IMPROVING STATISTICAL DOWNSCALING OF GENERAL CIRCULATION MODELS

by

Matthew Lee Titus

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
August 2010

DALHOUSIE UNIVERSITY

DEPARTMENT OF PHYSICS AND ATMOSPHERIC SCIENCE

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "IMPROVING STATISTICAL DOWNSCALING OF GENERAL CIRCULATION MODELS" by Matthew Lee Titus in partial fulfillment of the requirements for the degree of Master of Science.

Dated: August 4, 2010

Supervisors:

Ian Folkins

Richard Greatbatch

Readers:

Jinyu Sheng

Harold Ritchie

DALHOUSIE UNIVERSITY

DATE: August 4, 2010

AUTHOR: Matthew Lee Titus

TITLE: IMPROVING STATISTICAL DOWNSCALING OF GENERAL CIRCULATION MODELS

DEPARTMENT OR SCHOOL: Department of Physics and Atmospheric Science

DEGREE: M.Sc. CONVOCATION: October YEAR: 2010

To my children, Noah, Maddison and the generations of children to come. This is my contribution to the climate problem which is arguably the largest problem ever faced on this planet. This work is part of global shift in conciousness already in progress needed to ensure your future on planet Earth.

TABLE OF CONTENTS

List of Tables . . . . . **vii**

List of Figures . . . . . **viii**

Abstract . . . . . **x**

Acknowledgements . . . . . **xi**

Chapter 1 Introduction **1**

1.1 Overview of Climate Change 1

1.2 The Case for Downscaling 4

1.3 Motivation 5

1.4 Objective and Approach 10

Chapter 2 Mathematical and Statistical Basis **11**

2.1 Multiple Linear Regression 11

2.2 Confidence Intervals and Hypothesis Testing 17

2.3 Principal Component Regression 24

Chapter 3 Data Investigation **29**

3.1 The Predictands 29

3.2 Maximum Daily Temperature 29

3.3 Minimum Daily Temperature 31

3.4 Tmax and Tmin Differences 32

3.5 The Predictors 34

Chapter 4 Methodology and Results **39**

4.1 The Predictor Selection Process 39

4.2 Regression Development 42

4.3 Predictor Physics 50

4.4 Validation of Regression and Regression Results 54

4.5 CGCM3 Hindcasting 64
4.6 Future Projections 70
4.7 Alternative Future Projections 73
4.8 Future Projections Discussion 76

Chapter 5 Conclusions **79**
5.1 Summary of the Downscaling Process 79
5.2 Conclusions 80
5.3 Future Work 82

Bibliography **84**

LIST OF TABLES

Table 3.1 Tmax Seasonal cycle regression coefficients 32

Table 3.2 Tmin Seasonal cycle regression coefficients 32

Table 3.3 Predictor names 37

Table 4.1 Fall and winter predictor names 43

Table 4.2 Spring and summer predictor names 44

Table 4.3 Winter Tmin regression from NCEP information 46

Table 4.4 Winter Tmax regression from NCEP information 47

Table 4.5 Spring Tmin regression from NCEP information 47

Table 4.6 Spring Tmax regression for NCEP information 48

Table 4.7 Summer Tmin regression from NCEP information 48

Table 4.8 Summer Tmax regression from NCEP information 49

Table 4.9 Fall Tmin regression from NCEP information 49

Table 4.10 Fall Tmax regression from NCEP information 50

Table 4.11 Predictor physics for winter and spring 53

Table 4.12 Predictor physics for summer 53

Table 4.13 Predictor physics for fall 54

Table 4.14 Validation Tmin 55

Table 4.15 Validation Tmax 55

Table 4.16 Future Tmin results 72

Table 4.17 Future Tmax results 73

Table 4.18 Alternative future Tmin results 75

Table 4.19 Alternative future Tmax results 76

Table 4.20 Comparison of the projected change in annual mean by the 2080's . 78

Table 5.1 Comparison of SDSM Tmax daily anomaly predictions and the thesis method daily anomaly prediction of Tmax 82

LIST OF FIGURES

Figure 1.1 Contours of change in the annual temperature mean over the entire globe 2

Figure 1.2 How temperature distribution changes influence climate 3

Figure 1.3 Map of Atlantic Canada with CGCM3 grid over laid 4

Figure 1.4 Motivation for statistical downscaling 9

Figure 2.1 A visual example of linear regression 12

Figure 2.2 Grapical representation of the sum of squares 16

Figure 2.3 Z-distribution critical values for a 95 percent confidence interval . 19

Figure 2.4 Comparison of the Z and t-distributions 21

Figure 2.5 Plot of the F-distribution 22

Figure 2.6 Geometry of Principal Component Analysis 27

Figure 3.1 Observed Tmax from Shearwater NS. 31

Figure 3.2 Observed Tmin from Shearwater NS. 33

Figure 3.3 CGCM3 land-sea mask. 38

Figure 4.1 NCEP and CGCM3 surface specific humidity distribution comparison 42

Figure 4.2 Time series validation comparison for winter 57

Figure 4.3 Time series validation comparison for summer 58

Figure 4.4 Observed Tmin and Tmax versus NCEP predicted (downscaled) in winter. 60

Figure 4.5 Observed Tmin and Tmax versus NCEP predicted (downscaled) in spring. 61

Figure 4.6 Observed Tmin and Tmax versus NCEP predicted (downscaled) in summer. 62

Figure 4.7 Observed Tmin and Tmax versus NCEP predicted (downscaled) in fall. 63

Figure 4.8 Observed Tmin and Tmax versus CGCM3 predicted versus raw CGCM3 in winter. 65

Figure 4.9 Observed Tmin and Tmax versus CGCM3 predicted versus raw CGCM3 in spring. 66

Figure 4.10 Observed Tmin and Tmax versus CGCM3 predicted versus raw CGCM3 in summer. 67

Figure 4.11 Observed Tmin and Tmax versus CGCM3 predicted versus raw CGCM3 in fall. 68

Figure 5.1 Flow chart (part 1) overview of statistical downscaling 79

Figure 5.2 Flow chart (part 2) overview of statistical downscaling 80

ABSTRACT

Credible projections of future local climate change are in demand. One way to accomplish this is to statistically downscale General Circulation Models (GCM's). A new method for statistical downscaling is proposed in which the seasonal cycle is first removed, a physically based predictor selection process is employed and principal component regression is then used to train the regression. A regression model between daily maximum and minimum temperature at Shearwater, NS, and NCEP principal components in the 1961-2000 period is developed and validated and output from the CGCM3 is then used to make future projections. Projections suggest Shearwater's mean temperature will be five degrees warmer by 2100.

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank my supervisors, Dr. Ian Folkins, Dr. Richard Greatbatch, and Dr. Jinyu Sheng for their scientific guidance and patience. Although my work had many obstacles and debates, I have learned much in the process.

Thanks to Gary Lines of Environment Canada, and Environment Canada itself for support of this research. A special thanks to Dr. Keith Thompson who helped bring multivariate statistics to a practical level for my work.

I appreciate my friends and peers who helped me many times along the way, especially Eric Oliver, Faez Bakalian, and Jennifer Mecking. Last but not least, I thank my family who help keep me going, especially when things seem to be going nowhere.

CHAPTER 1

INTRODUCTION

1.1 Overview of Climate Change

Anthropogenic emissions of greenhouse gases (GHG's) and their climatic effect have been the subject of much debate within the scientific community in recent years. Although it is agreed that GHG's alone have a surface warming effect, that is not the only factor influencing our climate. We live on a planet that has an unstable climate system with highly non-linear positive and negative feedbacks. In order to make projections about future climate, General Circulation Models (GCM's) must be used. GCM's solve the full non-linear system of equations that govern the atmosphere for a prescribed forcing.

For the Intergovernmental Panel on Climate Change (IPCC) 2007 exercise, the projected change in global mean temperature was calculated from a suite of GCM's of various complexity. Based on the A2 scenario of the Special Report on Emission Scenarios (SRES - see below), global mean temperature is projected to be 2.0 to 5.4 degrees Celsius warmer by the end of the 21st century (*Solomon et al.*, 2007). As Figure 1.1 demonstrates, a global mean temperature change does not imply the globe will warm uniformly. Although most locations will likely warm, some locations show no warming at all or even cool slightly. The three emission scenario results shown in Figure 1.1 are SRES B1, A1B and A2 and clearly produce different results. The scenarios differ by population growth, economics and other variables and are story lines about how the future will unfold in terms of emissions of GHG's (*Nakicenovic and Swart*, 2000). The fact that the average temperature in the lowest eight kilometers of the troposphere has risen in the past four decades, snow and ice cover has decreased globally, and global ocean levels and heat content have risen are consistent with the view that climate change induced by anthropogenic forcing

is real.

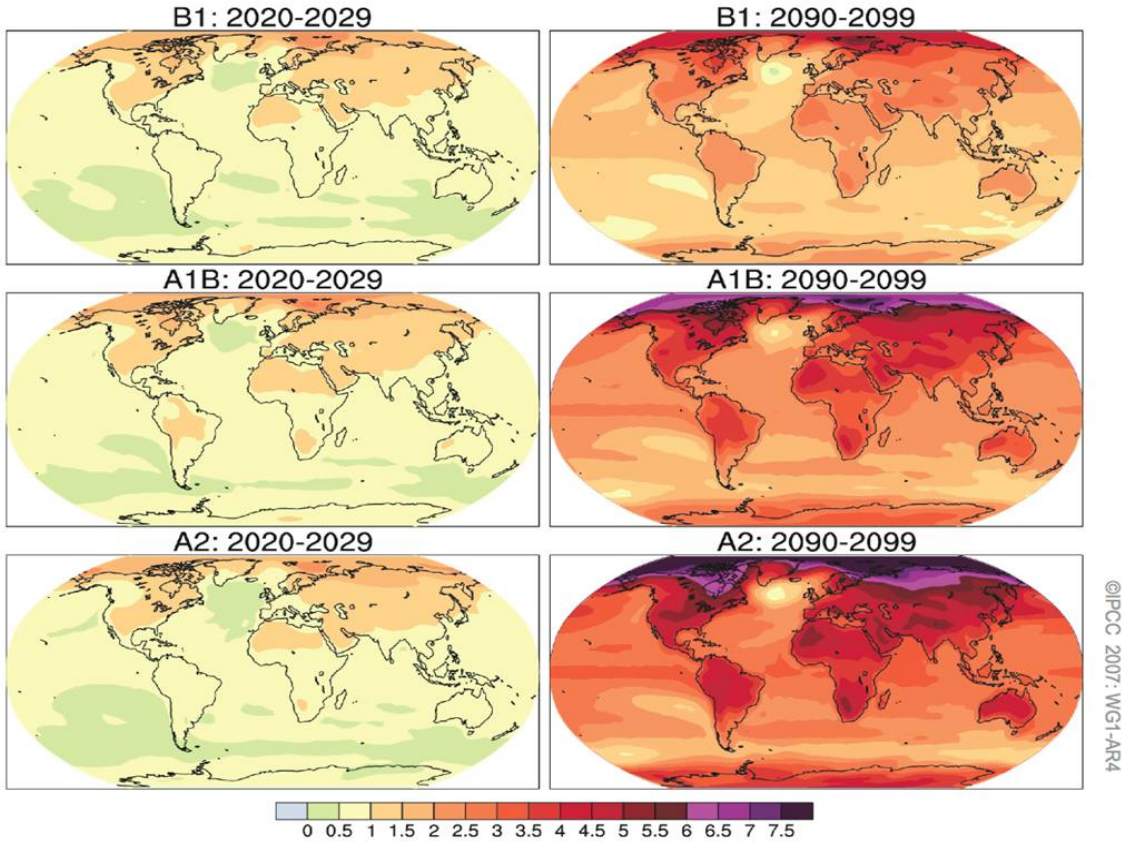

Figure 1.1: GCM ensemble average of change in annual mean temperature (degrees Celsius) for two different periods in the future for three different emission scenarios (SRES B1,A1B,A2). From *Solomon et al.*, 2007

A way to quantify the climate change associated with a warming planet is illustrated with basic statistics in Figure 1.2. An increase in the mean of a normal distribution of temperature, without change of variance, clearly results in warming temperatures. However, the variance of a distribution can also change. Instead of shifting the whole distribution, the variance controls the shape of the distribution. The shape of the distribution determines how much probability occurs in the tails, which is directly related to extreme events. Figure 1.2 suggests that, if the mean and the variance of temperature increase in a future climate, the result could be much more frequent extreme temperatures. The magnitude of the extremes could increase as well. Earth's climate could become drastically different compared to the climate that currently sustains our way of life. Humans can adapt to a slowly changing mean temperature. However, we can not easily adapt to

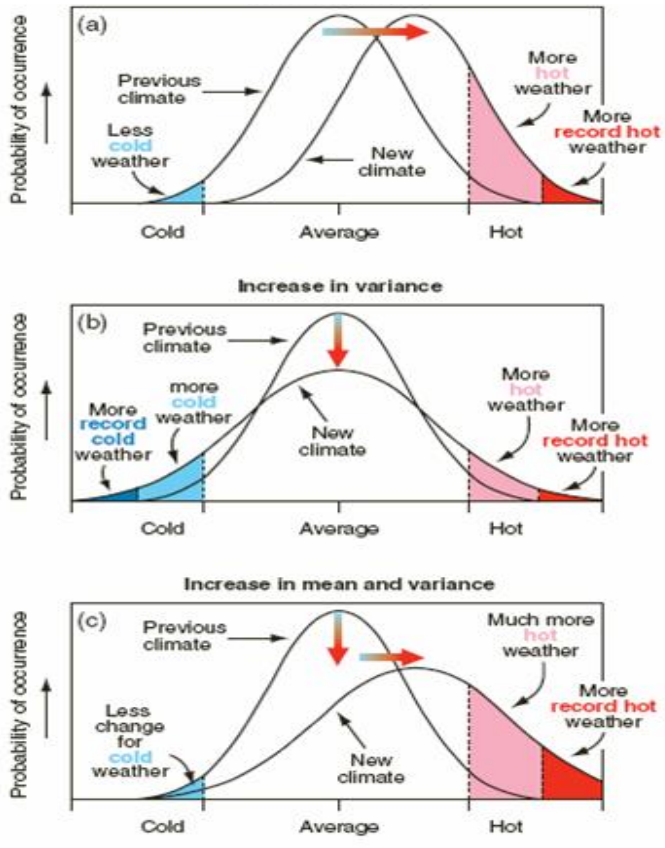

Figure 1.2: The figure demonstrates how normally distributed temperature, defined by the mean and variance, changes if the mean and/or variance changes. The top panel shows the climate change due to increase in mean. The middle panel shows how climate changes when the variance increases. The bottom panel demonstrates the change in climate when both the mean and variance increase. From *Solomon et al.*, 2007

frequently occurring new extreme events.

1.2 The Case for Downscaling

A rise in global mean temperature is not sufficient to say anything about future climate at a specific location on the globe. GCM's have a low spatial resolution, with a typical grid spacing of 300 by 400 km, as shown in Figure 1.3.

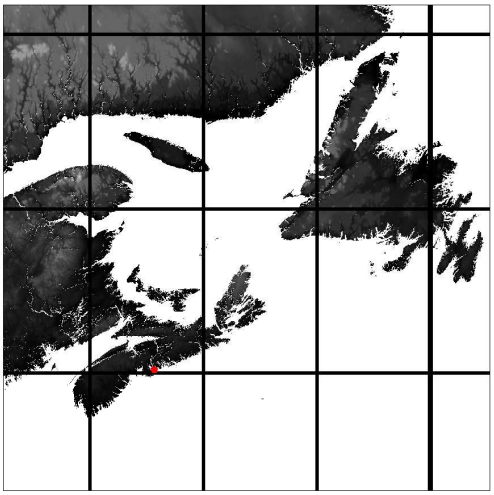

Figure 1.3: Grid boxes from a general circulation model (CGCM3), about 300 km by 400 km are plotted over Atlantic Canada. This thesis takes its observations from Shearwater Airport, NS (latitude 44.63N,longitude 63.5W) which is used as a proxy for Halifax. Shearwater airport (red dot) is about 4 km east of Halifax's downtown core.

GCM's are unable to resolve local climate forcings such as detailed topography or land-sea interfaces, which have a strong influence on climate at specific locations. The need for local climate change information leads to the following statement: "In order to best assess expected climate change impacts on a species, ecosystem or natural resource in a region, climate variables and climate change scenarios must be developed on a regional or even site-specific scale" (*Wilby et al.*, 2002). The coarse resolution of GCM's leads to another statement: "To provide these values, projections of climate variables must be downscaled from the general circulation model (GCM) results, utilizing either dynamical or statistical methods" (*Houghton et al.*, 2001).

One way to overcome the GCM resolution problem is dynamical downscaling. Dynamical downscaling inserts a high resolution regional model inside the coarse resolution global circulation model. An example is provided by the Canadian Regional Climate Model (CRCM; see *Laprise et al.*, 2003). The CRCM inserts a high resolution nested grid over the area of interest. The higher resolution model with grid length of about 50 km is then driven by time-dependent boundary conditions taken from the coarser resolution GCM. The higher resolution grid allows for better representation of the local forcing such as topography. The main advantage of dynamical downscaling is it allows projections of climate variables from the evolution of the full non-linear equations that govern the atmosphere. The main disadvantage is the vast amount of computational power required.

A second method is through Statistical Downscaling (SD). SD involves the development of a regression between observations of a local climate variable (predictand) and the larger scale atmospheric variables (predictors) over a specific site or region. SD is based on the view that local climate is forced by the large scale climate, and local forcings such as the land sea-interface. GCM derived predictors can then be used in the regression to make future projections. In statistical downscaling, the regression coefficients are developed from observations, and should therefore include realistic local effects. SD has been applied using many different regression procedures. These include linear regression (e.g. *Cheng et al.*, 2008), canonical correlation analysis (e.g.*von Storch et al.*, 1993) and artificial neural networks (e.g.*Schoof and Pryor*, 2001).

The main advantage of SD is the relatively low computational requirements. The major disadvantage of SD is the assumption that the regression will hold under future climatic conditions. The main argument for using SD has two parts. First, it has the benefit of the GCM predictors which evolve through the full atmospheric equations of motion responding to the prescribed forcing. Second, SD includes statistics through a regression that should have realistic local climate forcing contained within it, since it was derived from observations.

1.3 Motivation

This thesis was primarily motivated by large differences between the observed distributions of maximum and minimum daily temperature at Shearwater Nova Scotia airport compared to the distribution generated by the Canadian general circulation model version

3 (CGCM3) as shown in Figure1.4. Clearly the CGCM3 produces the distributions of temperature (Tmax and Tmin) differently from observations. The main goal of statistical downscaling is to produce a regression model using observed, large-scale atmospheric variables (predictors) to capture the distribution of the observed Tmax and Tmin. This carefully developed regression model allows use of the same predictors, in this case predictors produced from the CGCM3, to predict a more realistic distribution for Tmax and Tmin than is provided by the raw CGCM3 Tmax and Tmin distributions. The regression model can then be used to make future projections for Tmax and Tmin in a future climate using CGCM3 future predictors.

