. (2009) conducted a pilot study on evaluating the reliability of NARR data for hydrologic modelling. They applied NARR data to calibrate a hydrologic model and compared it to the calibration obtained with observed weather station data in northern Manitoba. In their study, the use of NARR data for hydrological modelling was found to be promising. Kim et al. (2008) conducted a study using NARR as a replacement for weather station data in the k-nearest neighbor resampling downscaling model developed in this report to evaluate the effect of climate change scenarios on the Winnipeg River Basin. A detailed description of the downscaling using NARR data can be found in Section 6.4.

4.4 CCCma CGCM3.1/T47

The Canadian Centre for Climate Modelling and Analysis (CCCma) is a division of the Climate Research Branch of the Meteorological Service of Canada. The CCCma conducts research in coupled and atmospheric climate modelling, sea-ice modelling, climate variability and predictability, the carbon cycle, and a number of other areas. The CCCma has developed a GCM named the Canadian Centre for Climate Modelling and Analysis (CCCma) Coupled Global Climate Model (CGCM). The CGCM is a combination of two components, an ocean model and an atmospheric circulation model. An in-depth description of the CCCma CGCM can be found in Flato et al. (2000).

To date, there have been three generations of the CGCM. The third generation model, CGCM3, is composed of a second generation atmospheric circulation model developed by the CCCma, the AGCM2, and the ocean component is an updated version based on the GFDL MOM1.1 code. Like other GCMs, the CGCM3 requires huge computational power to run. Originally, the CGCM3 was developed and ran on a NEC SX/6 vector supercomputer. Modifications were made to the model to allow it to be run on a distributed memory IBM computer system. This latter version, called the CGCM3.1, supplies the data available on CCCma's web site for the third generation CGCM.

A vast array of variables are available from the CGCM3.1. The variables are similar to the variables available form the NCEP/NCAR Reanalysis 1 data set. The data are available for download as daily data (in netCDF format) for a variety of emission scenarios, including:

- 20c3m: The IPCC 20^{th} Century experiment for years 1850-2000, available at daily time scale for 1961-2000.
- SRES A1B: The IPCC SRES A1B 720 ppm stabilization experiment for years 2001-2100.
- SRES B1: The IPCC SRES B1 550 ppm stabilization experiment for years 2001-2100.
- SRES A2: The IPCC SRES A2 experiment for years 2001-2100, initialized from the end of the 20C3M experiment.
- COMMIT: The IPCC committed experiment for years 2001-2100, initialized from the end of the 20C3M experiment with greenhouse gas concentrations remaining constant throughout the 21^{st} century.

- PICNTRL: The IPCC pre-industrial control experiment.

For the 20c3m run, model output for 2-D variables is available for the years 1850-2000, while model output for 3-D variables is available for the years 1961-1980, and 1981-2000. For the daily time scale of the future scenarios of A1B, B1, A2 and COMMIT, the 2-D variables are available for 2001 to 2100, while 3-D variables are only available for 2046-2065 and 2081-2100. The PICNTRL scenario has model output for the five variables sea level pressure, precipitation, maximum surface temperature, minimum surface temperature, and surface temperature for time slices of the years 1850-1950, 1951-2050, 2051-2150, 2151-2250, 2251-2350, 2351-2450, 2451-2550, 2551-2650, 2651-2750, and 2751-2850. All remaining 2-D variables and 3-D variables are available for the years 1961 to 2000. For all available data, the 3-D variables are available at the 200, 300, 400, 500, 600, 700, 850, 925, and 1000 mb geopotential heights.

During this study the scenarios used from the CGCM3.1/T47 model were the 20c3m (1961 to 2000), SRES A1B (2046-2065 and 2081-2100), SRES B1 (2046-2065 and 2081-2100) and SRES A2 (2046-2065 and 2081-2100)

Data for the CGCM3.1 are available in two different grid resolutions. The T47 version has a grid resolution of approximately 3.75 degrees in latitude and longitude (Figure 4.1), and 31 levels in the vertical. The T63 version has a grid resolution of approximately 2.8 degrees latitude and longitude and 31 levels in the vertical. At the time at this study, not all atmospheric variables were available for the T63 version. Therefore the CGCM3.1/T47 model supplied the GCM data used in this study.

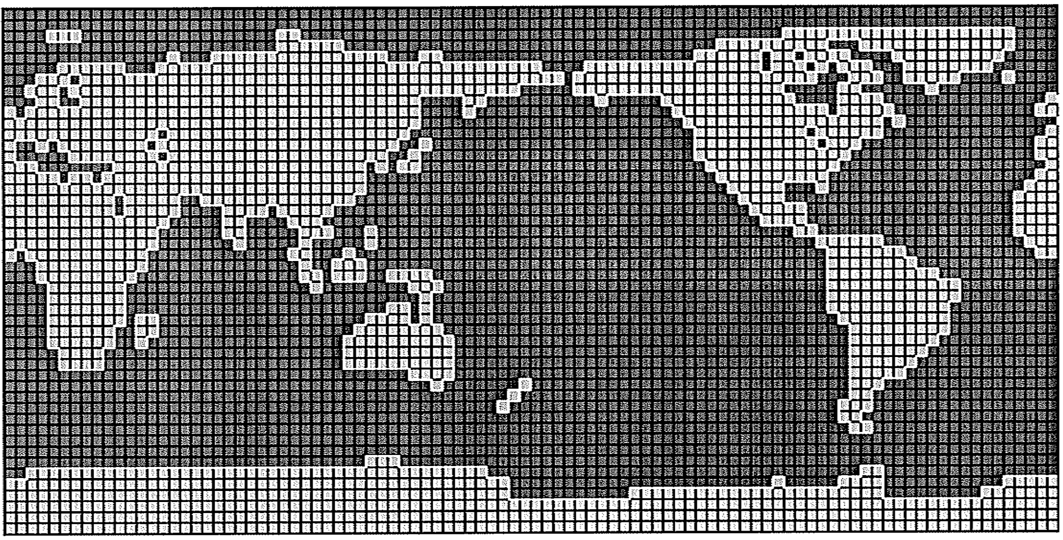


Figure 4.1: CGCM3.1/T47 grid.

Chapter 5

Data Analysis Results

5.1 Global Climate Model Validation

In this project, the GCM selected to provide the climate change scenario data was the CCCma's CGCM3.1/T47. This model was selected because daily data for a multitude of emission scenarios are readily accessible to download free of charge. Due to time constraints and the effort required to fully explore and downscale a GCM data set, only one model was selected. In the future, to develop a full ensemble of possible future climate scenarios, it is recommended that other models be adapted to study climate change in the Canadian Prairies.

When selecting a GCM for climate change assessment, one should have confidence that the output of the GCM is realistic. To address this, the present section will examine the means of the CGCM3.1/T47 data to be used in the downscaling model.

The mean values for NCEP/NCAR Reanalysis 1 and the CGCM3.1T/47 data grids for surface temperature, temperature at 500 mb, temperature at 850 mb, 850 mb geopotential height and 500 mb geopotential height are shown on Figure 5.1 to Figure 5.5. The same mean values of variables were also plotted for the GCM output of future climate scenarios to validate the GCM output is consistent with the general trends presented in the literature. As an example of the future GCM data, the results from the A2 scenario are included in each of the figures for the 2046 to 2065 and 2081 to 2100 time slices.

The results for surface temperature are shown on Figure 5.1. Cooler temperatures are present for the 20c3m data compared to the NCEP/NCAR Reanalysis 1 data in some areas. The two data sets deviate the most in the north east and southwest portions of the plots. In the middle of the study area, the 20c3m data matches quite well to the NCEP/NCAR Reanalysis 1 data. For the A2 data, as expected the 2046 to 2065 data was warmer than the 20c3m data, and the 2081 to 2100 data was warmer than the 2046 to 2065 data. Future surface temperatures in the CGCM3.1/T47 data are consistent with the literature reviewed in Section 2.

The results for temperature at the geopotential height of 850 mb are shown on Figure 5.2. The 20c3m data has a similar pattern of contours as NCEP/NCAR Reanalysis 1 data; however in the middle and southern regions of the study area, the 20c3m has warmer temperatures by as much as 2°C. This is a fairly large temperature magnitude of difference, the same as between the different emission scenarios. As in the surface temperature results, the temperatures continually increase with time.

The results for temperature at the geopotential height of 500 mb are shown

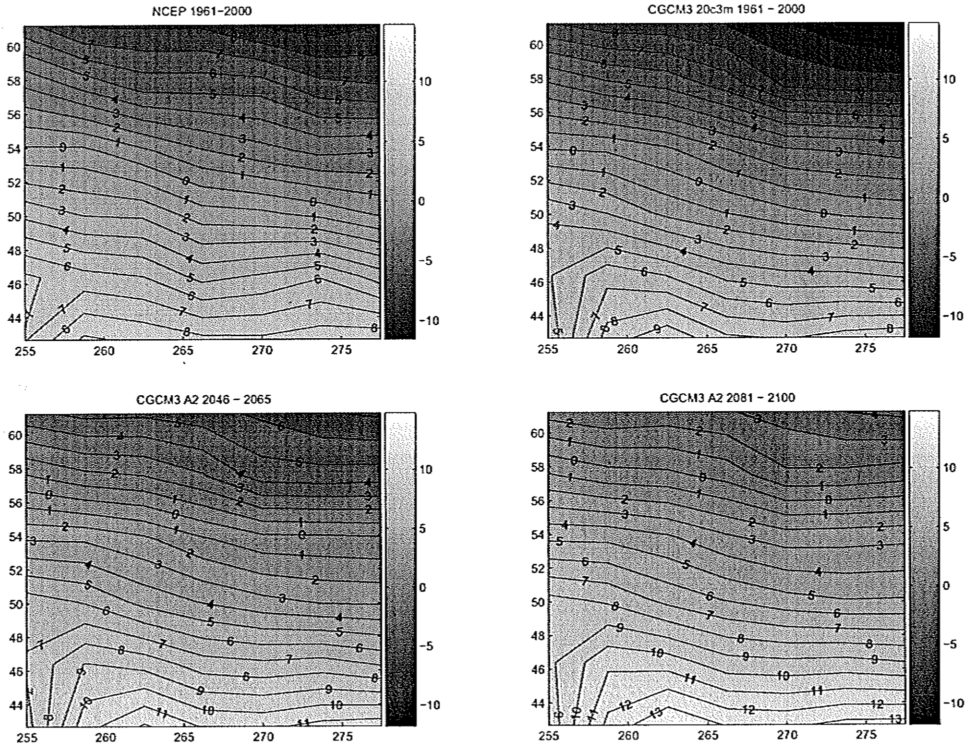


Figure 5.1: Mean values of surface temperature.

on Figure 5.3. The same magnitude of biases in temperature that were present at the 850 mb level are present at the 500 mb level. However, the 20c3m data is 1°C to 2.5°C cooler than the NCEP/NCAR Reanalysis 1 data. The largest biases are present in the southern region of the study area. The temperatures continually increase with time.

The results for the mean 850 mb geopotential height are shown on Figure 5.5. In general, the mean values of the 850 mb geopotential height are just above 1400 m above sea level. In the southern portion of the study area, the 20c3m and NCEP/NCAR Reanalysis 1 data match well, however in the northern areas the 20c3m 850 mb geopotential height is up to 20 m lower than the NCEP/NCAR Reanalysis 1 data. As temperatures increase, air expands and raises the distance

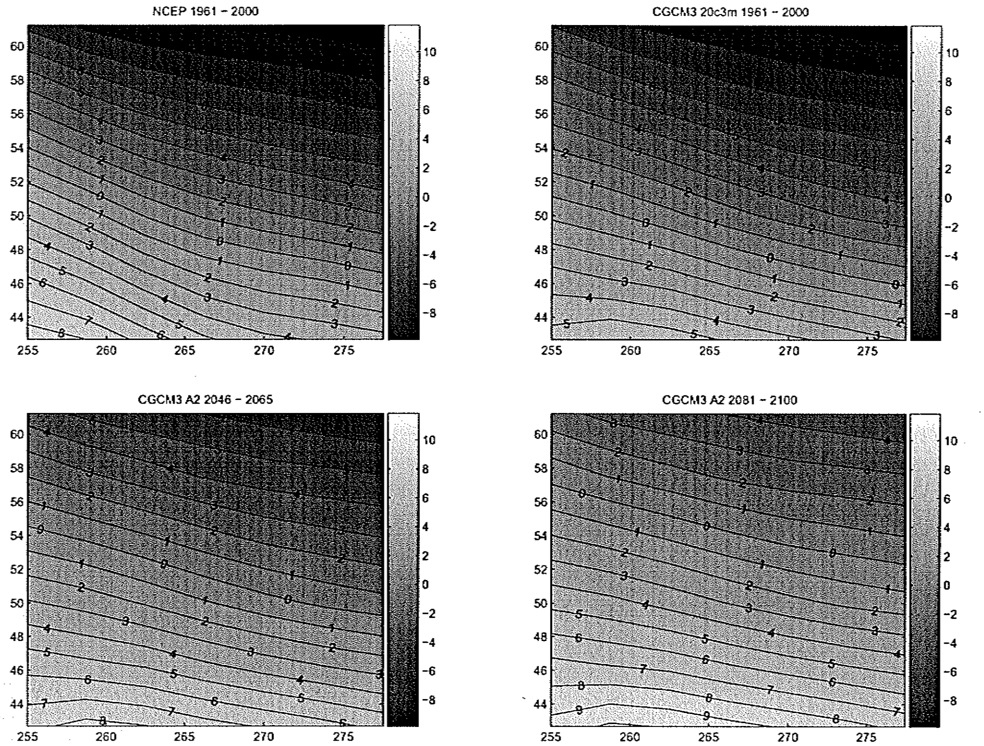


Figure 5.2: Mean values of temperature at the 850 mb geopotential height.

to geopotential height levels. In the A2 scenarios, the mean geopotential height increases slightly with time, which is consistent with expectations.

The results for the mean 500 mb geopotential height are shown on Figure 5.4. The mean values of the 500 mb geopotential height in the study area are generally around 5500 m above sea level. The 20c3m mean values are 25 to 50 m lower than the NCEP/NCAR Reanalysis 1 mean values. The difference between the two data sets for the current climate period are approximately equal to the difference between the 2046 to 2065 and 2081 to 2100 data sets of the A2 scenario.

The analysis of the mean values of the NCEP/NCAR Reanalysis 1 data and 20c3m scenario output from the CGCM3.1/T47 show that for the current period, the CGCM3.1/T47 has slight biases in the variables in many areas. Overall,

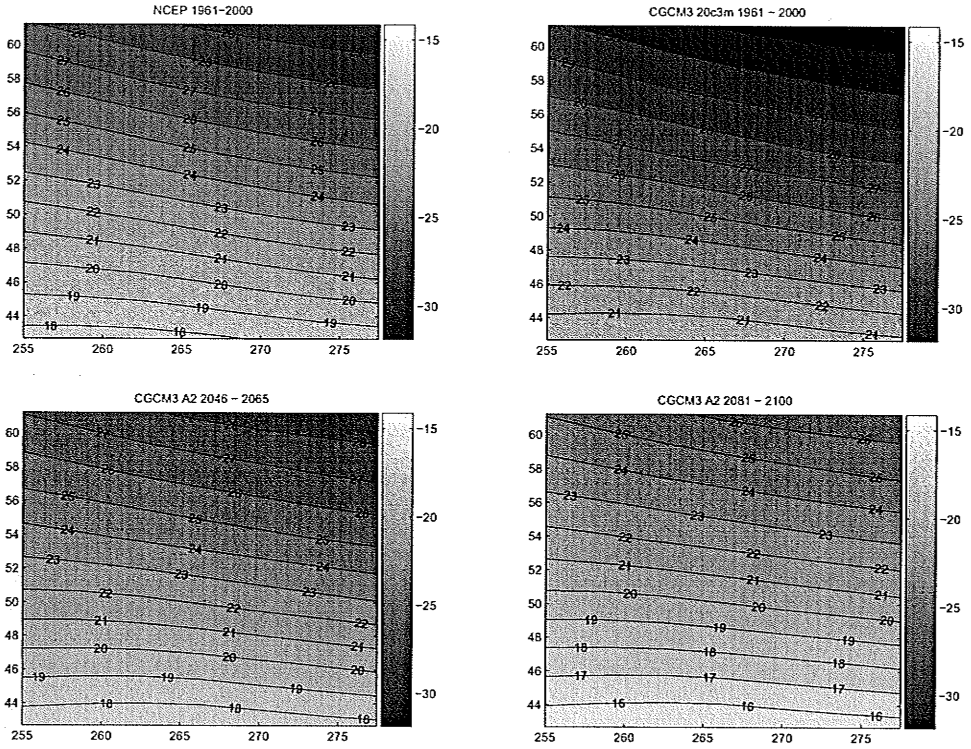


Figure 5.3: Mean values of temperature at the 500 mb geopotential height.

patterns in variables are reproduced very well, as seen in the contour lines of the various plots. The 20c3m output should be consistent with NCEP/NCAR Reanalysis 1 data. The biases in the CGCM3.1/T47 can be easily dealt with though standardization of the data. The NCEP/NCAR Reanalysis 1 and the 20c3m CGCM3.1/T47 output will be standardized so that each grid point has a mean of zero and a standard deviation of one.

When standardizing the GCM output from future emission scenarios, the biases present between these outputs and the 20c3m scenario must be preserved. These biases hold the information necessary to evaluate climate change trends. To preserve the bias, future scenarios will be standardized using the mean and standard deviation of the 20c3m scenario.

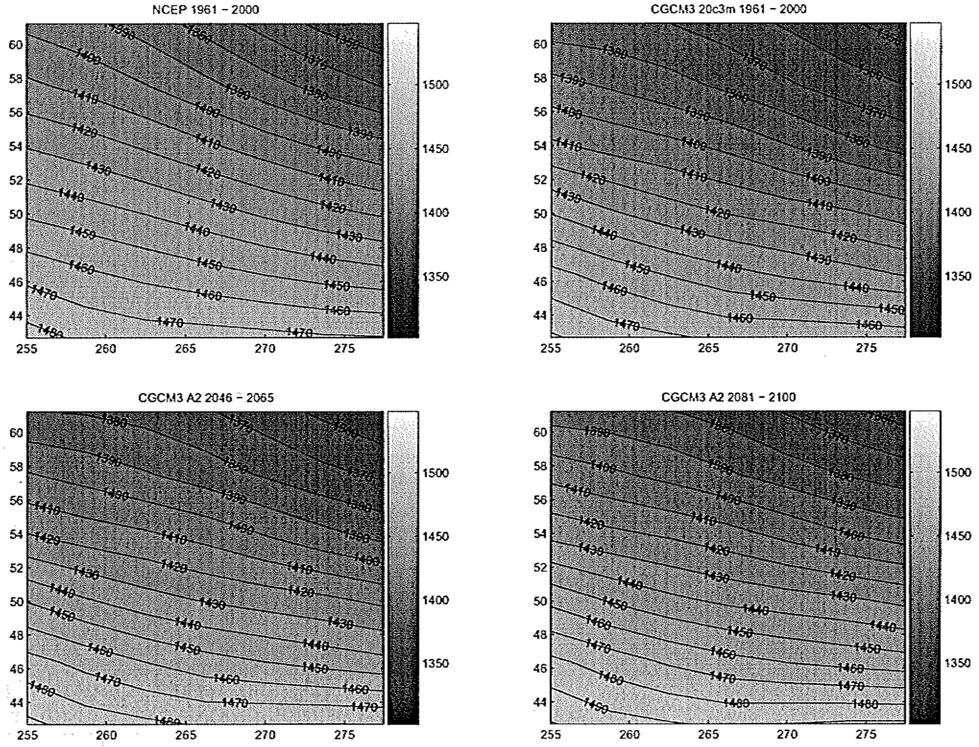


Figure 5.4: Mean values of 850 mb geopotential height.

To determine if biases are present in the temporal dimension of the 20c3m model output, the surface temperature data for Thompson, Manitoba, were extracted for both the 20c3m output and NCEP/NCAR Reanalysis 1 data and compared to the mean monthly observed temperature from Environment Canada's CDCD data set. The CGCM3.1/T47 and NCEP/NCAR Reanalysis 1 data were interpolated from the nearest grid points to the location of Thompson. The results of this exercise are shown on Figure 5.6.

Biases are present for both data sets, and these biases are not consistent throughout the year. For the 20c3m output, monthly biases range from underestimation by 1.6°C in the month of May, to an overestimation of 4.9°C in the month of December. In comparison, the bias of the NCEP/NCAR Reanalysis 1

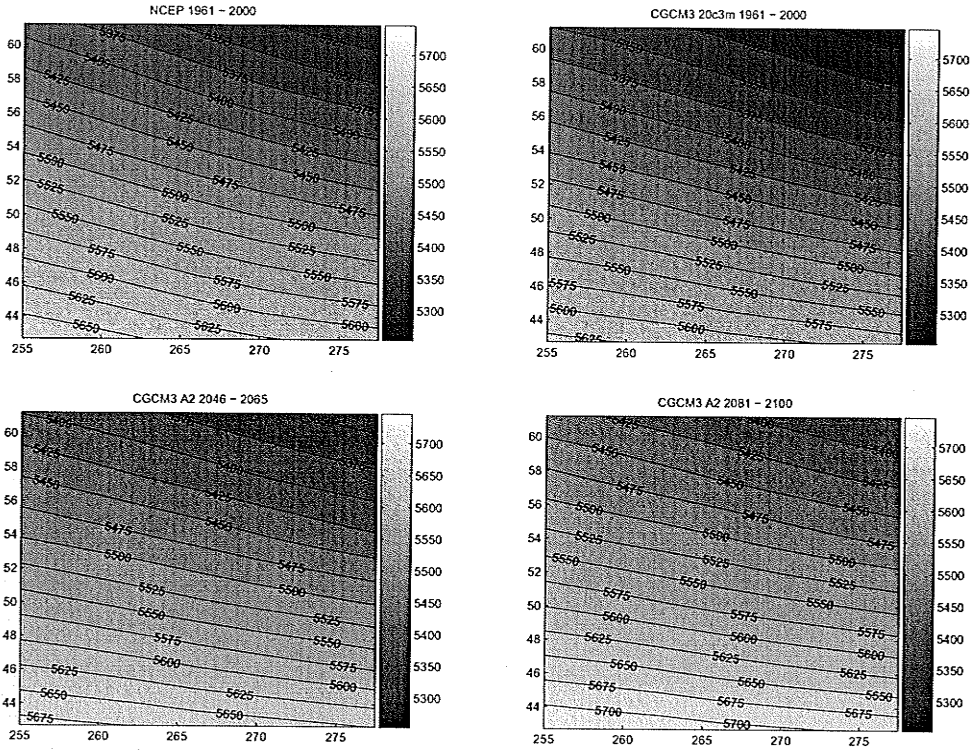


Figure 5.5: Mean values of 500 mb geopotential height

ranged from an underestimation of 1.7°C in the month April, to an overestimation of 2.1°C in the month of January. While biases are present in both data sets compared to the observed monthly means, large biases exist for the CGCM3.1/T47 20c3m output in the winter months of November to January.

To measure the significance of the biases in monthly means, t-tests were administered on the mean temperatures of July and December. NCEP/NCAR Reanalysis 1 monthly means were compared to the observed means, and CGCM3.1/T47 20c3m monthly means were compared to the observed means. Two months were selected to check, July and December, for a total of four t-tests. The null hypothesis $H == 0$, that the NCEP/NCAR Reanalysis 1 or CGCM3.1/T47 20c3m had the same mean values as the observed temperature was rejected for both data sets

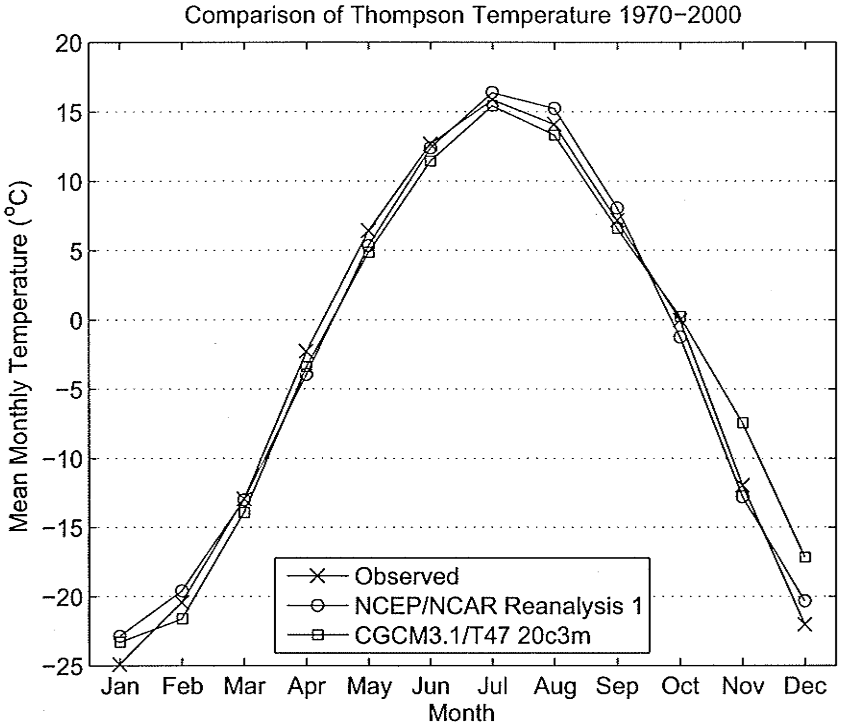


Figure 5.6: Mean monthly Thompson surface temperature comparison.

for July and December at the 1% significance level. This proves that the biases in NCEP/NCAR Reanalysis 1 and CGCM3.1/T47 20c3m output are indeed significant. To complicate the issue, the biases range greatly from month to month between overestimation and underestimation. Applying a global bias correction to the data would correct the annual average, but would actually amplify the bias of some months.

From the above discussion, it is obvious that any bias correction must account for the time-varying bias. When standardizing data, a daily value should be used for means and standard deviations. By using the daily statistics for each of the individual grid points, the resulting standardized data will be bias free in a spatial and temporal dimensions.

After bias correction by standardization, the CGCM3.1/T47 20c3m output should provide accurate representation of the current climate when downscaled. The downscaled results should better represent the monthly mean statistics than the original CGCM3.1/T47 grid points.

5.2 Circulation Pattern Classification

Circulation pattern (CP) classification was used to explore the relationships between large-scale and local climate variables, or more specifically, the ability of circulation patterns derived from geopotential height data to influence the occurrence and quantity of precipitation at a weather station.

Two methodologies of classification were applied, correlation-based classification and eigenvector-based classification, as described in the methodology section. In both applications the 500 mb geopotential height field over the area shown on Figure 5.7 was considered. The precipitation data were from Environment Canada's Thompson weather station. Thirty seven years of precipitation data were available from January 1, 1967 to December 31, 2003. NCEP/NCAR Reanalysis 1 data were retrieved for the same time period.

5.2.1 Correlation Classification

A Matlab function was created to perform the correlation-based map pattern classification methodology described in Section 3.1.4. The Pearson product-moment correlation threshold was set to 0.6, the middle of the range of 0.5 to 0.7 sug-

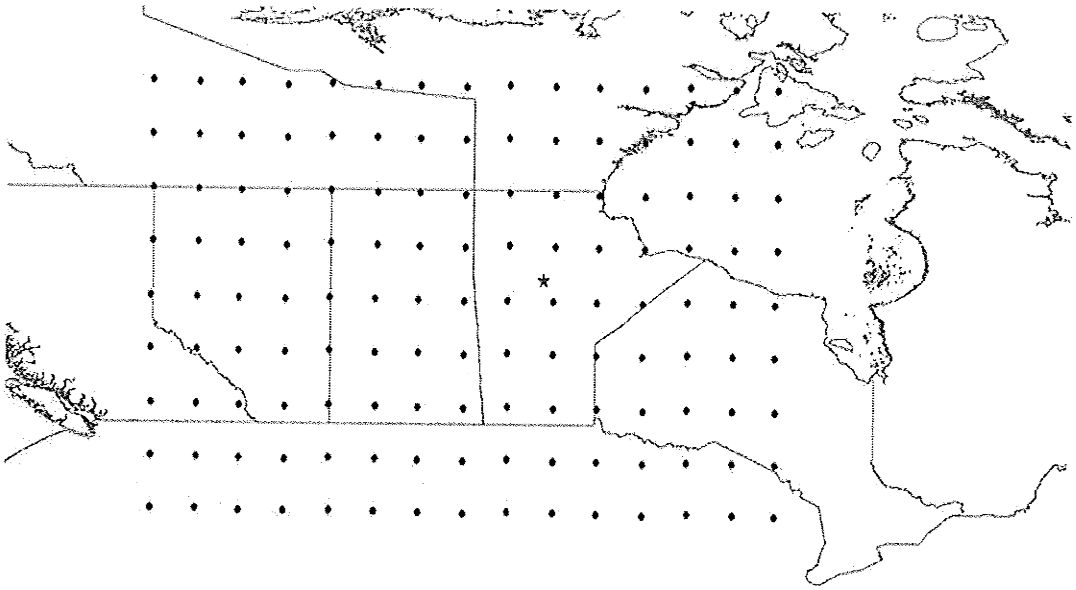


Figure 5.7: NCEP data grid and Thompson weather station.

gested by Yarnal (1993). Trial-and-error exploration led to ten circulation pattern categories.

The computation of the correlation matrix demands large amounts of computer memory. Without precautions for memory management during programming, the computational demand can limit the length of record used to find the key days. To make the classification possible for all 37 years of station data available, sparse matrices were used to store the position of correlations above the threshold value.

The classification algorithm was applied to the 37 years of geopotential height data. The selected key days are shown on Figure 5.8. Each key day shows a common pressure pattern that is significantly different from the other key days. Using these key days, each of the ten categories has a sufficiently large number of members.