Environment Canada has been using statistical downscaling since 2003 through software known as the Statistical Downscaling Model (SDSM) (*Wilby et al.*, 2002). This software, although very useful in providing relatively quick climate change projections at a location, has a number of limitations. The SDSM is largely a black box to the user with a problematic predictor selection process. The author of the SDSM alludes to this problem when he states that, "there is also a real problem to use or apply SD methods uncritically as black boxes, particularly when employing regression based modeling techniques" (*Wilby et al.*, 2004). Since the most important part of statistical downscaling is the development of the regression with observations, some of the limitations of SDSM are addressed in this thesis.

Since climate and atmospheric variables (predictands and predictors) have a large amount of variance in the seasonal cycle, it is suggested in this thesis that the seasonal cycle be removed from both the predictands and predictors before constructing the regression model. Without removing the seasonal cycle, the regression model focuses on fitting the seasonal cycle and not the day-to-day variability. However, for predicting extremes in daily maximum and minimum temperature in a future climate, it is the day-to-day variability that matters. Another issue to consider is the predictor selection process. In addition to having predictors with a strong correlation with the predictand, a physically based predictor selection process should be carried out. Since the goal is to use GCM predictors in the regression, it is essential to compare the observed predictor distributions with the GCM predictor distributions.

In general, these issues are not addressed in studies conducted in the past. However,

there are numerous studies that used statistical downscaling to get local climate projections. Noteworthy is the study on downscaling surface temperature in Central Europe by *Huth* et al. (2002). *Huth* et al. (2002) compared the statistical downscaling at sites in Europe using various methods, such as canonical correlation analysis, multiple linear regression, and singular value decomposition. *Huth* et al. (2002) found that the best temporal structure was achieved by stepwise multiple linear regression. Another important finding by *Huth* et al. (2002) was that using one circulation variable (vorticity for example) and one temperature related variable (geopotential for example) produced the best regression.

The authors of the SDSM conducted a study by comparing downscaling methods for rainfall (*Wilby et al.*, 1998). *Wilby et al.* (1998) compared weather generators, various regression techniques and artificial neural networks. *Wilby et al.* (1998) found that different methods produced significant differences in regression skill. *Wilby et al.* (1998) determined that the downscaling methods generally produced smaller changes in precipitation (historical vs. future) than the GCM. Since mainly atmospheric circulation variables were used as predictors in the downscaling, the hypothesis is that precipitation changes predicted by GCM's are not likely caused by circulation changes.

A study on temperature in northern Canada by *Gachon and Dibike* (2007), found that statistical downscaling models are able to capture the low frequency climate change signal. This was done with two different GCMs (HADCM3 and CGCM2). The SD models showed strong convergence in the timing and magnitude of the mean change, compared to observed.

Dibike et al. (2008) examined downscaled maximum and minimum daily temperatures in northern Canada. *Dibike et al.* (2008) found that all downscaling results reveal that the regression-based statistical downscaling methods driven by accurate GCM predictors are able to reproduce the climate regime over highly heterogeneous coastline areas of northern Canada.

These studies certainly give confidence that statistical downscaling can be a skillful tool to produce local climate projections. However, none of the previous studies tackles the problems mentioned with the SDSM (seasonal cycle and predictor selection). After an exhaustive literature search, no example was found that explicitly removes the average seasonal cycle in the historical and future periods, as proposed in this thesis. Again,

removal of seasonal cycle takes autocorrelation out of the data and forces the regression analysis to predict the variance in the anomalies instead of focusing on the deterministic part of the data. It was found that some downscaling studies use a mathematical technique, to account for autocorrelation, such as in *Cheng et al.* (2008). However, we argue that that explicit removal of the seasonal cycle is more physical and important for credible downscaling.

This thesis work also subjectively compares each NCEP predictor against the respective CGCM3 predictor to ensure they have similar distributions. This explicit distribution comparison has not been done previously. A more common technique discussed in the literature is just to remove surface and moisture variables, since some studies demonstrated that GCM's do not represent them properly (*Dibike et al.*, 2008). The SDSM for example, does not make any NCEP/GCM comparison. We will show in this thesis that comparison between observations and model predictor distributions is an essential step for downscaling, and must be done at all sites independently. Differences between the NCEP and CGCM3 predictor distributions will cause undesirable results when the CGCM3 predictors are used in the regression (trained with NCEP predictors). For clarity, undesirable means that a regression trained with predictor distibutions from NCEP will not give a good prediction of temperature (Tmax or Tmin), when using the same CGCM3 predictors, if the CGCM3 predictor distributions are very different from NCEP.

A final motivation for this thesis work comes from the need to address important scientific issues in the many technical documents on statistical downscaling. In house technical documents like *Lines et al.* (2005), *Pancura and Lines* et al. (2005), and *Swansburg et al.* (2005) use the SDSM for downscaling but do not consider any of the issues associated with it, especially to do with predictor selection. Although it cannot be proved the results in these technical documents are wrong, it can be said with certainty that they could be improved. Since so many researchers use statistical downscaling to get future projections to make decisions on adaptation, the credibility of the projections is most important. Credibility comes from a repeatable scientific method that addresses the issues discussed previously.

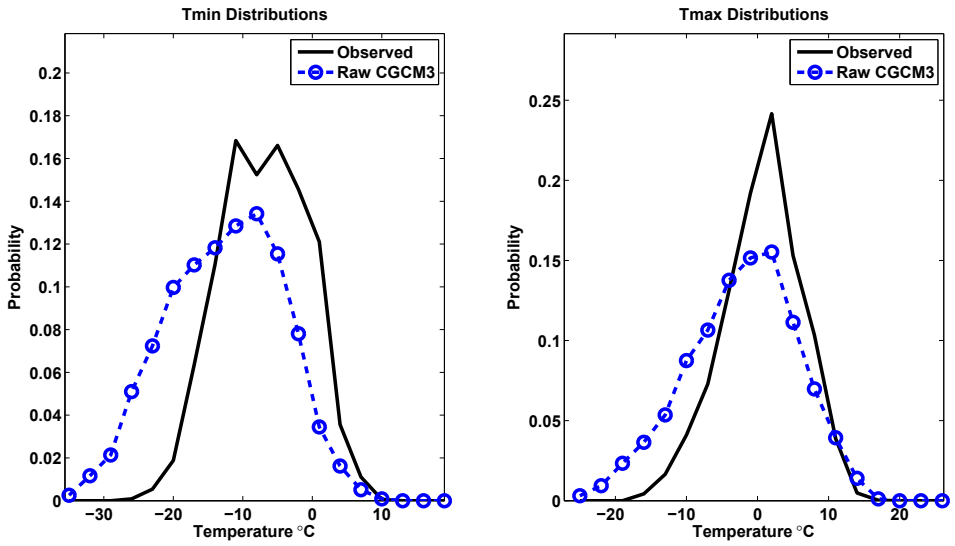

Figure 1.4: The left panel shows daily minimum temperature (Tmin) distributions from the raw CGCM3 (grid box containing Shearwater, NS) versus observations of Tmin at Shearwater Airport, both for the $1961 - 2000$ period. The right panel is similar to the the left panel except it is for maximum daily temperature (Tmax). Both panels are distributions for winter (DJF). Note the distributions are created by binning the data in three degree Celsius intervals, counting the occurrences in each bin, and finally dividing each of the total occurrences in each bin by the total number of observations to give the probability.

1.4 Objective and Approach

Accurate climate change projections are needed to make intelligent decisions about adaptation. The main objective of this thesis is to develop a method that is both statistically robust and has a physical motivation. Since there is no crystal ball to tell us about the future climate, it becomes even more important to have a scientifically defendable method to obtain the climate projections. Statistical downscaling that has a method with coherent defensible steps, is much more important than the projection itself. The focus of this thesis thus remains with the method rather than the future warming projection for Tmax and Tmin at Shearwater Nova Scotia. In other words, the goal is to develop the best regression model using observations in the historical period. It is not the primary objective to produce future projections. However, two methods for producing the future projections are explored. It is not stated with certainty which is best.

To address the problems noted in the motivation with SDSM and the literature in general, the following approach is used in this thesis work. Our approach begins with an exploratory analysis. This involves examination of the predictors in terms of the dominant physics in each season. An investigation of the predictand is made to decide what is deterministic (e.g. repeats annually) and should be removed before the regression model is developed. The predictor selection process is carried out by selecting predictors that have a physical relationship with the predictand. The steps necessary to choose predictors relevant for principal component analysis are carefully considered. It is also necessary to choose predictors that the GCM can represent in a reliable way.

The structure of this thesis begins with a review of the math and statistics necessary for SD in Chapter 2. Chapter 3 contains a discussion on the predictors and predictands used in the downscaling process. Chapter 4 discusses the main steps to develop the regression. Chapter 4 also contains a discussion on producing the future projections and the problems associated with producing them. Finally the thesis closes in Chapter 5, with a summary of the downscaling method and a comparison of the thesis downscaling results to SDSM results.

CHAPTER 2

MATHEMATICAL AND STATISTICAL BASIS

Multiple linear regression, hypothesis testing, and principal component analysis to accomplish credible statistical downscaling are reviewed in this chapter. The main focus here is on the the results rather than the derivations. A full treatment of the theory behind these techniques can be found in various textbooks referenced throughout the chapter.

2.1 Multiple Linear Regression

The simplest case of linear regression is a model that has one dependent variable (predictand, Y) and one independent variable (predictor, X). The desired regression model is in the form of a linear equation where the predicted value of Y at any time i ($\hat{Y}_i$) equals the Y intercept (β_o) plus the slope (β_1) multiplied by the value of X at time i (X_i) which is written as follows:

$$\hat{Y}_i = \beta_0 + \beta_1 X_i \tag{2.1}$$

In order to formulate an expression to get β_0 and β_1, a best fit line through the data is sought after. Figure 2.1 shows a scatter plot of one variable (Y) against another variable (X). A best fit line plotted through the data can be defined as the line that minimizes the sum of the squares of the errors (ϵ_i) where

$$\epsilon_i = Y_i - \hat{Y}_i. \tag{2.2}$$

Minimization of the sum of the squared errors is accomplished by taking the derivatives

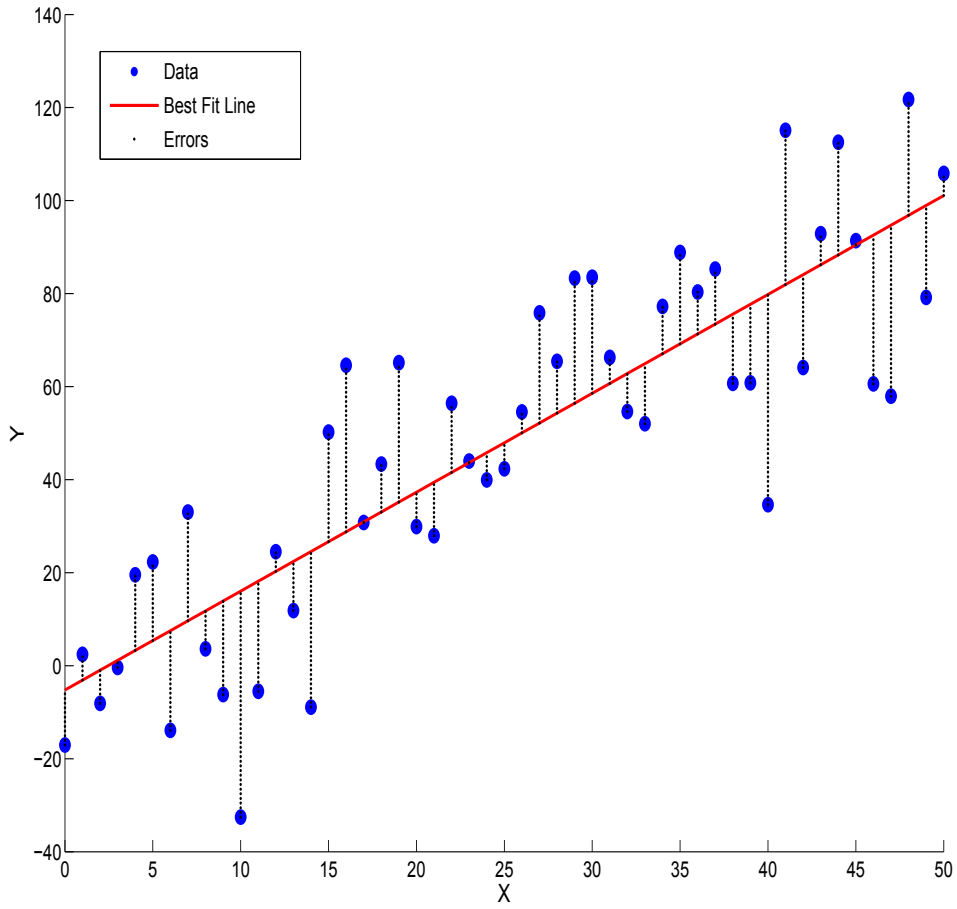

Figure 2.1: A scatter plot (blue) of the dependent (Y) variable versus the independent (X) variable highlights the linear relationship between Y and X. The red line is the best fit and is placed in such a location to minimize the sum of the length of the error lines (black) squared.

of the sum of the squared errors with respect to both β_0 and β_1 and setting them to zero.

$$\frac{\partial}{\partial\beta_0}(\sum_{i=1}^{n}\epsilon_i^2) = \frac{\partial}{\partial\beta_0}(\sum_{j=1}^{n}(Y_i - (\beta_0 + \beta_1 X_i))^2 = 0 \tag{2.3}$$

$$\frac{\partial}{\partial\beta_1}(\sum_{i=1}^{n}\epsilon_i^2) = \frac{\partial}{\partial\beta_1}(\sum_{j=1}^{n}(Y_i - (\beta_0 + \beta_1 X_i))^2 = 0 \tag{2.4}$$

This leads to two equations with two unknown parameters in β_o and β_1:

$$\sum_{i=1}^{n} Y_i = n\beta_o + \beta_1 \sum_{i=1}^{n} X_i \tag{2.5}$$

$$\sum_{i=1}^{n} X_i Y_i = \beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2 \tag{2.6}$$

From these two equations β_0 and β_1 can be solved for

$$\beta_0 = \frac{\widetilde{X}\widetilde{XY} - \widetilde{X^2}\widetilde{Y}}{\widetilde{X}\widetilde{X} - \widetilde{X^2}} \tag{2.7}$$

$$\beta_1 = \frac{\widetilde{X}\widetilde{Y} - \widetilde{XY}}{\widetilde{X}\widetilde{X} - \widetilde{X^2}} \tag{2.8}$$

where any variable (Q) with an overbar ($\widetilde{Q}$) is defined as

$$\widetilde{Q} = \frac{1}{n}\sum_{i=1}^{n} Q_i. \tag{2.9}$$

β_0 and β_1 fully characterize the best fit line defined in Eq. (2.1). This of course is a special case where one independent variable (predictor) and one dependent variable (predictand) are being used.

In general, there can be more than one predictand variable and more than one predictor variable. This leads to a multiple set of regression coefficients. There would be one set of coefficients for each predictand variable. It can be shown that the matrix equation for the regression coefficients $\boldsymbol{\beta}$ in general is

$$\boldsymbol{\beta} = (\mathbf{X^T X})^{-1}\mathbf{X^T Y}. \tag{2.10}$$

The bold font in Eq. (2.10) refers to a matrix. The predictor variables, $\mathbf{X}$ has $p + 1$ columns, where each column (except the first column) is a predictor variable. The first column is always a column of ones, which allows the regression to calculate the mean (β_o).

The predictor matrix ($\mathbf{X}$) has the form

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & \dots & X_{1p} \\ 1 & X_{21} & \dots & X_{2p} \\ & \vdots & & \\ 1 & X_{n1} & \dots & X_{np} \end{bmatrix} \tag{2.11}$$

Clearly $\mathbf{X}$ has dimensions of $n \times (p+1)$, where n is the number of measurements, and p is the number of predictors. $\mathbf{Y}$ is a matrix that has the same number of columns (q) as predictand variables. The dimensions of $\mathbf{Y}$ is $n \times q$, where n is the number of measurements and q is the number of predictands.

$$\mathbf{Y} = \begin{bmatrix} Y_{11} & \dots & Y_{1q} \\ Y_{21} & \dots & Y_{2q} \\ & \vdots & \\ Y_{n1} & \dots & Y_{nq} \end{bmatrix} \tag{2.12}$$

Following through with the dimensions in Eq. (2.10), $\boldsymbol{\beta}$ has dimensions of $(p+1) \times q$. In general $\boldsymbol{\beta}$ will have a set of regression coefficients for each predictand. In other words each column of $\boldsymbol{\beta}$ gives the associated regression coefficients (associated with each predictor variable) for each predictand variable.

For clarity, we consider a simple case in which the number of predictors (p) and predictands (q) is the same and equal to one. In this case $\mathbf{X}$ will have two ($p+1$) columns. $\mathbf{Y}$ will have one (q) column. This leads to $\boldsymbol{\beta}$ (Eq. (2.10)) having one column (q), containing two regression coefficients (β_o and β_1). Similarly if there were two predictands, $\mathbf{Y}$ would have two (q) columns. Assuming there is still one predictor variable p, then $\mathbf{X}$ would have two ($p+1$) columns. This would lead to $\boldsymbol{\beta}$ having two rows and two columns ($(p+1) \times q$). In other words two regression coefficients (β_o and β_1), for each of the two predictand variables.

The generalized case, where $\mathbf{Y}$ has more than one ($q > 1$) variables is known as multivariate linear regression. Multiple linear regression is a special case where $\mathbf{Y}$ has only one variable ($q=1$) in which case, $\boldsymbol{\beta}$ in Eq. (2.10) will have one column. Once $\boldsymbol{\beta}$ has been determined the multiple linear regression with one predictand and more than one predictor variables ($p > 1$) can be written as an extension of Eq. (2.1), where i refers to the individual elements in time and ranges from 1 to n.

$$\hat{Y}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + ... + \beta_p X_{ip} \tag{2.13}$$

The goal of any regression is to determine regression coefficients ($\boldsymbol{\beta}$) to allow $\hat{Y}$ (the predicted values) to explain (predict) optimally the variance in the predictand (Y) around the mean of Y ($\bar{Y}$). The variance in the predictand, the variance predicted by the regression and the error variance are written as follows:

Total Sum of Squares

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \tag{2.14}$$

Regression Sum of Squares

$$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \tag{2.15}$$

Error Sum of Squares

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{2.16}$$

The total variance (SST) can always be written as the sum of two variances. That is the variance of the regression (SSR) and the variance of what is left over which is the variance of the errors (SSE).

$$SST = SSR + SSE \tag{2.17}$$

This relationship can be proved mathematically by writing out Eq. (2.17) in terms of the mathematical definitions from Equations 2.14, 2.15 and 2.16,

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = SSR + SSE = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{2.18}$$

which can be rearranged to be

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}\left[(\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)\right]^2. \tag{2.19}$$

Further expansion and allowing the error term ϵ to be defined as in Eq. (2.2), Eq. (2.19) can be written as

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + 2\sum_{i=1}^{n}\epsilon_i(\hat{Y}_i - \bar{Y}) + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2. \tag{2.20}$$

An important property of linear regression is that the errors (ϵ_i) are uncorrelated with the predicted values ($\hat{Y}_i$). This means that the second term on the RHS in Eq. (2.20) is zero. Eq. (2.20) then becomes the desired relationship shown in Eq. (2.17). Eq. (2.17) can be viewed graphically for clarity. A contribution to the sum in SST, SSR and SSE for a

particular X is shown in Figure 2.2. Note the length of the lines in Figure 2.2, need to be squared to obey Eq. (2.17). The length's are actually the square root of SST, SSR, and SSE.