The goal of the classification was to divide days into categories based on their

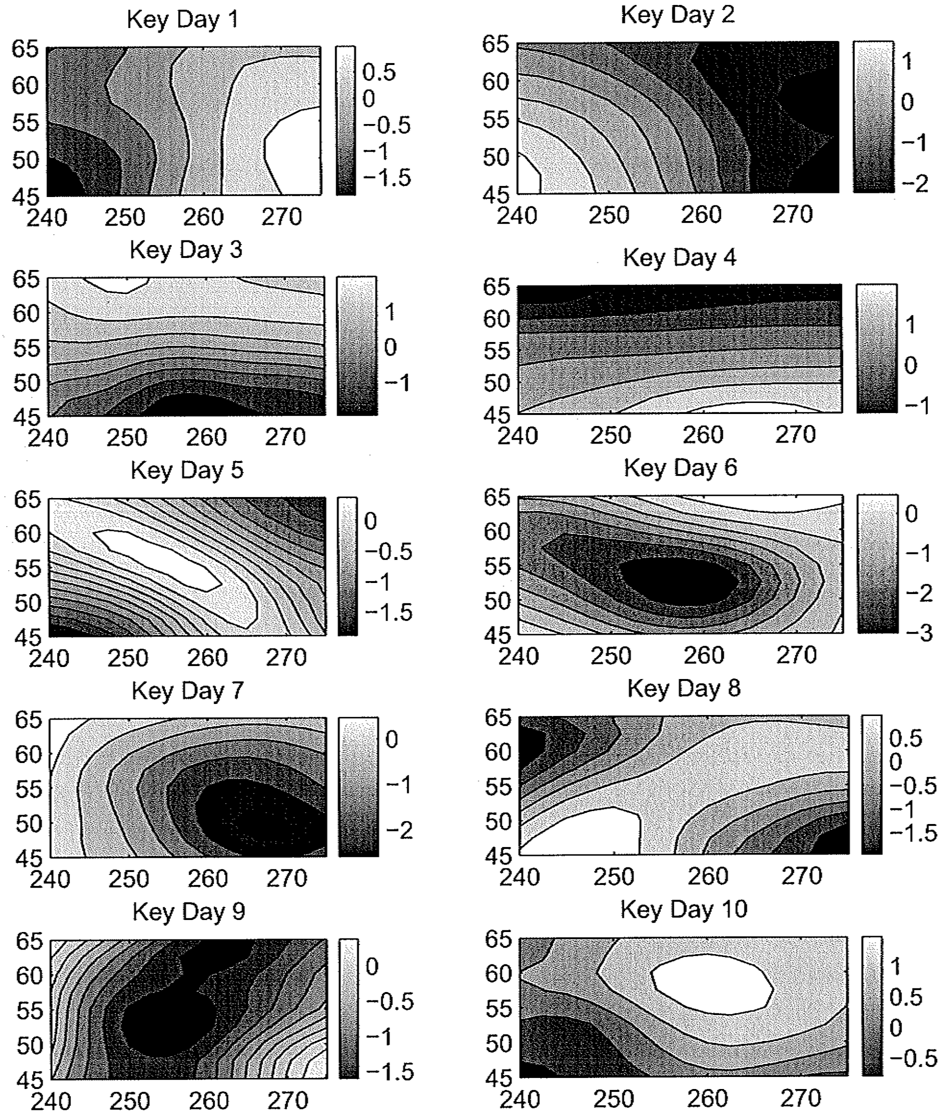


Figure 5.8: Key days of the correlation-based map-pattern classification.

Table 5.1: Correlation-based map-pattern classification statistics.

CP	P(CP)	P(Precip Occurs)	Mean (mm)	Std. (mm)	Skew.
1	0.14	0.44	4.41	6.45	2.89
2	0.09	0.31	2.47	3.82	3.67
3	0.07	0.47	4.37	6.53	3.60
4	0.12	0.27	2.61	4.20	3.30
5	0.10	0.39	2.88	4.83	4.92
6	0.09	0.29	2.00	2.95	3.15
7	0.15	0.48	5.19	7.86	2.93
8	0.09	0.42	2.55	3.50	3.37
9	0.07	0.27	2.51	4.74	5.73
10	0.08	0.42	3.87	6.56	4.26
Total		0.39	3.63	5.93	3.79

similarity to common pressure patterns. The potential usefulness of the division of days into categories is to also find patterns in the local precipitation based on these categories. The most important variable for hydrologic modelling is precipitation. The distribution of precipitation amounts and the probability of precipitation for the days in each category were calculated for the Thompson weather station. Table 5.1 shows the results for each of the CPs, including the probability of the CP occurring, the probability of precipitation, along with the first three moments of the distribution of daily precipitation accumulation.

Table 5.1 shows that each of these key statistics vary among the CP categories. On average, for any given day the probability of precipitation occurring is approximately 0.39. For CP 7 the probability of rain is 0.48, and for CP 9 it is 0.27. This is useful information when attempting to predict the occurrence of rainfall. The circulation patterns in geopotential height influence occurrence of precipitation at a weather station.

Each CP also has a different distribution of daily rainfall accumulation. The distribution statistics show that precipitation characteristics vary greatly between the CPs. For example, the days in the CP 7 category are likelier to have larger accumulations than the days in the CP 2 category.

5.2.2 Eigenvector Classification

A Matlab function was created to perform the eigenvector-based map pattern classification methodology in Section 3.1.5. The number of classification categories was set equal to ten categories, the same number used during the correlation-based classification.

The memory available to complete the clustering step was a factor in the method selection. When classifying the 37 years of data (13,505 days) simultaneously, Matlab (32-bit Windows Xp version) did not have enough memory to allow the variables to be stored during the calculations for the Ward and centroid method algorithms, even with many memory saving precautions taken. Therefore, the average linking method was used. The average linkage method proved to be time and memory efficient.

To view the physical meaning of the clusters, the mean of the grids of standardized geopotential heights in each category were calculated. The results are shown on Figure 5.9. The patterns derived from the eigenvector-based classification do not appear as unique as the key days of the correlation based classification, which is likely due to the fact that the key days are an actual single day of the NCEP/NCAR Reanalysis Data 1 data, while the data on Figure 5.9 are averages of hundreds or a few thousand days. The circulation patterns each show different

Table 5.2: Eigenvector-based map-pattern classification statistics

CP	P(CP)	P(Precip Occurs)	Mean (mm)	Std. (mm)	Skew.
1	0.07	0.38	4.44	6.66	2.77
2	0.01	0.62	7.44	10.20	2.46
3	0.11	0.31	2.68	4.68	4.95
4	0.13	0.34	2.49	3.95	3.53
5	0.30	0.45	3.86	6.11	3.61
6	0.01	0.32	3.81	6.17	2.96
7	0.02	0.24	2.38	4.12	3.08
8	0.03	0.34	2.62	4.00	2.77
9	0.23	0.35	3.72	6.13	4.09
10	0.09	0.39	3.65	5.22	3.07
Total		0.39	3.63	5.93	3.79

areas of high and low pressure systems. Some patterns are cyclonic, others anticyclonic, which results in each circulation pattern having a unique probability of precipitation occurrence and distribution of precipitation quantity.

The distribution of daily precipitation accumulations and the probability of precipitation occurrence for the days in each category were calculated for the Thompson weather station and the results are shown in Table 5.2.

As in the circulation pattern classification, different circulation patterns lead to higher (CP 2) or lower (CP 7) probabilities of precipitation occurrence. The distribution moments also vary between circulation patterns. CP 2, with a 0.01 probability of occurring, has a precipitation distribution with high mean and standard deviation, meaning this pattern may be associated with rare but heavy rainfall events.

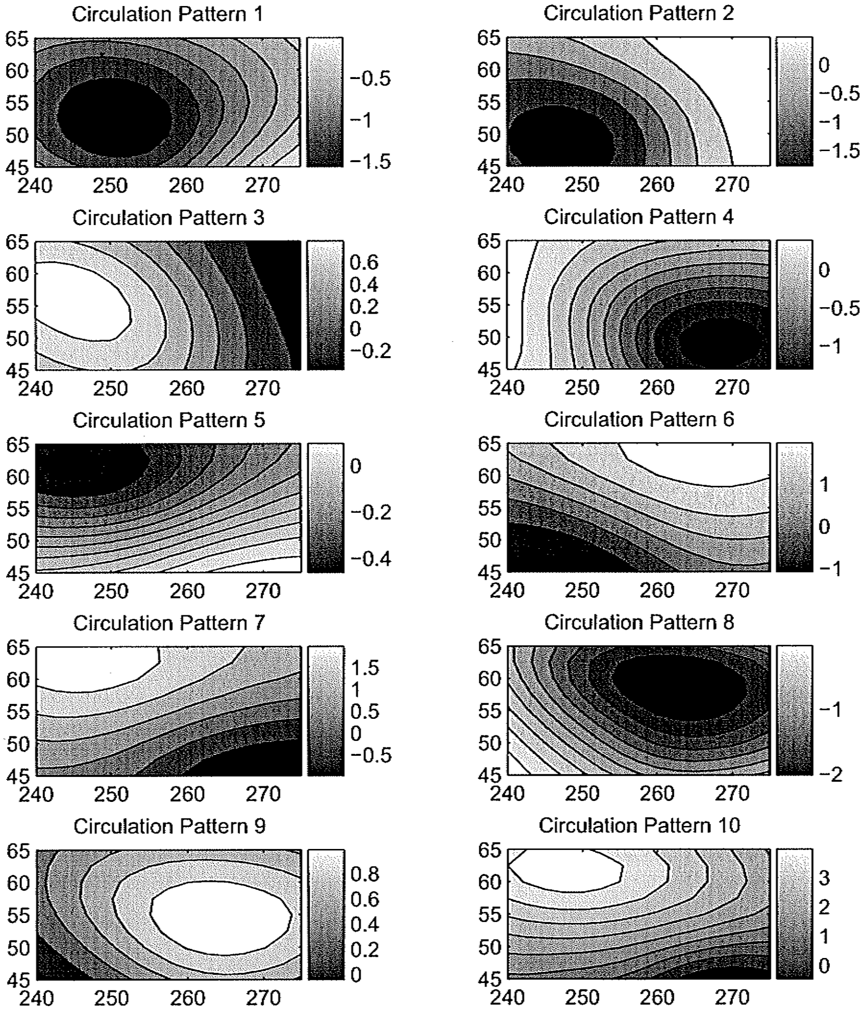


Figure 5.9: Mean values of the eigenvector-based map-pattern classification.

5.2.3 Circulation Pattern Classification Summary

Through circulation pattern classification, it was shown that patterns in the 500 mb geopotential height influence both the occurrence of precipitation and the distribution of precipitation quantities. Therefore, geopotential height fields can provide valuable information when included in the large-scale variables used in downscaling models. As a result of the preceding exercise, grids of 500 mb and 850 mb geopotential heights will be incorporated into the k-nn downscaling model developed in Chapter 6.

As a side note, although it is not in the scope of this project, a downscaling model could be derived based on the circulation pattern classification completed in this section. Section 2.3.2 provides more information on weather typing downscaling models.

5.3 Canonical Correlation Analysis

One of the difficulties of downscaling climate data is finding meaningful relationships between large-scale and local variables. There are many questions regarding the selection of large-scale variables. Canonical correlation analysis was used to evaluate the correlations between large variables and local variables, and more specifically to determine if large-scale temperature and geopotential height data are correlated to local temperature and precipitation observations.

The goal of exploring the large and local-scale variables was to justify the selection of variables by producing canonical variates with high correlations. It was assumed that since temperature usually changes gradually over large distances, the large-scale temperature should be highly correlated with local temperatures. Also, it was assumed that since circulation pattern classification showed that precipitation is related to geopotential height data, geopotential height data should be correlated to local precipitation occurrence.

For large-scale variables, the 850 mb and 500 mb geopotential heights will be used along with temperature at the surface, 850 mb geopotential height and 500 mb geopotential height. NCEP/NCAR Reanalysis 1 data for these variables were extracted for the grid shown on Figure 5.7. NCEP/NCAR Reanalysis 1 data were trimmed to match the temporal range of the station data. To remove seasonal influences in the data, each grid point was standardized using a daily mean and standard deviation smoothed using Fourier series. Since each of the data grids contain many grid points, principal component analysis was used to reduce the number of variables.

To determine the correlations with local temperature, the temperature data at the Thompson and The Pas Environment Canada weather stations were obtained for the years available at both stations, 1970 to 2000. Each station was standardized using a daily mean and standard deviation smoothed using Fourier series.

The results from the CCA with the first 24 principal components from the large-scale data and the daily temperature from the two weather stations is shown in Table 5.3. The weights from the first set of canonical variates for the temperature data, b_m, are approximately equal in magnitude and are both positive. Therefore, these sets of variate represent the magnitude of the temperature being above or below normal for the calender day. The canonical correlation of 0.93 for these two sets of variates demonstrates that the degree to which the stations are cooler or warmer than normal is explained very well. The second set of variates has weights of opposite sign applied to the two weather stations. The second set of variates therefore describes the temperature difference between the two stations. The canonical correlation of 0.56 demonstrates that the principal components also explain a portion of the temperature difference between the two stations.

The second application of CCA was to determine if correlations exist between the large-scale variables and station measurements of precipitation. Since precipitation is a stochastic process and the occurrence at a point is difficult to predict, more weather stations were added to the CCA. A total of six Environment Canada weather stations in Northern Manitoba were used, including Thompson, The Pas, Gillam, Grand Rapids Island Lake, Lynn Lake, and Norway House. The precipitation data were set to 1 or 0: days with precipitation greater than 0.2 mm were coded as 1, and days with less than 0.2 mm were codes as to 0. With six weather stations the results, shown in Table 5.4, can be difficult to interpret physically, however the first set of canonical variates provide some useful information. With all negative weights assigned to the weather stations, the first set of variates describes the occurrence of precipitation at all stations. Given the stochastic nature of rainfall, the canonical correlation of 0.58 demonstrates that the large-scale variables describe a large proportion of the local rainfall occurrence.

Table 5.3: Canonical correlation temperature analysis results.

	1	2
a_m	0.0853	-0.0008
	0.0453	-0.0095
	0.0127	-0.1131
	-0.0106	-0.0440
	0.0344	0.0671
	-0.0249	-0.0747
	-0.0758	-0.0394
	-0.0252	0.0091
	-0.0005	-0.0862
	0.1008	-0.0414
	0.0358	-0.0846
	0.0071	-0.0060
	-0.0349	-0.1263
	-0.0209	-0.1023
	0.0247	-0.0650
	0.0171	-0.0286
	-0.0355	0.0220
	-0.0625	0.0186
	-0.1038	-0.1538
	0.0180	0.1814
	-0.0186	0.1614
	0.0382	-0.0125
	-0.0664	0.0822
	0.0513	-0.1333
	0.4125	-0.2342
b_m	0.5701	-1.8595
	0.4542	1.8938
r_{Cm}	0.9269	0.5591

Table 5.4: Canonical correlation precipitation analysis results.

	1	2	3	4	5	6	7
a_m	0.0168	0.0035	0.0413	-0.0067	0.0087	0.0310	0.0284
	-0.0682	-0.0269	0.0414	-0.0171	-0.0367	0.0040	0.0100
	0.0094	0.0284	0.0483	0.0437	-0.0128	-0.0141	-0.0496
	-0.0162	-0.0124	0.0102	-0.0411	0.0173	-0.0306	0.0159
	-0.0675	0.1689	-0.0522	0.0430	-0.0115	-0.0214	-0.0267
	-0.0219	0.0327	-0.0971	0.0061	0.0108	0.1132	0.0178
	-0.0884	0.0017	-0.1292	-0.0236	-0.0692	-0.1346	0.0858
	0.0830	0.0166	-0.0623	0.0489	0.0516	-0.0100	0.0653
	-0.0729	-0.0642	-0.0240	0.1687	0.0730	0.0650	0.0645
	-0.0149	0.0127	-0.0120	-0.1638	0.1923	-0.0886	-0.0720
	-0.1569	-0.0571	0.1325	0.1287	0.0339	-0.1182	-0.0413
	0.0731	-0.1296	-0.0726	-0.0306	-0.0399	-0.1065	0.1009
	-0.0993	-0.1115	0.0481	0.0520	0.1912	0.0059	-0.0803
	0.2580	0.0865	0.0647	-0.0465	-0.1060	-0.1103	-0.0282
	0.0706	0.1160	0.0719	0.0039	-0.1799	-0.1237	-0.0661
	0.1679	-0.0855	-0.0504	0.2452	-0.0414	-0.0573	-0.1903
	0.0378	-0.2501	-0.1742	0.1208	0.0641	0.0631	-0.1787
	0.0371	0.1149	-0.1662	0.1125	0.1377	0.0984	0.0484
	-0.1039	-0.0784	-0.1096	-0.1307	0.0624	-0.0132	-0.3415
	-0.1639	0.2596	0.0690	-0.1047	0.0855	0.1176	-0.2020
	0.0559	-0.0466	-0.0556	-0.3544	-0.2624	0.1577	-0.1657
	0.1035	0.1191	0.0942	-0.0744	0.4368	-0.3555	0.0360
	-0.0026	-0.0482	-0.1089	-0.2435	0.1152	0.1894	-0.2916
	-0.0529	0.2538	0.1155	0.0046	0.1276	0.2022	0.2096
b_m	-0.3947	0.0712	0.3234	0.1975	1.5566	1.1917	1.7612
	-0.6351	1.1588	0.4175	-1.7662	-0.2694	-0.6398	-0.5364
	-0.2610	-0.8271	-0.6697	-0.8963	-1.4780	-0.2109	1.5834
	-0.6867	0.8045	-1.5168	1.3325	-0.5419	0.0234	-0.3330
	-0.5445	-0.3749	1.4551	1.2211	-1.0922	-0.7980	-0.1591
	-0.0418	-0.3953	0.4585	-0.3993	-0.4290	2.0457	-1.5434
	-0.5287	-1.0613	-0.4739	-0.1575	1.6316	-1.2331	-0.9464
r_{Cm}	0.5775	0.3389	0.2613	0.0971	0.0776	0.0686	0.0553

Chapter 6

k-Nearest Neighbor Resampling Results

6.1 Single-Site Application

The first application of the nearest neighbor downscaling model was designed for a weather station at Thompson, Manitoba. This application serves as a simplified pilot study of the k-nn model. The lessons learned from this application will be important for the next steps of applying the k-nn to multiple sites in the Nelson River basin and the Winnipeg River basin.

6.1.1 Data

In the application presented here, GCM data was downscaled to produce time series of minimum daily temperature, maximum daily temperature, and daily accumulated precipitation for the weather station at Thompson, Manitoba, Canada

(55°48'N, 97°51'W). Thirty seven years (1967-2003) of data were available for the Thompson weather station. Therefore the historical record from which to resample was limited to this time period.

The NCEP/NCAR Reanalysis 1 supplied the historical atmospheric data and the Canadian Daily Climate Data (CDCD) data set provided the historical station data for the Thompson weather station. The CCCma coupled GCM, CGCM3.1/T47, was selected to provide the simulation data for a 20th century control run (20c3m) and the IPCC SRES A2 climate change scenario. Five runs of 40 years (1961-2000) were available from the 20c3m experiment, and three runs of 20 years (2081-2100) were available for the A2 model run.

To adequately capture the large-scale circulation patterns, a large spatial area was selected over western Canada. The average surface temperature, 500mb temperature, 850mb temperature, 500mb geopotential height, and 850mb geopotential height variables were used as the large-scale variables. The grids for the NCEP/NCAR Reanalysis 1 and CCCma data sets have slightly different resolutions, 2.5° × 2.5° and 3.75° × 3.75° , respectively. To make the data sets consistent, the NCEP/NCAR Reanalysis 1 data was linearly interpolated onto the CCCma grid points. The data cover the region on Figure 6.1 and consist of 60 data points. Canonical correlation analysis and circulation pattern classification were used in Chapter 5 to establish that relationships exist between the large-scale variables and the Thompson weather station data.

NCEP/NCAR Reanalysis 1 data and the 20c3m experiment data were standardized using the mean and standard deviation from each data set to remove

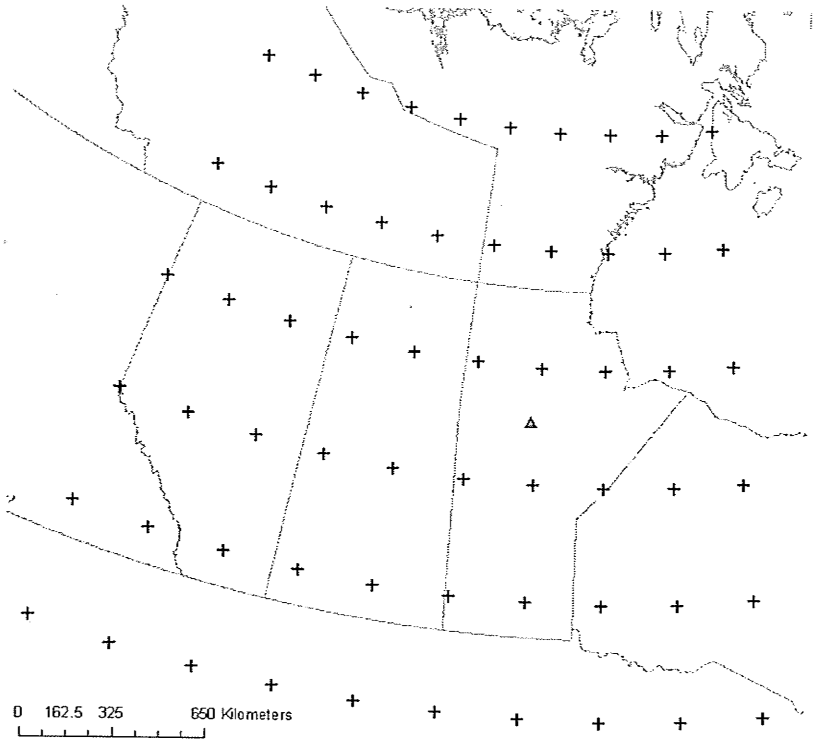


Figure 6.1: CGCM3.1/T47 data grid and Thompson weather station.

slight biases between the data sets for the current climate. The A2 scenario data were standardized using the means and standard deviations from the 20c3m data to preserve the biases created in the model due to changed atmospheric loadings.

With each of the five climate variable grids containing 60 data points, the number total number of data to compare between NCEP/NCAR Reanalysis data and GCM data totaled 300. Since a high degree of correlation exists spatially in each variable, and also between variables, principal component analysis was used to reduce the number of variables in the feature vector by removing redundant information (Gangopadhyay et al., 2005; Buishand and Brandsma, 2001; Young, 1994). The first 24 principal components were retained and explain over 96% of the variation contained in the original data sets. The eigenvectors calculated from the

NCEP/NCAR Reanalysis data were used to calculate the principal components for the GCM data sets to maintain the same modes of variation explained by the principal components of the GCM data as the NCEP/NCAR Reanalysis principal components.

6.1.2 Model set-up and optimization

Although the model does not require parameterization of specific relationships between large-scale and local variables, some components of the model should be adjusted to optimize the ability of the model to estimate the station data. The number of neighbors to retain, k, the window size, W, and the weights w_i can all be adjusted to improve model performance.

To optimize the model, a cross-validation method was set up in which the model was used to simulate the historical station data. The NCEP/NCAR Reanalysis 1 data for one year was considered as simulation data and removed from the historical record. Station data were then generated for this year of NCEP/NCAR Reanalysis 1 data. This process was repeated for each of the 37 years of data.

An objective must be specified to optimize the model. Since estimation of both temperature and precipitation are important, the objective considered here was to optimize the correlation between the estimated and observed daily temperature anomaly, and the correlation of estimated and observed accumulated winter precipitation (October to April) at the annual scale. Precipitation was limited to the winter season because the heavy convective rainfall that occurs in the summer was found to be difficult to predict as it is a local phenomenon rather than driven by synoptic scale climate. Furthermore, yearly runoff in the Nelson River and most other Canadian river systems is largely dominated by spring melt water from accumulated winter precipitation. The objective function for this application is

$$\max f(W, k, w_i) = |\rho_u| + |\rho_v|, \tag{6.1}$$

where u refers to maximum temperature anomaly and v refers to winter precipitation accumulation.

The model was initialized using equal weighting to each principal component and k set to retain only the most similar nearest neighbor. The window width, W, was optimized using trial and error. From Figure 6.2 it was observed that a window size of 21 days leads to the maximum model performance. The weighting vector, w, was optimized using the Matlab Optimization Toolbox. The software employed a gradient line search optimization methodology to optimize the objective function.

The number of nearest neighbors to retain was varied and the best results during the cross-validation were obtained when only the first nearest neighbor was retained. However, to encourage variability in the selection process when simulating with GCM data, k was set to retain the ten nearest neighbors. Increasing the number of nearest neighbors only slightly affected the validation results.

The model was able to reproduce the time series of minimum and maximum temperature quite well, with correlations of 0.93 and 0.95 respectively for the scatter plots on Figure 6.3. The stochastic nature of rainfall occurrence made daily prediction of precipitation difficult, especially for convective storms in sum-

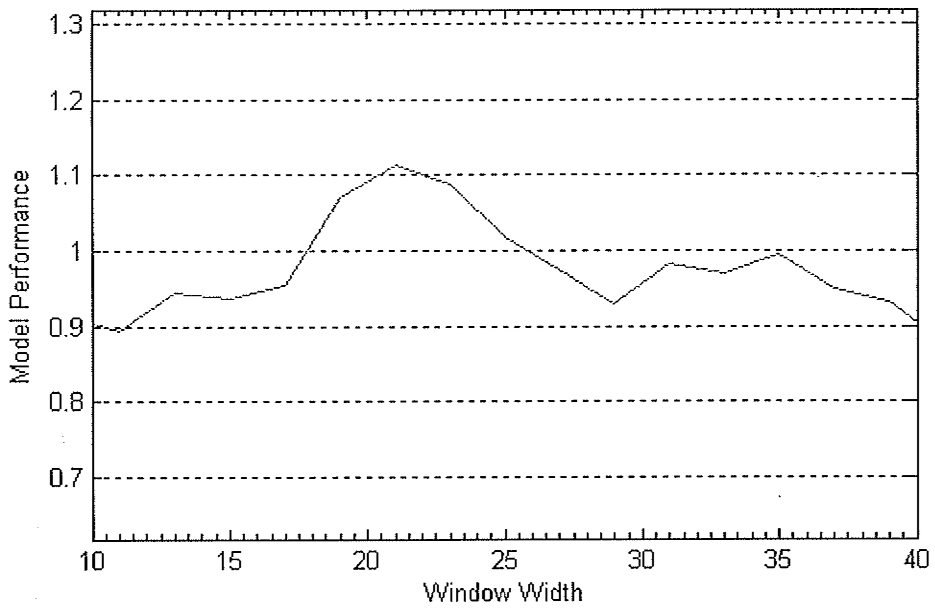


Figure 6.2: Objective function optimization for W.

mer months. The model was able to adequately capture season trends in winter accumulation with a correlation of 0.65 for the data shown on Figure 6.4. The ability of the model to simulate the winter precipitation storage is important, as the spring melt is the most important feature of the annual hydrograph. In some years, the model significantly over-estimated or under-estimated the accumulation of precipitation, such as in 1981, 1986, and 1997. In most years the model is able to capture the seasonal trends in precipitation quite well. Table 6.1 shows the model produced similar mean amounts of winter and annual precipitation and also contained similar amounts of variation. The year 1979 is omitted due to numerous missing data entries in the observed record.

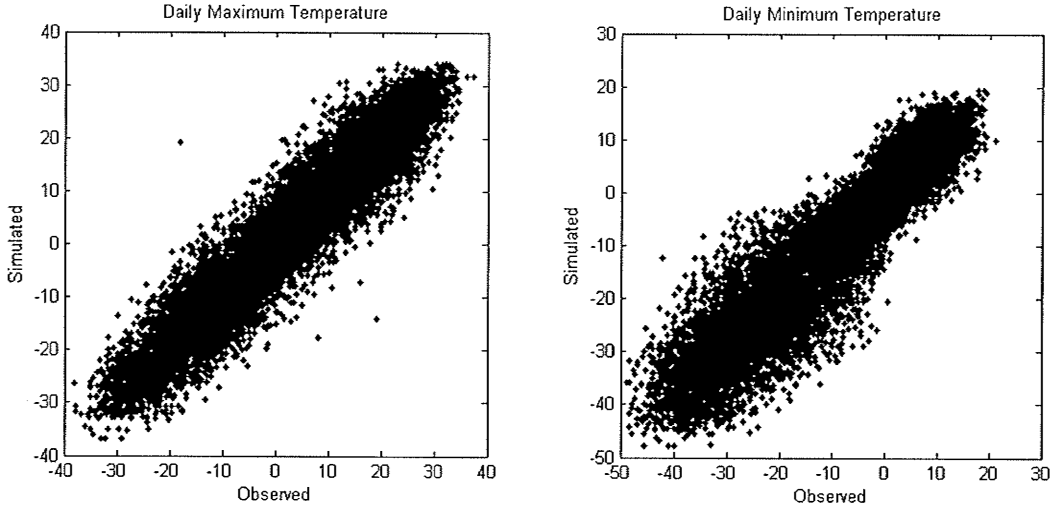


Figure 6.3: Validation scatter plots of simulated and observed daily minimum and maximum temperature at Thompson.

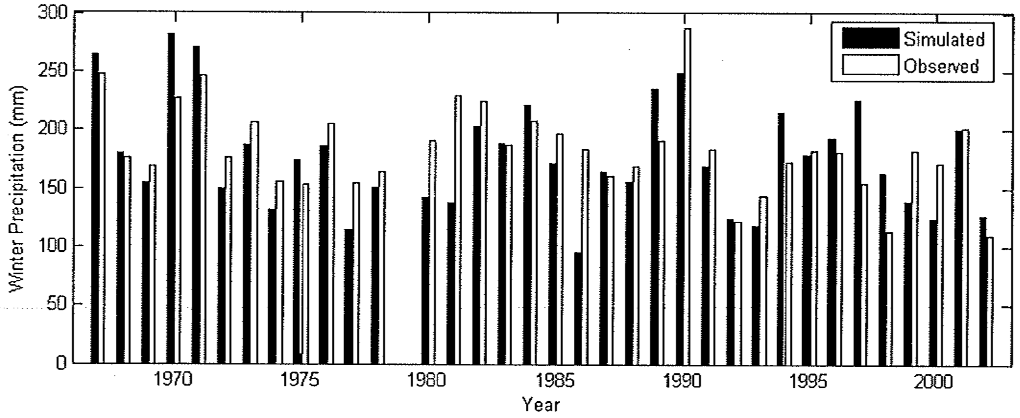


Figure 6.4: Annual accumulation of winter precipitation at Thompson weather station.

Table 6.1: Cross-validation precipitation statistics.

	Observed (mm)	Simulated (mm)
Winter mean	185	176
Winter standard deviation	38	46
Annual mean	512	504
Annual standard deviation	85	89

6.1.3 Model Application

The model was employed with data generated by the CGCM3.1/T47 to evaluate changes in temperature and precipitation.

Monthly averages of maximum temperature and accumulated precipitation are shown on Figure 6.5. As expected, the downscaled 20c3m (1961-2000) GCM runs produced mean temperatures very close to the observed data. The SRES A2 scenario (2081-2100) produced downscaled temperatures 3 to 5°C warmer in summer and 5 to 8°C warmer in the winter. The increase in temperature experienced in the spring and fall seasons will shorten the winter season and reduce the length of time precipitation is able to be stored as snow.

The 20c3m experiment led to downscaled precipitation results that slightly under-estimate monthly precipitation accumulation. The A2 scenario led to precipitation that was similar to the observed and 20c3m data, except for the months of June and July that had significantly less precipitation.

The statistics for winter and annual precipitation in Table 6.2 show that both GCM simulations provide atmospheric conditions that lead to reduced precipitation at the weather station. The 20c3m experiment underestimates winter precipitation by 20% and annual precipitation by nearly 10%. The A2 scenario leads to a

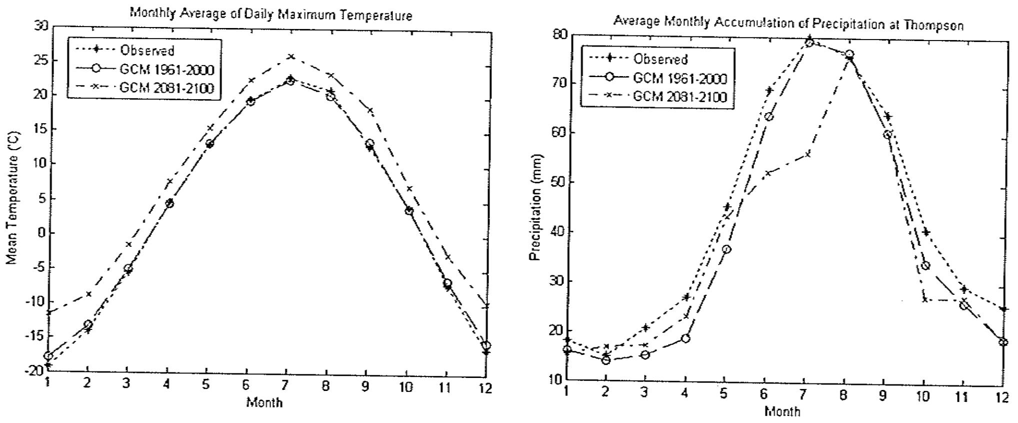


Figure 6.5: Monthly averages of maximum temperature and precipitation accumulation for observed and GCM simulated scenarios.

Table 6.2: Observed and downscaled Thompson precipitation.

	Observed (mm)	20c3m (mm)	A2 (1981-2100) (mm)
Winter mean	185	148	144
Winter standard deviation	38	34	27
Annual mean	512	475	430
Annual standard deviation	85	99	72

decrease of 22% in the winter season and an annual decrease of 16%. Although the 20c3m experiment led to an underestimation of precipitation, the further reduction in precipitation in the A2 scenario simulation shows that future precipitation may decrease at the Thompson weather station.

The most significant decrease in future precipitation occurs in the month of July (Figure 6.5). A more detailed investigation was made into the precipitation of July and August. Frequency distributions for these two months are shown on Figure 6.6. The observed and 20c3m (1961-2000) simulations experience similar frequency trends in both months. For the month of July, the frequency distribution for the downscaled SRES A2 (2081-2100) simulation shows an increase in the frequency of dry days and a reduction in frequency for all rainfall events, causing a significant

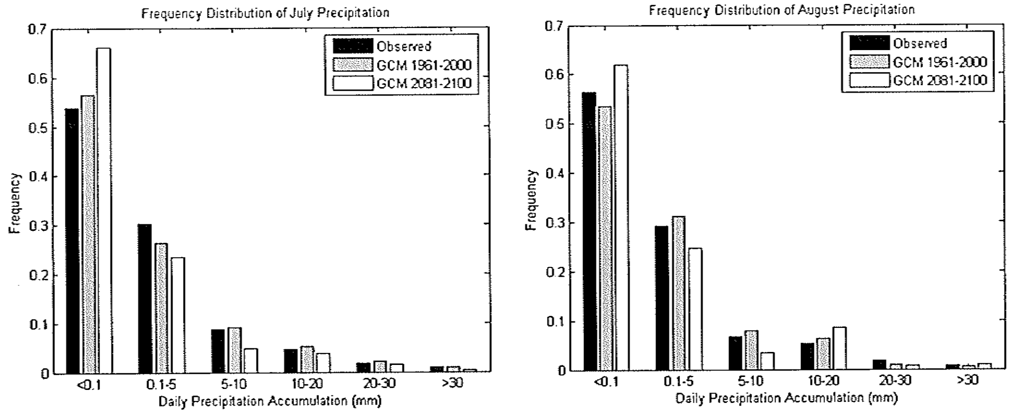


Figure 6.6: Monthly frequency distribution of daily precipitation accumulation.

decrease in mean accumulation for July. A less exaggerated increase in dry days occurs in August, where there is a decrease of small events less than 10 mm, but an increase in events with more than 10 mm. The increase in larger events offsets the increase in dry days and maintains the monthly mean accumulation at its current level. Cool temperatures are usually experienced during rainy days; therefore the decrease in small events may be attributed to warmer future temperatures. Since convective rainfall requires warmer temperatures, the frequency of large events in the future may not decrease to the same extent as small events.

The combination of increased temperature, shortened winter season, and reduced precipitation will certainly lead to changes in streamflow.

6.1.4 Discussion

A k-nearest neighbor resampling algorithm was developed to generate data at the Thompson weather station by downscaling large-scale atmospheric data. Optimization of some model parameters was necessary to improve the model performance. Atmospheric data generated by the CGCM3.1/T47 were used as input to generate weather data for climate change scenarios. The downscaled A2 scenario (2081-2100) resulted in a future climate at Thompson that is expected to be warmer throughout the year and slightly drier.

The next application of the k-nn model will be to downscale GCM data to produce weather variables at both the Thompson weather station and an addition weather station at The Pas, Manitoba. The lessons learned from this pilot application that will improve the methodology in the next application are:

- Large-scale grids of temperature and geopotential height atmospheric variables can be downscaled to adequately reproduce historical daily temperature and seasonal precipitation trends.
- Optimization of the model parameters k, w_i, and W can be used to improve the ability of the model to reproduce historical climate.
- A k-parameter equal to one resulted in the best model performance. However, it is recommended to use a larger k, $k = 10$ for example, to increase model variability while generating downscaled GCM data.
- Historical temperature is reproduced well without significant optimization. Future optimization should focus on precipitation.

The work presented for this single site application of k-nn was formulated into a conference paper presenting the optimization methodology for the model and also displaying the ability of the k-nn model to downscale GCM data (Lee and Rasmussen, 2007).

6.2 Nelson River Multi-site Application

In this application, the k-nn model will be used to generate variables for hydrological modelling in the Nelson River Drainage Basin. The hydrological model that will use the downscaled data is the SLURP hydrological model. The SLURP model requires the following variables at the daily time scale:

- Mean temperature,
- Depth of precipitation,
- Relative humidity, and
- Solar radiation or bright sunshine hours.

6.2.1 Data

The same large-scale input variables that were used in the single-site application will be used for this application. The input variables are the average surface temperature, 500mb temperature, 850mb temperature, 500mb geopotential height, and 850mb geopotential height over the grid shown on Figure 6.1.

The historical record to sample from was limited to 31 years, from 1970 to 2000. These were the years where all variables were available for both of the Thompson (55°48'N, 97°51'W) and The Pas (53°58'N, 100°6'W) weather stations. In total, eight variables will be downscaled simultaneously in this application, daily mean temperature, precipitation, relative humidity and bright sunshine hours.

The NCEP/NCAR Reanalysis 1 data for the large-scale variables were retrieved and prepared for downscaling by standardizing the data and applying principal component analysis as in the single-site application.

For CGCM3.1/T47 GCM data, five GCM model runs of the 20c3m scenario for the period of 1961-2000 are available from the Canadian Centre for Climate Modelling and Analysis (CCCma) website for downloading, but only three model runs of the A2 scenario were available. Two additional GCM runs for future emission scenarios were made available by the CCCma since the single-site application was completed. In this application, the additional GCM scenarios of SRES B1 and SRES A1B were downscaled for the time slices of 2046 to 2065 and from 2081 to 2100. It total, five scenario runs from the CGCM3.1/T47 were available. Therefore, five runs of forty years each were available for the 20c3m scenario, for a total of 200 years of data, and five runs of twenty years were available for a total of 100 years of data for each of the time slices for each future scenario.

The GCM data were standardized and transformed into principal components. As in the previous application, the future scenarios were standardized using the means and standard deviations from the 20c3m model runs. This preserves the bias or trends between the model simulated current climate (20c3m), and the climate under the different emission scenarios (SRES A2, A1B, and B1). To be consistent between data sets, the transformation of the GCM data into principal components was done using the eigenvectors of the NCEP/NCAR Reanalysis data.

6.2.2 Model Setup and Optimization

The model parameters consisting of the window width, W, the number of nearest neighbors to retain, k, and the weighting vector in the distance calculation, w_i, were optimized using the cross-validation methodology developed in the single-site application of the k-nn model.

In the single-site application, the objective function was specified as the correlation between the estimated and observed daily temperature anomaly, and the correlation of estimated and observed accumulated winter precipitation. In this application, more weather station variables are available to use in an objective function. It was found in the single-site optimization that by using temperature data as a significant part of the large-scale climate variable input, good results for downscaling historical temperature were achieved relatively easily. Precipitation is a key input variable for hydrological modelling and more difficult to accurately downscale than temperature. The SLURP model is not as sensitive to changes in relative humidity or bright sunshine hours input variables as it is to changes in precipitation. For these reasons the objective function in this application was focused on optimizing the downscaling of precipitation.

The objective function was specified as the average over the two stations' mean root mean square error (RMSE) of the estimated winter precipitation,

$$\min f(W, k, w_i) = \frac{\mathrm{RMSE}_{\text{The Pas Winter Precip}} + \mathrm{RMSE}_{\text{Thompson Winter Precip}}}{2}. \qquad (6.2)$$

A cross-validation was set up where the model would take one year of the NCEP/NCAR Reanalysis data as simulation data and use the other years as a historical record to resample from. This process was repeated for all 31 years of the historical record. The objective function was then calculated using the simulated station variables and the historical station variables.

The window width was optimized manually by adjusting W. From Figure 6.7 it can be seen that the optimum window width was 25 days. The larger window width compared to the optimum window width of 21 days for the single site application is likely due to the shorter historical record available for the multi-site application. The historical record was 37 years for the single-site, but only 31 years for the multi-site application.

The Matlab Optimization Toolbox was used to optimize the weight vector, w. With the large number of variables in the feature vector, the optimization of the w_i vector was a time consuming process, even with the use of software.

The objective function improved from 47.0 to 34.7 by adjusting the w vector, a decrease of 26%. After optimization, the model was able to capture trends in the winter precipitation at the weather stations. Figure 6.8 shows the optimization results for the Thompson weather station. RMSE's of 39.9 mm and 30.6 mm, and correlations of 0.41 and 0.68, were achieved for the simulation of winter precipitation accumulations for The Pas and Thompson.

The simulated temperature data was checked to ensure the model was able to adequately estimate the historical temperature data, despite temperature not being a criterion in the objective function. The correlation of observed and simulated

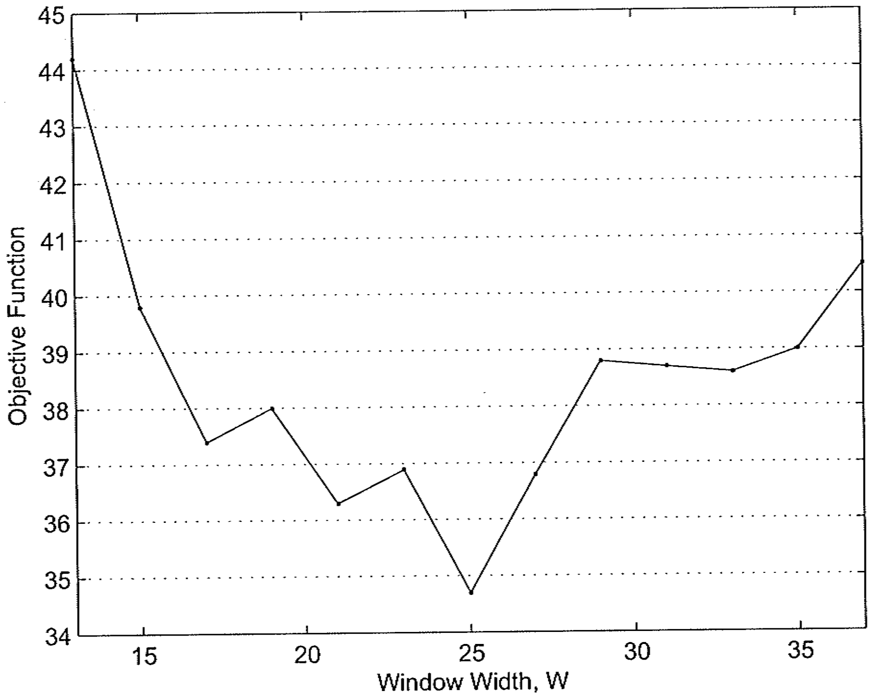


Figure 6.7: Optimization of W for Nelson River

daily temperature was 0.95 for both The Pas and Thompson weather stations. A plot of the estimated and observed temperatures at the Thompson weather station is shown on Figure 6.9. The temperature data were adequately simulated.

Relative humidity and solar radiation, the other variables simulated, also showed adequate correlation to the observed variables. At Thompson, the simulated relative humidity had a 0.54 correlation with the observed relative humidity, and the simulated bright sunshine hours had a 0.51 correlation with the observed bright sunshine hours. At The Pas the simulated relative humidity had a 0.47 correlation with the observed relative humidity, and the simulated bright sunshine hours had a 0.43 correlation with the observed bright sunshine hours. The correlations are much lower than the temperature correlations, however this may not be a criti-

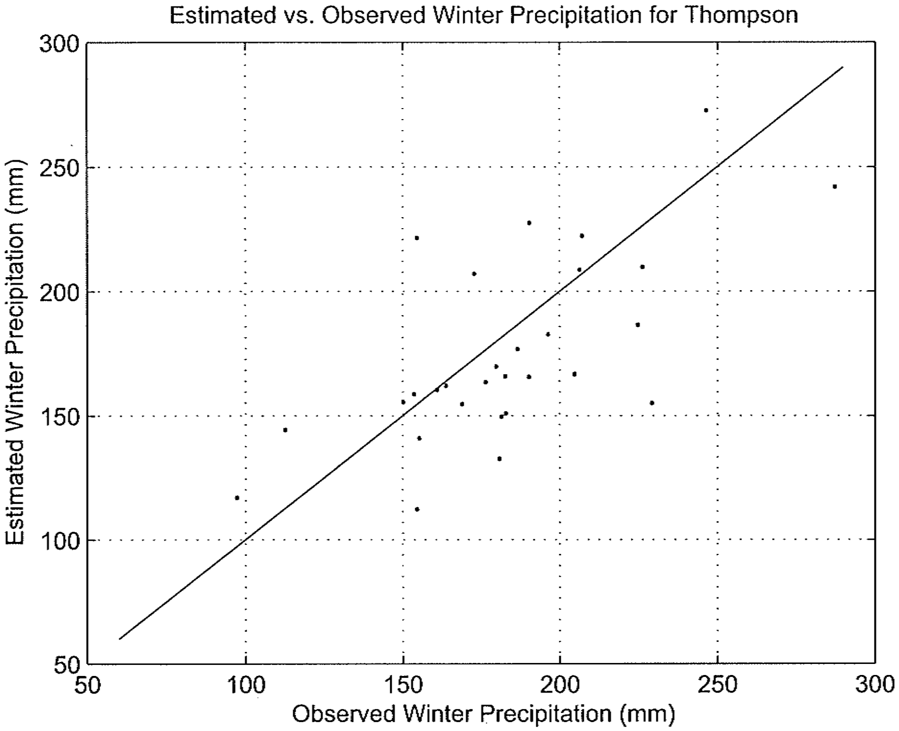


Figure 6.8: Thompson winter precipitation cross-validation results.

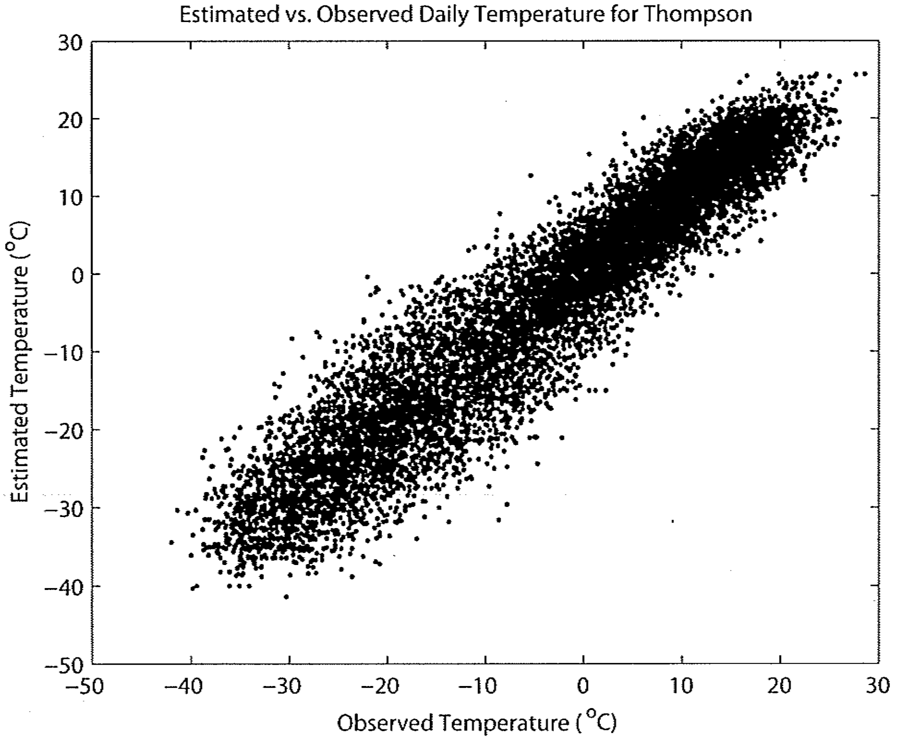


Figure 6.9: Thompson daily temperature cross-validation results.

cal concern since the SLURP hydrological model is not overly sensitive to these variables. Therefore these correlations are considered adequate.

6.2.3 Downscaling Results

20c3m Control Run

With the k-nn model simulating the four input variables for Thompson and The Pas with acceptable level of performance, it was then used to generate downscaled GCM data by using large-scale variables from the CGCM3.1/T47 GCM data as input to the model.

The first scenario downscaled by the k-nn model was the 20c3m scenario. The GCM runs for this scenario represent the atmospheric composition for the period from 1961 to 2000. The downscaled variables from this scenario should lead to monthly mean averages that are similar to the observed data, and also to the simulated station variables downscaled from the historical NCEP/NCAR Reanalysis 1 data.

Figure 6.10 shows that the mean monthly temperatures of the station data simulated from the downscaled NCEP/NCAR Reanalysis 1 and 20c3m are an excellent fit to the observed monthly mean temperatures. Figure 6.11, Table 6.5 (page 155), and Table 6.6 (page 155) show that the mean monthly precipitation results of downscaled NECP/NCAR Reanalysis 1 and 20c3m are similar to the single-site application. The annual distribution of rainfall is reproduced well when precipitation data is downscaled from the 20c3m scenario. The downscaled results from NCEP/NCAR Reanalysis 1 data showed similar precipitation volumes as

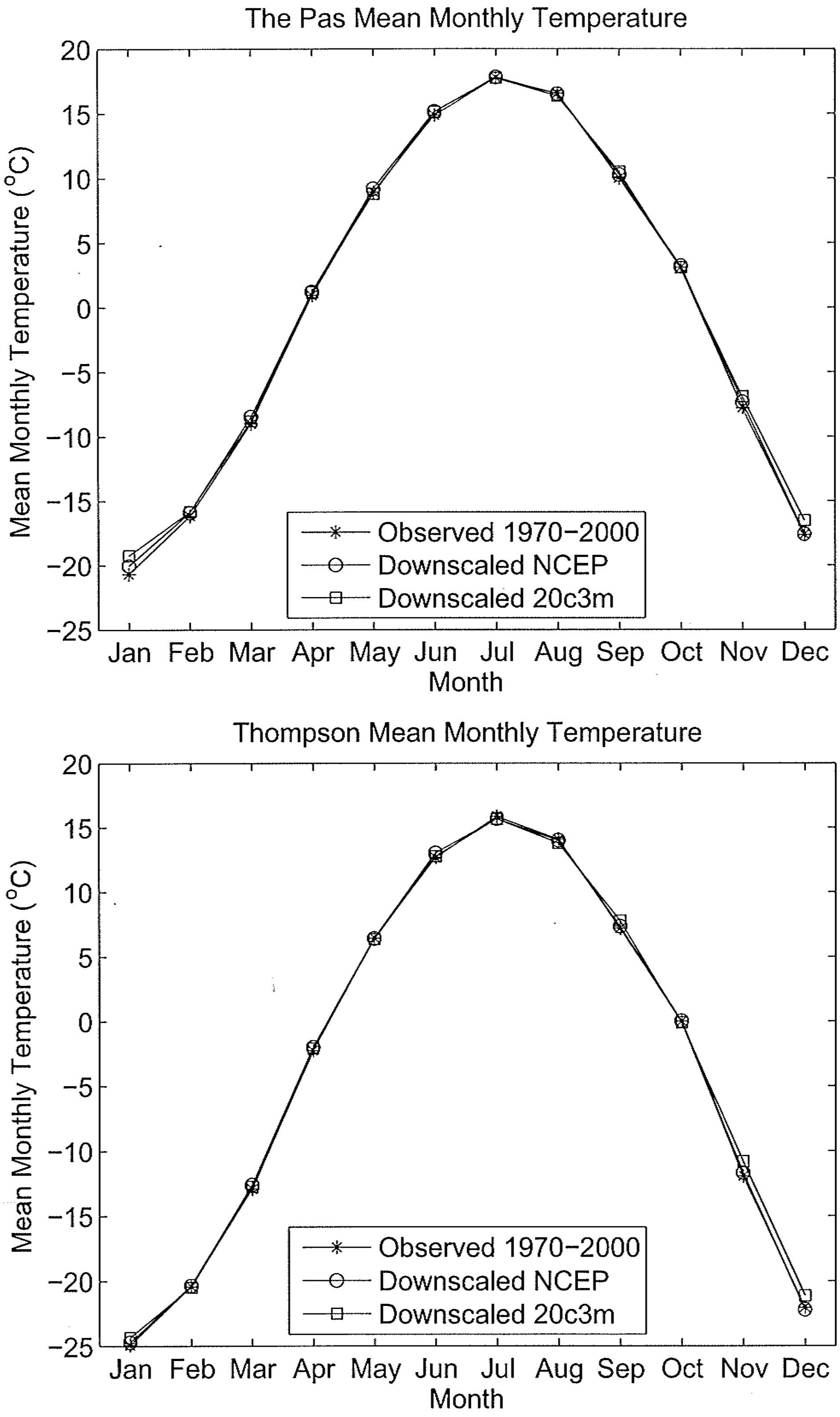


Figure 6.10: Optimization results for mean monthly temperatures.

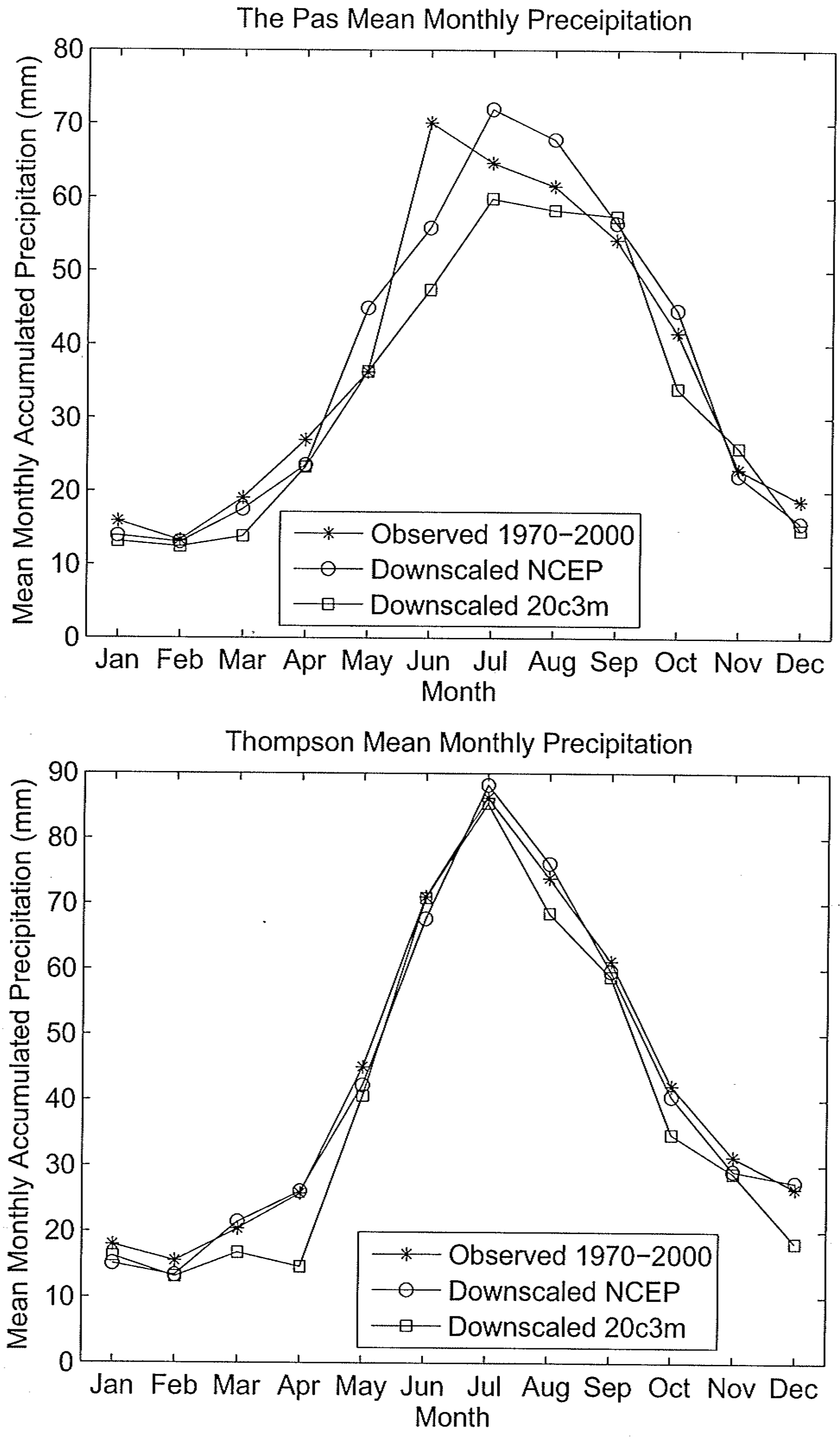


Figure 6.11: Optimization results for mean monthly precipitation.

the observed record, an improvement compared to the single-site application. On average, the annual precipitation accumulation for the 20c3m scenario compared to the observed record was underestimated by 10%, and winter precipitation was underestimated by 17%. With the complexity of the rainfall processes, GCMs, and downscaling models, it is difficult to ascertain the origin of the underestimation of precipitation. Although the accumulations from the 20c3m scenario do not match the observed record, valuable information can be derived using the 20c3m scenario as the baseline to which the results of future scenarios can be compared. The bias between the future scenarios compared to the 20c3m scenario, rather than the observed record, will provide more useful trend information.

The *k*-nn model was then used to downscale the CGCM3.1/T47 data for the A2 (2046-2065), A2 (2081-2100), A1B (2081-2100), A1B (2081-2100), B1 (2081-2100), and B1 (2081-2100) scenarios.

Future Temperature

The results for downscaled temperature can be seen for The Pas on Figure 6.12, and Table 6.3 (page 154), and for Thompson on Figure 6.13 and Table 6.4 (page 154).