The *explained variance* is the fraction of observed variance that is explained by the regression. Mathematically, this is the ratio of the regression sum of squares (SSR) to the total sum of squares (SST).

$$r^2 = \frac{SSR}{SST} \tag{2.21}$$

This expression is directly related to the correlation between the predicted values from the regression and the predictand values used to create the regression. It can be shown that the correlation squared is the explained variance shown in Eq. (2.21).

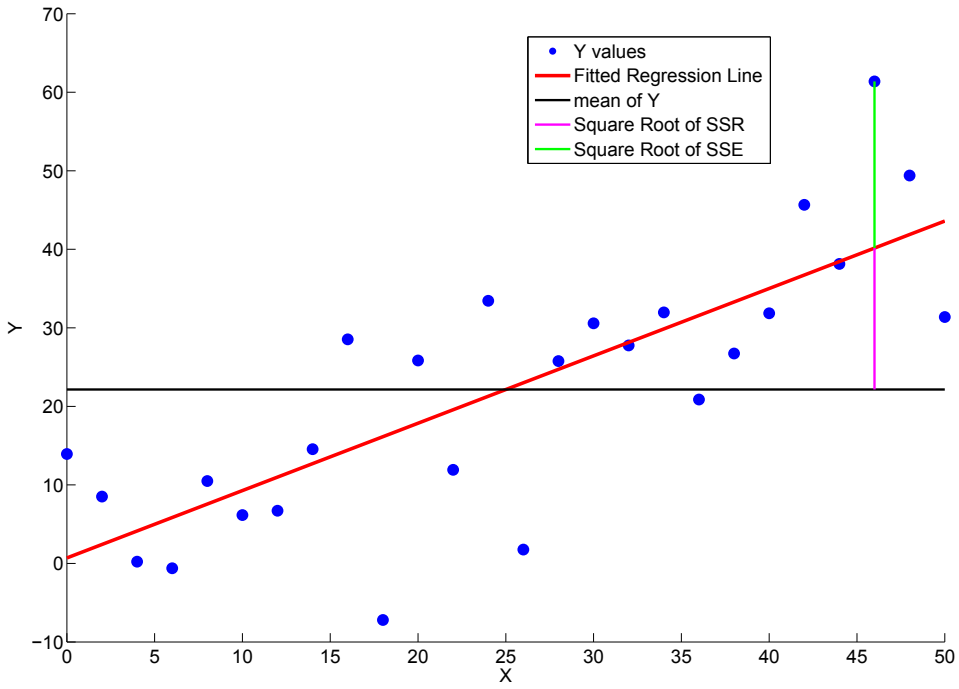

Figure 2.2: A scatter plot (blue) of the dependent (Y) variable versus the independent (X) variable highlights the linear relationship between Y and X. The mean of Y is shown in black. The contribution to the total sum of squares (SST) is the length of the magenta line squared (SSR) and the length of the green line squared (SSE) added together. The magenta line length represents the square root of the contribution to SSR. The green line length is the square root of the contribution to SSE.

It is important that the predictors are linearly related to the predictand (linearity). Linearity ensures a linear line can be fit through the predictand and is a major assumption of linear regression. Furthermore linear regression is based on three other major assumptions. They are *normality of errors*, *homoscedasticity* and *independence of errors*. Once the line of best fit ($\hat{Y}$) has been found, each assumption should be checked to ensure it has not been violated. The first assumption (*normality of errors*) can be checked by plotting a histogram of the errors to ensure they are approximately normally distributed. The second assumption (*homoscedasticity*) can be verified by plotting the errors against each predictor used in the regression. The plot should look random and not have any pattern, such as larger errors for larger predictor values. Finally, *independence of errors* could be validated by a plot of errors versus time. Again the errors should be random and not be all positive or negative.

Data containing autocorrelation may produce a regression that violates the independence of errors assumption. Autocorrelation means that data at different points in time are correlated with each other. Thus some future value could be predicted from past values. In other words, realisations at different times are not independent and this causes problems with the *independence of errors* assumption. A good example of this is hourly temperature data. The rise and fall of solar radiation during the day causes temperatures to increase over half the day and then decrease for the other half. Removal of the diurnal cycle from the hourly observations may be necessary before fitting a regression.

A good general introduction to regression can be found in "Applied Statistics for Scientists and Engineers" (*Levine et al.*, 2001). A more advanced mathematical explanation can be found in "Applied Multivariate Statistical Analysis" (*Johnson and Wichern*, 2002).

2.2 Confidence Intervals and Hypothesis Testing

In any statistical analysis, the need arises to make statements about a population (assumed to be infinite in size) of data with limited information, the main limitation being that you only have a sample of the larger population. The two main population parameters of interest are the mean (μ) and the variance (σ^2). Inferences can be made by assuming the data are normally distributed. Data that are normally distributed have a Gaussian (bell

shaped) probability curve, defined by an exponential function.

$$P(X) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\left(\frac{1}{2}\right)\left[\frac{X-\mu}{\sigma}\right]^2\right) \tag{2.22}$$

The probability density function in Eq. (2.22) for a variable X is defined by the mean (μ) and the standard deviation (σ) of X. Since a normal distribution is symmetric about the mean, the area below and above the mean both represent fifty percent of the probability. In reality, various datasets have many combinations of means and variances. This makes finding probabilities tedious because the various combinations of means and variances define different probability density functions. A simplification can be made by standardizing the data set. Removing the mean of X from each X value and then dividing by the standard deviation of X, transforms each X value (X_i) into a Z value (Z_i).

$$Z_i = \frac{X_i - \mu}{\sigma} \tag{2.23}$$

The Z values define the standardized normal distribution known as the Z-distribution. The Z-distribution is a normal distribution with a mean of zero and a standard deviation of one, and is shown in Figure 2.3.

Inference about the mean of a population is chosen to introduce the idea of confidence intervals. A 95 percent confidence interval implies that of all possible samples of size n, 95 percent will have the true mean (μ) within their confidence intervals. The percent level of confidence is written mathematically as $(1-\alpha)$x100, where α is the total probability that μ will not lie inside the confidence interval. α of 0.05 is usually chosen in practice which would be a 95 percent confidence interval. In general, it is desirable to have the sample mean at the center point of the confidence interval. Since a normal distribution is 2-sided as shown in Figure 2.3, the probability of the actual mean being above or below the confidence interval is α divided by two. Therefore there would be a 2.5 percent chance of the actual mean either being above or below the confidence interval window.

To find the confidence interval associated with the population mean μ, from the sample means ($\bar{X}$), we first define Z for the sample means. This is done by replacing X with $\bar{X}$ and σ with $\frac{\sigma}{\sqrt{n}}$ in Eq. (2.23) where n is the sample size.

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{2.24}$$

The $\sqrt{n}$ is necessary because σ is the standard deviation of the sample means. The sample means will have less variance than the actual data due to the averaging used to obtain the

sample mean. As n becomes large, the confidence interval will shrink to the limit that the sample mean becomes the true mean as can be seen in Eq. (2.25). Z is now effectively measuring how many standard deviations the sample mean is from the true mean. Eq. (2.24) can be used to express the 95 percent confidence interval for the population mean. As discussed, the Z-distribution is two sided with the true mean in the center. The critical values shown in Figure 2.3 are used for Z and are associated with the ninety five percent confidence interval. This means the confidence interval will include the true population mean (μ) ninety five percent of the time.

$$\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}} \tag{2.25}$$

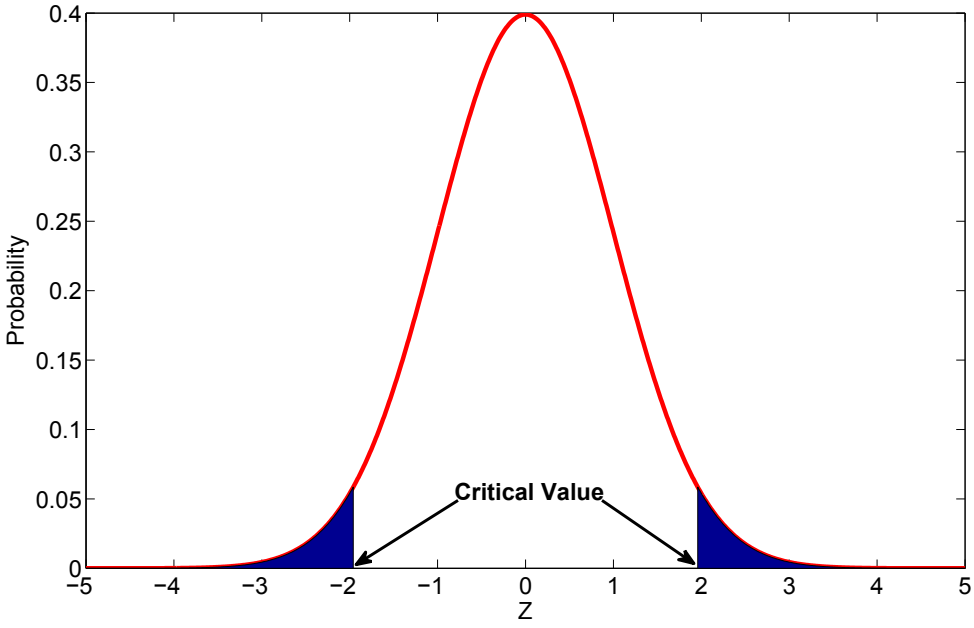

Figure 2.3: A standardized normally distributed probability function. The arrows point to the critical values of Z. They are -1.96 and 1.96 for a 95 percent confidence interval. The critical values define the shaded regions which encompass α/2 (0.025) of the probability. Values in these regions are outside the confidence interval. The shaded regions are also known as the region of rejection in hypothesis testing.

Eq. (2.25) assumes that σ is known. In reality, σ is usually unknown. Therefore another confidence interval has to be developed to allow for the fact that only the sample standard deviation (s) is known. The distribution used in this case is called the Student's t-distribution. This t-distribution for a given sample size (n) has more area in the tails and

less in the center than the normal distribution as shown in Figure 2.4. The t-distribution tends toward the normal distribution as the sample size increases. Mathematically the confidence interval statement is the same as Eq. (2.25), where Z has been replaced by the t-distribution with $n-1$ degrees of freedom with n being the sample size and σ is replaced by the sample standard deviation (s). Note t_{n-1} is the critical value of the t-distribution associated with the ninety five percent confidence interval similar to the Z-distribution (See Figure 2.3). Figure 2.4 demonstrates the difference between the Z-distribution and the t-distribution. A ninety five percent confidence interval in a t-distribution with a sample of six (n) values (5 degrees of freedom, $n-1$) dictates t-values must lie between -2.57 and 2.57, as discussed in Figure 2.4.

$$\bar{X} - t_{n-1}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1}\frac{s}{\sqrt{n}} \tag{2.26}$$

where t_{n-1} equals 2.57 in this case.

In practice, the t-distribution is well approximated by the normal distribution once the sample size becomes large. This is because the sample variance (s^2) approaches the population variance (σ^2). This is typically the case in statistical downscaling. Therefore equation 2.25 can be easily applied without the complication of the t-distribution in most statistical downscaling applications.

The idea of confidence intervals leads directly to hypothesis testing. Hypothesis testing is a good statistical inference technique for asking questions about the similarity of population parameters. In statistical downscaling the question one should ask is: Is the predicted historical distribution statistically similar to what was observed? Hypothesis testing is one way of answering this question. In hypothesis testing you formulate the question in terms of the null hypothesis and the alternative hypothesis. As in confidence intervals, Z in equation (2.24) can be used to formulate a hypothesis test to compare the means of the two distributions. If $\bar{X}$ is the mean of the sample distribution, μ is the true mean, and σ^2 is the true variance, then a 95 percent hypothesis test can be constructed. This means that there is 95 confidence that the null hypothesis will not be rejected when it is true and should not be rejected. The null hypothesis (H_o) is known as the status quo. In terms of the mean, it says the sample mean is statistically the same as the true mean. The alternative hypothesis (H_1) will be the opposite case, which is the sample mean is not equal to the true mean.

$$Null \Longrightarrow H_o : \bar{X} = \mu \tag{2.27}$$

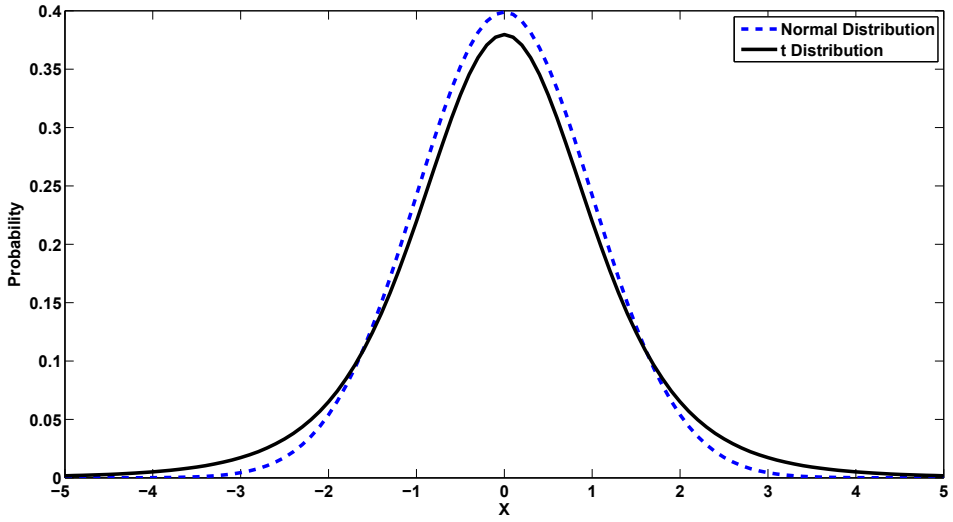

Figure 2.4: The normal distribution with a mean of zero and standard deviation of one (dashed) versus the t distribution (solid) of the data with 5 degrees of freedom (sample size $n = 6$). Note the t distribution has more area in the tails and less in the main body than the normal distribution. Note t_{n-1} has critical values of -2.57 and 2.57 in this case, for a ninety five percent confidence interval.

$$Alternative \Longrightarrow H_1 : \bar{X} \neq \mu \tag{2.28}$$

In confidence intervals, the critical values of Z are -1.96 and 1.96 for a 95 percent confidence interval as in Figure 2.3. If Z in Eq. (2.24) is less than -1.96 or greater than 1.96 the null hypothesis will be rejected. Similarly, if Z is greater than or equal to -1.96 and less than or equal to 1.96 then we accept the null hypothesis. Again, a 95 percent confidence interval is associated with an α of 0.05 which is divided equally in the tails of the distribution shown in Figure 2.3. The shaded regions in Figure 2.3 are the regions that the null hypothesis is rejected and the remaining unshaded area is where the null hypothesis would be accepted. In this thesis, hypothesis testing will be used to compared the mean of a predicted distribution ($\bar{X}$) against the observed mean (μ) to see if the Z value found from Eq. (2.24) will pass the ninety five percent confidence hypothesis test. If the null hypothesis is accepted, it suggests the means of the observed distribution and the predicted distributions are statistically the same.

Another important question is whether the variances of two distributions are statistically the same. If the data are normally distributed then the ratio of the variances is

known to follow the F-distribution (shown in Figure 2.5). The F-distribution is ratio of the variances of the two distributions being compared.

$$F = \frac{S_1^2}{S_2^2} \tag{2.29}$$

Again, a 95 percent confidence interval ($\alpha = 0.05$) defines the critical values in the F-distribution as shown in Figure 2.5. This is almost identical to the hypothesis test for the mean, the difference being the F-distribution is used instead of the Z distribution. The hypothesis test can be stated as follows.

$$Null \Longrightarrow H_o : \sigma_1^2 = \sigma_2^2 \tag{2.30}$$

$$Alternative \Longrightarrow H_1 : \sigma_1^2 \neq \sigma_2^2 \tag{2.31}$$

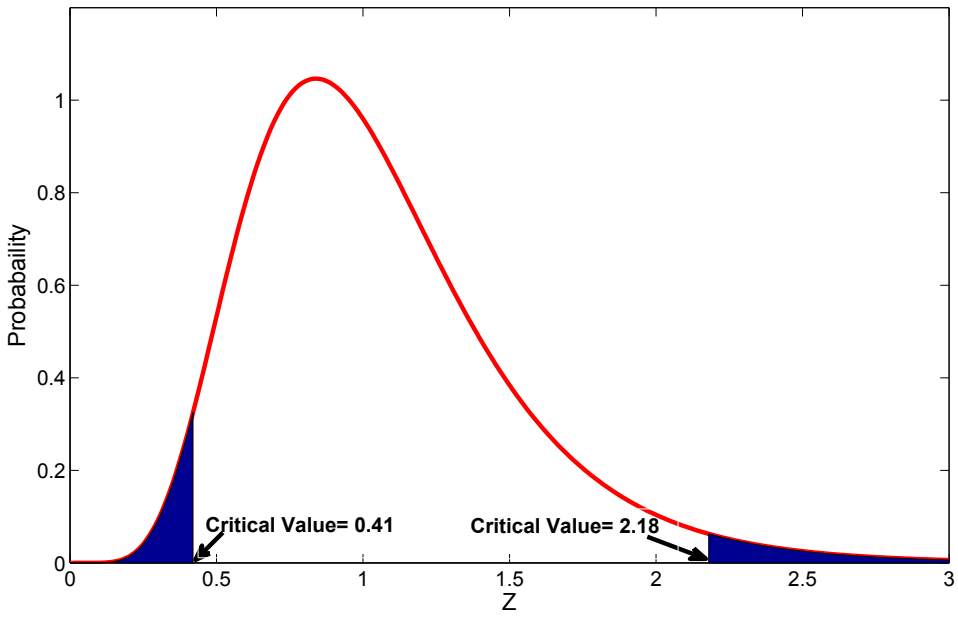

Figure 2.5: The F-distribution plotted with both the lower (F_L,0.41) and upper (F_U,2.18) critical values for two samples with 17 and 39 degrees of freedom respectively. Values of F from Eq. (2.29) greater than the F_U or less than F_L will prompt a rejection of the null hypothesis in favor of the alternative in Equations 2.30 and2.31. The shaded blue areas each have a probability of 2.5 percent and are known as the regions of rejection. The unshaded region is the region of non-rejection.

This thesis will use the F-test for variance to see if the ratio of the two distribution

variances (Eq. (2.29)) produces an F value within the confidence limits (see Figure 2.5), the two distributions being the observed distribution and the predicted distribution. If the null hypothesis can be accepted, the variances are said to be statistically the same.

One of the best references for the topic of statistical inference, including confidence intervals, and hypothesis testing is “Applied Statistics for Scientists and Engineers” (*Levine et al.*, 2001).

2.3 Principal Component Regression

Climate data can be overwhelming because there is so much in terms of space and time that it is difficult to make sense of it. Another issue is that there is often not a long enough record to say anything statistically significant about trends and patterns (*Wunsch*, 1999). Principal Component Analysis (PCA) is a powerful data analysis technique to find the independent linear combinations of variables within a data set that explain the variance in the data.