For the 2046-2065 time slice, all future scenarios show year-round increases in mean monthly temperature at both sites. Winter increases range from 1.7°C in the B1 scenario to 3.6°C in the A2 scenario. The A2 and A1B scenarios show more warming than the B1 scenario. Summer temperatures also increase in all scenarios, however the increases are not as large as for the winter months. The increases ranged from less than 0.5°C in the B1 scenario to approximately 1°C in

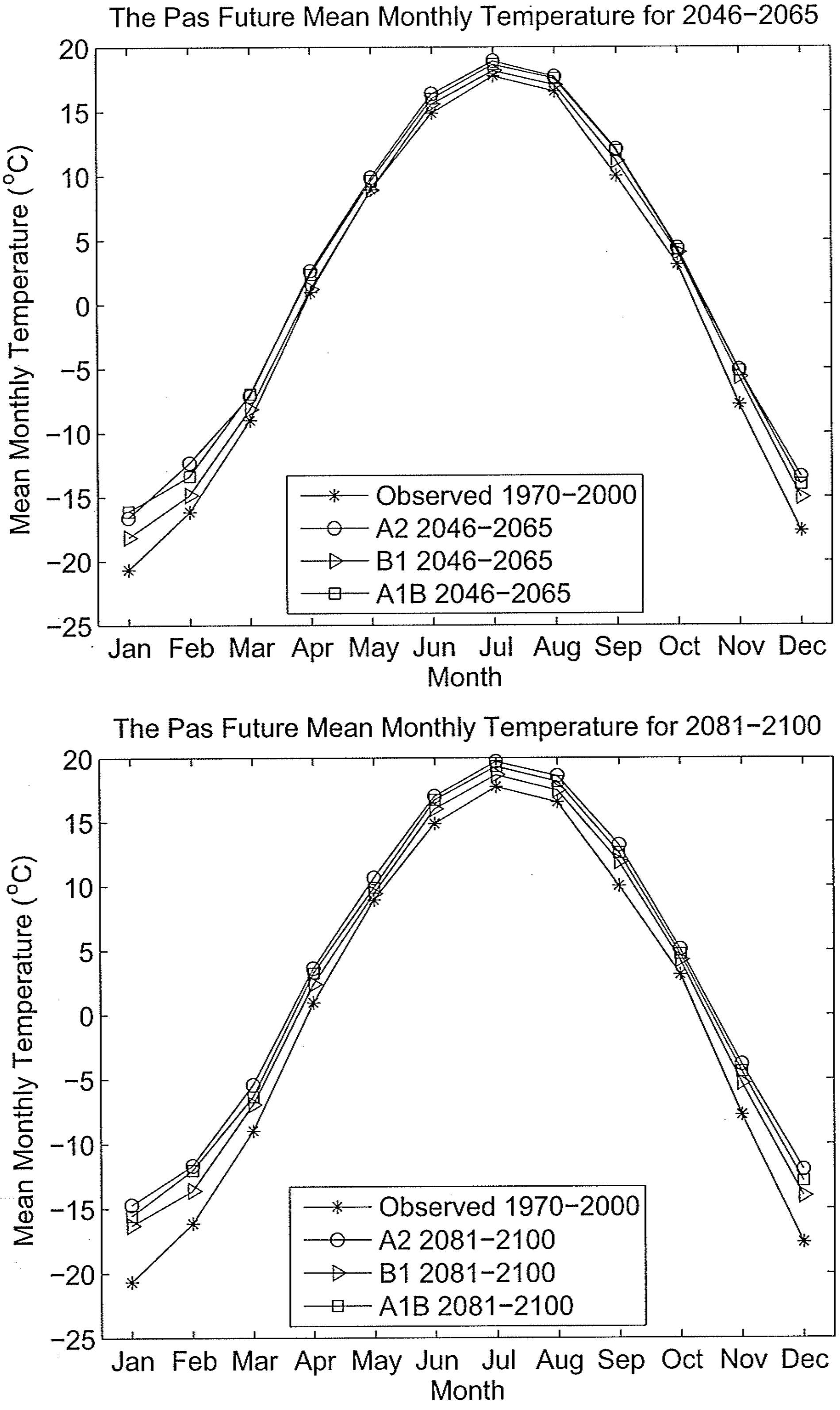


Figure 6.12: Future mean monthly temperature at The Pas.

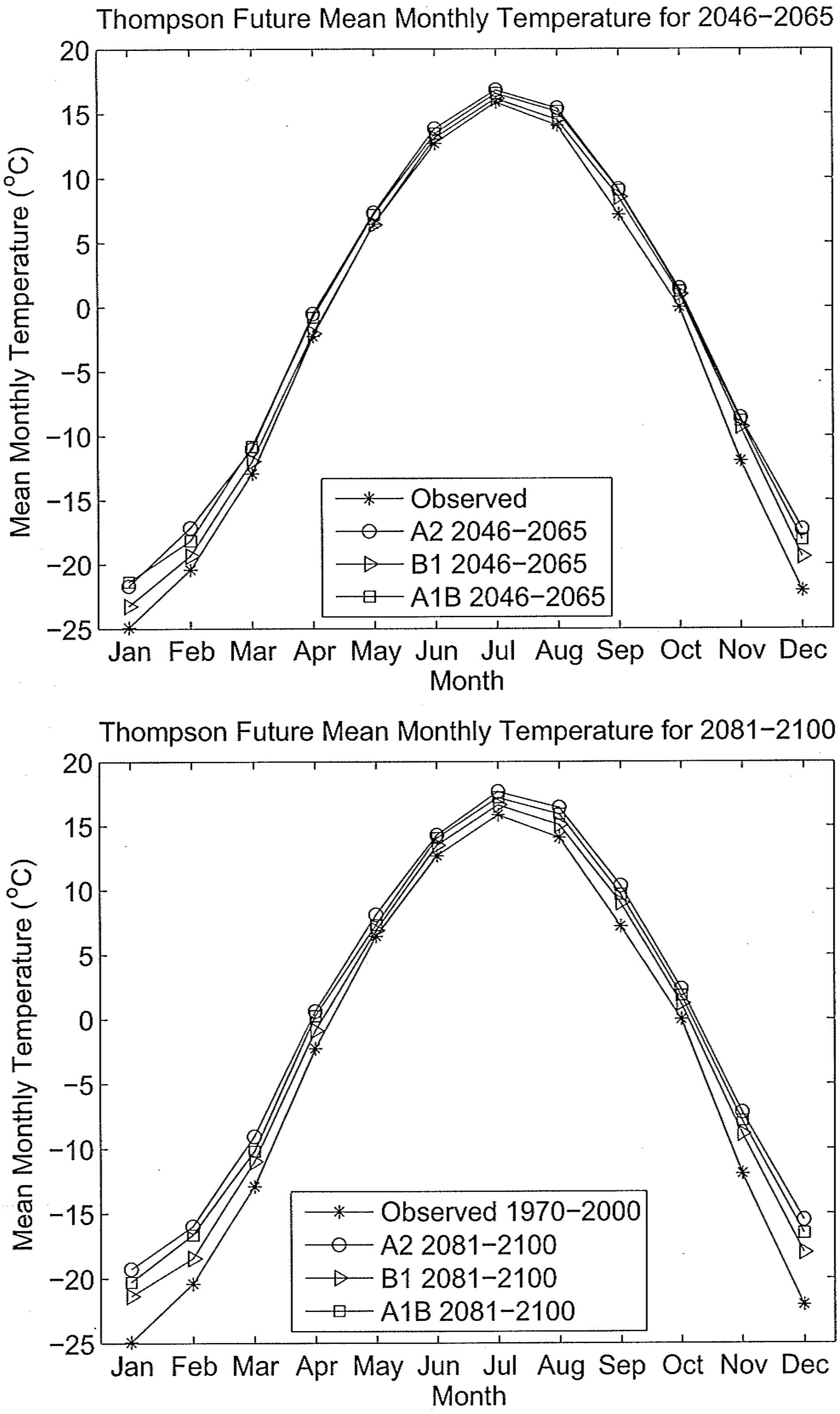


Figure 6.13: Future mean monthly temperature at Thompson.

the A2 scenario.

Temperature rises are more dramatic for the 2081-2100 time slice. Winter temperature increases compared to the 20c3m scenario range from 3.0°C in the B1 scenario to up to 5.5°C in the A2 scenario.

Future Precipitation

The results for downscaled mean monthly precipitation can be seen for The Pas on Figure 6.14 and for Thompson on Figure 6.15, and as annual and winter accumulations in Table 6.5 (page 155) and Table 6.6 (page 155).

As expected, the trends in precipitation are not as clear as in temperature. This is likely due to the random and complex nature of rainfall occurrence, particularly for large storms. However, in most scenarios, for both The Pas and for Thompson, the annual precipitation accumulations decrease. The B1 scenario shows the smallest changes; the B1 2046-2065 data is the only scenario in which annual precipitation did not decrease compared to the 20c3m scenario. Figure 6.13 and Figure 6.12 show that most of the decrease in precipitation occurred during the summer months. The maximum decrease in precipitation at the annual scale occurred in the A2 scenario, where, during the 2081-2100 time slice, The Pas saw a decrease of 23%, and Thompson saw a decrease of 12% compared to the 20c3m scenario.

The annual accumulations over winter months are shown in Table 6.6 (page 155). The A2 and A1B were slightly drier than the B1 scenario.

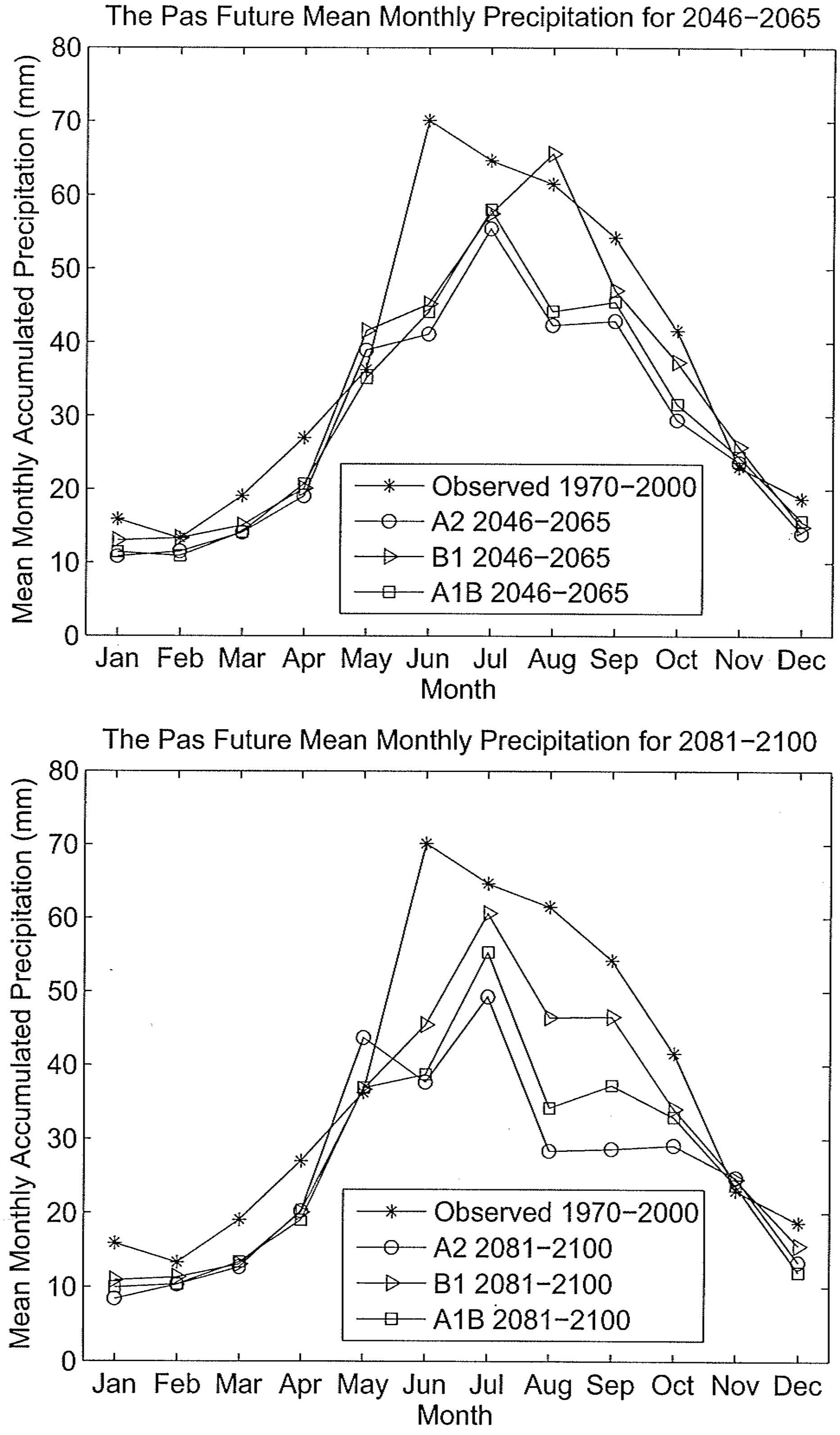


Figure 6.14: Future mean monthly precipitation at The Pas.

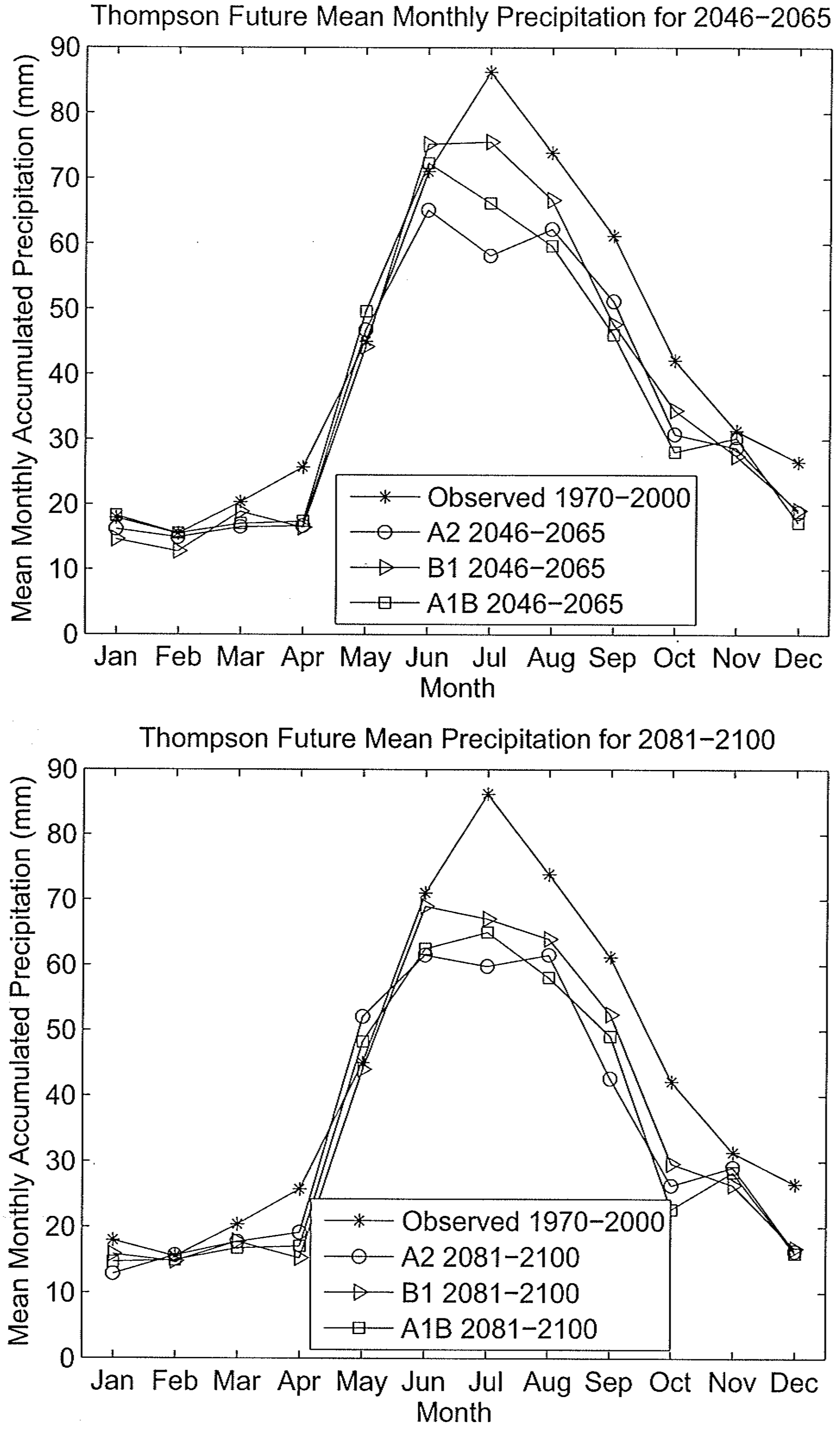


Figure 6.15: Future mean monthly precipitation at Thompson.

6.3 Winnipeg River Multi-site Application

The Winnipeg River is a large western-flowing river originating from Lake of the Woods near the City of Kenora, Ontario, and discharging into Lake Winnipeg in Manitoba. This river is approximately 235 km long and its drainage basin covers approximately 150,000 km^2 in Ontario, Manitoba, and northern Minnesota. The Winnipeg River is significant to hydroelectric power production in Manitoba. Five hydroelectric dams are on the Winnipeg River, and the River also is an important contributor to the total flow of the Nelson River and the hydroelectric dams on the Nelson River.

The methodology presented in the Nelson River multi-site application was repeated for two Environment Canada weather stations in the Winnipeg River drainage basin. Data was downscaled using the k-nn model for the Redlake and Sioux Lookout weather stations. The Redlake weather station (51°4'N, 93°47'W) is close to the Troutlake River Basin, and the Sioux Lookout weather station (50°7'N, 91°54'W) is close to the Sturgeon River drainage basin. The k-nn model was again employed to downscale four variables for the two stations simultaneously, allowing hydrological modelling of the two basins for climate change assessment. These two sub-basins of the Winnipeg River basin were selected as a representative model for the larger Winnipeg River Basin because they are two of the few sub-basins that are unregulated and also have an adequate amount of historical climate and streamflow data.

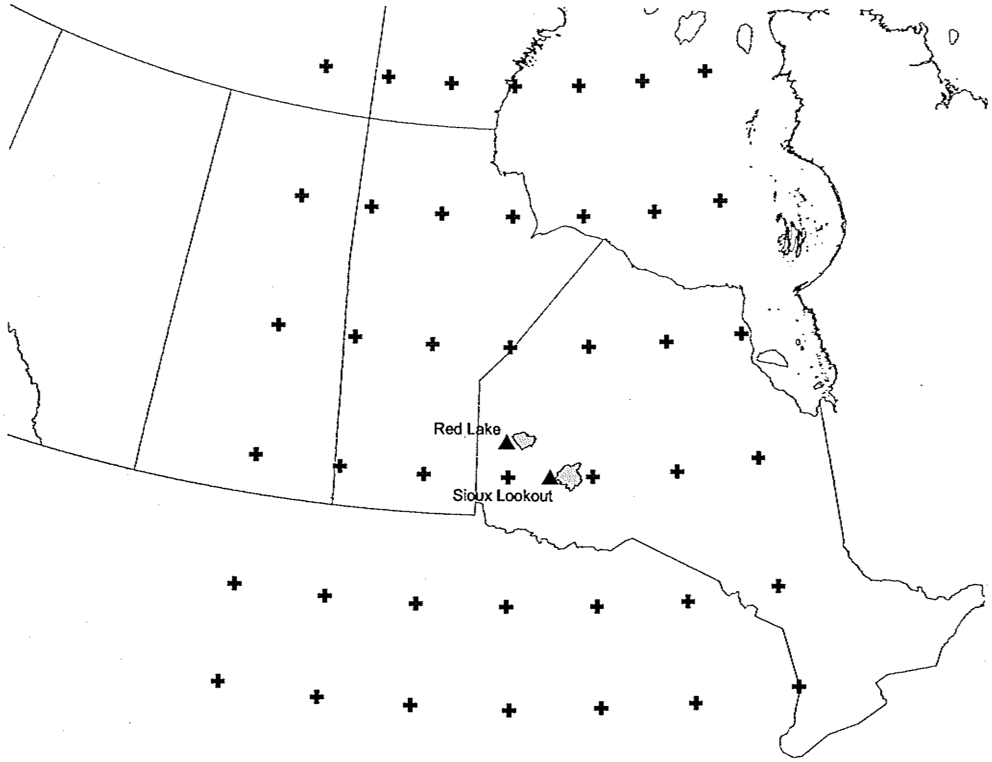


Figure 6.16: Large-scale variable grid for Winnipeg River applications.

6.3.1 Data

The large-scale variables used in the downscaling are the same as those used in the Nelson River applications: average surface temperature, 500mb temperature, 850mb temperature, 500mb geopotential height, and 850mb geopotential height. However, the variables were redownloaded and reprocessed to center the large-scale grid over the Winnipeg River basin as shown on Figure 6.16.

The grid in the Winnipeg River applications was selected to be slightly smaller than the grid used in the Nelson River applications. The grid covering the Winnipeg River basin measured 6 × 7 CGCM3.1/T47 grid points, spaced at approximately 3.75° × 3.75° latitude-longitude. The NCEP/NCAR Reanalysis 1 data were downloaded for a slightly larger area and then interpolated onto the CGCM3.1/T47 grid points.

Since a high degree of spatial correlation existed in the data, principal component analysis was used to reduce the number of variables in the feature vector by removing redundant information (Gangopadhyay et al., 2005; Buishand and Brandsma, 2001; Young, 1994). The first 17 principal components were retained and explain over 96% of the variation contained in the original data sets. The eigenvalues associated with the first thirty principal components, as well as the cumulative percent of explained variance, are shown on Figure 6.17. For the larger grid used in the Nelson River applications, 24 principal components were required to capture 95% of the original variance. The reduction in the number of variables made a noticeable reduction in the data processing time.

The weather stations had a good length of record for daily temperature, precipitation and relative humidity. Both of the weather stations had data from 1965 to 2004, for a total length of 40 years to be used as a historical record to resample from. However, neither station had solar radiation or bright sunshine hours data available over the 1965 to 2004 time period. This problem of missing variables was overcome by first downscaling the other three variables for the two stations. The solar radiation was then generated using a simplified resampling model. A nearest neighbor model, using the six already downscaled variables as a feature vector, was used to resample solar radiation data from NARR data grid points close to the weather stations. The generation of solar radiation will be discussed later in the presentation of the downscaling results.

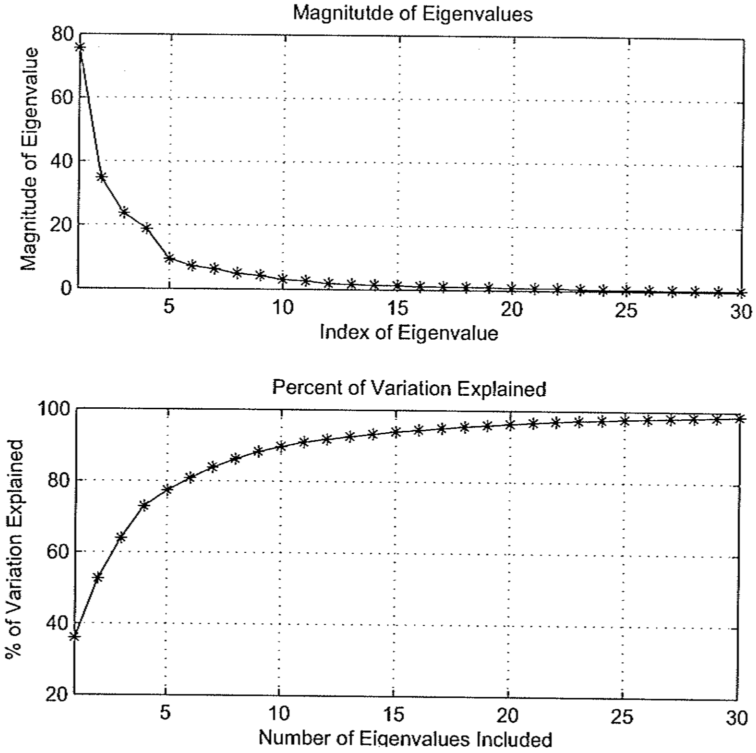


Figure 6.17: Eigenvalues and percent variation explained for Winnipeg River basin application.

All the available scenarios and model runs from the CGCM3.1/T47 were downscaled, including the 20c3m (1961-2000) control run and future scenarios B1, A2, and A1B (2046-2065; 2081-2100). With five model runs available, a total of 200 years of data were downscaled for the 20c3m runs, 100 years for each future scenario for the time slice 2046 to 2065 and 100 years for each future scenario for the time slice 2081 to 2100.

6.3.2 Model Setup and Optimization

The model parameters, i.e. the window width, W, the number of nearest neighbors to retain, k, and the weighting vector in the distance calculation, w, were optimized using the same cross-validation methodology as in the Nelson River application.

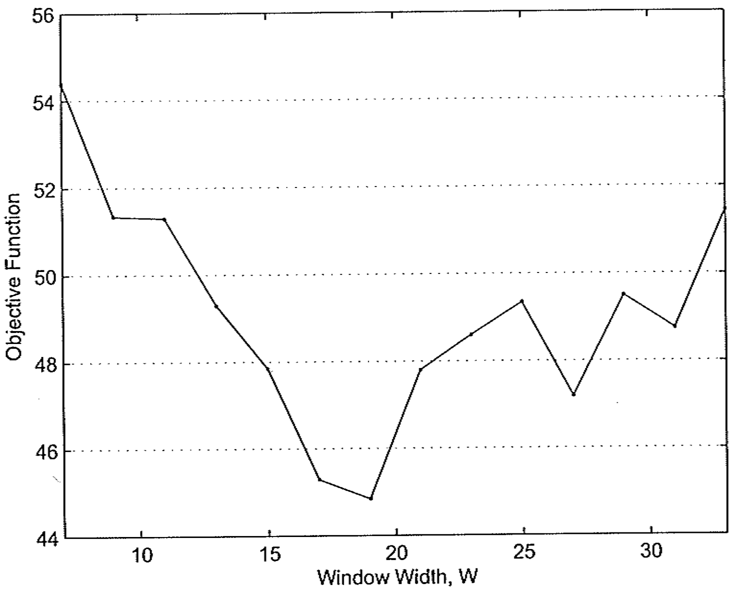


Figure 6.18: Objective function optimization for W.

The objective function was specified as the average of the two stations' root mean square error (RMSE) of the estimated winter precipitation accumulation,

$$\min f(W, k, w_i) = \frac{\mathrm{RMSE}_{\text{Redlake Winter Precip}} + \mathrm{RMSE}_{\text{Sioux Lookout Winter Precip}}}{2}. \quad (6.3)$$

The typical cross-validation was set up where the model takes one year of NCEP/NCAR Reanalysis 1 data as simulation data and uses the other years as a historical record to resample from. This process was repeated for the 40 years of the historical record available for the Redlake and Sioux Lookout weather stations. The objective function was then calculated using the simulated station variables and the historical station variables.

The window width was optimized by manually adjusting the W parameter.

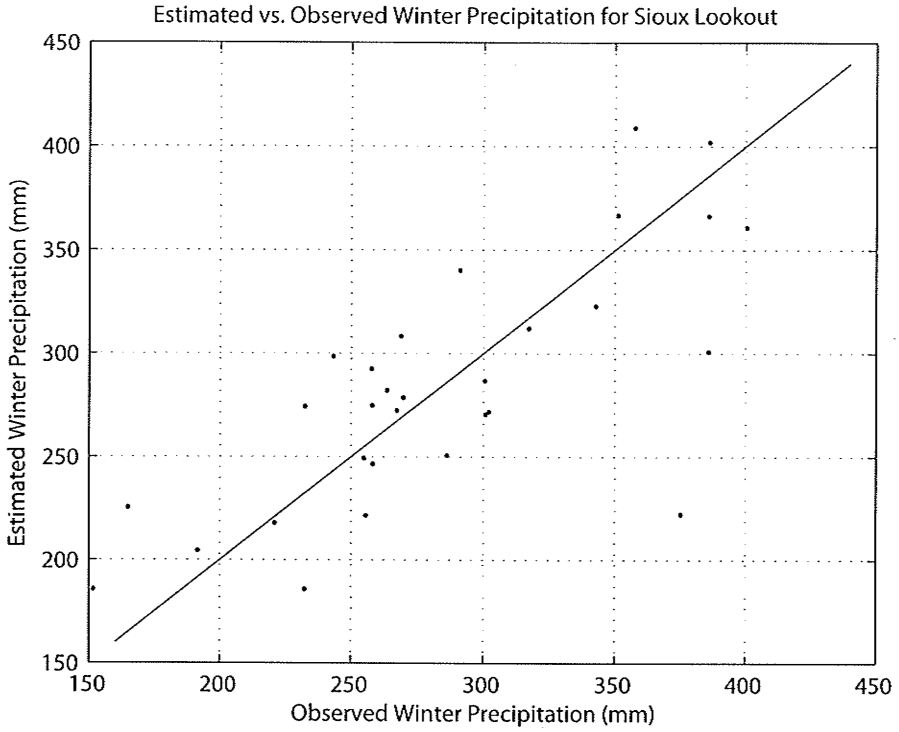


Figure 6.19: Sioux Lookout winter precipitation accumulation cross-validation results.

From Figure 6.18 it can be seen that the optimum window width is 19 days. An interesting observation is that in the applications the window width appears to be directly related to the length of the historical record available to resample from. The window width of 19 days for a historical period of forty years is smaller than the window widths of 21 and 25 days that were optimal for historical records of 31 and 37 years, respectively.

The Matlab Optimization Toolbox was used to optimize the w vector. In this application, the w vector consisted of weights for 17 variables. In the Nelson River applications, 24 weights had to be optimized in the w vector. The reduced length of the feature vector reduced the overall time required for optimization.

The optimization process started with equal weight given to all variables. By

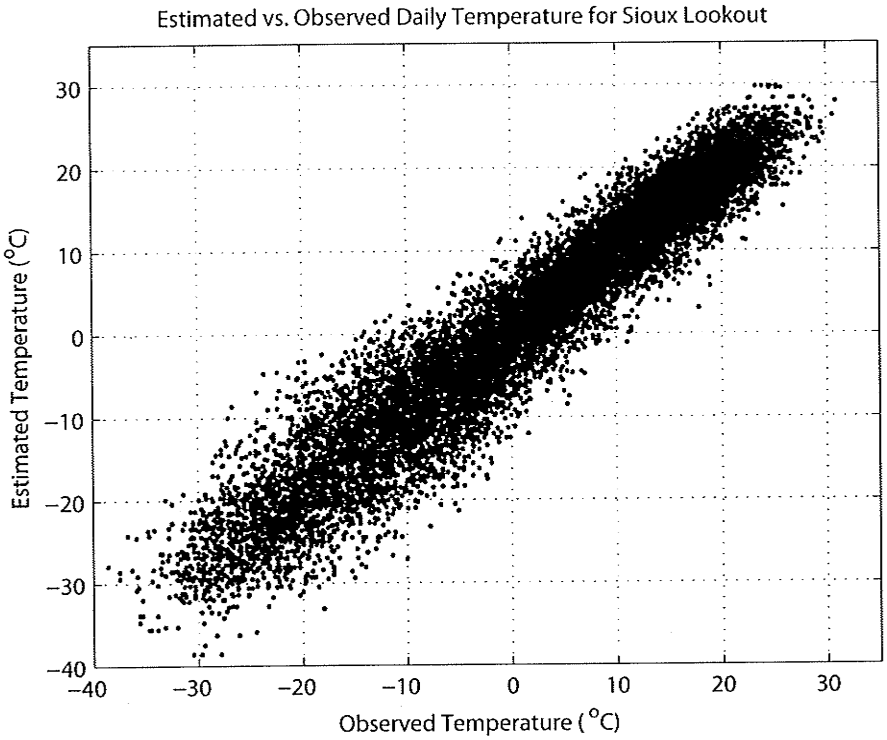


Figure 6.20: Sioux Lookout daily temperature cross-validation results.

adjusting the w vector, the objective function improved from 49.2 to 44.1, a decrease of 10.4%. The decrease in the objection function was not as large as in the Nelson River multi-site application, but the optimization of the feature vector did improve the model an appreciable amount.