In statistical downscaling, the researcher usually has time series of multiple atmospheric variables at one point. Atmospheric variables have a large covariance among themselves. This causes problems in terms of using the predictors in a linear regression, since independent predictors are needed. Principal component analysis transforms the multivariate predictor set into a new predictor set known as principal components. The principal components will all be linearly independent, and thus have zero covariance between them.

The first step is to take the multivariate predictor set and transform it into Z-scores (see Eq. (2.23)). This means subtracting the mean from each predictor value respectively and dividing each predictor value by its standard deviation. This has the desirable effect of making the predictors non-dimensional. This is an important step in statistical downscaling because atmospheric variables have different units. Once each predictor has been transformed into its respective Z-score, a matrix ($\mathbf{Z}$) of all the predictors Z-scores measured over time can be constructed.

$$\mathbf{Z} = \begin{bmatrix} Z_{11} & \dots & Z_{1j} & \dots & Z_{1p} \\ Z_{21} & \dots & Z_{2j} & \dots & Z_{2p} \\ \vdots & & & & \\ Z_{i1} & \dots & Z_{ij} & \dots & Z_{ip} \\ \vdots & & & & \\ Z_{n1} & \dots & Z_{nj} & \dots & Z_{np} \end{bmatrix} \tag{2.32}$$

Each column j (valued one through p) of $\mathbf{Z}$ is a predictor (Z-score of geopotential height for example) measured over all times ($i = 1$ through $i = n$). The rows (i) of $\mathbf{Z}$ correspond to measurements of the Z-scores (j) at a given time (i). Note that i equals n just refers to the last row in $\mathbf{Z}$, where n is the total number of predictor measurements in time. For example, if column one (j=1) of $\mathbf{Z}$ is the Z-score values of geopotential height,

then the value in any particular row (i, $i = 1...n$) will be the measurement of geopotential height (Z-score) at time i. To avoid confusion, it should be noted that when referring to a particular predictor, the reference is to all the realisations of that predictor over time.

The next step is to calculate the correlation matrix ($\mathbf{R}$) of $\mathbf{Z}$. Each element of the correlation matrix can be determined with the following equation:

$$R_{jk} = \frac{1}{n}\sum_{i=1}^{n}\frac{(Z_{ij} - \mu_{Z_j})(Z_{ik} - \mu_{Z_k})}{\sigma_{Z_j}\sigma_{Z_k}} \tag{2.33}$$

where the means (μ) and standard deviations (σ) are zero and one respectively for Z-scores. The index j and k in R_{jk} refers to the correlation between the predictors (Z-scores) that exist in columns j and k of $\mathbf{Z}$ (see Eq. (2.32)). Again, the sum over i is referring to the sum over time of predictor measurements (rows in $\mathbf{Z}$). Note that n in Eq. (2.33) refers to the total number of measurements in time. If for example there are 25 columns in $\mathbf{Z}$ (twenty five predictors), then both indices j and k would be valued one through twenty five. Using Eq. (2.33), each element of the correlation matrix $\mathbf{R}$ can be found. It should be noted that the correlation matrix is symmetric (R_{jk} equals R_{kj} except for k=j). The diagonal values of $\mathbf{R}$ are the elements when j equals k, which is the variance of a predictor. Note that the diagonal is all ones, as expected for Z-scores.

$$\mathbf{R} = \begin{bmatrix} 1 & R_{12} & \dots & R_{1p} \\ R_{21} & 1 & \dots & R_{2p} \\ & \vdots & & \\ R_{p1} & R_{p2} & \dots & 1 \end{bmatrix} \tag{2.34}$$

It is easy to verify that the correlation matrix ($\mathbf{R}$) is identical to the covariance matrix. In Eq. (2.33), the mean (μ) and standard deviation (σ) are zero and one respectively because the means and standard deviations are from Z-scores. The off-diagonals of the correlation matrix ($\mathbf{R}$), derived from $\mathbf{Z}$, will be non-zero and vary between one and minus one. This is because the predictors are not independent and have covariance between them. Thus the goal of the PCA is to create a predictor set that has a correlation matrix for which all elements are zero everywhere, except along the diagonal. The diagonal values will be the variances of each new independent predictor.

As previously stated, $\mathbf{R}$ is a symmetric matrix, which means it is square, and switching the rows and columns yields an identical matrix. In mathematical terms, $\mathbf{R}$ is equal to

the transpose of $\mathbf{R}$ ($\mathbf{R} = \mathbf{R^T}$). A result from linear algebra says, given a real symmetric $p \times p$ matrix $\mathbf{R}$, there exists an orthogonal matrix $\mathbf{E}$ such that

$$\mathbf{E}^T\mathbf{R}\mathbf{E} = \mathbf{\Lambda} \tag{2.35}$$

where $\mathbf{\Lambda}$ is a diagonal matrix with all elements equal to zero except for the diagonal elements. An orthogonal matrix has rows that are orthogonal (perpendicular) to each other. The columns are also orthogonal to each other. Finally, the transpose of an orthogonal matrix, is equivalent to the inverse of the matrix ($\mathbf{E^{-1}} = \mathbf{E^T}$). The diagonal values are known as the eigenvalues, which quantify how important the eigenvectors are in explaining the variance in the data set. $\mathbf{E}$ is a matrix in which all the columns are the orthonormal eigenvectors of $\mathbf{R}$. All the eigenvectors are orthogonal to each other, and have unit length. In other words, Eq. (2.35) transforms the predictor (Z-scores) correlation matrix ($\mathbf{R}$) into a diagonal matrix ($\mathbf{\Lambda}$). Again, the diagonal matrix ($\mathbf{\Lambda}$) is a new correlation matrix which is zero everywhere off the diagonal. The eigenvector matrix ($\mathbf{E}$), has size $p \times p$, where each column (eigenvector) is made up of a linear combination of the original predictors. New independent predictor variables (principal components, PC's) can be found by projecting the original predictors (used to create $\mathbf{R}$) onto the eigenvectors in $\mathbf{E}$. This is done at each observation time making the PC's uncorrelated since they are derived from a projection onto the orthogonal eigenvectors.

It is common in atmospheric science to exclude the eigenvectors associated with little variance in the data set. This essentially removes the noise in the data set and also reduces its dimensions. However, in statistical downscaling, a common mistake made is to drop the principal components associated with the smallest variance eigenvectors before developing the regression. Since the PCA only deals with the variance within a data set, it does not give any information on which predictors are good for predicting the predictand (the predictand has not been used yet). As a result, there is no guarantee that the PC's associated with the largest eigenvalues are the best predictors for a given predictand. Therefore, in principal component regression it is essential to keep all the principal components and determine which are the best predictors.

The principal components $\mathbf{P}$ themselves are created by projecting the original predictors (Z scores) onto the orthonormal eigenvectors as follows:

$$\mathbf{P} = \mathbf{Z}\mathbf{E} \tag{2.36}$$

Note that $\mathbf{Z}$ is the multivariate predictor set in Eq. (2.32). $\mathbf{P}$ turns out to be of the same dimensions of $\mathbf{Z}$ as expected. Graphically, the eigenvectors represent a coordinate transformation. The eigenvector associated with the most variance (largest λ) points in the direction of the largest variance within the data set. The second eigenvector associated with the second largest λ points in the direction of the next largest variance constrained to be orthogonal to the first eigenvector and so on. This can be easily visualized in two dimensions, as in Figure 2.6. The same thing occurs for p dimensions, but is impossible to visualize.

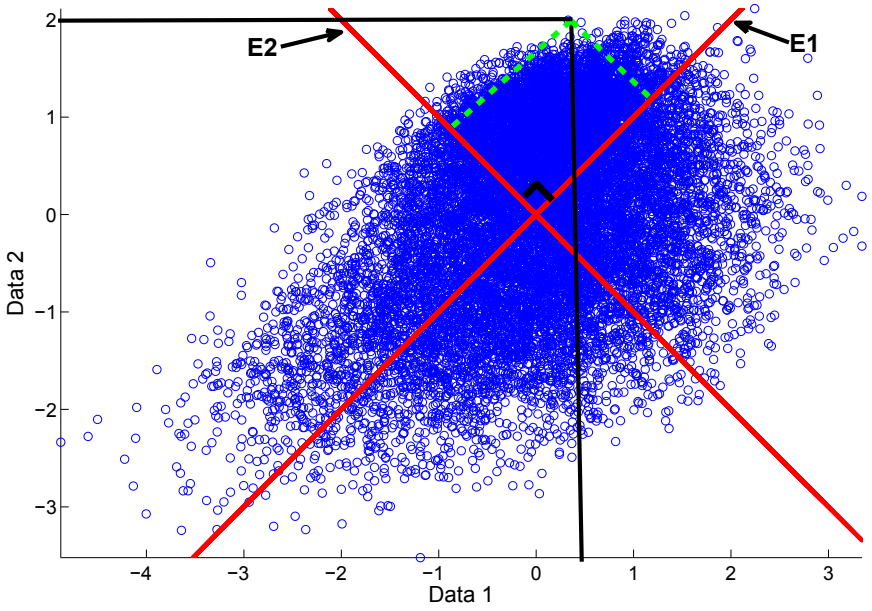

Figure 2.6: A scatter plot of two data sets (Z scores) is plotted in blue. It is clear that there is a linear relationship between the two variables. The original coordinates are shown by the black lines. The eigenvectors found from the covariance matrix of the data are plotted in red. The first eigenvector (E1) points in the direction of largest variance of the data and the second eigenvector (E2) is orthogonal to the first. The PC's are the new coordinates of data with respect to the eigenvectors. The PC's are found by projection of the data onto the eigenvectors shown in green.

Principal components are very useful in terms of regression, since they are orthogonal and covariance among the predictors has been eliminated. The regression coefficients ($\boldsymbol{\beta}$) requires the inverse of a matrix to be calculated (see Eq. (2.10)). Since an inverse requires taking the determinant, predictors that have a large covariance will result in a matrix that

is ill-conditioned for taking the inverse. This will inflate the regression coefficients and result in over fitting. Consider a special case in which two predictors which are exactly the same are used. The least squares estimate will not ignore one of the predictors. It will actually split the predicted variance between the two regression coefficients. The first coefficient will be large and the second one will be large in the opposite sense to compensate. This is the main reason why principal component regression (linearly independent predictors) is so desirable.

If it turns out that the smallest eigenvalue PC's are the best predictors then it is necessary to be cautious. The PC's that explain the least variance in the data set can represent interdependencies or conserved quantities in the original data. Since PC's are a linear combination of the original predictors, sometimes when looking at the eigenvector associated with the smallest λ, you will see two or three of the variable weightings add to zero. This essentially means that these three variables are roughly a conserved quantity since the PC's variance is almost zero. This also suggests there is a linear dependence in the data set and you should consider removing one of the variables that makes up that eigenvector. The PC's explaining the lowest variance in the data set is discussed further in Chapter 4 in relation to the thesis downscaling.

A good reference on the topic discussed in this section is the book entitled "Applied Mutivariate Statistical Analysis" (*Johnson and Wichern*, 2002).

CHAPTER 3

DATA INVESTIGATION

Statistical Downscaling (SD) often uses large data sets to determine the regression coefficients from observations over a historical period. Once the best regression coefficients are found, the same suite of predictors (same as reanalysis predictors which we call observations) from a General Circulation Model (GCM) can be used in the regression to make future projections of the predictand. In multivariate statistical analysis it is important to become familiar with the data sets which can help with interpretation of the results. The sources of the data and an investigation of the data are the essence of this chapter.

3.1 The Predictands

The climate variables chosen for making the projections in this thesis are maximum and minimum daily temperatures (Tmax and Tmin, respectively). Forty years ($1961 - 2000$) of homogenized maximum and minimum temperatures (*Vincent et al.* (2002)) were taken from Shearwater International Airport shown in Figure 1.3 of Chapter 1. Homogenized implies the data have been corrected for instrument and location changes. Any years that were leap years required the removal of February 29. Therefore all years will have 365 days which makes the analysis less complicated. The total number of predictand values is 14600 days for both Tmax and Tmin.

3.2 Maximum Daily Temperature

Observed Tmax has a positive linear trend of 0.047°C per year (i.e. 0.00013°C per day) in the $1961-2000$ period. This trend leads to an increase in the mean maximum temperature

of about $1.9°$C over the forty year period between $1961 - 2000$. It is unknown how much of this warming can be attributed to anthropogenic forcing, since climate varies due to natural forcing as well. Several patterns of temporal variability can be seen in a plot of Tmax (Figure 3.1). The largest variance in the data is the seasonal cycle which is the long repeating sinusoidal like pattern that repeats every year. There is also temporal variability at lower frequencies (longer period), with typical periods longer than a year, which is known as the interannual variability. Finally the high frequency variability can be seen as the quickly fluctuating pattern around the seasonal cycle, which can be considered as the weather.

A power spectrum of Tmax reveals a dominate peak at a period of one year. This means most of the observed variance in Tmax comes from the annual cycle. The seasonal cycle is a combination of the annual cycle and other harmonics with periods less than one year. The annual cycle (first harmonic) is the most important part of the seasonal cycle. The 6-month and 4-month harmonics are very small in comparison to the annual cycle. Going beyond three harmonics is not necessary because they have little influence. The seasonal cycle has a similar amplitude to the annual cycle, but it is distorted from the shape of a true sine wave. The origin of the harmonics are of interest, but beyond the scope of this work. The annual cycle is caused by our orbit around the sun. Less certain is the origin of the six and four month harmonics.

From a statistical downscaling point of view, the average seasonal cycle in the $1961 - 2000$ period is deterministic in comparison with variability at other frequency bands. Thus the seasonal cycle should be removed before developing the regression. This is because the regression will try to predict the most variance in the data. If the seasonal cycle is not removed, the regression will try to predict the seasonal cycle instead of the weather related variability. The average seasonal cycle plotted in Figure 3.1 can be determined using the following regression of sines and cosines with the associated regression coefficients shown in Table 3.1. The regression coefficients ($\mu, \beta_1 ... \beta_6$) reveal the importance of each harmonic.

$$S(t) = \mu + \beta_1 sin(\omega t) + \beta_2 cos(\omega t) + ... + \beta_5 sin(3\omega t) + \beta_6 cos(3\omega t) \qquad (3.1)$$

Note that ω is defined as 2π divided by 365 days, and $t = 0$ represents January 1st 1961 in this thesis.

The bottom panel of Figure 3.1 shows the variance of Tmax throughout the year. This

was done by calculating the variance of each day throughout the forty years $(1961-2000)$. For example, the value on the vertical axis associated with the first value on the x-axis in the bottom panel of Figure 3.1 is the variance of the forty values of January 1st in the $1961-2000$ period. The observed variance of Tmax is much higher in winter than summer, which is expected, since baroclinicity makes weather stronger in the winter than the summer. Since the main function of weather is to transport heat and momentum, it makes sense to have larger Tmax swings in winter than summer. This also fits in with our experience as the large swings in weather and temperature occur in the winter months in Atlantic Canada.

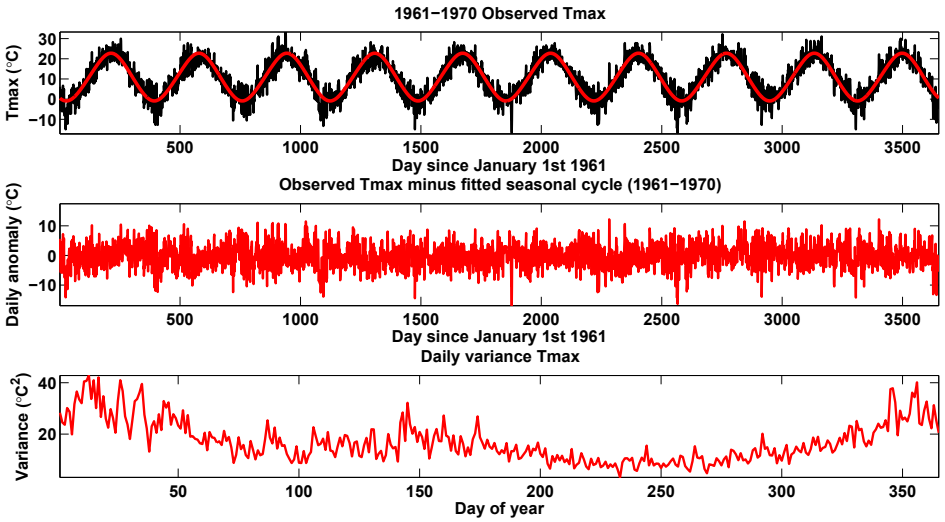

Figure 3.1: The top panel shows 10 years (1961-1970) of observed Tmax from Shearwater Airport, NS. Shown in black is the observed data. A fitted three harmonic seasonal cycle (fitted over 1961-2000) is plotted over the observed data in red. The middle panel is the daily anomaly. This is found by subtracting the red line from the black line in the top panel. The bottom panel gives the variance in Tmax for each day of the year over the forty years of data (1961-2000).

3.3 Minimum Daily Temperature

Observed Tmin from $1961-2000$ also has a linear trend of $0.058°$C per year (i.e. $0.00016°$C per day). This leads to a warming of the average minimum temperature of $2.3°$C over the forty year period. Again it is not clear how much of this warming is due to anthropogenic

Variable	Coefficient
μ	10.8249
β_1	-10.1219
β_2	-5.6479
β_3	-0.0966
β_4	0.1851
β_5	-0.0197
β_6	-0.2274

Table 3.1: Numerical regression coefficients for the seasonal cycle of observed Tmax. Units of degress Celsius

Variable	Coefficient
μ	3.0680
β_1	-9.4681
β_2	-5.7860
β_3	-0.4993
β_4	0.0566
β_5	-0.2056
β_6	-0.7212

Table 3.2: Numerical regression coefficients for the seasonal cycle of observed Tmin. Units of degress Celsius

forcing. The same patterns of temporal variability including the seasonal cycle, interannual variability, and the weather can be seen in the Tmin data as in Tmax. A seasonal cycle with three harmonics can be removed via Eq. (3.1), similar to Tmax with the associated regression coefficients shown in Table 3.2.

3.4 Tmax and Tmin Differences

As mentioned previously, the observed Tmax and Tmin both have positive linear trends defined over the $1961 - 2000$ period. These trends dictate Tmin has warmed more than Tmax over the forty year period. The minimum mean (Table 3.2) is lower than the maximum mean (Table 3.1) as expected. The seasonal cycle regression coefficients associated with the annual cycle (β_1 and β_2) from Table 3.1 and Table 3.2 are dominant (much larger) compared the other regression coefficients. The Tmax seasonal cycle has a similar amplitude to the Tmin seasonal cycle as would be expected.

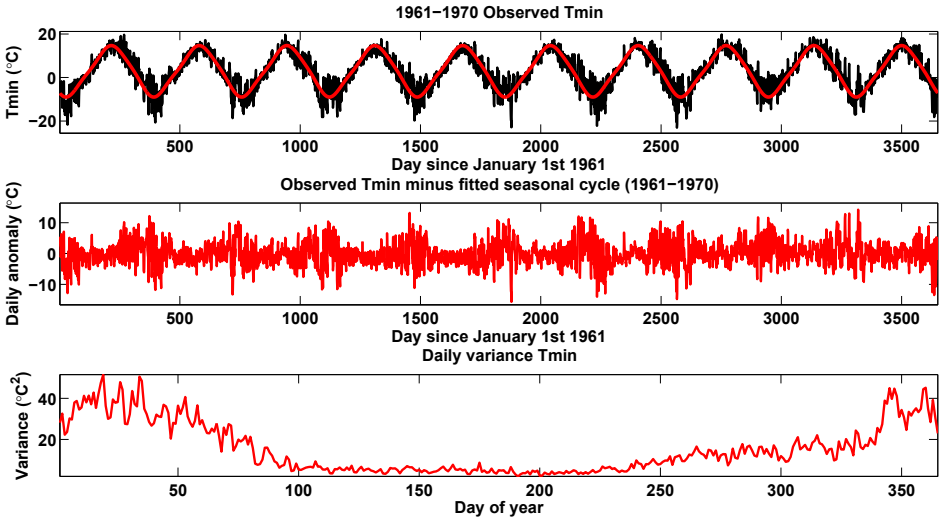

Figure 3.2: The top panel shows 10 years (1961-1970) of observed Tmin from Shearwater Airport, NS. Shown in black is the observed data. A fitted three harmonic seasonal cycle (fitted over 1961-2000) is plotted over the observed data in red. The middle panel is the daily anomaly. This is found by subtracting the red line from the black line in the top panel. The bottom panel gives the variance in Tmin for each day of the year over the forty years of data (1961-2000).

The daily variance shown in Figure 3.1 and 3.2 reveals that Tmin varies more than Tmax in the winter, but varies less than Tmax in summer. In winter, the sun is a secondary factor to advection when it comes to temperature. The amount of sun reaching the ground and Tmax are anticorrelated, on a daily time scale. Sunny days in winter are usually associated with cold advection and falling temperatures. Atlantic Canada is subject to frequent rapidly intensifying low pressure systems and strong high pressure systems in winter. Intensifying low pressure systems have stronger cold advection than warm advection. At night, cold advection leads to cold clear nights. Cloudless skies, caused by the subsidence associated with cold advection, leads to more radiative cooling (positive feedback). Winter Tmax usually occurs in the daytime where the sun has little influence due to snow cover. Also, radiative heating from the sun in winter makes the boundary layer unstable and induces low clouds which is a negative feedback on Tmax. These are some reasons why Tmax does not vary as much in winter as Tmin. In summer, the influence of large scale advection on day-to-day weather is relatively weak. However, the relative influence of the amount of sun reaching the ground on day-to-day weather is stronger. At night in summer, Tmin is forced by radiative cooling which is usually slow and steady. Tmax is defined by the amount of radiation, cloud cover and soil moisture. These again are some of the reasons Tmax varies more in summer than Tmin.

3.5 The Predictors

The predictors used in this thesis were created by an independent source referenced below. The predictors were all downloaded from the Canadian Climate Change Scenarios Network (CCCSN) website (www.cccsn.ca). No choice in the predictors to be downloaded was available. There are two types of predictor data sets downloaded for the statistical downscaling in this thesis. The first is reanalysis data produced from the National Center for Environmental Prediction (NCEP, *Kistler and Kalnay*, (2001)) and the second from a GCM of choice, in this case the Canadian General Circulation Model Version 3 (CGCM3). Twenty five predictors (see Table 3.3) from NCEP for 1961 to 2000 were obtained from the CCCSN website (www.cccsn.ca). Details on the NCEP predictors creation are found in *Gachon et al.* (2008). Note that the geostrophic winds are used as predictors instead of the real winds. It was determined in *Gachon et al.* (2008) that the CGCM3 was poor at producing a realistic real wind distribution compared to NCEP.

However, the CGCM3 was skillful in producing the geostrophic wind distribution. Since the goal of this thesis work is to use CGCM3 predictors in the regression, it was necessary to use the geostrophic winds as predictors instead of the real winds.

A quick review of the major steps in creating these predictors follows (*Gachon et al.* (2008)). The predictors, which we refer to as NCEP predictors, are loosely referred to as observations since they are used to train the downscaling regression. However, NCEP predictors are not really observations. They come from running a general circulation model (resolution 2.5°lat $\times$ 2.5° lon) with observations assimilated into the model. It is assumed that NCEP predictors are representative of the state of a predictor at a given time and space. Since the goal of this thesis is to use the CGCM3 predictors in the trained regression, the NCEP predictors were interpolated onto the larger (3.75°lat x 3.75° lon) CGCM3 Gaussian grid. This is necessary to ensure the predictors are representative of the same geographic space. A daily predictor value was derived by averaging the 6 hourly (four per day) values. The reason for the daily average predictors is because Tmax and Tmin have one daily value, and the same number of measurements in time of the predictors and predictand is required for regression. One further step of standardizing the predictor values (Z scores) with respect to the 1961 to 1990 means ($\bar{X}$) and standard deviations (σ) was done with the following expression:

$$Z_i = \frac{X_i - \bar{X}_{61-90}}{\sigma_{61-90}} \tag{3.2}$$

The 1961 to 1990 period was chosen as the standardization period because it was considered a base climate period used in other SD work. The Z-scores are created separately for each data set (CGCM3 and NCEP). This is a necessary step to remove any biases between the predictors from CGCM3 and NCEP. Finally the predictors had the seasonal cycles removed from each predictor via Eq. (3.1) to obtain the daily anomalies.

The same set of twenty five predictors (see Table 3.3) from the CGCM3 were also used. Again these predictors were downloaded from www.cccsn.ca. More details on there creation can be found in *Gachon et al.* (2008). The CGCM3 is a general circulation model with the same ocean component as the CGCM2 (*Flato and Boer*, 2001)). However, the CGCM3 uses a substantially updated atmospheric component known as the AGCM3 (*McFarlane et al.* (2006)), of which the major improvements are in the treatment of land processes, water vapor transport and cumulus parameterization. The CGCM3 has a resolution of 3.75°lat $\times$ 3.75° lon with 31 vertical levels. Predictors from the CGCM3 are

daily mean variables. The daily value is achieved by averaging the 6 hourly (4 per day) values similar to NCEP. Daily values from 1961-2100 were standardized using Eq. (3.2) where $\bar{X}$ and σ are the mean and standard deviation of the respective CGCM3 predictors from $1961 - 1990$. The CGCM3 predictors were divided into two parts. The historical CGCM3 predictors are the predictors from $1961-2000$ and the future CGCM3 predictors refers to the 2001-2100 period. Again the seasonal cycle for the historical period was removed via Eq. (3.1) from each of the predictors as discussed previously.

Removal of the seasonal cycle from the future predictors needs to be treated more carefully. The mean (μ) removed via in Eq. (3.1) was added back on to the future predictors due to the fact the predictors were standardized with respect to the $1961 - 1990$ means and standard deviations via Eq. (3.2). Using a $1961 - 1990$ standarization period will create a non-zero mean in the future predictor Z-scores, since the future predictors likely have a different mean than the $1961-1990$ reference period. Removing the mean (via Eq. (3.1)) from the future predictors would eliminate the trend in the predictors maintained by the consistent standardization period. It is important to retain the trend in the predictors if the trained regression is to be used with future predictors. The predictor trends are what contains the information on the climate change signal and allows for projections of Tmax and Tmin. This is discussed further in chapter 4. For full details on the predictor creation see the document entitled Predictor Datasets derived from the CGCM3.1 T47 and NCEP/NCAR Reanalysis (*Gachon et al.* (2008)) on the Canadian Climate Change Scenarios website (www.cccsn.ca).

Surface processes like heat flux in the CGCM3 depend on whether its numerical grid is an ocean box or a land box. Figure 3.3 shows the land-sea scheme from the CGCM3. The detailed geography in the figure is for reference, but does not represent the actual geography seen by the CGCM3. The CGCM3 would have much less geographical detail due to its coarse resolution. The main purpose of Figure 3.3 is to be able to determine how specific geography and locations are represented by the course grid of the CGCM3. Shearwater, Nova Scotia for example, is the chosen site to be downscaled and is represented by an ocean box within the CGCM3. In fact, all of Nova Scotia is in an ocean box.

Number	Name
1	Mean sea level pressure
2	1000hpa wind speed
3	1000hpa zonal wind
4	1000hpa meridional wind
5	1000hpa vorticity
6	1000hpa wind direction
7	1000hpa divergence
8	500hpa wind speed
9	500hpa zonal wind
10	500hpa meridional wind
11	500hpa vorticity
12	500hpa wind direction
13	500hpa divergence
14	500hpa geopotential height
15	850hpa wind speed
16	850hpa zonal wind
17	850hpa meridional wind
18	850hpa vorticity
19	850hpa wind direction
20	850hpa divergence
21	850hpa geopotential height
22	500hpa specific humidity
23	850hpa specific humidity
24	1000hpa specific humidity
25	Surface mean temperature

Table 3.3: The names of the twenty five predictors used from both NCEP and the CGCM3. It should be noted that the winds are the geostrophic winds.

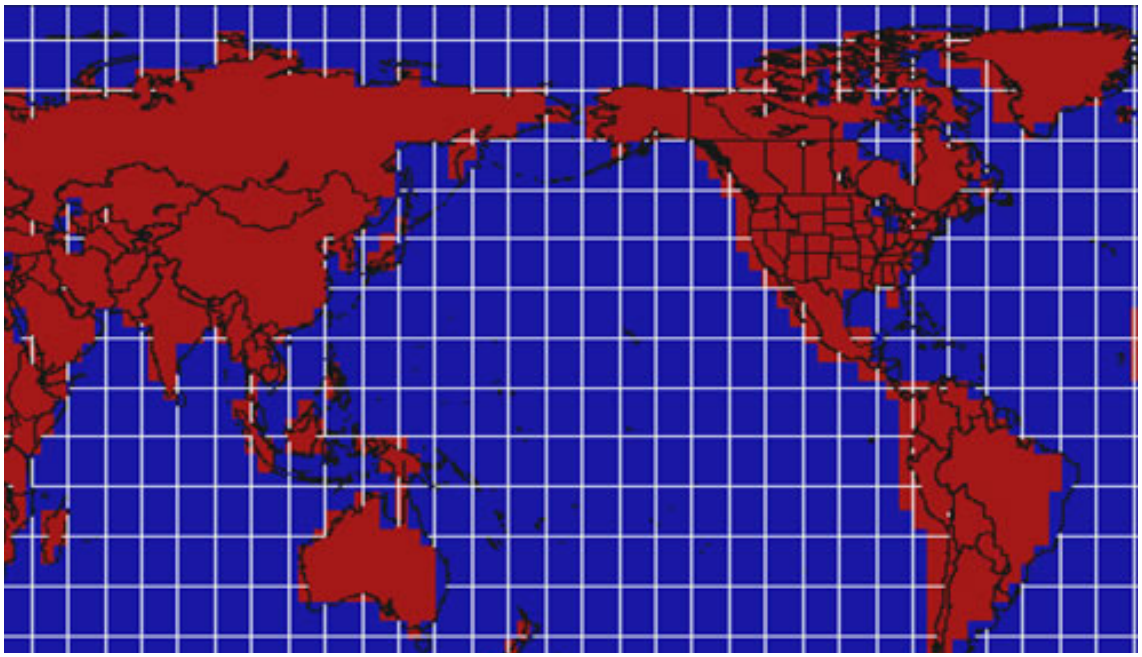

Figure 3.3: A mercator projection of the planet is shown. Overlaid is the lat-lon grid (square boxes). The geographic locations that the CGCM3 sees as land is in red. The blue represents where the CGCM3 has an ocean box.

CHAPTER 4

METHODOLOGY AND RESULTS

The steps in the statistical downscaling process and the associated results are the subject of this chapter. The first step in any statistical downscaling process is to choose predictors through an effective predictor selection process. The second step is the development of the regression using historical observations (NCEP predictors and observed temperature). This involves a validation process and a check that the final regression makes physical sense. Once the best historical regression is created from observations, the same predictors from the CGCM3 are used in the regression to ensure the statistical properties of the historical predictand distribution can be captured. Finally, the future CGCM3 predictors are used in the regression to make future projections of the predictand distributions.

4.1 The Predictor Selection Process

As described in Chapter 3, there are two predictor sets used in this study. The first is the daily NCEP predictors from $1961 - 2000$. The second is the daily CGCM3 predictors which is divided into the historical ($1961 - 2000$) and future ($2001 - 2100$) predictors. The process begins with the 25 predictors shown Table 4.1. Each had their seasonal cycles removed to produce the daily anomaly. We also eliminated predictors that are either not useful or redundant.

The goal of principal component analysis is to find the directions of greatest variance, so it makes sense to remove total wind speed and direction in favor of the zonal (U) and meridional (V) wind, since they are mathematically directly related to each other as follows:

$$Speed = |\vec{U}| = \sqrt{U^2 + V^2} \tag{4.1}$$

$$Direction = \theta = \arctan\left(\frac{V}{U}\right) \tag{4.2}$$

The total wind speed and direction are predictors at three levels (500 hpa, 850 hpa and surface). Removing them amounts to a reduction of six predictors from the initial 25, leaving 19 predictors.

It was found that the divergence and the meridional wind speed had a correlation of negative one at all three levels. This is because the winds are geostrophic. It can be easily verified that the geostrophic meridional wind, and its divergence are related by a constant on a β plane (*Holton*,2004) as follows

$$\vec{\nabla} \cdot \vec{U_g} = \left(\frac{-\beta}{f_o}\right) V_g. \tag{4.3}$$

The constant is the derivative of the Coriolis parameter with latitude ($\beta = \frac{\partial f}{\partial y}$) divided by the Coriolis parameter at a given latitude (f_o).

Since the goal is to develop a regression, either the meridional wind or the divergence is redundant. From a prediction point of view they are identical and have identical predictive power. Failure to remove one of them would lead to inflation of the regression coefficients as discussed in Chapter 2. It was decided to remove the divergence at all three levels (500 hpa, 850 hpa and surface). This amounts to a further reduction of three predictors, for a total of 16 remaining candidate predictors.

One final important step was taken to ensure the predictors are appropriate for the downscaling process. Predictors should be chosen based on both their physical relevance to the predictand, and their accurate representation by the climate model used for the climate change simulation (*Wilby and Dawson*,2004). Validation of climate model output (such as the CGCM2 and HadCM3) has shown that GCM's have problems representing the distribution of near surface variables, such as specific humidity and temperature compared to the NCEP reanalysis (*Dibike et al.*,2008).

It is known that different physics operate in different seasons, and the GCM may have different biases in different seasons. Therefore, the regression analysis was done in each season independently. The remaining sixteen predictors from the CGCM3 and NCEP for the historical period ($1961 - 2000$) were separated into the following seasons; winter (DJF), spring (MAM), summer (JJA), fall (SON). Next, predictor distributions were viewed in terms of their Probability Density Function (PDF). Each predictor (CGCM3 vs

NCEP) in each season was subjectively compared. It is not meaningful to compare directly the time series of the CGCM3 and NCEP predictors because the CGCM3 is not data assimilated and cannot reproduce the so-called observed predictors from NCEP. However, the GCM should be capable of representing the distribution of NCEP predictors. CGCM3 variables that were viewed to have major differences in distributions compared to NCEP were omitted as a predictor, to reduce the set further.

For clarity, Figure 4.1 shows a Probability Density Function (PDF) of surface specific humidity in summer (JJA) from both NCEP and the CGCM3. NCEP specific humidity is shown in black and the CGCM3 specific humidity is shown in red. The CGCM3 has a very different distribution of surface specific humidity than NCEP reanalysis and, therefore, is not useful for statistical downscaling. It should be noted that any PDFs included in this chapter are created from binning the data. This means the data was separated into bins that were three degrees Celsius wide over the range of the data. The probability on the vertical axis (of all PDF plots) refers to the number of occurrences within a particular bin divided by the total number of observations. The exception is Figure 4.1, where the binning is done in intervals of 0.3 of a Z-score, since this figure is the PDF of predictor Z-scores.

Two sets of predictors were obtained, one for spring and summer and the other for fall and winter. It should be noted that after filtering (comparing NCEP/CGCM3) the surface predictors had large differences in their PDF's, and needed to be removed. This is consistent with the previous study made by *Dibike et al.*(2008) who indicated that GCM's do not reproduce surface variables in a realistic manner, compared to observed. Table 4.1 for fall and winter, and Table 4.2 for spring and summer, show the original 25 predictors on the left. The right column in each table is the final subset after removing the total wind speed, wind direction, and divergence at all three levels. The right column also includes the reduction following the subjective distribution comparison of the NCEP and CGCM3 predictors, for the years $1961 - 2000$.

It is interesting to note that surface mean temperature is not a valid predictor for downscaling in this case. Even though it is a good predictor in reality (Tmin, Tmax and Tmean have a high daily correlation), it cannot be included because the CGCM3 does not reproduce it realistically, compared to NCEP. Although the best regression is sought after, predictive power is not the only consideration. If variables that have predictive power

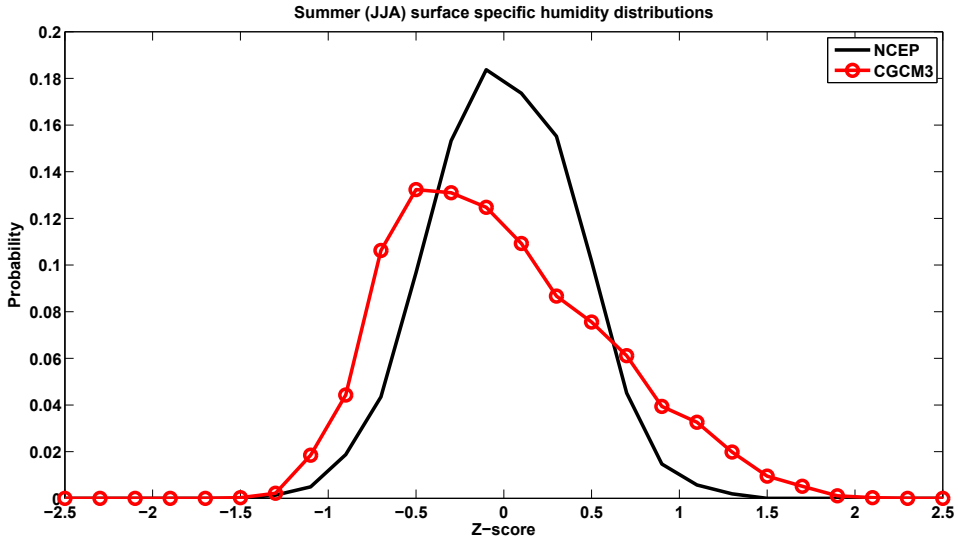

Figure 4.1: Probability density function (NCEP, black and CGCM3, red) of surface specific humidity in summer (JJA) between 1961 and 2000. Note that the distributions are from the CGCM3 grid box that is over Shearwater, NS. The NCEP distribution is from the NCEP predictors for Shearwater that were mapped onto the CGCM3 grid as discussed in Chapter 3. It should be noted also the predictors are Z-scores described in Chapter 3.

were included despite the ability or inability of the GCM to represent them, then the projections are not credible. This is because the downscaled Tmax or Tmin generated using NCEP predictors will be different from the downscaled Tmax or Tmin generated using CGCM3 predictors, if NCEP and CGCM3 have different predictor distributions.