Satisfactory results were achieved after optimization. The model was able to reproduce seasonal trends in the winter precipitation at both stations. Better results were achieved for the Sioux Lookout station than for the Redlake station. Figure 6.19 shows the estimated and observed winter precipitation accumulations for the Sioux Lookout after optimization. RMSE's of 50.0 mm and 38.3 mm, and correlations of 0.60 and 0.75, were achieved for the simulation of winter precipitation accumulations for Redlake and Sioux Lookout.

Although the other simulated variables were not included in the optimization, it is important that the model also simulate these variables adequately. The simulated temperature data were checked and correlations at the daily time scale of 0.97 were found for both the Sioux Lookout and Redlake stations. A scatter plot of simulated and observed daily temperatures at Sioux Lookout is shown on Figure 6.20. The results show daily temperature is adequately simulated.

Correlations for estimated and observed relative humidity at the daily time scale were 0.50 for both stations. As in the Nelson River application, relative humidity was not simulated with as high correlation as daily temperature. However, since the SLURP model is most sensitive to precipitation and temperature, the results obtained for relative humidity were judged to be adequate in this application as well.

The numbers of neighbors, k, to retain for resampling was set to ten. In the previous model setups, k was found to be optimal when the resampling was limited to only the single nearest neighbor. However, to increase the variability in downscaled GCM data, the number of neighbors was set to ten.

6.3.3 Downscaling Results

20c3m Control Run

The first step was to downscale the CGCM3.1/T47 20c3m control run and compare the results to the observed weather statistics. Figure 6.21 shows that the observed monthly mean temperature is in excellent agreement with the monthly averages of temperature simulated by downscaling both NCEP/NCAR Reanalysis 1 data and

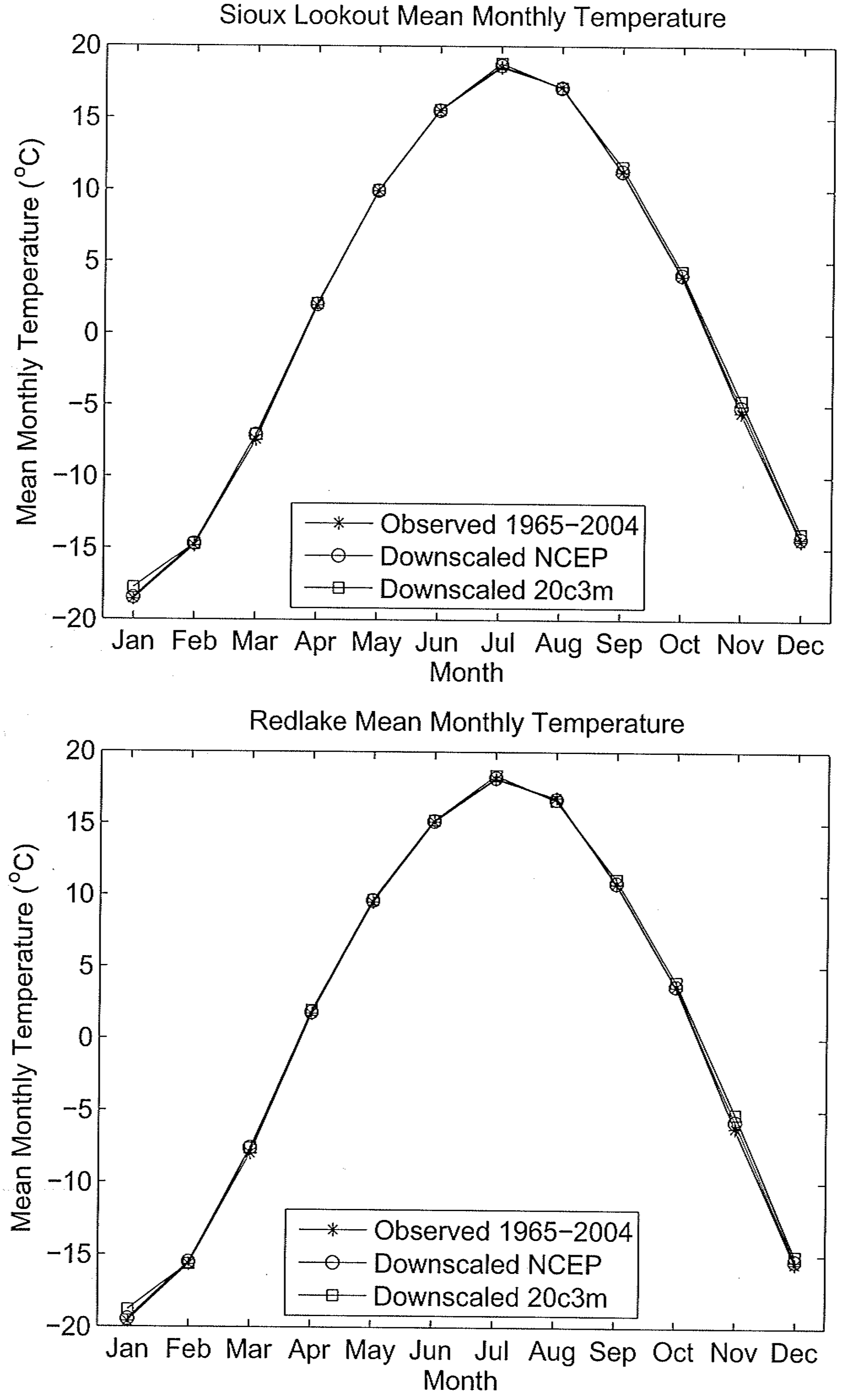


Figure 6.21: Optimization results for mean monthly temperatures.

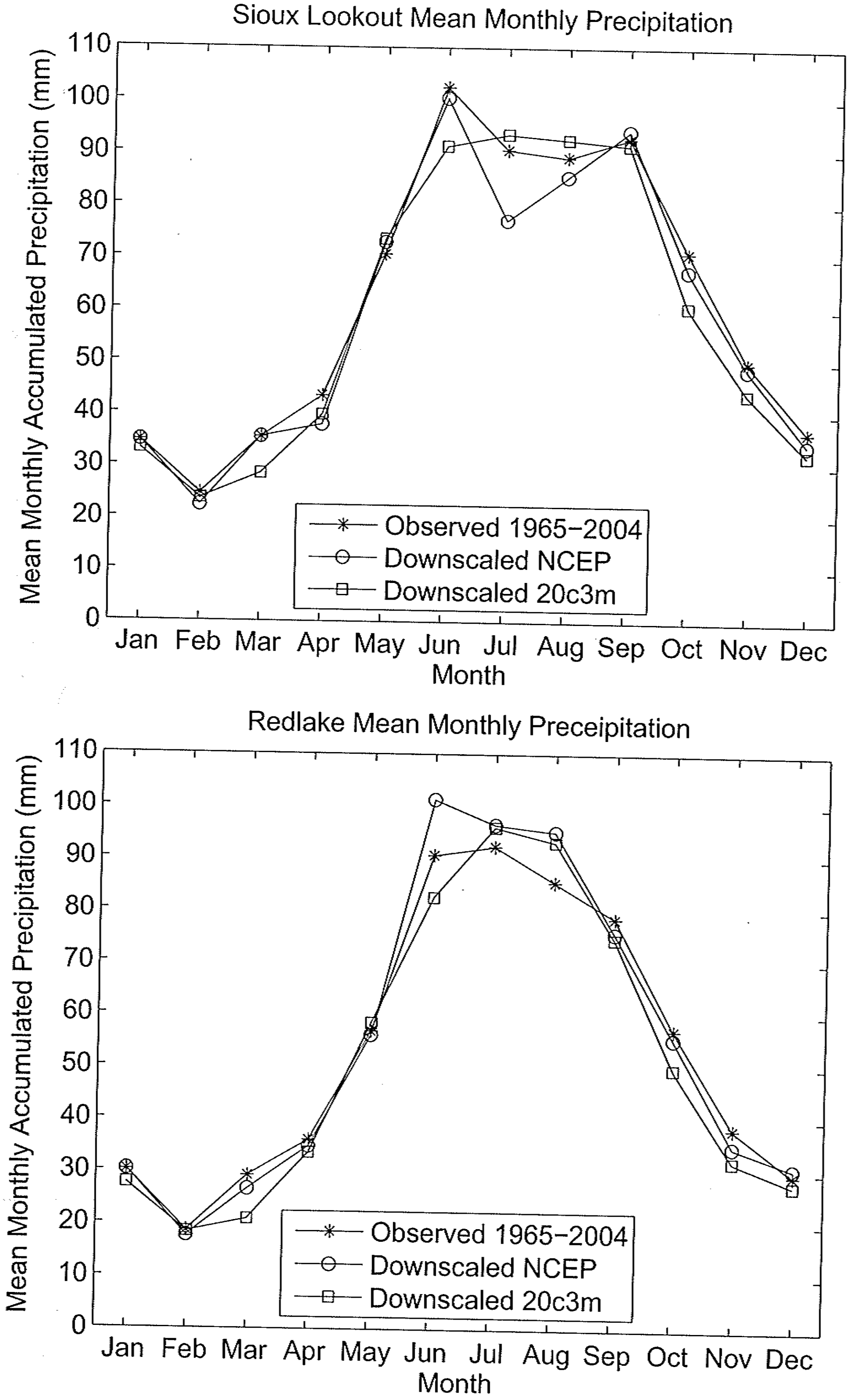


Figure 6.22: Optimization results for mean monthly Winnipeg River Station precipitation.

the 20c3m GCM data. There are no biases in the downscaled 20c3m and downscaled NCEP/NCAR Reanalysis 1 data compared to the observed temperature data at the monthly scale.

Figure 6.22 shows that the downscaled NCEP/NCAR Reanalysis 1 and 20c3m data have a good match for the average monthly accumulation of rainfall.

The annual precipitation was slightly underestimated by the 20c3m scenario. The mean annual observed precipitation at Sioux Lookout was 741 mm, while the 20c3m downscaled precipitation had a annual average accumulation of 704, a difference of 5%. Similarly, the annual precipitation was underestimated by 5% for the Sioux Lookout station. The winter precipitation was slightly more underestimated: 15% for Redlake and 10% for Sioux Lookout. The agreement between observed and downscaled 20c3m data annual and winter precipitation accumulations is better than the results for the Nelson River multi-site application.

The k-nn model was then used to downscale the GCM data for the A2, A1B, and B1 scenarios for the future time slices of 2046 to 2065 and 2081 to 2100.

For the downscaling application in the Winnipeg River Basin, solar radiation or bright sunshine hours data were not available. Since one of these variables is required as input to the SLURP hydrological model, an alternative source of downscaled solar radiation or bright sunshine hours data was necessary. Solar radiation data from two NARR data points close to the Redlake and Sioux Lookout weather stations were used as sources of data to generate solar radiation data for the downscaled scenarios. The period of overlap between the NARR data and the historical station data is from 1979 to 2004. A simplified nearest neighbor resampling model was set up to generate solar radiation data for the downscaled CGCM3.1/T47 and NCEP/NCAR Reanalysis 1 data sets by searching for days in the historical record based on a feature vector of the previously downscaled temperature, precipitation, and relative humidity variables.

Since the SLURP model is not as sensitive to solar radiation as temperature and precipitation, the resampling procedure was designed much simpler and weights were assigned based on judgement rather than robust optimization. The three variables in the feature vector were daily temperature (°C), daily occurrence of precipitation (1 if precipitation greater than or equal to 0.2 mm, 0 if precipitation is less than 0.2 mm), and relative humidity. Weights were assigned to the variables as 1, 10 and 0.1, respectively, meaning almost all weight was placed on temperature and the occurrence of rainfall. The resampling of days was not restricted to a moving window, and temperature data was not deseasonalized by standardizing using the seasonal means and standard deviations. Using temperature data with seasonal influences restricted the resampling to days in the same season as the simulation day. A Euclidean distance metric was employed and the nearest neighbor ($k = 1$) was retained as the day to resample solar radiation from. This model proved to be time efficient and capitalized on the advantages of the nearest neighbor model, such as preserving the correlation between the historical solar radiation and temperature, precipitation and relative humidity variables.

Future Temperature

The results for downscaled temperature can be seen for Redlake on Figure 6.23, and Table 6.7 (page 156), and for Sioux Lookout on Figure 6.24 and Table 6.8 (page 156).

For the time slice 2046 to 2065 all future scenarios show increases in temperature throughout the year. The B1 scenario shows the least warming, with temperature increases of less than 3.0°C in the winter and 2.0°C in the summer compared to the 20c3m scenario.

For the time slice of 2081 to 2100 all future scenarios showed increases in temperature compared to the 2046 to 2065 time slice. However, the temperature increases were not as significant as the increases between the 20c3m and 2046 to 2065 time slice. The A2 scenario showed only slight warming between the two time slices, and in some months showed slight cooling of up to 0.9°C. The A2 scenario was the warmest scenario in the late winter months of January, February and March, while the A1B was the warmest in the summer months and early winter months of October, November and December. The B1 scenario remained the coolest of the future scenarios, only increasing 3.1 to 3.9°C in the winter and 2.2 to 2.4°C in the summer. Compared to the 20c3m scenario, the largest winter temperature increases were around 6.1°C at Sioux Lookout and 6.4°C at Redlake in the A2 scenario, and the largest increases in the summer temperatures were around 3.6°C at Sioux Lookout and 3.8°C at Redlake in the A1B scenario.

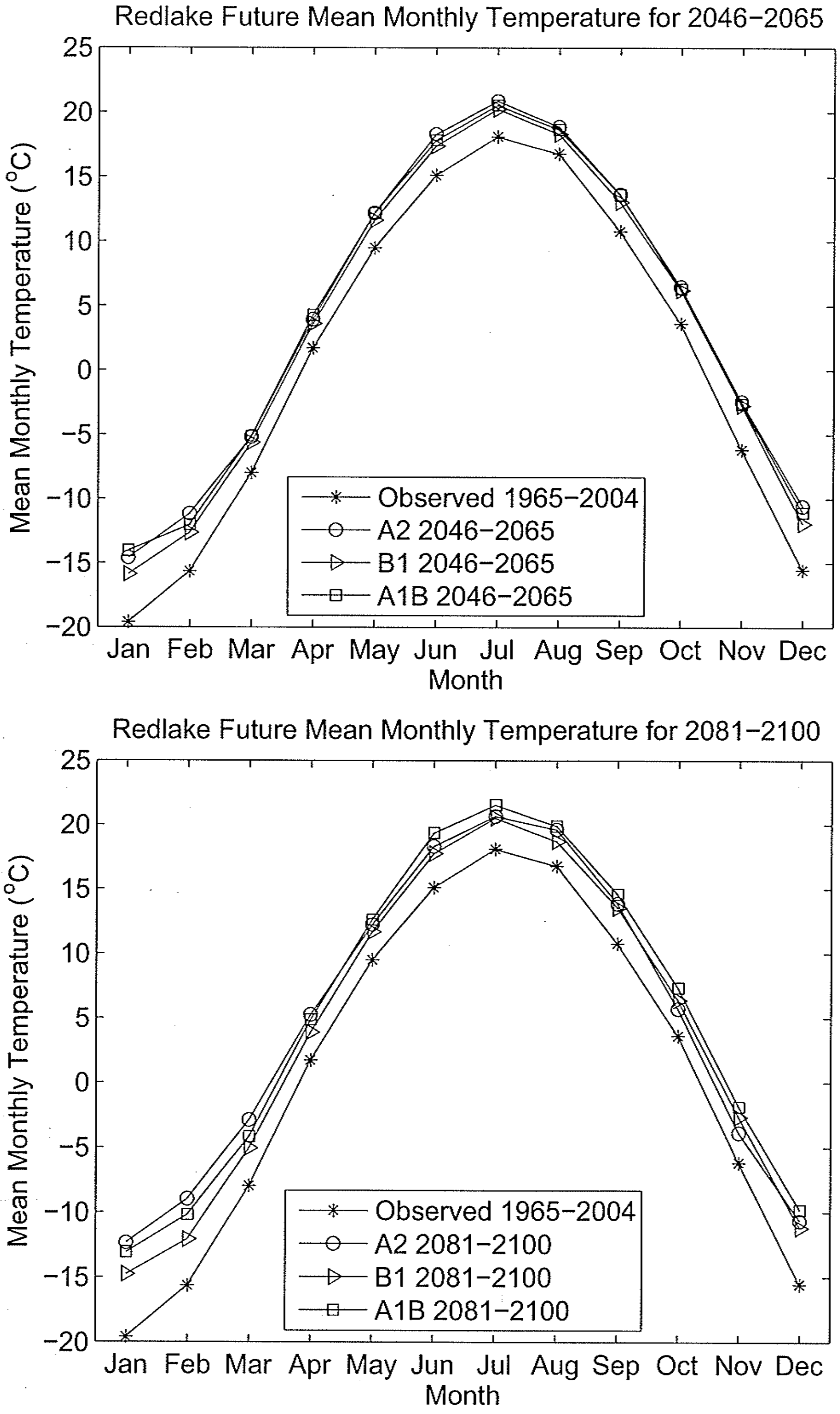
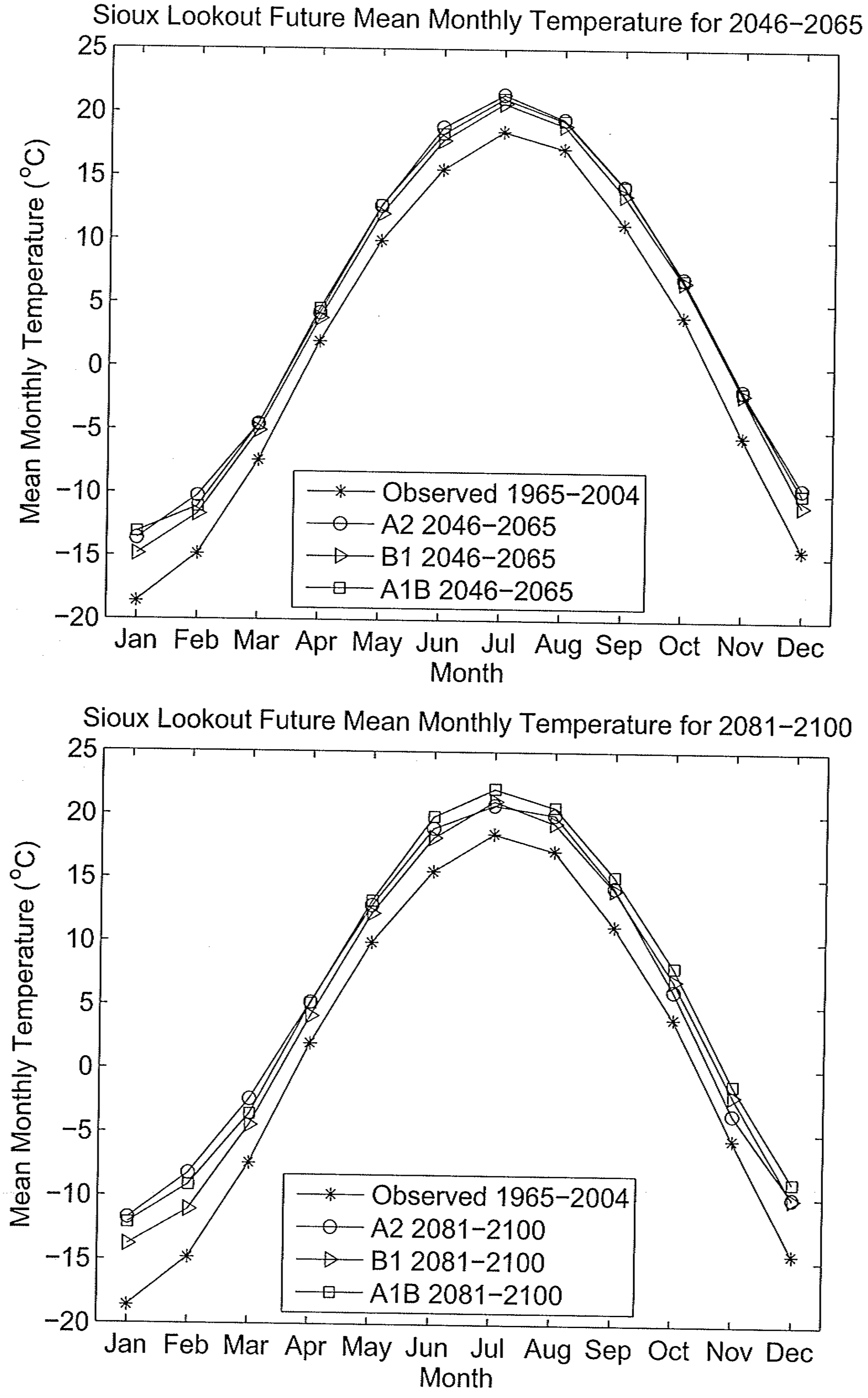


Figure 6.23: Future mean monthly temperature at Redlake.


Sioux Lookout Future Mean Monthly Temperature for 2046–2065

Sioux Lookout Future Mean Monthly Temperature for 2081–2100

Figure 6.24: Future mean monthly temperature at Sioux Lookout.

Future Precipitation

The results for downscaled precipitation as monthly mean accumulations can be seen for Redlake on Figure 6.25 and for Sioux Lookout on Figure 6.26 , and as annual and winter accumulations in Table 6.9 (page 157) and Table 6.10 (page 157).

As in the Nelson River application, the monthly mean precipitation accumulations are much more variable than the monthly temperature means. While the Nelson River application resulted in decreases in annual and winter precipitation, the results for this application suggest that precipitation will remain close to the present or slightly increase at the Redlake and Sioux Lookout stations. For the 2046 to 2065 time slice at the annual scale, Sioux Lookout's precipitation ranges from remaining almost the same in the B1 scenario to an increase of 2% in the A1B scenario. Redlake's annual precipitation does not change more than 0.1% for any scenario compared to the 20c3m run. Winter precipitation shows increases at both stations for all scenarios with the maximum increase of 14% occurring at Sioux Lookout for the A2 scenario.

For the 2081 to 2100 time slice, almost all scenarios lead to further increases in annual precipitation with the exception of a slight decrease of 2% at the Redlake station for the A1B scenario. The largest increases for 2081 to 2100 period are experienced with the A2 scenario, where precipitation increases 10% at the Redlake station and 9% at the Sioux Lookout station compared to the 20c3m run. Winter precipitation changes are more varied, with small decreases for the A2 scenario and increases of up to 13% for the B1 scenario compared to the 20c3m run.

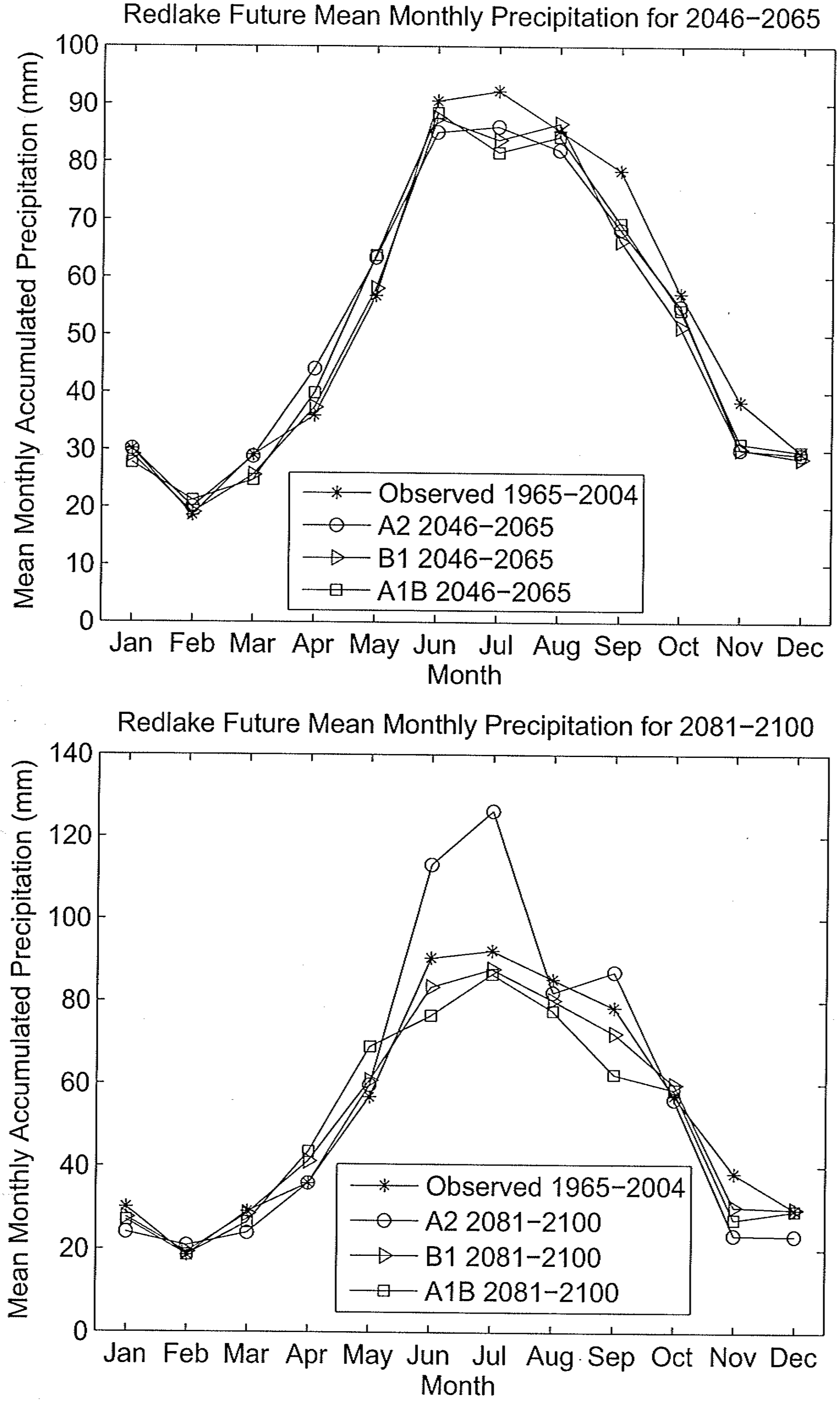


Figure 6.25: Future mean monthly precipitation at Redlake.

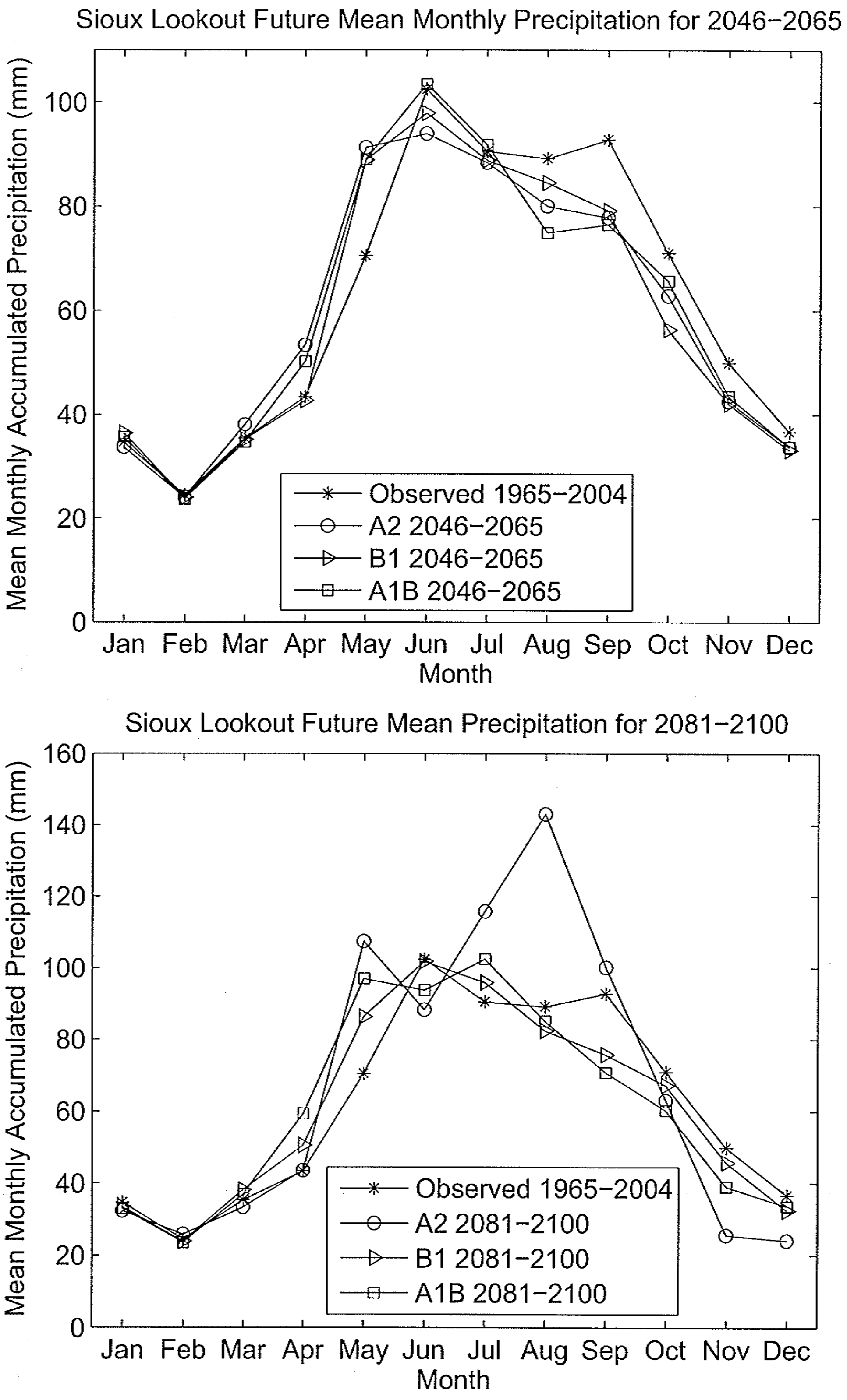


Figure 6.26: Future mean monthly precipitation at Sioux Lookout.