4.2 Regression Development

As discussed in Chapter 2, it is desirable to have independent predictors for a regression. Following the mathematics in Section 2.3, the final predictor subsets from NCEP (Tables 4.1 and 4.2) were transformed into their principal components (PC's). This resulted in twelve PC's for fall and winter and ten PC's for spring and summer. The principal components were the predictors used to train the regression in each season, instead of the predictors themselves. Once the principal components were determined, their correlation (see the discussion in next paragraph) with both Tmax and Tmin for each season was determined to see which PCs were useful predictors. The results of the PC analysis and regression training are shown in Tables 4.3 through 4.10. The regression training was

Number	Original Predictors	Final subset Predictors
1	Mean sea level pressure	Mean sea level pressure
2	1000hpa wind speed	1000hpa zonal wind
3	1000hpa zonal wind	1000hpa meridional wind
4	1000hpa meridional wind	1000hpa vorticity
5	1000hpa vorticity	500hpa zonal wind
6	1000hpa wind direction	500hpa meridional wind
7	1000hpa divergence	500hpa vorticity
8	500hpa wind speed	500hpa geopotential height
9	500hpa zonal wind	850hpa zonal wind
10	500hpa meridional wind	850hpa meridional wind
11	500hpa vorticity	850hpa vorticity
12	500hpa wind direction	850hpa geopotential height
13	500hpa divergence	
14	500hpa geopotential height	
15	850hpa wind speed	
16	850hpa zonal wind	
17	850hpa meridional wind	
18	850hpa vorticity	
19	850hpa wind direction	
20	850hpa divergence	
21	850hpa geopotential height	
22	500hpa specific humidity	
23	850hpa specific humidity	
24	1000hpa specific humidity	
25	Surface mean temperature	

Table 4.1: The names of the original twenty five predictors used from both NCEP and the CGCM3 and the reduced subset used in the downscaling process for fall and winter. Note the winds are the geostrophic winds.

Number	Original Predictors	Final subset Predictors
1	Mean sea level pressure	Mean sea level pressure
2	1000hpa wind speed	1000hpa zonal wind
3	1000hpa zonal wind	1000hpa meridional wind
4	1000hpa meridional wind	500hpa zonal wind
5	1000hpa vorticity	500hpa meridional wind
6	1000hpa wind direction	500hpa vorticity
7	1000hpa divergence	500hpa geopotential height
8	500hpa wind speed	850hpa zonal wind
9	500hpa zonal wind	850hpa meridional wind
10	500hpa meridional wind	850hpa vorticity
11	500hpa vorticity	
12	500hpa wind direction	
13	500hpa divergence	
14	500hpa geopotential height	
15	850hpa wind speed	
16	850hpa zonal wind	
17	850hpa meridional wind	
18	850hpa vorticity	
19	850hpa wind direction	
20	850hpa divergence	
21	850hpa geopotential height	
22	500hpa specific humidity	
23	850hpa specific humidity	
24	1000hpa specific humidity	
25	Surface mean temperature	

Table 4.2: The names of the original twenty five predictors used from both NCEP and the CGCM3 and the reduced subset used in the downscaling process for spring and summer. Note the winds are the geostrophic winds.

done on the data from $1961 - 1990$.

Correlation between a predictor X (one of the developed PC's from the original predictors) and a predictand Y (Tmax or Tmin) is defined mathematically as:

$$R_{XY} = \frac{1}{n} \sum_{i=1}^{n} \frac{(X_i - \mu_X)(Y_i - \mu_Y)}{\sigma_X \sigma_Y} \tag{4.4}$$

where i refers to an individual element and σ refers the standard deviation of X and Y. n refers to the total number of observations. Recall that the means (μ_X and μ_Y) have already been removed from X and Y. The predictors are Z-scores (then their seasonal cycle was removed from the Z-score) so by definition their means are zero. The seasonal cycle was removed from the predictands (which are not Z-scores) which removed their means as well. Since the means of the predictands and predictors are zero, Eq. (4.4) is even simpler because the means drop out (are zero).

A critical correlation value was chosen, such that if the correlation between the predictor (PC) and predictand (Tmax or Tmin) was greater than or equal to that value, then that predictor was used in the regression. If the correlation was less than that critical correlation value, that predictor was not used in the regression analysis. This is known as a correlation cut-off. The absolute value of 0.1 was chosen as the correlation cut-off, based on a parameter known as γ^2 after *Thompson and Sheng* (1997). Gamma squared defined below is the ratio of the variance in the errors of the regression (See Eq. (2.2)) to the variance in the predictand. Referring to Section 2.1, it can be easily seen that gamma squared is equivalent to SSE in Eq. (2.16) divided by SST in Eq. (2.14). The terms defining Eq. (2.17) show that gamma squared is equal to one minus the explained variance of the regression ($\gamma^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$). Gamma squared is a measure of regression accuracy and is defined mathematically as follows:

$$\gamma^2 = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i)^2} \tag{4.5}$$

where i is the index referring to each time measurement of the observations (Y_i) and the prediction ($\hat{Y}_i$). Again, n is the total number of observations. Also, it should be noted the mean is assumed to be zero in Eq. (4.5), which is true in this thesis work. Clearly if the variance in the errors is zero (numerator of Eq. (4.5)), then the regression can predict the predictand perfectly. In other words, as Gamma squared decreases, the prediction accuracy increases. Gamma squared was calculated using various correlation cut-offs

PC	PC percent explained variance	Correlation	Regression Coefficients (β)
			β_0 = 0.03
1	35	0.12	0.32
2	26	0.54	1.66
3	22	-0.14	-0.48
4	6	0.39	2.57
5	5	0.06	
6	3.1	0.15	1.16
7	1.3	0.05	
8	0.74	0.32	5.80
9	0.53	-0.28	-5.58
10	0.2	-0.01	
11	0.1	0.07	
12	0.03	0.27	22.09

Table 4.3: The percentage of total variance explained by each of the twelve principal components in the original twelve possible predictors (NCEP) used in the regression analysis for winter Tmin. Each PCs correlation with Tmin is also shown in winter. The last column on the right shows the regression coefficients (units of degrees Celsius), using the principal components. Note that PCs with a correlation less than 0.1 were not used in the regression and therefore have no regression coefficient displayed.

and it was subjectively determined that an absolute value correlation of 0.1 allowed a sufficient accuracy. That is, increasing the number of predictors used in the regression did not decrease the gamma squared significantly. The values of gamma squared for each regression in the training period are shown in Section 4.4 (Tables 4.14 and 4.15).

In Tables 4.3 through 4.10, the percentage of total variance explained in the original predictor data set by each principal component used in the regression and its correlation with both Tmax and Tmin are shown. Note that the percent explained variance of the PCs sums to one hundred percent since the original predictors and the new predictors (PCs) have the same information. Note that only predictors (PCs) that a had an absolute value correlation of 0.1 (determined from γ^2) or greater with Tmax and Tmin have an associated regression coefficient in Tables 4.3 through 4.10. PCs that had a lower correlation were not used in the regression.

As discussed at the end of Section 2.3, caution should to be taken when using the PCs that explain the least variance in the predictor set in the regression. Some of these PCs can be representative of interdependicies among the predictors. It is well known that these PCs can sometimes lead to over fitting the regression. In each season, the PC that explained

PC	PC percent explained variance	Correlation	Regression Coefficients (β)
			β_0 = -0.05
1	35	0.23	0.55
2	26	0.62	1.72
3	22	0.07	
4	6	0.40	2.28
5	5	0.11	0.66
6	3.1	0.21	1.55
7	1.3	0.01	
8	0.74	0.20	3.25
9	0.53	-0.23	-4.07
10	0.2	-0.01	
11	0.1	0.08	
12	0.03	0.18	13.41

Table 4.4: The percentage of total variance explained by each of the twelve principal components in the original twelve possible predictors (NCEP) used in the regression analysis for winter Tmax. Each PCs correlation with Tmax is also shown in winter. The last column on the right shows the regression coefficients (units of degrees Celsius), using the principal components. Note that PCs with a correlation less than 0.1 were not used in the regression and therefore have no regression coefficient displayed.

PC	PC percent explained variance	Correlation	Regression Coefficients (β)
			β_0 = -0.09
1	31	-0.29	-0.55
2	29	-0.01	
3	24	-0.31	-0.63
4	6	0.31	1.32
5	3.8	-0.17	-0.79
6	2.9	-0.02	
7	2.3	0.20	1.22
8	0.6	0.40	4.86
9	0.3	0.01	
10	0.1	0.03	

Table 4.5: The percentage of total variance explained by each of the ten principal components in the original ten possible predictors (NCEP) used in the regression analysis for spring Tmin. Each PCs correlation with Tmin is also shown in spring. The last column on the right shows the regression coefficients (units of degrees Celsius), using the principal components. Note that PCs with a correlation less than 0.1 were not used in the regression and therefore have no regression coefficient displayed.

PC	PC percent explained variance	Correlation	Regression Coefficients (β)
			β_0 = -0.11
1	31	-0.23	-0.52
2	29	-0.16	-0.33
3	24	0.13	0.36
4	6	0.31	1.63
5	3.8	-0.12	-0.74
6	2.9	-0.06	
7	2.3	0.21	1.55
8	0.6	0.35	5.04
9	0.3	0.01	
10	0.1	0.12	2.84

Table 4.6: The percentage of total variance explained by each of the ten principal components in the original ten possible predictors (NCEP) used in the regression analysis for spring Tmax. Each PCs correlation with Tmax is also shown in spring. The last column on the right shows the regression coefficients (units of degrees Celsius), using the principal components. Note that PCs with a correlation less than 0.1 were not used in the regression and therefore have no regression coefficient displayed.

PC	PC percent explained variance	Correlation	Regression Coefficients (β)
			β_0 = -0.09
1	31	0.18	0.32
2	29	0.30	0.53
3	22	0.01	
4	5.5	-0.15	-0.67
5	4.5	-0.22	-0.99
6	3.8	-0.29	-1.54
7	3	-0.03	
8	0.5	0.42	5.99
9	0.4	0.02	
10	0.3	0.05	

Table 4.7: The percentage of total variance explained by each of the ten principal components in the original ten possible predictors (NCEP) used in the regression analysis for summer Tmin. Each PCs correlation with Tmin is also shown in summer. The last column on the right shows the regression coefficients (units of degrees Celsius), using the principal components. Note that PCs with a correlation less than 0.1 were not used in the regression and therefore have no regression coefficient displayed.

PC	PC percent explained variance	Correlation	Regression Coefficients (β)
			β_0 = -0.09
1	31	-0.07	
2	29	0.12	0.35
3	22	-0.46	-1.62
4	5.5	-0.14	-0.99
5	4.5	-0.11	-0.79
6	3.8	-0.30	-2.56
7	3	-0.16	-1.50
8	0.5	0.28	6.41
9	0.4	-0.03	
10	0.3	0.03	

Table 4.8: The percentage of total variance explained by each of the ten principal components in the original ten possible predictors (NCEP) used in the regression analysis for summer Tmax. Each PCs correlation with Tmax is also shown in summer. The last column on the right shows the regression coefficients (units of degrees Celsius), using the principal components. Note that PCs with a correlation less than 0.1 were not used in the regression and therefore have no regression coefficient displayed.

PC	PC percent explained variance	Correlation	Regression Coefficients (β)
			β_0 = -0.03
1	36	0.04	
2	26	0.55	1.27
3	20	-0.13	-0.30
4	6	0.33	1.50
5	5	0.07	
6	3	0.08	
7	2	0.04	
8	0.8	0.05	
9	0.6	0.43	6.67
10	0.3	-0.04	
11	0.2	-0.06	
12	0.1	-0.28	-17.90

Table 4.9: The percentage of total variance explained by each of the twelve principal components in the original twelve possible predictors (NCEP) used in the regression analysis for fall Tmin. Each PCs correlation with Tmin is also shown in fall. The last column on the right shows the regression coefficients (units of degrees Celsius), using the principal components. Note that PCs with a correlation less than 0.1 were not used in the regression and therefore have no regression coefficient displayed.

PC	PC percent explained variance	Correlation	Regression Coefficients β
			β_0 = -0.05
1	36	0.12	0.22
2	26	0.52	1.15
3	20	0.15	0.38
4	6	0.41	1.89
5	5	0.17	0.97
6	3	0.07	
7	2	0.05	
8	0.8	-0.01	
9	0.6	0.38	5.68
10	0.3	0.04	
11	0.2	0.05	
12	0.1	-0.22	-13.43

Table 4.10: The percentage of total variance explained by each of the twelve principal components in the original twelve possible predictors (NCEP) used in the regression analysis for fall Tmax. Each PCs correlation with Tmax is also shown in fall. The last column on the right shows the regression coefficients (units of degrees Celsius), using the principal components. Note that PCs with a correlation less than 0.1 were not used in the regression and therefore have no regression coefficient displayed.

the least predictor set variance was checked in this thesis work. If an interdependence was observed in the eigenvector, then a predictor was removed, and the regression was trained again. We found that this made no difference in the regression results suggesting this step is not necessary, which is consistent with the finding in *Huth* (2002). *Huth* (2002) found that the PC explaining the least variance in the predictor set did not cause over fitting if used in the regression. In this research, an example occurs in winter where the PC that explains the least predictor set variance (see Tables 4.3 and 4.4) is used in the regression. The associated eigenvector shows that mean sea level pressure along with 500 hpa and 850 hpa geopotential height sum to zero. This PC has little variance and sums to a constant (zero), which suggests there is a relationship among the variables. However, removing mean sea level pressure and doing the regression training again yielded the same results.

4.3 Predictor Physics

As mentioned earlier, it is an important step to check whether the most important predictor in the regression has a known physical relationship with the predictand. This step

helps give confidence in the skill of the regression analysis. The highest correlated PC with the predictand (Tmax or Tmin) was examined in each season in this thesis. Since principal components are a linear combination of the original predictors, the importance of a particular predictor in making up a particular PC comes from the size of the coefficient or weight of that predictor in the linear combination making up that PC. Tables 4.11 through 4.13 show the correlation of the highest correlated PC with both Tmax and Tmin in each season. Also shown are the most important original predictors making up that PC chosen from their associated weights. Since the PCs are non-dimensional, an individual PC's associated weights from the original predictors squared must sum to one (similar to the components of a vector). Their actual weights can be interpreted as their relative importance in constructing that PC.

The largest fraction of the variance (Tmax or Tmin) is accounted for by the same PC in winter (Table 4.11). This PC explains about forty percent of the total variance in Tmax and roughly thirty percent of the total variance in Tmin. The dominant original predictors in this PC are meridional wind at all three levels, which is expected since temperature advection from the north/south causes Tmax and Tmin at Shearwater to decrease/increase in winter.

Before discussing the predictor physics in summer, caution is needed when considering physical processes, like the sea-breeze circulation. The predictor winds used in this thesis are geostrophic and therefore do not directly contain the sea-breeze. If we were using the real winds, it is likely that the effect of the sea-breeze on surface temperature would show up in the regression. As discussed in Chapter 3, when the predictors were created, it was determined that the real winds could not be used as predictors. This is due to the fact that the distribution of real wind produced by the CGCM3 is significantly different from the real wind distribution from NCEP in the grid box corresponding to Shearwater. The large grid-spacing in the GCM is likely to be the culprit of this major difference. This makes the GCM incapable of resolving boundary layer effects like the sea-breeze which have a large influence on the real wind. Similarly the large grid-spacing likely does not resolve the large accelerations from baroclinic storms that put the atmosphere out of geostrophic balance and influences the real wind in the upper atmosphere. We can anticipate that onshore geostrophic winds in the summer should bring down day time temperatures, due to the ocean influence. Offshore winds will bring day time temperatures

up, due to sensible heating as air parcels cross the land mass. This is similar to the sea-breeze, but is not actually the sea-breeze due to the fact the predictor winds used in the regression are geostrophic.

In spring, the same leading PC is found for both Tmax and Tmin (Table 4.11). The dominant predictor explains roughly ten to fifteen percent of the total variance in Tmax and Tmin. It is not surprising that this predictor is dominated by geopotential height, since it is related to thickness (difference in geopotential height between two pressure surfaces). The connection between thickness and mean layer temperature is easy to derive from hydrostatic balance. However, it should be noted that this PC does not explain a very large fraction of the observed variance in either Tmax or Tmin in spring. Temperature in spring depends also on other factors, such as ocean temperature and cloud cover, which are not considered here.

Summer is the only season that has a different leading PC for Tmax and Tmin (Table 4.12). Tmin depends on a PC similar to that which is important in spring. Roughly fifteen percent of the total Tmin variance is explained by geopotential height (related to thickness), which is not a large fraction of the total variance. Again Tmin in summer, which is similar to spring, is probably related to predictors not considered here. Tmax in summer has a leading PC that explains over twenty percent of the total variance in Tmax. This PC is negatively correlated with Tmax. In other words as the PC values go up, Tmax goes down and vice-versa. The combination of zonal and meridional wind with a negatively correlated PC is directly related to the influence of onshore/offshore wind on temperature. For example, as it can be seen in Table 4.12 a northerly 850 hpa wind (negative meridional wind, northerly) decreases the PC's value (all else being equal) which means Tmax would increase. The opposite is true for a southerly wind (positive) where Tmax would decrease. The zonal wind has a negative weighting. Therefore an easterly wind (negative) will increase the value of the PC which suggests Tmax would go down. A similar analysis shows that a westerly wind (positive zonal anomaly) increases Tmax. In summary, the sun makes a strong land/ocean temperature contrast, which is why onshore/offshore winds are more important in the day than night in influencing temperature.

Finally the leading PC for Tmax and Tmin in fall (Table 4.13) is very similar to winter. The same PC is important for both Tmax and Tmin. This PC is dominated by meridional wind, which suggest again that temperature advection is the most important forcing in

Winter Tmin and Tmax	Spring Tmin and Tmax
Tmin correlation = 0.54 Tmax correlation = 0.62	Tmin correlation = 0.40 Tmax correlation = 0.35
Important associated weights surface meridional wind = 0.52 850hpa meridional wind = 0.57 500hpa meridional wind = 0.48	*Important associated weights* 500hpa geopotential height = 0.81 500hpa vorticity = 0.41

Table 4.11: List of predictors with the largest weighting (greater than 0.35) associated with the leading PC in winter and spring. The leading PC is the PC that accounts for the largest fraction of total variance in the predictand. The correlation between the predictand and the leading PC is shown also.

Summer Tmin	Summer Tmax
Tmin correlation = 0.42	Tmax correlation = -0.46
Important associated weights 500hpa geopotential height = 0.76 500hpa vorticity = 0.38	*Important associated weights* 500hpa meridional wind = 0.46 850hpa meridional wind = 0.42 850hpa zonal wind = -0.41 surface zonal wind = -0.40

Table 4.12: List of predictors with the largest weighting (greater than 0.35) associated with the leading PC in summer. The leading PC is the PC that accounts for the largest fraction of total variance in the predictand. The correlation between the predictand and the leading PC is shown also.

the fall (and winter). Again this makes sense because temperature advection is the main influence when solar insolation is reduced.

Fall Tmin and Tmax
Tmin correlation = 0.55 Tmax correlation = 0.52
Important associated weights surface meridional wind = 0.52 850hpa meridional wind = 0.58 surface meridional wind = 0.52

Table 4.13: List of predictors with the largest weighting (greater than 0.35) associated with the leading PC in fall. The leading PC is the PC that accounts for the largest fraction of total variance in the predictand. The correlation between the predictand and the leading PC is shown also.

4.4 Validation of Regression and Regression Results

Validation is an important step in any regression analysis. We assess the performance of the newly developed statistical downscaling method in predicting Tmax and Tmin at Shearwater during two periods: the training period between 1961 and 1990 and the validation period between 1991 and 2000. The trained regression (trained on $1961-1990$ data) was used with the predictor data (PC's) from $1991-2000$ to predict Tmax and Tmin for that period $(1991-2000)$ in each season. The predicted Tmax and Tmin were then correlated with what was observed in the $1991-2000$ period to ensure the regression had predictive skill. This is known as cross validation, which confirms the regression can predict Tmax and Tmin data that is independent of the training period. Tables 4.14 and 4.15, one for Tmin and one for Tmax, show the percent explained variance of the regression found from training the regression in the $1961-1990$ period. The Tables also show the percent explained variance of the independent validation period $(1991-2000)$. Again, as discussed in Section 2.1, the percent explained variance is the square of the correlation multiplied by 100, between the predicted Tmax or Tmin in a given period and what was observed during the same period. Also shown in Tables 4.14 and 4.15 is the regression accuary of the training and validation periods, measured by gamma squared (γ^2) as discussed in Section 4.2.

Season	Training Period	Validation Period	γ_1^2	γ_2^2
$Winter$	78	77	0.22	0.23
$Spring$	51	49	0.49	0.51
$Summer$	45	45	0.55	0.55
$Fall$	69	68	0.31	0.32

Table 4.14: The percentage of total variance explained by the regression for Tmin during the training period (1961-1990, second column) and the validation period (1991-2000, third column). γ_1^2 and γ_2^2 shows the gamma squared (see Section 4.2) associated with the training period (γ_1^2) and the validation period (γ_2^2).

Season	Training Period	Validation Period	γ_1^2	γ_2^2
$Winter$	75	74	0.25	0.26
$Spring$	40	40	0.60	0.60
$Summer$	45	46	0.55	0.54
$Fall$	71	69	0.29	0.31

Table 4.15: The percentage of total variance explained by the regression for Tmax during the training period (1961-1990, second column) and the validation period (1991-2000, third column). γ_1^2 and γ_2^2 shows the gamma squared (see Section 4.2) associated with the training period (γ_1^2) and the validation period (γ_2^2).

As shown in Tables 4.14 and 4.15, the explained variance and γ^2 in the training and validation period are almost identical in each season. Therefore the developed regression is capable of predicting Tmax and Tmin with predictors independent of the training period. One final step of validating the regression assumptions was done. As discussed in Chapter 2, three major assumptions in linear regression are *Normality of Errors*, *Homoscedasticity* and *Independence of Errors*. In each season, the distribution of the regression errors was checked to make sure it looked normal. The errors were plotted against each predictor in the regression to ensure no patterns were observed. Also, the errors were plotted versus time to ensure no pattern was observed. From there it was determined that the regressions developed in the historical period between Tmax and Tmin and the NCEP predictors were not overfit and had no major violations of the assumptions on which they were based. Tables 4.14 and 4.15 show that the best regression occurs in the winter season, with fall being a close second. Spring and summer have significantly less regression accuracy than fall and winter. Spring and summer require the use of different predictors than the ones considered in the predictor selection process to get a higher explained variance.

The explained variances in Tables 4.14 and 4.15 show the regression itself does not predict one hundred percent of the total variance in the daily anomalies (predictand). To address this problem, a step is taken to try and account for the variance not explained by the regression. This is done by multiplying the predicted daily anomaly values by an inflation factor. This inflation factor is defined as the reciprocal of the correlation (R) between the observed daily anomaly and the predicted daily anomaly (*Huth*,2002)

$$I = \frac{1}{R}. \tag{4.6}$$

The inflation factor (I) can be easily found for Tmin and Tmax for all seasons with the explained variances shown in Tables 4.14 and 4.15. For example, the explained variance for winter Tmax is 0.78. This means the correlation between observed Tmax and predicted Tmax in winter is 0.88 (i.e. $\sqrt{0.78}$). Therefore the inflation factor is 1.14 (i.e. $\frac{1}{0.88}$). Similarly spring, summer, and fall have inflation factors of 1.40, 1.49 and 1.20 respectively.

It needs to be stressed that the inflation does not improve the regression fit. It can be shown that if gamma squared were calculated after inflation was completed, gamma squared would increase compared to the non-inflated prediction. Inflation is just a technique to allow the prediction to have a similar variance to what was observed. This issue is discussed in *Karl et al.* (1990) where they discuss the application of an inflation factor reduces the prediction accuracy, but ensures the variance of the prediction matches observations. Since the prediction error on any given day is not important, but the correct variance of the projections for climate studies are important, inflation is desirable for statistical downscaling.

Once the predictions for Tmax and Tmin have been inflated, a time series plot can be made to assess the performance of the statistical downscaling. Figures 4.2 and 4.3 are plots of two years of the time series for Tmax and Tmin. The first year is part of the training period (1990), the second year is part of the validation period (1991). These two plots show similar skills in both the training and validation period when predicting the predictand. Along with the values in Tables 4.14 and 4.15, these plots (Figures 4.2 and 4.3) are pretty convincing that the regresion has predictive skill.

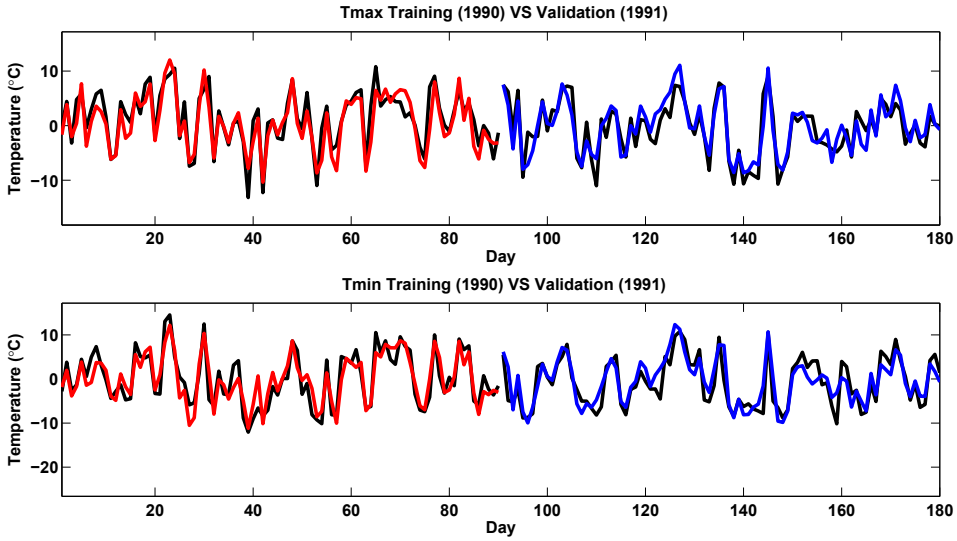

Figure 4.2: Top panel shows time series for two years $(1990 - 1991)$ of observed Tmax daily anomalies for winter (black). The inflated NCEP prediction of the daily anomaly for the last year of the training period (1990) is shown in red. The blue line represents the inflated NCEP prediction of the daily anomaly for first year of the validation period (1991). The bottom panel is identical to the top except it is for Tmin

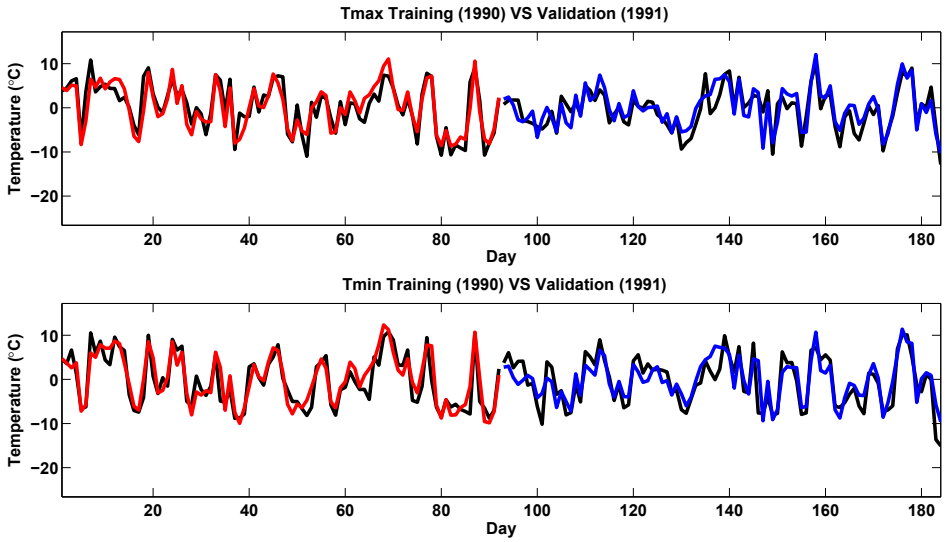

Figure 4.3: The top panel shows time series for two years $(1990-1991)$ of observed Tmax daily anomalies for summer (black). The inflated NCEP prediction of the daily anomaly for the last year of the training period (1990) is shown in red. The blue line represents the inflated NCEP prediction of the daily anomaly for first year of the validation period (1991). The bottom panel is identical to the top except it is for Tmin

To recover the total predicted values of Tmax and Tmin, the observed average seasonal cycles (see Chapter 3) needs to be added back on to the inflated prediction of the daily anomalies. Before doing that, hypothesis testing was employed to check that means and variances of the inflated predictions (daily anomalies) were similar to that of the observed daily anomalies. A Z-test (see Section 2.2) for the mean and an F-test (see Section 2.2) for the variance were used to test if the first two moments (mean and variance) of NCEP predicted Tmax and Tmin daily anomalies in each season were statistically the same as what was observed from $1961-2000$. In all four seasons, for both Tmax and Tmin, a ninety five percent hypothesis test was passed for both the mean and the variance. In other words, the developed regressions (NCEP predictors) with associated inflation factors are capable of predicting the statistical properties of the observed distributions. However it is noteworthy that the inflation factor was needed to pass the hypothesis test for variance.

As the explained variance of the regression goes down, the inflation factor goes up. Large inflation factors are required for spring and summer. As mentioned earlier, spring and summer depend on other factors (predictors) not considered in this thesis. Although the CGCM3 and NCEP predictor distributions were checked to ensure they are similar

(Section 4.1), they are not identical. Seasons that have large inflation factors could magnify relatively small predictor distribution differences, when the CGCM3 predictors are used in the NCEP trained regression. Predictor distribution differences could lead to undesirable results in your projections as previously discussed. This is why the comparison between NCEP and the CGCM3 distributions is essential.

Finally, adding back on the observed average seasonal cycles removed previously to the inflated predictions, the following four PDFs (Figures 4.4,4.5,4.6,4.7) are plotted for inspection. Shown are PDFs of the NCEP predicted (downscaled) total Tmax and Tmin against what were observed in the $1961 - 2000$ period. These predictions (NCEP predictions) are created from predicting the daily anomaly via the regression then inflating the variance and finally adding back on the observed average seasonal cycle. Note that NCEP predicted refers to the regression predicted Tmax or Tmin using the NCEP predictors.

The NCEP predicted total Tmax and Tmin using the newly developed statistical downscaling method have similar PDFs the observed Tmax and Tmin PDFs, especially in terms of the general shape. This is true for all seasons for both Tmax and Tmin. The main caveat is that spring and summer require large inflation factors to do this. Although the shape comparisons appear similar, close inspection reveals differences around the center peak of the PDFs for Tmin (Figure 4.4). This also occurs in the other seasons for both predictands. The peakedness of a PDF is related to the fourth moment known as kurtosis. A way to address this is to have a more objective predictor selection process that considers the higher moments of the distributions. The distributions could be more objectively compared through hypothesis testing of the first four moments of the data.

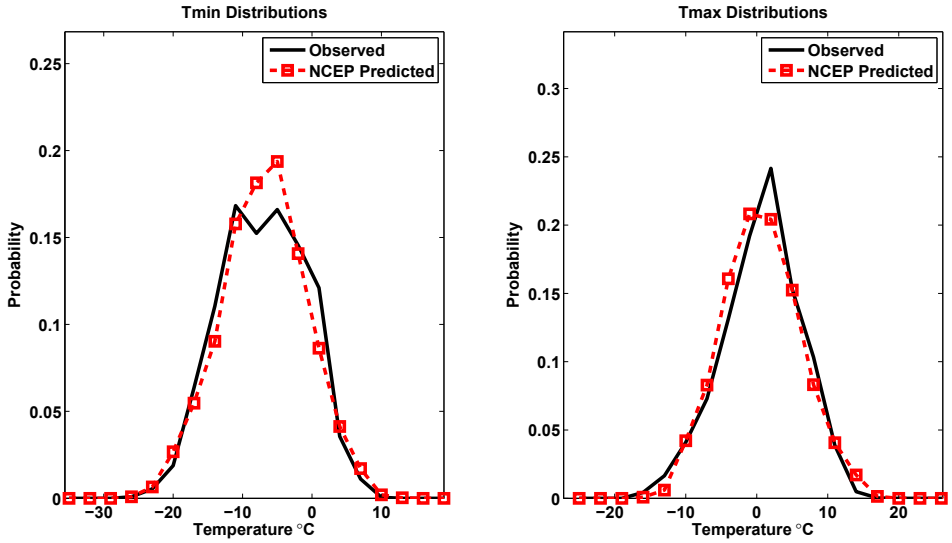

Figure 4.4: Probability density functions of observed and NCEP predicted in winter at Shearwater, NS. Tmin is in left panel, Tmax is in right panel.

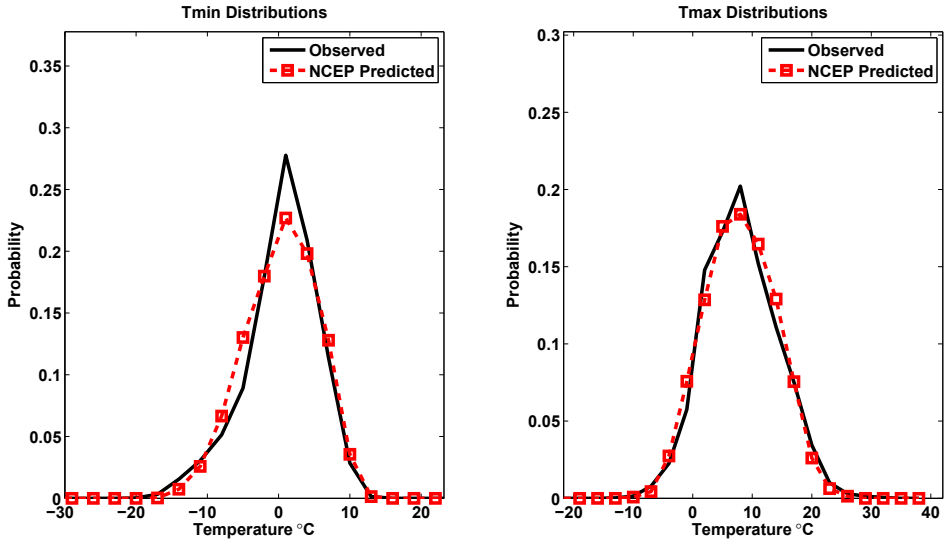

Figure 4.5: Probability density functions of observed and NCEP predicted in spring at Shearwater, NS. Tmin is in left panel, Tmax is in right panel.

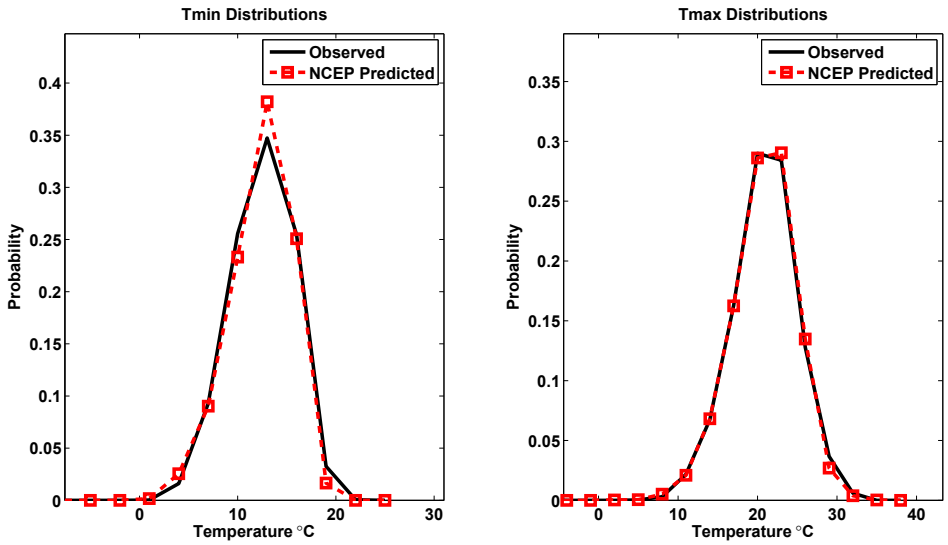

Figure 4.6: Probability density functions of observed and NCEP predicted in summer at Shearwater, NS. Tmin is in left panel, Tmax is in right panel.

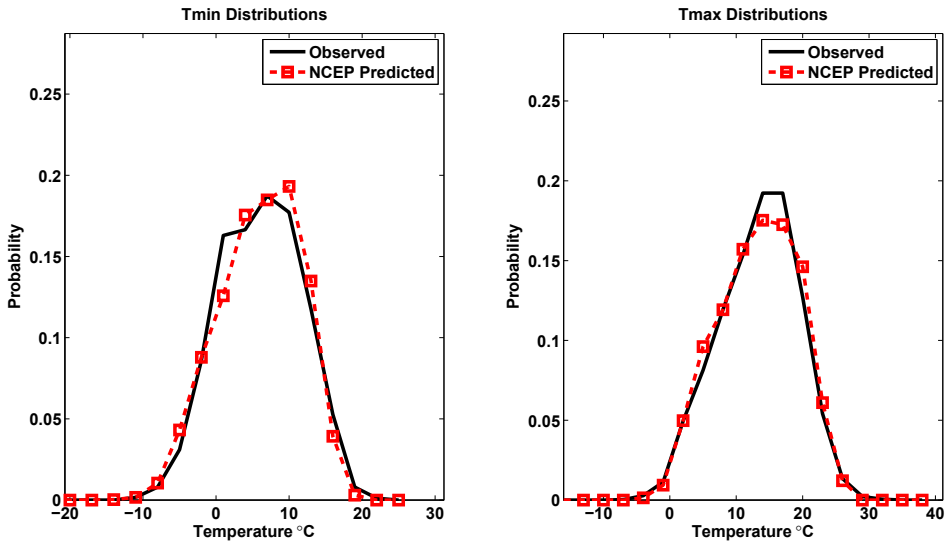

Figure 4.7: Probability density functions of observed and NCEP predicted in fall at Shearwater, NS. Tmin is in left panel, Tmax is in right panel.