6.4 Winnipeg River NARR Application

One of the difficulties in assessing the potential effects of climate change in Canada is the sparsity of station data. This is particularly evident in northern Canada and other sparsely populated areas. A possible replacement for the non-existent station data is North American Regional Reanalysis (NARR) data. With a grid resolution of 32 km, multiple NARR data points are available in close proximity to any watershed. If NARR data can be substituted into the downscaling methodology in the place of regular Environment Canada CDCD weather station variables, it would facilitate assessment of watersheds in areas with few or no weather stations.

In this application of the *k*-nn model, the utility of NARR data as a substitute for weather station data will be explored. NARR data is used to supply the historical measurements of temperature, precipitation, relative humidity and solar radiation. A limitation of NARR data is the relatively short temporal record available compared to many weather stations. NARR data is currently only available from January 1979 to December 2004.

6.4.1 Data

Five NARR grid points as shown on Figure 6.27 provided the historical observation data for this application. Three grid points were selected in the Sturgeon River basin, and two were selected from the Troutlake River Basin. The weather variables required by the SLURP hydrological model, daily temperature, precipitation, relative humidity and solar radiation, were extracted for these grid points.

The length of the NARR data was 26 years, from January 1, 1979, to Dec 31, 2004. An advantage of using NARR data is that since the data is a reanalysis data set, the data set is complete and there are no missing entries.

The large-scale data in this application remained unchanged from the previous multi-site Winnipeg River Drainage Basin application. The large-scale data was comprised of the 850 geopotential height, the 500 mb geopotential height, surface temperature, temperature at 850 mb geopotential height and temperature at the 500 mb geopotential height. NCEP/NCAR Reanalysis 1 data for the same time period as the NARR data was interpolated on to the CGCM3.1/T47 grid shown on Figure 6.16. The 6 × 7 GCM data grid contains 42 data points, approximately 2/3 the size of the 60-point grid used for the Nelson River Drainage Basin applications. Although this grid is smaller than the grid used for the Nelson River Drainage Basin applications, it was used for the Winnipeg River multi-site application and led to comparable model performance with reduced computational time.

The large-scale data sets were reduced in size by principal component analysis. The 42 grid points for each of the five data sets, for a total of 210 variables, were reduced to seventeen summary variables. The first seventeen variables were withheld as the amount of variance explained by including further principal components diminishes quickly after this point. The next variables have eigenvalues less than one, meaning that an original single grid point would explain more variance than each of these eigenvalues. The seventeen principal vectors explain more than 96% of the variance of the original 210 variables.

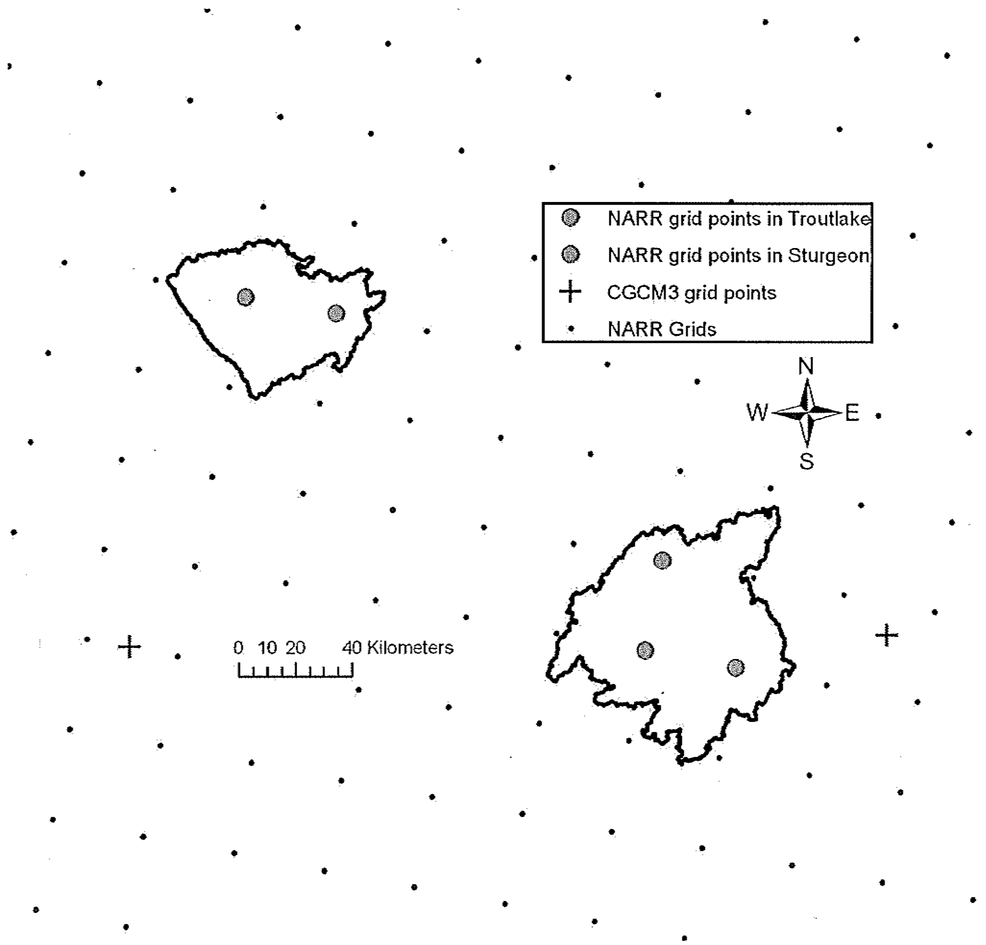


Figure 6.27: NARR grid point locations.

6.4.2 Model Setup and Optimization

The model does not require parameterization of specific relationships between large-scale and local variables, however some components of the model should be adjusted to optimize the ability of the model to estimate the station data. As in the previous applications, the number of neighbors to retain, k, the window size, W, and the weights w can all be adjusted through optimization to improve model performance.

In the application using NARR data, the model was optimized using the same cross-validation method as the Nelson River and Winnipeg River weather station applications. The model was used to simulate the historical NARR data for all five grid points simultaneously from January 1979 to December 2004. The NCEP/NCAR Reanalysis 1 data for one year was considered as simulation data and removed from the historical record. NARR data was then generated for this year of NCEP/NCAR Reanalysis 1 data. This process was repeated for each of the 26 years of data.

The optimization objective function was specified as the root mean square error for the accumulated winter (October to April) precipitation averaged over the five data points. Temperature was not included into the objective function as good results were observed for temperature validation statistics in applications using only precipitation as the objective. The objective function was specified as

$$\min f(W,k,w_i) = \frac{\sum_{i=1}^{5} \mathrm{RMSE}_{\mathrm{NARR}_i\,\mathrm{Winter\,Precip}}}{5}. \tag{6.4}$$

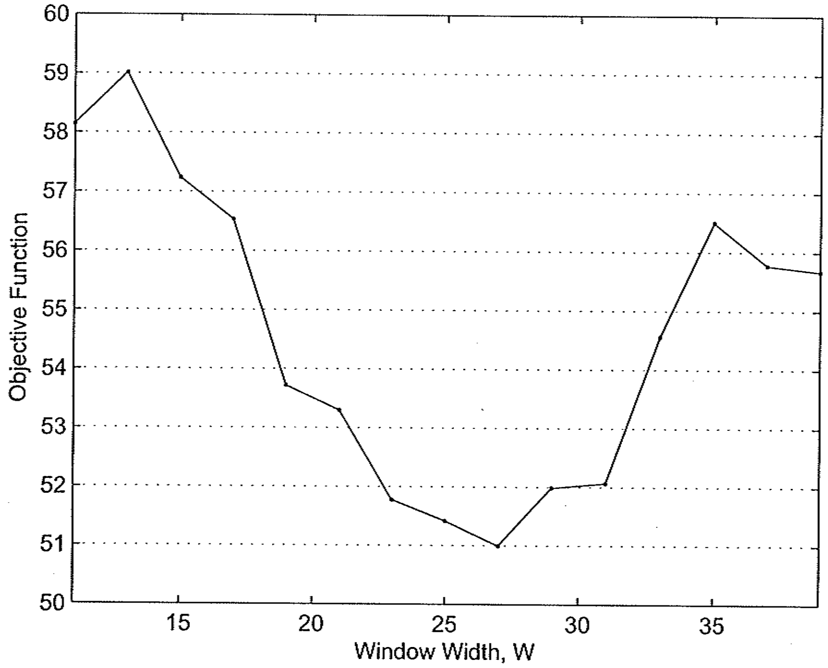


Figure 6.28: Objective function optimization for W.

The model was initialized by assigning equal weight to each principal component variable and k set to retain only the first nearest neighbor. The window width was optimized by manually adjusting W. From Figure 6.28 it was observed that a window size of 27 days leads to the maximum model performance.Combined with the other applications, there is a relationship between the size of search window, W, and the length of the historical record to resample from.

The weighting vector, w, was optimized using the Matlab Optimization Toolbox as in previous applications. The Matlab Optimization Toolbox optimized the w vector using a steepest descent line search method. For the model initialized with equal weight to all feature vector variables, the average root mean square error among the five grid points was 51.0 mm. The optimization procedure improved the average root mean square error to 46.4 mm, a decrease of 9.0%. As a byproduct of optimizing the root mean square error, the average correlation of the simulated winter accumulations to the observed winter accumulations for the five grid points improved from 0.44 to 0.68.

As in similar applications, to encourage variability in the selection process when simulating with GCM data, k was set to retain the ten nearest neighbors.

The model was able to adequately reproduce the observed local weather variables. Figures 6.29 and 6.30 display the results for one of the five grid points from downscaling the historical large-scale climate variables to reproduce the historical local weather. Figure 6.29 shows the estimated and observed daily temperature. The estimated and observed accumulated precipitation over the winter months are shown on Figure 6.30. A correlation of 0.96 was achieved for daily temperature. Relative humidity and solar radiation were also well reproduced, with average correlations at the daily time scale of 0.65 and 0.82, respectively. The results for relative humidity and solar radiation were much better than those for the previous applications.

At the monthly time scale the model is able to reproduce the monthly means of temperature and accumulated precipitation. Figure 6.31 shows that the downscaled NCEP/NCAR Reanalysis 1 data matches the mean monthly temperature of the NARR data almost perfectly. Figure 6.32 shows the downscaled NCEP/NCAR Reanalysis 1 data was also able to reproduce the trends in monthly mean precipitation accumulation.

The optimization results were similar to those achieved for applications resampling from weather station measurements. This demonstrates that NARR data

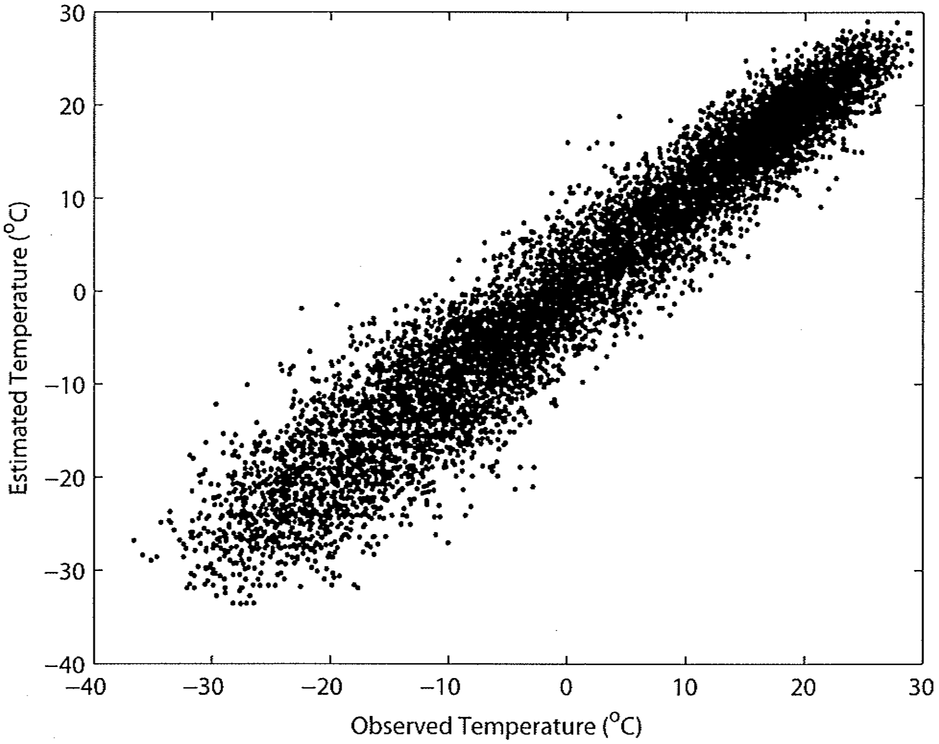


Figure 6.29: Troutlake River 2 daily temperature cross-validation results.

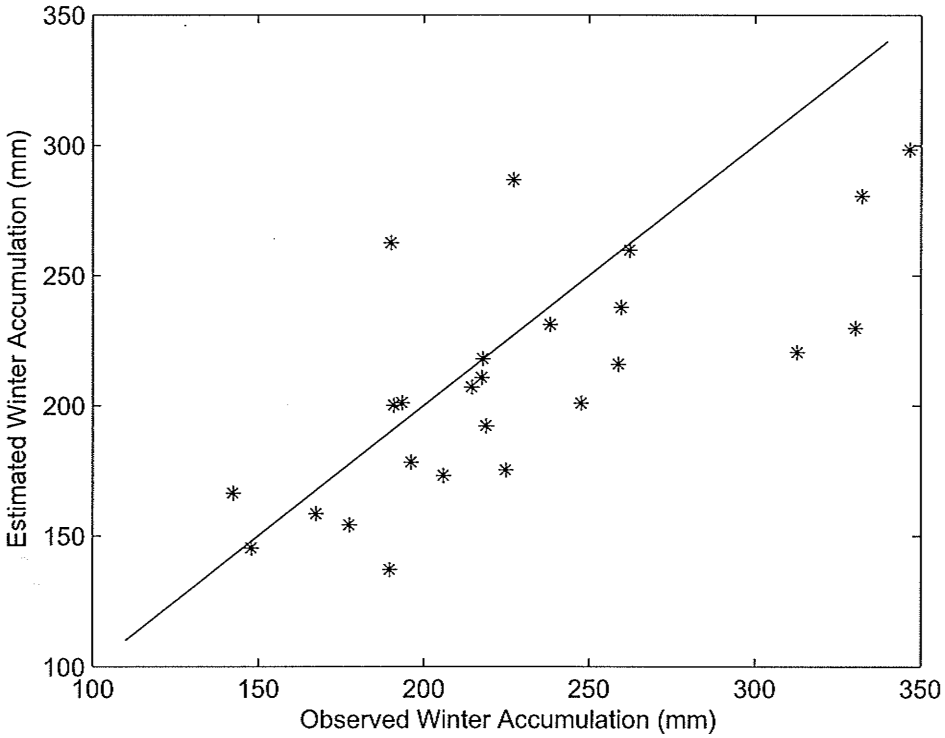


Figure 6.30: Optimization results for winter precipitation accumulation (Troutlake River 2).

can be substituted for weather station data in the k-nn downscaling model.

6.4.3 Downscaling Results

20c3m Control Run

With the k-nearest neighbor model optimized and showing good performance during cross-validation, the model was used to downscale 20c3m control run and the IPCC SRES A2, B1, and A1B climate change scenarios. The k-nn model used the large-scale climate variables provided by the CGCM3.1/T47 to resample days from the historical NARR data set.

First, the 20c3m model runs were downscaled to compare the control run of the GCM to the historical NARR data statistics. The monthly mean temperature and accumulated precipitation for one of the Troutlake grid points averaged over the five 20c3m GCM model runs are shown on Figures 6.29 and 6.30, along with monthly averages from the NARR data and downscaled NCEP/NCAR Reanalysis 1 data. The downscaled 20c3m data show excellent agreement with the other two data sets.

The first two rows of Tables 6.16 (page 161) and 6.17 (page 161) show that the downscaled 20c3m data slightly underestimates precipitation. The underestimation is minor, less than 10% for winter precipitation and less than 5% at the annual scale. The precipitation accumulation downscaled from the 20c3m scenario compare better to the observed NARR data than the results typical of the station data applications.

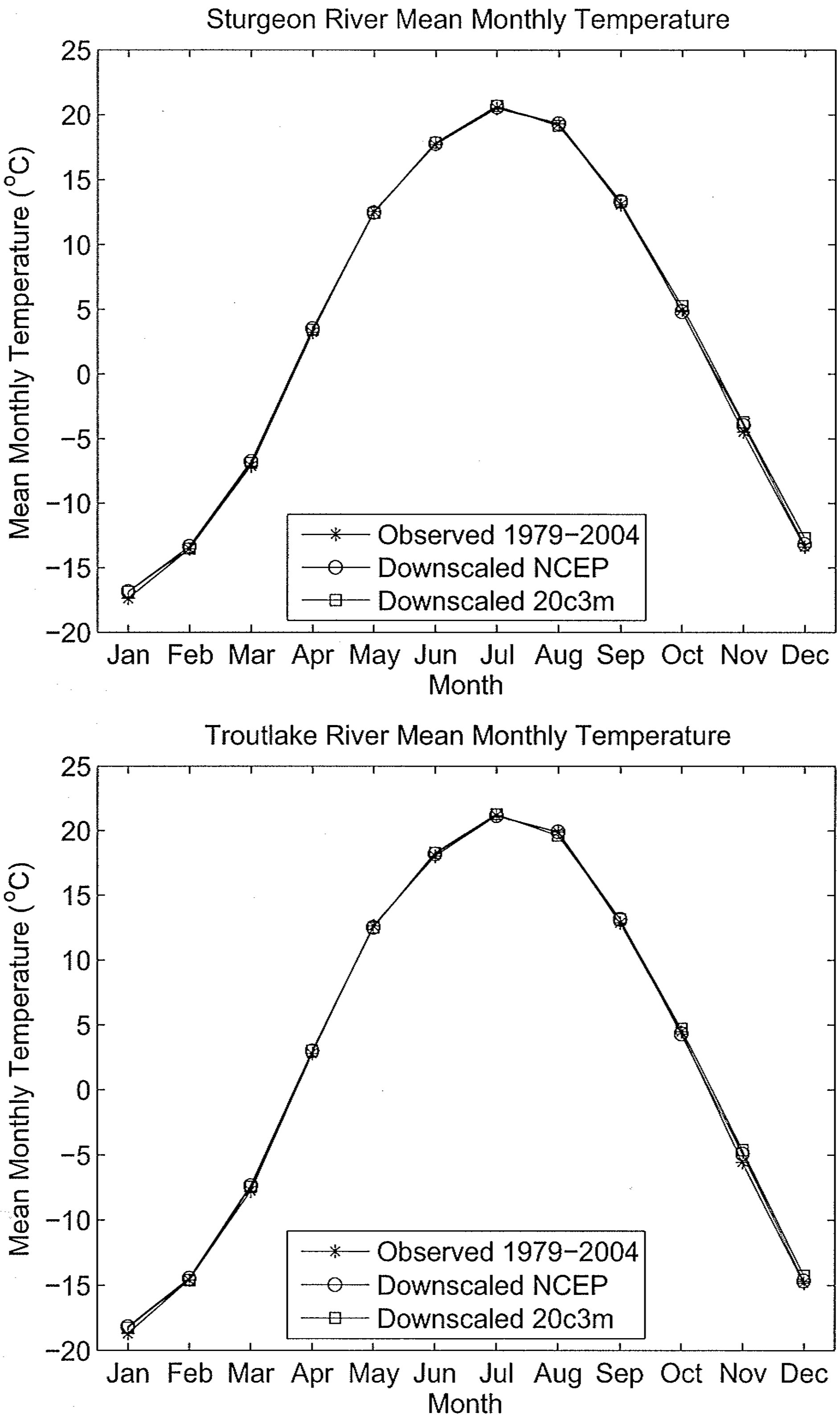
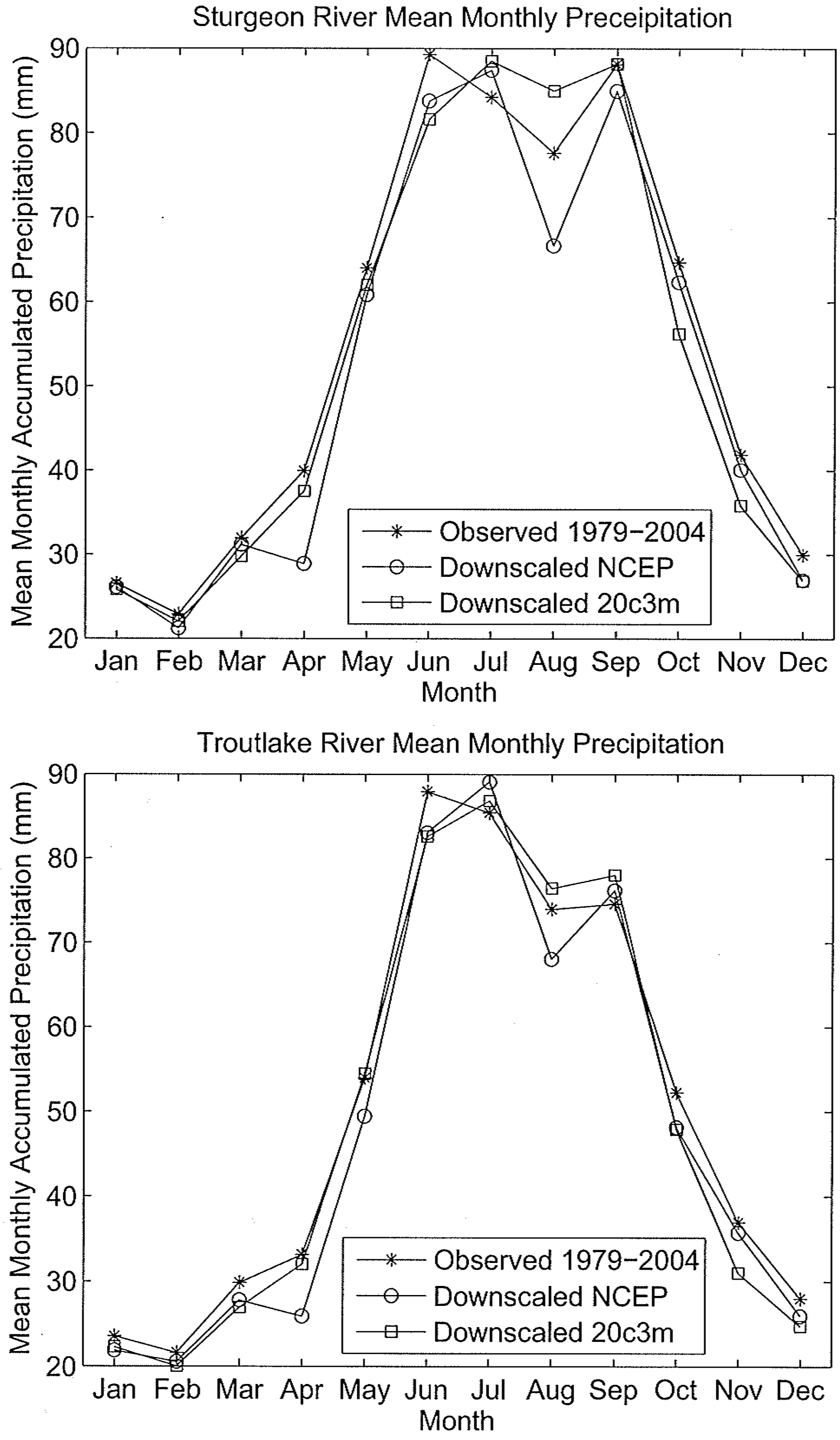


Figure 6.31: Optimization results for mean monthly temperature.



Sturgeon River Mean Monthly Preceipitation

Troutlake River Mean Monthly Precipitation

Figure 6.32: Optimization results for mean monthly precipitation.

Future Temperature

The results obtained for downscaled temperature data were consistent with the results obtained in the other applications. The future monthly mean temperatures from two of the five grid points can be viewed on Figure 6.33 for Sturgeon River, and on Figure 6.34 for Troutlake River. All scenarios exhibit various degrees of warming. The B1 scenario has the least amount of warming. At all stations either the A2 and A1B scenario has the largest amount of warming, depending on the location of the grid point and the time slice (see Tables 6.11 to 6.15 on pages 158 to 160). Temperature increases range up to 5.9°C in the winter of the A2 scenario, while the B1 scenario shows the smallest increases of 2°C in the summer and 4°C in the winter.

Future Precipitation

The trends in future precipitation varied greatly between the different emission scenarios. Tables 6.16 (page 161) and 6.17 (page 161) show the mean annual and winter precipitation accumulations for the different downscaled scenarios. For the A2 scenario, precipitation increases at all data grid point locations, with the maximum increase occurring at a Troutlake River location where the 2081 to 2100 time slice experiences 17% more annual precipitation than the 20c3m scenario. The B1 scenario has similar precipitation as the 20c3m scenario; the future precipitation is within just a few percentage points higher or lower at each grid point. At the annual time scale, the A1B scenario is slightly drier at the three Sturgeon River grid points and slightly wetter at the two Troutlake River grid points. However,

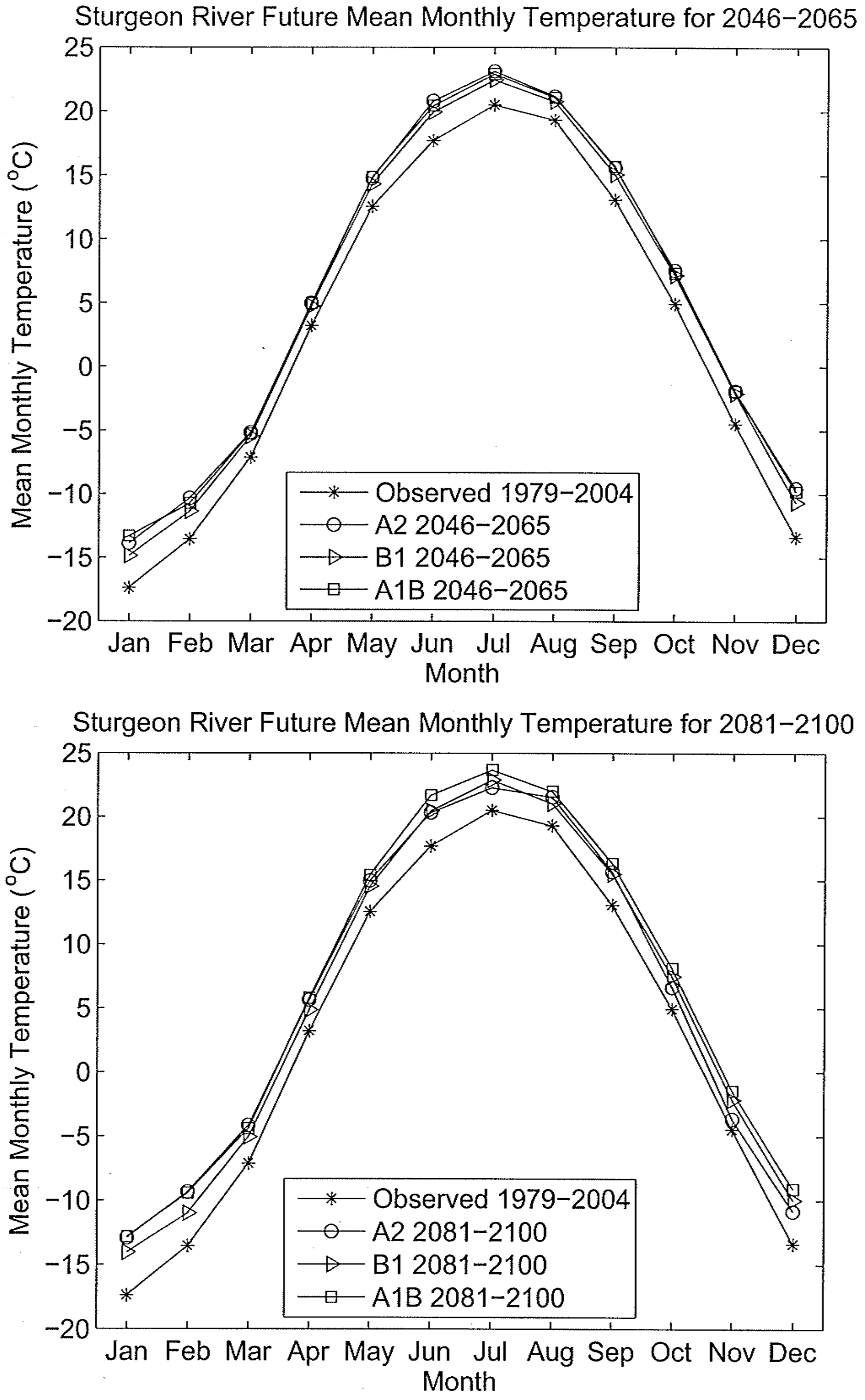


Figure 6.33: Future mean monthly temperature in Sturgeon River Basin.

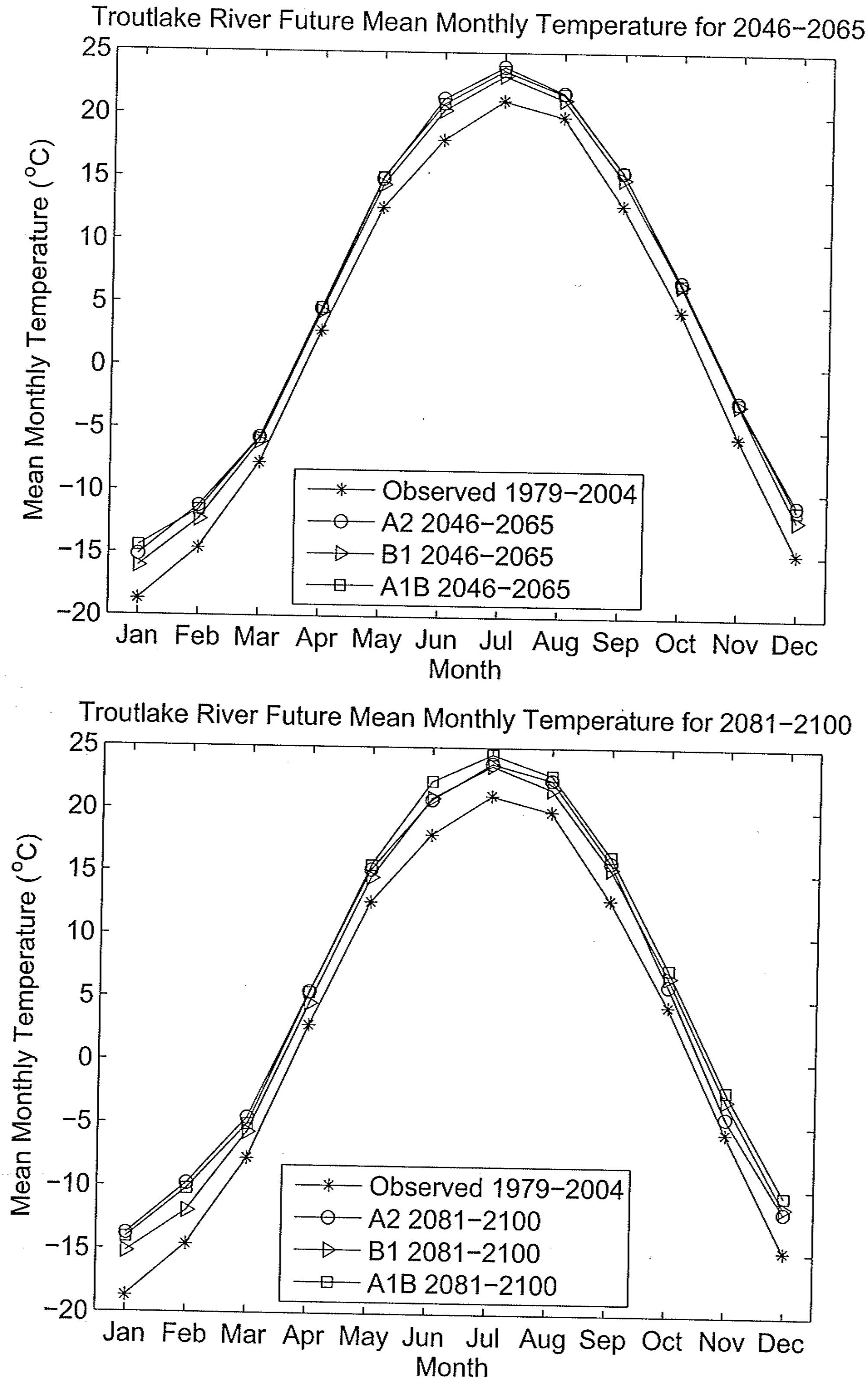


Figure 6.34: Future mean monthly temperature in Troutlake River Basin.

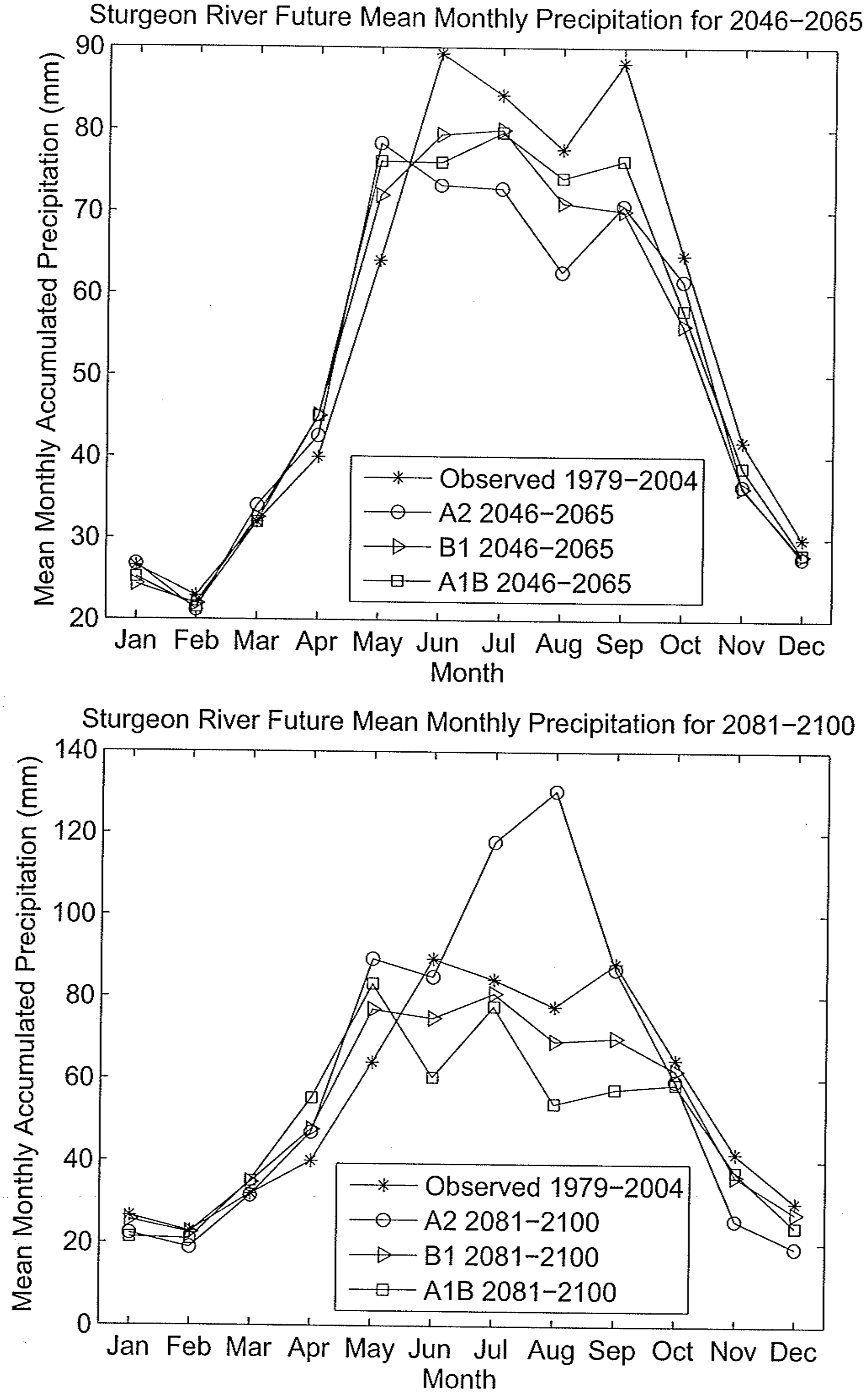


Figure 6.35: Future mean monthly precipitation accumulation in Sturgeon River Basin.

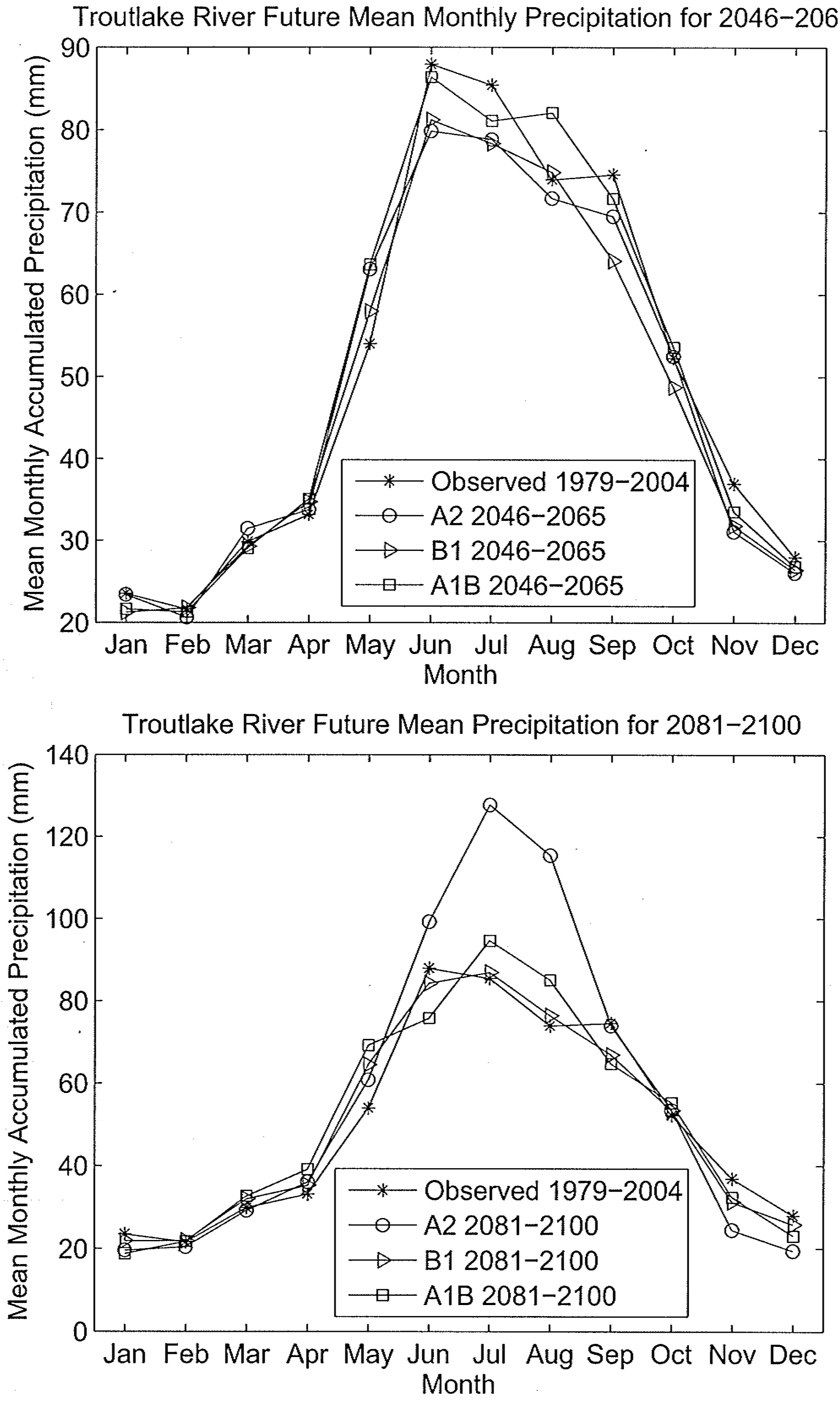


Figure 6.36: Future mean monthly precipitation accumulation in Troutlake River Basin.

during the winter season, all five grid points including the Sturgeon River locations show slight increases in precipitation.

The future monthly distribution of rainfall compared to the historical NARR data can be viewed for a Sturgeon River grid point on Figure 6.35, and for a Troutlake River grid point on Figure 6.36. The monthly distributions are relatively similar to the historical distributions, except that the A2 scenario shows an increase in summer precipitation at both locations during the 2081 to 2100 time slice.

6.4.4 Discussion of NARR Application

The downscaled data generated in the this application will be used to test if downscaled GCM data using NARR data is an acceptable replacement for weather station data. Therefore, the main objective of this application of the k-nn downscaling model was to determine if the model can downscale large-scale GCM data with the NARR data serving as a substitute for weather station data. The results obtained showed that the model can use the same optimization procedure as was developed using station data and similar performance can be achieved. Downscaling using NARR data has very promising potential for use where station data are not available. Downscaling using NARR data would be useful in many of the sparsely populated areas of Canada, particularly northern regions, where weather station data are generally not available in either quality or quality.

Table 6.3: Observed and downscaled mean monthly temperature at The Pas (°C).

Scenario	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Observed (1970-2000)	-20.7	-16.2	-9.0	0.9	8.9	14.9	17.7	16.5	10.0	3.1	-7.8	-17.6
20c3m (1961-2000)	-19.2	-15.8	-8.8	1.2	8.8	15.1	17.7	16.4	10.5	3.1	-6.9	-16.5
A2 (2045-2046)	-16.6	-12.3	-7.1	2.6	9.9	16.3	18.9	17.7	12.1	4.4	-5.1	-13.4
A2 (2081-2100)	-14.7	-11.7	-5.5	3.6	10.6	17.0	19.7	18.6	13.2	5.1	-3.9	-12.0
B1 (2045-2046)	-18.1	-14.8	-8.1	1.2	8.9	15.6	18.2	17.1	11.2	4.0	-5.6	-15.0
B1 (2081-2100)	-16.3	-13.6	-7.0	2.4	9.4	16.0	18.7	17.5	11.8	4.2	-5.3	-14.0
A1B (2045-2046)	-16.1	-13.4	-7.0	2.4	9.6	16.0	18.7	17.6	11.9	4.2	-5.2	-14.0
A1B (2081-2100)	-15.6	-12.0	-6.4	3.2	9.9	16.7	19.3	18.1	12.6	4.7	-4.4	-12.9

Table 6.4: Observed and downscaled mean monthly temperature at Thompson (°C).

Scenario	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Observed (1970-2000)	-24.9	-20.4	-12.9	-2.3	6.4	12.7	15.9	14.1	7.2	-0.0	-11.9	-22.0
20c3m (1961-2000)	-24.4	-20.5	-12.7	-2.1	6.4	12.8	15.7	13.8	7.8	-0.1	-10.8	-21.1
A2 (2045-2046)	-21.7	-17.1	-11.0	-0.5	7.3	13.8	16.8	15.4	9.1	1.4	-8.6	-17.3
A2 (2081-2100)	-19.3	-16.0	-9.1	0.6	8.1	14.3	17.6	16.4	10.3	2.3	-7.2	-15.5
B1 (2045-2046)	-23.2	-19.3	-12.0	-2.0	6.4	13.1	16.2	14.5	8.5	1.0	-9.3	-19.4
B1 (2081-2100)	-21.4	-18.5	-11.0	-0.9	6.9	13.5	16.6	15.1	9.0	1.2	-8.9	-18.0
A1B (2045-2046)	-21.3	-18.2	-10.8	-0.8	7.1	13.4	16.6	15.22	9.1	1.2	-8.8	-18.1
A1B (2081-2100)	-20.3	-16.7	-10.2	0.2	7.3	14.1	17.2	15.9	9.7	1.8	-7.9	-16.5

Table 6.5: Observed and downscaled mean annual precipitation accumulation (mm).

Scenario	The Pas	Thompson
Observed (1970-2000)	446	512
20c3m (1961-2000)	397	467
A2 (2045-2046)	344	426
A2 (2081-2100)	307	415
B1 (2045-2046)	397	453
B1 (2081-2100)	366	433
A1B (2045-2046)	356	438
A1B (2081-2100)	324	413

Table 6.6: Observed and downscaled mean winter precipitation accumulation (mm).

Scenario	The Pas	Thompson
Observed (1970-2000)	159	185
20c3m (1961-2000)	137	143
A2 (2045-2046)	123	143
A2 (2081-2100)	119	137
B1 (2045-2046)	140	144
B1 (2081-2100)	130	137
A1B (2045-2046)	129	144
A1B (2081-2100)	121	130

Table 6.7: Observed and downscaled mean monthly temperature at Redlake (°C).

Scenario	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Observed (1965-2004)	-19.6	-15.7	-8.0	1.7	9.5	15.1	18.1	16.8	10.8	3.6	-6.2	-15.5
20c3m (1961-2000)	-18.8	-15.6	-7.6	1.9	9.7	15.2	18.3	16.6	11.1	4.0	-5.2	-15.0
A2 (2045-2046)	-14.6	-11.2	-5.2	3.9	12.2	18.3	20.9	18.9	13.6	6.5	-2.4	-10.6
A2 (2081-2100)	-12.3	-9.0	-2.9	5.3	12.2	18.3	20.7	19.6	13.9	5.7	-3.9	-10.7
B1 (2045-2046)	-15.9	-12.6	-5.6	3.6	11.6	17.4	20.2	18.3	13.0	6.2	-2.7	-11.9
B1 (2081-2100)	-14.8	-12.1	-5.1	3.9	11.7	17.8	20.5	18.7	13.5	6.4	-2.7	-11.2
A1B (2045-2046)	-14.0	-12.0	-5.1	4.3	12.2	17.8	20.5	18.7	13.6	6.3	-2.6	-11.1
A1B (2081-2100)	-13.1	-10.2	-4.2	4.9	12.6	19.4	21.5	19.9	14.6	7.3	-1.8	-9.8

Table 6.8: Observed and downscaled mean monthly temperature at Sioux Lookout (°C).

Scenario	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Observed (1965-2004)	-18.6	-14.8	-7.4	2.0	9.9	15.5	18.5	17.2	11.2	4.0	-5.5	-14.4
20c3m (1961-2000)	-17.8	-14.7	-7.1	2.1	10.0	15.6	18.8	17.1	11.6	4.4	-4.6	-13.9
A2 (2045-2046)	-13.6	-10.2	-4.6	4.2	12.7	18.8	21.4	19.5	14.3	7.0	-1.8	-9.5
A2 (2081-2100)	-11.7	-8.2	-2.4	5.2	12.8	18.9	20.8	20.0	14.3	6.2	-3.5	-10.0
B1 (2045-2046)	-14.9	-11.7	-5.1	3.8	12.0	17.8	20.7	19.0	13.5	6.7	-2.1	-10.9
B1 (2081-2100)	-13.8	-11.1	-4.5	4.1	12.2	18.2	21.1	19.4	14.1	7.0	-2.0	-10.0
A1B (2045-2046)	-13.1	-11.1	-4.6	4.6	12.7	18.3	21.1	19.4	14.2	6.9	-1.9	-9.9
A1B (2081-2100)	-12.1	-9.2	-3.6	5.1	13.2	19.9	22.1	20.6	15.2	8.0	-1.2	-8.8

Table 6.9: Observed and downscaled mean annual precipitation accumulation (mm).

Scenario	Sioux Lookout	Redlake
Observed (1965-2004)	741	649
20c3m (1961-2000)	704	615
A2 (2045-2046)	720	622
A2 (2081-2100)	802	676
B1 (2045-2046)	709	604
B1 (2081-2100)	734	622
A1B (2045-2046)	723	617
A1B (2081-2100)	735	603

Table 6.10: Observed and downscaled mean winter precipitation accumulation (mm).

Scenario	Sioux Lookout	Redlake
Observed (1965-2004)	290	247
20c3m (1961-2000)	262	210
A2 (2045-2046)	289	238
A2 (2081-2100)	248	208
B1 (2045-2046)	270	221
B1 (2081-2100)	292	237
A1B (2045-2046)	287	228
A1B (2081-2100)	286	232

Table 6.11: Observed and downscaled mean monthly temperature at Sturgeon River 1 (°C).

Scenario	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Observed (1970-2000)	-17.4	-13.6	-7.1	3.2	12.5	17.7	20.5	19.3	13.1	4.9	-4.5	-13.4
20c3m (1961-2000)	-16.9	-13.5	-6.9	3.4	12.5	17.8	20.7	19.2	13.3	5.3	-3.7	-12.7
A2 (2045-2046)	-13.9	-10.3	-5.2	5.0	14.8	20.8	23.1	21.2	15.6	7.6	-1.9	-9.5
A2 (2081-2100)	-12.9	-9.3	-4.2	5.6	15.0	20.3	22.3	21.5	15.7	6.6	-3.7	-10.8
B1 (2045-2046)	-14.9	-11.4	-5.5	4.7	14.3	19.9	22.5	20.8	15.1	7.2	-2.1	-10.6
B1 (2081-2100)	-14.1	-11.0	-5.1	4.9	14.5	20.5	22.9	21.0	15.5	7.5	-2.6	-10.0
A1B (2045-2046)	-13.3	-10.7	-5.3	5.0	14.9	20.4	22.9	21.1	15.7	7.4	-1.9	-9.8
A1B (2081-2100)	-12.8	-9.4	-4.4	5.8	15.4	21.7	23.7	22.0	16.3	8.1	-1.5	-9.1

Table 6.12: Observed and downscaled mean monthly temperature at Sturgeon River 2 (°C).

Scenario	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Observed (1970-2000)	-17.5	-13.7	-7.3	3.0	12.4	17.5	20.3	19.2	12.9	4.7	-4.6	-13.4
20c3m (1961-2000)	-17.0	-13.7	-7.1	3.2	12.3	17.7	20.5	19.0	13.1	5.0	-3.9	-12.8
A2 (2045-2046)	-14.0	-10.4	-5.3	4.7	14.6	20.6	22.9	21.0	15.4	7.4	-2.0	-9.6
A2 (2081-2100)	-12.9	-9.4	-4.3	5.4	14.9	20.2	22.0	21.4	15.5	6.5	-3.8	-10.8
B1 (2045-2046)	-14.9	-11.5	-5.6	4.5	14.2	19.8	22.2	20.6	14.8	7.0	-2.2	-10.7
B1 (2081-2100)	-14.1	-11.1	-5.2	4.6	14.4	20.3	22.7	20.9	15.3	7.3	-2.3	-10.0
A1B (2045-2046)	-13.4	-10.8	-5.4	4.7	14.7	20.3	22.7	20.9	15.5	7.2	-2.0	-9.8
A1B (2081-2100)	-12.8	-9.5	-4.5	5.5	15.3	21.5	23.4	21.8	16.1	8.0	-1.5	-9.1

Table 6.13: Observed and downscaled mean monthly temperature at Sturgeon River 3 (°C).

Scenario	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Observed (1970-2000)	-17.7	-13.9	-7.4	2.9	12.3	17.4	20.2	18.9	12.7	4.6	-4.8	-13.7
20c3m (1961-2000)	-17.3	-13.9	-7.2	3.1	12.2	17.5	20.3	18.8	13.0	4.9	-4.0	-13.1
A2 (2045-2046)	-14.2	-10.6	-5.4	4.6	14.5	20.4	22.7	20.8	15.2	7.2	-2.1	-9.8
A2 (2081-2100)	-13.1	-9.5	-4.3	5.4	14.8	20.1	22.0	21.3	15.4	6.3	-3.9	-11.0
B1 (2045-2046)	-15.2	-11.7	-5.7	4.4	14.0	19.6	22.1	20.3	14.7	6.8	-2.4	-11.0
B1 (2081-2100)	-14.3	-11.3	-5.4	4.5	14.2	20.1	22.5	20.6	15.1	7.1	-2.4	-10.3
A1B (2045-2046)	-13.6	-11.0	-5.5	4.7	14.6	20.1	22.5	20.7	15.3	7.0	-2.2	-10.1
A1B (2081-2100)	-13.0	-9.7	-4.7	5.4	15.1	21.3	23.3	21.6	16.0	7.8	-1.7	-9.4

Table 6.14: Observed and downscaled mean monthly temperature at Troutlake River 1 (°C).

Scenario	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Observed (1970-2000)	-18.7	-14.6	-7.8	2.8	12.6	18.0	21.2	19.9	12.9	4.5	-5.6	-14.8
20c3m (1961-2000)	-18.3	-14.6	-7.5	3.0	12.5	18.3	21.3	19.6	13.1	4.7	-4.6	-14.3
A2 (2045-2046)	-15.2	-11.2	-5.8	4.6	14.9	21.3	23.9	21.8	15.5	6.8	-2.7	-11.0
A2 (2081-2100)	-13.8	-9.8	-4.6	5.4	15.2	20.8	23.7	22.4	15.9	6.1	-4.3	-11.8
B1 (2045-2046)	-16.1	-12.3	-6.1	4.3	14.4	20.4	23.2	21.3	15.0	6.6	-2.9	-12.1
B1 (2081-2100)	-15.2	-12.0	-5.7	4.5	14.6	20.9	23.5	21.7	15.4	6.8	-2.9	-11.3
A1B (2045-2046)	-14.5	-11.6	-5.8	4.6	15.0	20.9	23.5	21.7	15.5	6.6	-2.8	-11.3
A1B (2081-2100)	-14.1	-10.2	-5.1	5.4	15.6	22.3	24.5	22.8	16.4	7.4	-2.2	-10.5

Table 6.15: Observed and downscaled mean monthly temperature at Troutlake River 2 (°C).

Scenario	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Observed (1970-2000)	-18.5	-14.6	-7.9	2.6	12.1	17.5	20.4	19.3	12.4	4.1	-5.8	-14.8
20c3m (1961-2000)	-18.1	-14.6	-7.6	2.8	12.0	17.7	20.5	19.1	12.7	4.4	-4.9	-14.2
A2 (2045-2046)	-14.9	-11.2	-5.9	4.4	14.4	20.7	23.1	21.2	15.2	6.7	-2.8	-10.9
A2 (2081-2100)	-13.4	-9.8	-4.7	5.2	14.7	20.2	22.8	21.7	15.5	5.9	-4.4	-11.6
B1 (2045-2046)	-15.9	-12.3	-6.2	4.1	14.0	20.0	22.4	20.7	14.6	6.4	-3.1	-12.0
B1 (2081-2100)	-15.0	-11.9	-5.8	4.3	14.1	20.3	22.8	21.1	15.1	6.7	-3.1	-11.3
A1B (2045-2046)	-14.3	-11.6	-5.9	4.5	14.6	20.3	22.8	21.1	15.2	6.5	-2.9	-11.2
A1B (2081-2100)	-13.7	-10.2	-5.1	5.2	15.1	21.6	23.7	22.2	16.1	7.3	-2.3	-10.3

Table 6.16: Observed and downscaled mean annual precipitation accumulation for NARR application (mm).

Scenario	S1	S2	S3	T1	T2
Observed (1970-2000)	661	669	652	601	618
20c3m (1961-2000)	639	648	631	583	601
A2 (2045-2046)	708	613	601	582	593
A2 (2081-2100)	733	741	726	680	695
B1 (2045-2046)	617	627	610	571	585
B1 (2081-2100)	628	637	624	601	612
A1B (2045-2046)	631	636	624	606	617
A1B (2081-2100)	586	585	579	613	614

Table 6.17: Observed and downscaled mean winter precipitation accumulation for NARR application (mm).

Scenario	S1	S2	S3	T1	T2
Observed (1970-2000)	256	265	251	224	228
20c3m (1961-2000)	235	243	229	205	210
A2 (2045-2046)	250	259	246	219	225
A2 (2081-2100)	224	232	223	203	208
B1 (2045-2046)	245	253	240	214	220
B1 (2081-2100)	257	265	252	222	230
A1B (2045-2046)	248	256	245	221	228
A1B (2081-2100)	253	260	249	223	229

Chapter 7

k-Nearest Neighbor Model Discussion

7.1 Model Evaluation

The goal of this section is to demonstrate the value of downscaling raw GCM output with the k-nn model. The model is first evaluated in terms of its ability to reduce the bias in the GCM at weather station locations for the current climate. Then the temperature and precipitation data generated by the k-nn model for the various emission scenarios for the 2081 to 2100 time slice is compared with the range of temperature and precipitation trends simulated by a variety of GCMs for the region containing the weather stations.

7.1.1 Comparison to Raw GCM Data

One of the primary goals of statistical downscaling is to improve the quality of raw GCM output. A statistical downscaling model should reproduce statistics such as mean monthly temperature or mean monthly precipitation accumulations better than the raw GCM data.

To determine if the k-nn model is an improvement compared to the raw CGCM3.1 data, the downscaled raw GCM data for the 20c3m emission scenario were extracted for the grid points nearest to the Thompson and Sioux Lookout weather stations. Theoretically, the raw and downscaled GCM data should reproduce the monthly mean data at the weather stations, although some degree of bias is expected.

The comparison between the downscaled and raw CGCM3.1 for temperature at the two weather stations is shown on Figure 7.1. It is clearly demonstrated that for mean monthly temperature the downscaled data is much closer to the observed means than the raw GCM data. The CGCM3.1 underestimates temperature most of the year, although it overestimates temperature in the early winter months. The downscaled data also tends to overestimate in the winter months, but the biases of the downscaled data are much smaller throughout the entire year for both stations. In most months the downscaled data is within 0.5°C of the observed monthly mean.

The comparison between the downscaled and raw CGCM3.1 precipitation data at the two weather stations is shown on Figure 7.2. At the Thompson weather station the CGCM3.1 grid cell underestimates the monthly precipitation for most months. During the months of June through to September the underestimation is

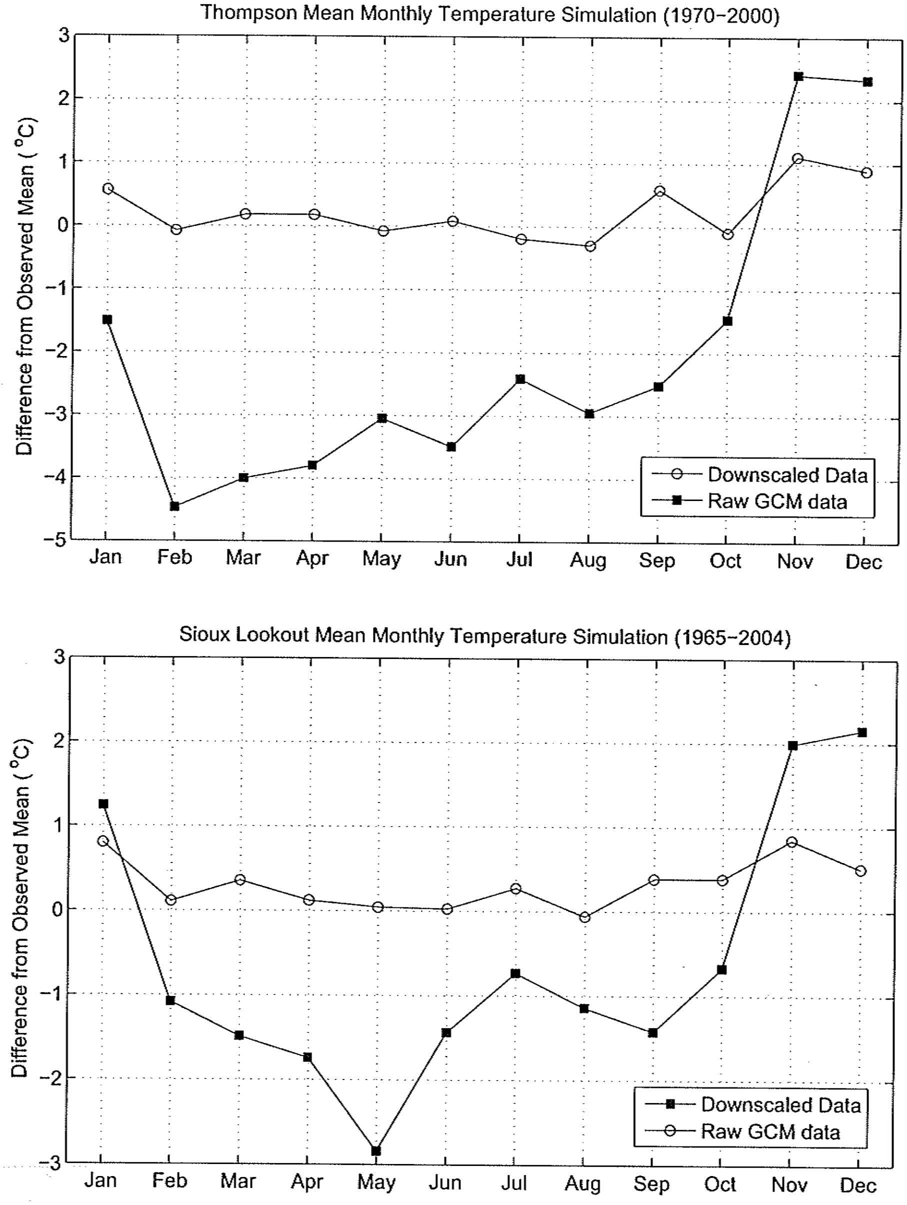


Figure 7.1: Comparison of downscaled and raw CGCM3.1 temperature data.

large. The downscaled data also underestimates precipitation throughout the year, however the large underestimation in the GCM data are significantly reduced in the downscaled data. The annual observed precipitation at Thompson is 512 mm. The annual mean for the CGCM3.1 grid cell is 455 mm for the 20c3m scenario while the downscaled 20c3m scenario data has an annual mean of 467 mm. By downscaling the GCM, the annual underestimation of precipitation was reduced from 11% to 9%. At the Sioux Lookout weather station location the CGCM3.1 20c3m scenario data underestimates precipitation for all months. Overall, the downscaled data also underestimates precipitation but provides better results than the raw GCM data. Similar to the Thompson station location, the downscaled data reduces large biases present in the GCM data during the late summer months. For the mean monthly precipitation in September, an underestimation of nearly 30 mm by the CGCM3.1 is reduced to almost zero in the downscaled data. The annual observed precipitation at Sioux Lookout is 741 mm. The annual mean for the CGCM3.1 grid cell is 640 mm for the 20c3m scenario while the downscaled 20c3m scenario data has an annual mean of 704 mm. By downscaling the GCM, the annual underestimation of precipitation was reduced from 14% to 5%.

The above discussion demonstrates the ability of the k-nn model to improve the output of CGCM3.1 data. The downscaled data better simulates mean monthly statistics, especially for temperature. Biases are present in downscaled precipitation data, however the biases are less than those present in the GCM data.

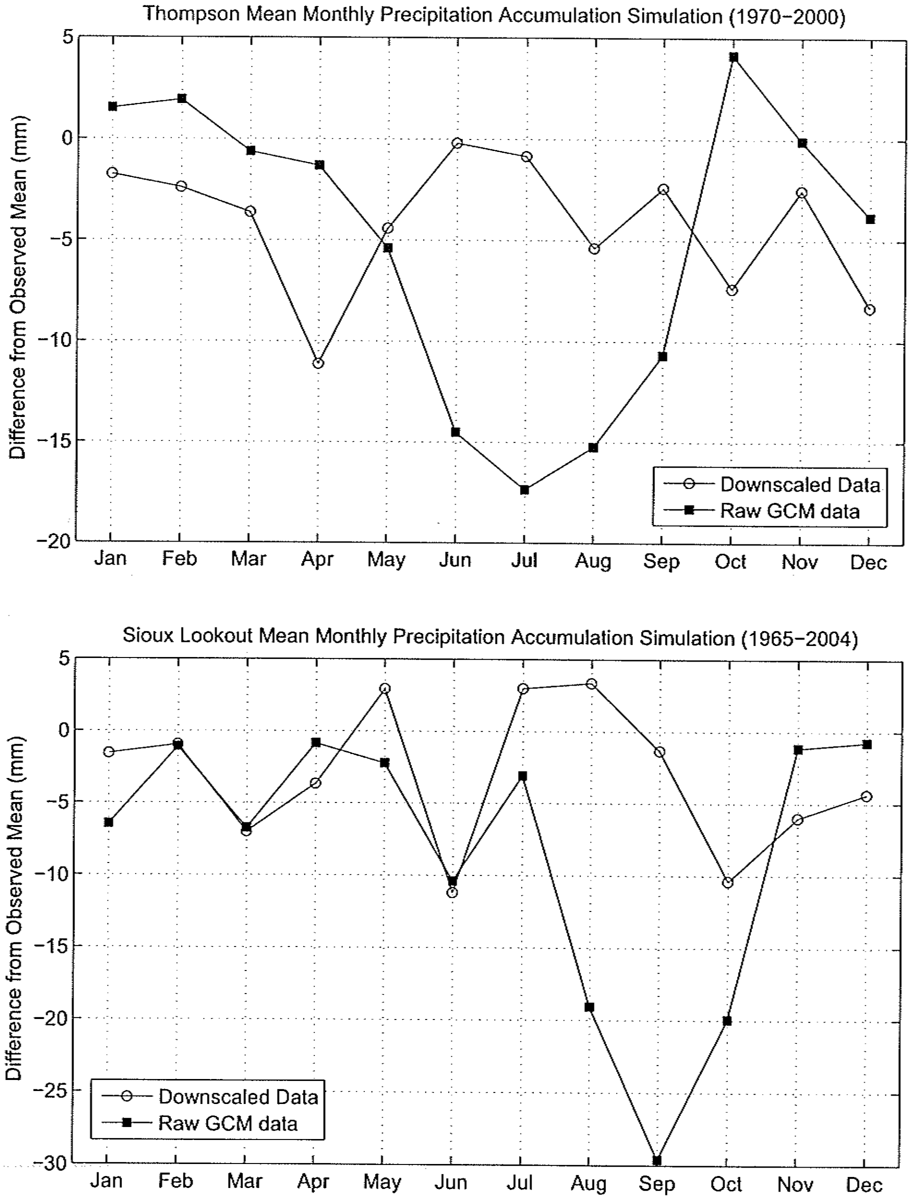


Figure 7.2: Comparison of downscaled and raw CGCM3.1 precipitation data.

7.1.2 Comparison to Other Simulations of Future Climate

The future climate is unknown, making it difficult to determine the validity of the data generated by downscaling GCM output based on the future emission scenarios. The best available comparison is to other simulations of future climate. Sauchyn and Kulshreshtha (2008) summarized the simulations made by a variety of GCMs for multiple emission scenarios for the Prairie Provinces of Canada (see Tables 2.1 and 2.2 on page 11). Chiotti and Lavender (2008) provide a similar summary for the western portion of Ontario. These summaries are for large geographical areas, but the k-nn downscaling model should produce similar trends to the raw GCM simulation data.

For temperature in the forest region of the Prairie Provinces, GCM simulated temperature increases ranged from 2.3 to 10.8°C for the 2080s time horizon depending on the different emission scenarios and GCMs. The temperature predictions for Thompson and The Pas fit within this range. For western Ontario, temperature simulations of GCMs show increases from 2.7 to 11.8°C for the 2080s time horizon (Chiotti and Lavender, 2008). The simulations for Red Lake and Sioux Lookout are within this range.

For precipitation in the forest region of the Prairie Provinces, increases range from 2% to 25% for the 2080s horizon (Sauchyn and Kulshreshtha, 2008). Depending on the different emission scenarios, during the Nelson River application in northern Manitoba precipitation data downscaled by the k-nn model for the 2081 to 2100 time period decreased from 7% to 12% at Thompson weather station, and decreased from 8% to 23% at The Pas weather station. For the multi-site application of the k-nn algorithm in the Nelson River drainage basin produced drier future climates than the GCMs summarized by Sauchyn and Kulshreshtha.

For western Ontario, precipitation simulated by the GCMs showed increases of 5% to 23% for the 2080s time horizon (Chiotti and Lavender, 2008). For the Winnipeg River application in western Ontario, the Sioux Lookout precipitation data downscaled by the k-nn model for the 2081 to 2100 time period showed increases of 4% to 14%, while the Redlake changes ranged from a decrease of 3% to an increase of 10%. Some of the downscaled scenarios were below the range of GCMs presented by Chiotti and Lavender (2008), however, for the most part the downscaled precipitation data were similar to GCM simulations.

A possible explanation to the simulation of drier future climates for Thompson and The Pas than simulated by raw GCM data may be drawn from the analysis of the downscaled precipitation data in the single-site application for Thompson. In general, days with precipitation tend to be overcast and cooler. Downscaling of the A2 scenario in the single-site application resulted in an increase of the frequency of dry days compared to the 20c3m scenario. Since temperatures increase substantially in future scenarios, the model may be less inclined to resample cool wet days. This may result in the simulation of a drier climate as temperature increases.

7.1.3 Evaluation Conclusion

As expected of statistical downscaling models, the k-nn model removes much of the bias present in the CGCM3.1 for the 20c3m scenario. Aside from the possibility that precipitation is underestimated for future scenarios in the Nelson River applications, the k-nn downscaling model simulations are consistent with the range of future climates simulated by a variety of GCMs. Overall, the GCM output downscaled with the k-nn model is an improvement over the raw GCM output.

7.2 k-Nearest Neighbor Application Recommendations

The following section is a summary of the methodology applied for downscaling GCM data using a k-nearest neighbor model. Recommendations based on the lessons learned are also made where appropriate. Section 3.2 provides a literature review of k-nn modelling.

7.2.1 Feature Vector Selection

The feature vector, D_t, is used to compare the simulated day to historical days. The selection of large-scale variables contained in the feature vector is an important decision in the modelling process. The spatial extent that the feature vector variables represent is also an important factor.

Multivariate statistical analysis methods can be utilized to aid in deciding the composition of the feature vector. In the applications presented in this report, canonical circulation analysis and circulation pattern classification analysis were used to identify the existence of relationships between the large-scale and local-scale climate variables.

The length of the feature vector is an important consideration. Longer feature vectors lead to greater computation time, especially during the highly iterative optimization process. It was found in this study that the 6 × 7 grid and 6 × 10 grid of GCM data gave similar results during cross-validation. However, after principal component analysis the smaller grid required a smaller feature vector and led to much faster computation times. It is recommended to use statistical analysis, or small scale pilot applications, to determine the minimum grid region required to maintain adequate cross-validation results.

7.2.2 Finding the k-Nearest Neighbors

The selection of the nearest neighbors was dependent on three parameters: window width, the weighting vector and the number of neighbors retained. Optimization of an objective function was used to determine the best values of these parameters.

Since large-scale and local temperature variables were highly correlated, very good results were obtained during cross-validation for downscaling of daily temperature. Therefore the objective functions focussed on maximizing the models' performance related to precipitation. Measuring the performance of the models for downscaling precipitation on a daily scale was difficult since the results were highly sensitive to large one-day precipitation events. It was decided to use precipitation statistics determined on a seasonal scale to limit the influences of the large one day events. Since most of the large precipitation events occur during the summer season, it was decided to use only the winter season including the months from October to April. During these months, precipitation is driven by large synoptic scale weather systems rather than convective storms, making the precipitation occurrence easier to downscale accurately. The winter precipitation is also important to the hydrological cycle of the study area due to the storage of precipitation as snow and release as spring melt water. For these reasons, the recommended objective function is to maximize the performance of the model to downscale winter precipitation. This was done by defining the objective function for cross-validation as the root mean squared error of each season's simulated and observed accumulated precipitation.

Since the window width, W, is limited to discrete numbers over a relatively small range, it was easily optimized by manually adjusting it to determine the optimum width. In the different applications of the model, it was found that W was directly related to the length of historical records available to resample from. The summary of the optimum window width and the number of years in the historical record used can be viewed in Table 7.1 or on Figure 7.3. The relationship is almost perfectly linear. The relationship can be expressed as

$$W = 42 - 0.57N \tag{7.1}$$

where N is the number of years in the historical record.

Table 7.1: Summary of historical record length and optimal W.

Application	**Record Length** (years)	**Optimal** W (days)
Single-site Nelson River	37	21
Multi-site Nelson River	30	25
Multi-site Winnipeg River	40	19
NARR Winnipeg River	26	27

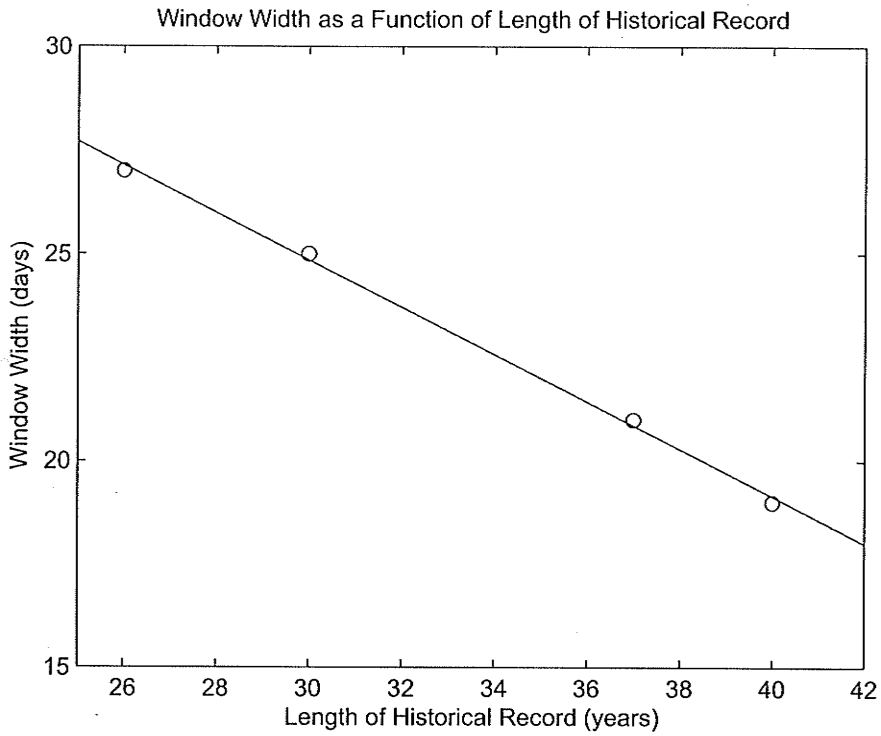


Figure 7.3: Scatter plot of W vs. number of years in the historical record.

The weighted Euclidean distance was used to calculate the distance between the simulated and historical feature vectors,

$$\delta_{tu} = \sqrt{\sum_{i=1}^{n} w_i \left(v_{ti} - v_{ui}\right)^2} \tag{7.2}$$

where n is the number of variables in the feature vector, and w_i are the weight given to the variables of v_i. The w vector can be optimized to minimize an objective function. With a large number of variables, optimization was complex and the use of computer software was necessary. The Matlab Optimization Toolbox minimized the objective functions using a steepest descent line search method. Since the process was highly iterative, and each iteration involved simulating between 26 and 40 years of data, optimization was a time consuming process.

Cross-validation showed that retaining the single nearest neighbor ($k = 1$) led to the best objective function results. However, to promote variability in the modelling process, a larger k is recommended during the downscaling of GCM data. A k-value equal to ten was used in the applications presented in this report. The performance of the model was not significantly affected by using a larger k-value.

7.2.3 Choosing a Neighbor

Once the k nearest neighbors were determined, the next step was to resample one of the neighbors. A decreasing kernel density function was used to assign weight to the first ten nearest neighbors. To save computation time, the decreasing density

$$p_j = \frac{1/j}{\sum_{i=1}^{k} 1/i}, \tag{7.3}$$

where p_j is the probability that day j is resampled, was employed rather than a density function dependent on calculated distances.

Once the neighbor was selected, the desired variables were retrieved from the historical weather record. This process was repeated for each simulation day. Multiple variables were resampled at once during the applications presented. For instance, in the NARR application twenty variables were resampled at one time.

Chapter 8

Conclusion

The most efficient way to conclude the study is to review if the original objectives set before work began were met through the course of the project. The objectives of this report were to:

1. Review climate change principles,
2. Review downscaling techniques,
3. Explore GCM data,
4. Explore relationships in large-scale and local weather variables, and
5. Develop and apply a k-nn downscaling model.

Each of these objectives was completed successfully. This section will review the accomplishments made towards the above objectives.

8.1 Climate Change and Statistical Downscaling

In Chapter 2 climate change was discussed in terms of its global significance as well as how global changes could impact the Canadian Prairies. On the Canadian Prairies, temperatures will likely rise up to 9°C, while precipitation changes could range between a decrease of 6% to an increase of 29% (Sauchyn and Kulshreshtha, 2008). One of the greatest threats of climate change to the prairies is its impacts to water resources.

GCMs are large computer models used to forecast long periods of weather over the entire globe. To model the Earth's complex climate systems in a reasonable amount of time the models use coarse grid resolutions, typically one to five degrees in latitude and longitude. These models are used to simulate future climate dependent on greenhouse gas emission scenarios specified by the IPCC. The coarse resolution of the GCMs make their output impractical for direct application in water resources. Therefore the GCM output must be post-processed before it is utilized in hydrological models. The post-processing consists of downscaling the data by dynamic or statistical downscaling. Dynamic downscaling involves nesting a finer resolution regional climate model (RCM) within a GCM. Statistical downscaling methods use the statistical relationships that exist between large-scale and local climate variables to downscale GCMs.

A variety of statistical methods can be used to downscale GCM output. Common categories of models include transfer functions, weather typing, and weather generators. The methodology and past applications of nearest neighbor resampling was thoroughly researched and reported on as another possible statistical downscaling model. Nearest neighbor resampling is a less common method to downscale data, but its relative simplicity and flexibility are advantages compared to the other methodologies.

8.2 GCM Data Analysis

Several aspects of the large-scale and local climate variables were explored. Both spatial and temporal biases were identified in the CGCM3.1/T47 20c3m control run data compared to NCEP/NCAR Reanalysis 1 and weather station data. Circulation pattern classification demonstrated geopotential height data contain useful information for predicting the occurrence and depth of daily rainfall. Canonical correlation analysis demonstrated that the combination of large-scale temperature data at multiple levels and geopotential height data can describe much of the variation in temperature and precipitation processes at weather stations.

Two important conclusions were reached during the data exploration exercises. The first is that data must be standardized using a daily mean and standard deviation to remove bias. Secondly, data exploration identified relationships between the large-scale and local climate variables which leads to confidence that the large-scale data grids selected are appropriate to downscale temperature and precipitation data at weather stations.

8.3 *k*-Nearest Neighbor Downscaling Results

The *k*-nearest neighbor model was used in four downscaling applications. The results from the applications demonstrate that *k*-nearest neighbor is a viable methodology to downscale GCM data.

The first pilot application of the methodology was to downscale the CGCM3.1/T47 output to generate temperature and precipitation at the Thompson weather station. During this application an optimization methodology was developed and it was demonstrated that the downscaling model could reproduce historical weather variables. It was determined that precipitation was the most difficult and important variable to optimize, as historical temperature is reproduced well without significant optimization.

The next undertaking was a multi-site application in the Nelson River Basin. In this application, the variables required for the SLURP hydrological model, temperature, precipitation, solar radiation, and relative humidity, were downscaled for the locations of the Thompson and The Pas weather stations for multiple climate change scenarios generated by the CGCM3.1/T47. The *k*-nn model was optimized using only statistics for precipitation as this variable is both the most important and most difficult to model. This approach was found to not have a negative effect on the model's ability to reproduce daily temperature, solar radiation, or relative humidity. The inclusion of data from two weather stations slightly improved the optimization performance compared to the single site application. Temperatures were found to increase throughout the 21^{st} century for all future emission scenarios. Overall, the A2 and A1B showed more warming than the B1 scenario.

This is consistent with global trends simulated by the Intergovernmental Panel on Climate Change (IPCC, 2007). Precipitation was found to be underestimated by the 20c3m scenario compared to the observed record. When comparing the 20c3m scenario to the future GCM scenarios, a decrease in all future scenarios except for the B1 2046-2065 scenario were observed. Precipitation decreased more in the later half of the 21^{st} century. The precipitation results were near the lower bounds of future precipitation simulations in literature.

In the next application of the k-nearest neighbor downscaling model, GCM data was downscaled for hydrological modelling in the Winnipeg River Basin at Redlake and Sioux Lookout weather stations. The GCM data in this application was centered over the weather stations with a slightly smaller grid than was used in the Nelson River application. This resulted in a smaller feature vector and faster optimization of the w vector, without a reduction in model performance. For future temperature, increases occurred in all future scenarios. As in the Nelson River Drainage Basin, the A2 and A1B scenarios were warmer than the B1 scenario. Maximum increases in temperature were up to 6.4°C at Redlake for the A2 scenario from 2081 to 2100. The results for precipitation showed that the Winnipeg River Basin could expect to see precipitation accumulations increase up to 14%. The data showed that variable future winter precipitation, with some scenarios showing increases of 13% and some showing decreases of almost 5% compared to the downscaled 20c3m scenario.

The fourth application of the k-nn model was to downscale GCM data with NARR data replacing station data. NARR data could serve as an alternative to station data for climate change assessments in areas such as northern Canada that lack weather station data. The utility of NARR data in this capacity was tested in this application. The results for the application with NARR data were consistent with the results obtained in the Winnipeg River weather station application. All emission scenarios show increases in future temperature. The B1 scenario shows the least amount of warming and the A1B and A2 show the most warming. Precipitation varied between scenarios, either slightly increasing or remaining relatively the same compared to historical means. Different emission scenarios had similar warming trends when using the station data and the NARR data for downscaling. In both applications, precipitation either remains approximately the same or increases moderately. This application demonstrated that NARR data can be a viable replacement for station data in climate change assessments. The application also demonstrated the advantage the k-nn model has of downscaling many variables at one time. Four variables were downscaled at each of the five NARR grid points, for a total of twenty variables produced simultaneously.

8.4 Challenges and Significance of Research

The objective of applying a k-nn model to downscale GCM data offered many challenges. Climate change modelling, assessment of climate change impacts, and downscaling of GCM are all relatively new branches of science and engineering. Before the project was started, much energy was spent acquiring expertise in climate change, multivariate statistics, advanced numerical methods, and Matlab programming.

The organization and processing of climate data requires significant time and attention. GCM data had to be validated to inspect that it was functioning properly in the study region. The pre-processing of data before use in the downscaling model included application of complex standardization and principal component analysis. The volume of data climate data used in the study was large. In total, thousands of years of climate data were processed and downscaled; with each day represented by many large grids of GCM or NCEP/NCAR reanalysis data, and weather station or NARR data.

Another challenge in the development and application of the *k*-nn was the lack of previous applications of the methodology to downscale GCM data. The software packages that are available for other methods such as weather generators are not available for *k*-nn. While the premise of the method is relatively simple, the successful building of the model required developing Matlab code to run the model. Optimization of a *k*-nn model was not present in any of the literature found. The development of an optimization procedure was one of the greatest contributions of the research and should be used in future applications of the nearest neighbor resampling model. An optimization procedure similar to the one presented in this work could also be useful where nearest neighbor resampling is used in an application other than downscaling climate data.

The research presented is a valuable contribution to climate change assessment research and to the study of climate change on the Canadian Prairies. As climate change concerns continue to grow, assessments of many systems that interact with the environment and climate will be required. This may include infrastructure, agriculture, forestry, ecosystems, and of course water resources. With the uncertainty involved in simulating future climate, the best approach is to use as many tools as possible to develop a full ensemble of possible futures. The research demonstrated that a k-nn downscaling model can successfully downscale large-scale climate. The k-nn downscaling model should be included in future climate change assessments along side the other common downscaling models.

References

Brandsma T. and T.A. Buishand (1998) Simulation of extreme precipitation in the Rhine basin by nearest-neighbor resampling, Hydrology and Earth System Sciences, 2(2-3), 195-209.

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

Caya, D. and R. Laprise (1999) A semi-implicit semi-lagrangian regional climate model: The Canadian RCM, Monthly Weather Review, 127, 341-362.

Chiotti, Q. and B. Lavender (2008) Ontario; in From Impacts to Adaptation: Canada in a Changing Climate 2007, edited by D.S. Lemmen, F.J. Warren, J. Lacroix, and E. Bush, Government of Canada, Ottawa, Ontario, 227-274.

Choi, W., A. Moore, and P. F. Rasmussen (2007) Evaluation of temperature and precipitation data from NCEP-NCAR global and regional reanalyses for hydrological modeling in Manitoba, Proceedings of 18th Canadian Hydrotechnical Conference, Winnipeg, Canada, 10 pages.

Choi, W., S.J. Kim, P.F. Rasmussen, and A.R. Moore (2009) Use of the North American Regional Reanalysis for hydrological modelling in Manitoba, Canadian

Water Resources Journal, 34(1), 17-36.

Chong, E. and Z. Stanislaw (2001) An Introduction to Optimization, John Wiley & Sons, Inc., New York, NY.

Flato, G.M., G.J. Boer, W.G. Lee, N.A. McFarlane, D. Ramsden, M.C. Reader, and A.J. Weaver (2000) The Canadian Centre for Climate Modelling and Analysis Global Coupled Model and its climate, Climate Dynamics, 16, 451-467.

Gangopadhyay, S., M. Clark, and B. Rajagopalan (2005) Statistical downscaling using k-nearest neighbors, Water Resource Research, 41(2), 1-23.

Hughes, J.P. and P. Guttorp (1994) A class of stochastic models for relating synoptic atmospheric patterns to regional hydrological phenomenon, Water Resources Research, 30(5), 1535-1546.

Hughes, J.P., P. Guttorp, and S. Charles (1999) A non-homogeneous hidden Markov model for precipitation occurrence, Applied Statistics, 28(1), 15-30.

IPCC (Intergovernmental Panel on Climate Change) (2000) Emission Scenarios: A Special Report of Working Group III of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, UK.

IPCC (Intergovernmental Panel on Climate Change) (2001),Climate Change 2001: Synthesis Report, Summary for Policymakers.

IPCC (Intergovernmental Panel on Climate Change) (2007) Climate Change 2007: Synthesis Report. Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, Pachauri, R.K and A. Reisinger(eds.)], IPCC, Geneva, Switzerland, 104 pages.

Johnson, R.A. and D.W. Wichern (1988) Applied Multivariate Statistical Analysis, 2nd Ed., Prentice-Hall, New Jersey.

Kalnay, E. , M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, A. Leetmaa, R. Reynolds, M. Chelliah, W. Ebisuzaki, W.Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, and J. Wang (1996) The NCEP/NCAR 40-Year Reanalysis Project, Bulletin of the American Meteorological Society, 77, 437-471.

Kim, S. J., W. Choi, and P.F. Rasmussen (2007) Calibration of the SLURP hydrological model using North American Regional Reanalysis (NARR) data, Proceedings of 18th Canadian Hydrotechnical Conference, Winnipeg, Canada, 10 pages.

Kim, S. J., M. Lee, W. Choi, and P.F. Rasmussen (2008) Utilizing North American Regional Reanalysis for climate change impact assessment on water resources in Central Canada, Proceedings of International Water Resources Association World Water Congress, Montpellier, France, 14 pages.

Environment Canada (2000) Western Canadian Daily Climate Data CD-ROM.

Lall, U., B. Rajagopalan, and D.G. Tarboton (1996) A nonparametric wet/dry spell model for resampling daily precipitation, Water Resources Research, 32(9), 2803-2823.

Lall, U. and A. Sharma (1996) A nearest neighbor bootstrap for resampling hydrologic time series, Water Resource Research, 32(3), 679-693.

Lee, M. and P.F. Rasmussen (2007) Downscaling of GCM data in the Nelson River drainage basin by nearest neighbor resampling, Proceedings of 18th Canadian Hydrotechnical Conference, Winnipeg, Canada, 10 pages.

Lemmen, D.S. and F.J. Warren (2004) Climate Change Impacts and Adaptation: A Canadian Perspective, Government of Canada, Ottawa, ON, 174 pages.

Meehl, G.A., W.M. Washington, C.A. Ammann, J.M. Arblaster, T.M.L. Wigleym, and C. Tebaldi (2004). Combinations of natural and anthropogenic forcings in twentieth-century climate, Journal of Climate, 17, 3721-3727.

Mesinger, F., G. DiMego, E. Kalnay, K. Mitchell, P.C. Shafran, W. Ebisuzaki, D. Jovi, J. Woollen, E. Rogers, E.H. Berbery, M.B. Ek, Y. Fan, R. Grumbine, W. Higgins, H. Li, Y. Lin, G. Manikin, D. Parrish, and W. Shi (2006) North American Regional Reanalysis, Bulletin of the American Meteorological Society, 87, 343360.

Nicks, A.D. and G.A. Gander (1994) CLIGEN: A weather generator for climate inputs to water resource and other models, Proceedings of the Fifth International Conference on Computers in Agricultural Engineering, Orlando, Florida, American Society of Agricultural Engineering.

Nocedal, J. and S. Wright (1999) Numerical Optimization, Springer Series in Operations Research, Springer Science + Business Media, Inc. New York, NY.

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

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

Sauchyn, D. and S. Kulshreshtha (2008) Prairies; in From Impacts to Adaptation: Canada in a Changing Climate 2007, edited by D.S. Lemmen, J. Lacroix and E. Bush, Government of Canada, Ottawa, Ontario, 275-328.

Semenov, M.A. and R.J. Brooks (1999) Spatial interpolation of the LARS-WG stochastic weather generator in Great Britain, Climate Research, 11, 125-136.

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

Sharpley, A.N. and J.R. Williams (1990) EPIC-Erosion Productivity Impact Calculator, Model Documentation, US department of Agriculture, Technical Bulletin No. 1768, 235 pages.

Wilby, R.L., C.W. Dawson, and E.M. Barrow (2002) SDSM - a decision support tool for the assessment of regional climate change impacts, Environmental and Modelling Software, 17, 145-157.

Wilby R.L. and T.M.L. Wigley (1997) Downscaling general circulation model output: A review of methods and limitations, Progress in Physical Geography, 21(4), 530-548.

Wilks, D.S. (1995) Statistical Methods in the Atmospheric Sciences: An Introduction, American Press.

Wójcik, R. and T.A. Buishand (2003) Simulation of 6-hourly rainfall and temperature by two resampling schemes, Journal of Hydrology, 273(1-4), 29-80.

Yarnal, B. (1993) Synoptic Climatology in Environmental Analysis, Belhaven Press.

Yates, D., S. Gangopadhyay, B. Rajagopalan, and K. Strzepek (2003) A technique for generating regional climate scenarios using a nearest-neighbor algorithm, Water Resources Research, 39(7), SWC71-SWC715.

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