4.5 CGCM3 Hindcasting

Once the regression coefficients are determined from the observations (ie. NCEP re-analysis in this thesis), the CGCM3 predictors can be used in the regression analysis. Since a careful predictor selection process was used to ensure the CGCM3 predictors have similar distributions to the NCEP distributions, the CGCM3 predictors should be capable of hindcasting Tmax and Tmin with similar skill to NCEP. Since it is assumed that NCEP eigenvectors are the true directions of variance, the first step is to project the CGCM3 predictors onto the NCEP derived eigenvectors in each season. This gives the CGCM3 version of the principal components to be used in the regression analysis. Next, the NCEP seasonally derived regression coefficients were used with the CGCM3 PC's to predict Tmax and Tmin in each season. The predicted Tmax and Tmin are then multiplied by the appropriate inflation factor derived in the previous section (Section 4.3) from the NCEP data. Finally the observed average seasonal cycle which includes the annual mean removed previously from the observations (see Chapter 3) is added to the prediction. The result is the total Tmin and Tmax values predicted from the CGCM3 predictors.

As previously mentioned, it is not meaningful to compare time series of Tmax and Tmin predicted using predictors from the CGCM3 with observed Tmax an Tmin because the CGCM3 is not data assimilated. As a result, we compare the PDFs of the CGCM3 predicted and observed Tmax and Tmin in each season. Figures 4.8, 4.9, 4.10, and 4.11 are plotted by season for $1961 - 2000$. The plots of Tmax and Tmin show the CGCM3 predicted (downscaled) Tmax and Tmin distributions (PDFs) using the CGCM3 PC's compared with the actual observed distribution. Also shown is the distribution predicted by the raw CGCM3 without any downscaling.

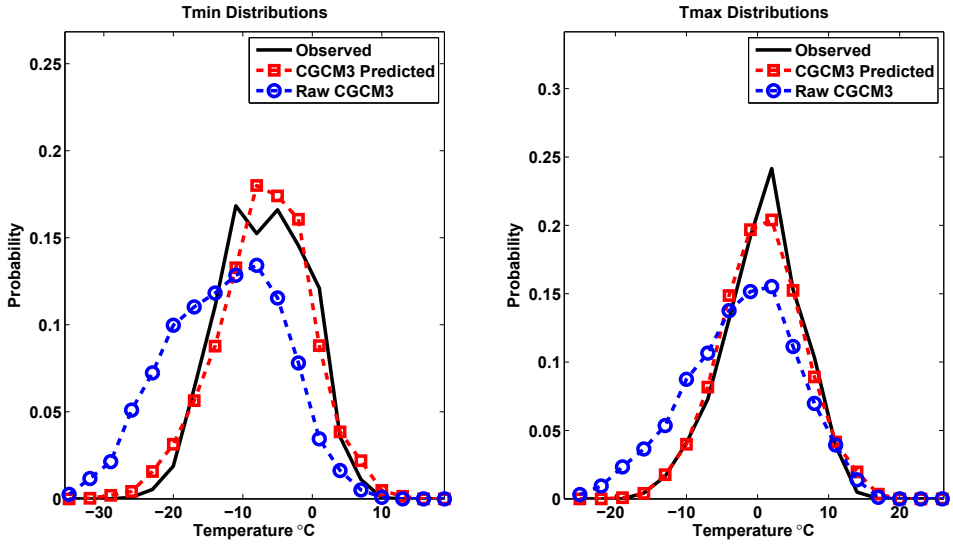

Figure 4.8: Winter Probability Density Function (PDF) of the observed distribution from $1961 - 2000$ (black), the PDF from the CGCM3 predicted (downscaling prediction) for the same period (red) and the PDF from the raw CGCM3 for the same period (blue). The left panel is Tmin and the right panel is Tmax, both for Shearwater, NS.

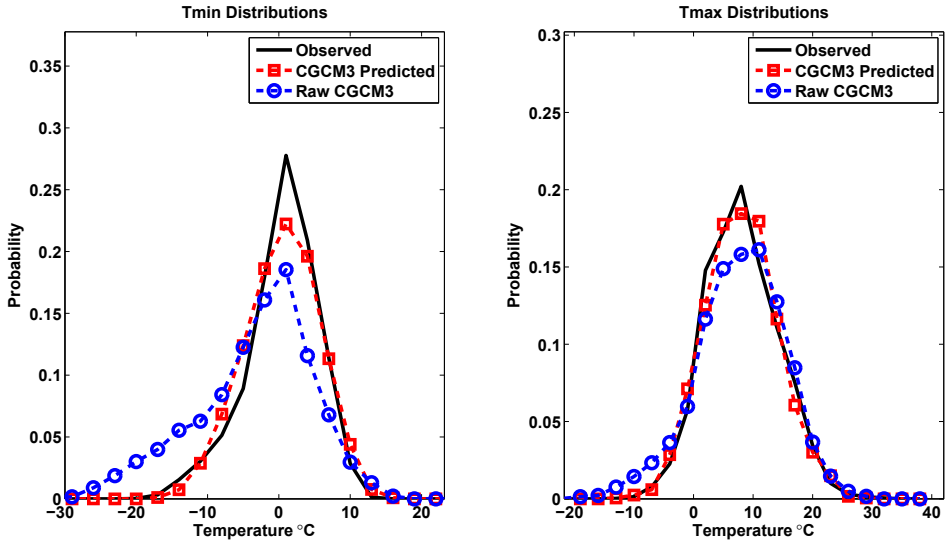

Figure 4.9: Spring Probability Density Function (PDF) of the observed distribution from $1961 - 2000$ (black), the PDF from the CGCM3 predicted (downscaling prediction) for the same period (red) and the PDF from the raw CGCM3 for the same period (blue). The left panel is Tmin and the right panel is Tmax, both for Shearwater, NS.

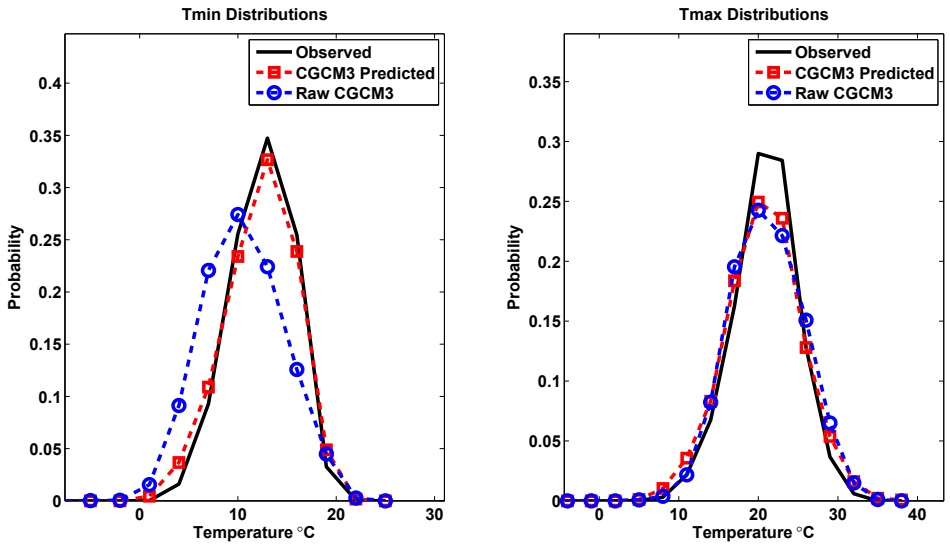

Figure 4.10: Summer Probability Density Function (PDF) of the observed distribution from $1961 - 2000$ (black), the PDF from the CGCM3 predicted (downscaling prediction) for the same period (red) and the PDF from the raw CGCM3 for the same period (blue). The left panel is Tmin and the right panel is Tmax, both for Shearwater, NS.

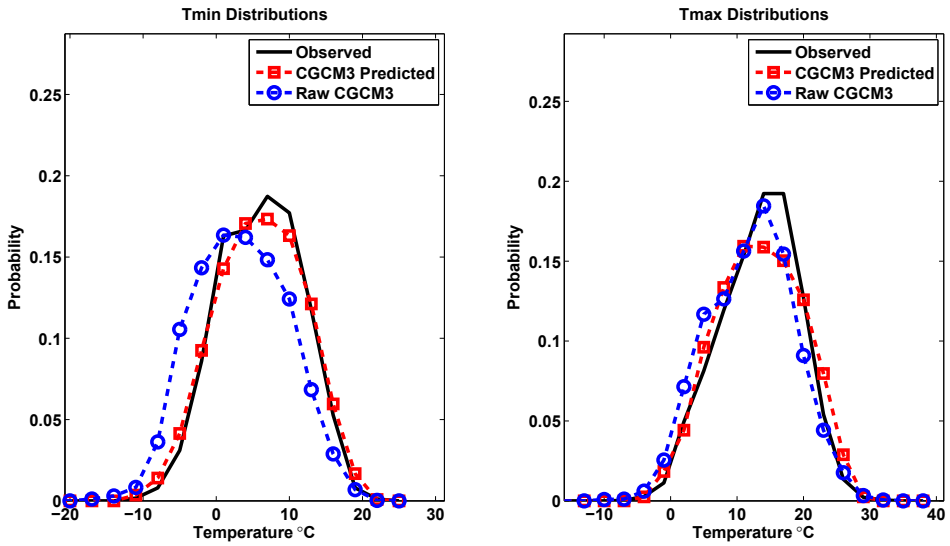

Figure 4.11: Fall Probability Density Function (PDF) of the observed distribution from $1961 - 2000$ (black), the PDF from the CGCM3 predicted (downscaling prediction) for the same period (red) and the PDF from the raw CGCM3 for the same period (blue). The left panel is Tmin and the right panel is Tmax, both for Shearwater, NS.

A subjective comparison (Figures 4.8, 4.9, 4.10, 4.11) of the CGCM3 predicted (downscaled), observations and the raw CGCM3 prediction PDFs reveals the following: In general, the statistical downscaling (CGCM3 predicted) is closer to observed than the raw CGCM3. The largest improvement occurs for Tmin in winter, where the downscaled version of Tmin has a very different shape than the raw CGCM3 Tmin, and is much closer to the observed Tmin. The downscaled Tmax PDF in winter shows marked improvement as well compared to the raw CGCM3. The spring Tmin downscaled version is also much closer to observed than the raw CGCM3. Tmax in spring shows improvement as well from the downscaling; however the raw CGCM3 appears to do a better job for Tmax than Tmin when comparing with observed Tmax. Tmin in summer shows improvement from the downscaling as well. It is interesting to note that the summer Tmax downscaling shows little improvement. In other words, the raw CGCM3 seems to be doing a good job with summer Tmax compared to observations. Again, caution needs to be taken with the accuracy of the downscaling in spring and summer because the inflation factor is large (regression has low explained variance). Finally, the downscaling in fall shows marked improvement for Tmin, but CGCM3 predicted Tmax only shows marginal improvement

since the raw CGCM3 appears to do well when compared to observed.

These PDFs may not be a fair comparison for the downscaling accuracy, since the average observed seasonal cycle is included in the observed data, and the same average observed seasonal cycle has been added onto the CGCM3 predicted inflated daily anomaly (downscaling prediction). Similarly, the raw CGCM3 temperature prediction includes the seasonal cycle predicted by the raw CGCM3 which is different from the observed average seasonal cycle. For a fair comparison, the variance of the daily anomalies (Tmax and Tmin) was compared with hypothesis testing (F-test, see Chapter 2) in each season. Since the daily anomaly is what the regression predicts, the observed anomaly and the inflated CGCM3 predicted anomaly should be compared with the raw CGCM3 anomaly. To get the raw CGCM3 anomaly, a seasonal cycle was fit to the raw CGCM3 (same as described in Chapter 3) and then removed.

After doing the hypothesis testing (F-test, see Section 2.2), which compares the variances of the two predicted anomalies (CGCM3 predicted, and raw CGCM3) with the observed anomalies, the same result as the initial subjective comparison was found. That is: compared to observed, the statistical downscaling CGCM3 prediction anomaly has a variance statistically the same as observed. However, the raw CGCM3 anomaly does not. This is true with ninety five percent confidence (19 times out of 20). This result certainly suggests that the downscaling does lead to an improvement in the representation of Tmax and Tmin compared to observed. From here we are now ready to use the regression with future predictors from the CGCM3 to make future projections.

In closing this section, we argue that variance is not the only consideration, when trying to predict a distribution's shape. The hypothesis testing employed above considers only the variance. The inflation factor (see Section 4.4) forces the CGCM3 predicted variance to match the observed variance. The only way to eliminate the inflation factor is to have predictors that explain one hundred percent of the variance in the predictand. The higher moments of the data (skewness and kurtosis) are not addressed by the inflation factor. In fact, they are not considered explicitly in this thesis. It may be useful in the future to explicitly consider them to create an objective method of comparing the GCM/NCEP predictor distributions in the predictor selection process to replace the subjective approach taken here (see Section 4.1 . This could improve the shape of the predicted distribution, which is the main goal of statistical downscaling, as discussed in the motivation (See

Section 1.3).

4.6 Future Projections

Once it was clear the predictors from the CGCM3 were capable of hindcasting the historical Tmax and Tmin, future projections can be generated using future $(2001-2100)$ CGCM3 predictors. As discussed in Chapter 3, all the predictors were normalized with respect to the $1961-1990$ means and standard deviations. This consistent standardization period preserves the trend in the predictors, since all predictors have the same mean removed from them. However, part of the downscaling process required the removal of the seasonal cycle. The removal of the fitted seasonal cycle was done carefully in the future, as to not remove the future means.

Similar to Chapter 3, a seasonal cycle was fitted to each future predictor $(2001-2100)$. This was done via a regression of sines and cosines over the 100 years of future data $(2001-2100)$ based on Eq. (3.1). The seasonal cycle fit also included the mean (μ) of the future 100 year period as in Eq. (3.1). Note that the mean of the future 100 years is not zero because the data are normalized with respect to the $1961-1990$ period (See Chapter 3). After the determined future seasonal cycle, including μ, was found and removed, μ was added back on to the resulting predictor anomalies. Note, removal of the seasonal cycle and the mean in the historical period did not have this problem because it was normalized with respect to that period (giving a zero mean). However, future predictors have a non zero mean with respect to the $1961-1990$ reference period. Thus to preserve the trend in the predictors, the future predictor means have to be added back on after the seasonal cycle was found and removed. This allows continuity of the trend in the predictors.

It is expected that the future predictors will have a different mean in the future than the historical mean. This change in the mean is part of the climate response to the forcing imposed in the model. The future period $(2001-2100)$ was split into three periods (tridecades), which are $2011-2040$, $2041-2070$ and $2071-2100$. The CGCM3 predictors from each of these periods were used in the regression for Tmax and Tmin in each future period. In each season, the CGCM3 predictors were projected onto the NCEP derived eigenvectors to make the future PC's and then used in the regression for Tmax or Tmin. The average observed seasonal cycles removed in the $1961-2000$ period were added

back on to the future inflated predicted anomalies to give the total Tmax or Tmin in the future. The mean and standard deviation of the prediction (including the addition of the observed seasonal cycle) for each season and period is shown in Tables 4.16 and 4.17. It should be noted that the change in the mean from the $1961 - 2000$ reference period is calculated from the CGCM3 predictors using the regression model. Also shown are the historical ($1961 - 2000$) mean and standard deviation (including the seasonal cycle) for each season. The change in mean can be found through the difference between the future period mean and the historical mean. The mean and standard deviation in Tables 4.16 and 4.17 are defined as usual; the mean is the arithmetic mean of the temperatures in that period and the standard deviation is the square root of the variance of the temperature in that period.

Period	Season	μ	σ	Historical μ	Historical σ
$2011-2040$	Winter	-5.7	6.40		
$2041-2070$	Winter	-4.2	6.31	-6.96	6.11
$2071-2100$	Winter	-2.0	5.74		
$2011-2040$	Spring	1.6	5.40		
$2041-2070$	Spring	2.7	5.29	0.44	5.06
$2071-2100$	Spring	3.9	5.38		
$2011-2040$	Summer	13.6	3.59		
$2041-2070$	Summer	15.2	3.67	12.54	3.08
$2071-2100$	Summer	16.9	3.69		
$2011-2040$	Fall	7.4	5.89		
$2041-2070$	Fall	8.1	5.79	6.10	5.58
$2071-2100$	Fall	9.7	6.15		

Table 4.16: Results for each period in each season for Tmin. The mean of the predicted future distribution and its standard deviation are shown. Also shown are the historical observed distribution mean and standard deviation. The units in this table are in degrees Celsius.

Period	Season	μ	σ	Historical μ	Historical σ
$2011-2040$	Winter	1.6	5.67		
$2041-2070$	Winter	2.7	5.70	0.67	5.47
$2071-2100$	Winter	4.3	5.35		
$2011-2040$	Spring	9.3	6.32		
$2041-2070$	Spring	10.8	6.08	8.16	6.07
$2071-2100$	Spring	12.6	6.39		
$2011-2040$	Summer	22.3	4.42		
$2041-2070$	Summer	24.1	4.40	20.97	3.99
$2071-2100$	Summer	25.9	4.40		
$2011-2040$	Fall	14.6	6.38		
$2041-2070$	Fall	15.4	6.31	13.34	5.91
$2071-2100$	Fall	17.1	6.58		

Table 4.17: For each period in each season for Tmax, the mean of the predicted future distribution and its standard deviation are shown. Also shown are the historical observed distribution mean and standard deviation. The units in this table are in degrees Celsius.

4.7 Alternative Future Projections

In this method, the regression coefficients determined in the past (Section 4.2) were again used with the future CGCM3 predictors. However, the future and historical predictand/predictors were detrended before the regression was developed. In the historical period the results of the regression development are all identical to the results from the method described in the main method of this thesis (all the work previous to this). This is because the trends in the 1961-2000 period are small, so removing them makes little difference. However, there are significant differences when moving into projection mode. The detrended future predictors have lost their information on how their mean changes with time since that is defined by the removed trend. In this case, the regression captures the shape of the distribution only. The shift in the mean is taken directly from the trend in the grid box from the raw CGCM3. The mean of each future year is found by taking the observed mean of the historical distribution ($1961-2000$) and following the trend taken from the raw CGCM3 to the year in question. The base year is 1980 which is the middle point of the historical distribution. In mathematical terms, the future mean is defined as:

$$\mu_n = \mu_H + \delta_*(n - 1980) * 365 \tag{4.7}$$

where μ_n is the mean of the year n in question, μ_H is the observed mean of the

historical($1961 - 2000$) distribution for maximum or minimum daily temperature. δ is the slope of the trend line (from the raw CGCM3, $1961 - 2100$), which is 0.00013 degrees Celsius per day for Tmax and 0.00016 degrees Celsius per day for Tmin. This is multiplied by the number of days since 1980 (the middle of the historical distribution). Note the shift in mean is calculated from the trend in daily temperature (Tmax,Tmin) from the raw CGCM3 to ensure it shifts in a smooth manner. The numerical means and standard deviations for this alternative methodology can be found in Tables 4.18 and 4.19. Note that the means for the future periods in Tables 4.18 and 4.19 are actually the means for the center years of the periods. For example, the mean for the $2011 - 2040$ period is actually the mean of the year 2025. The means are shown as the mean for a tri-decade for easy comparison to the results of the main method described in this thesis, shown in Tables 4.16 and 4.17. The mean and standard deviation in the following two tables (4.18, 4.19) are defined as follows; the mean is the mean of the center year of the tri-decades found from Eq. (4.7) in degrees Celsius. The standard deviation is the square root of the variance of the temperature in that period (including the seasonal cycle).

Period	Season	μ	σ	Historical μ	Historical σ
$2011-2040$	Winter	-4.3	6.46		
$2041-2070$	Winter	-2.6	6.42	-6.96	6.11
$2071-2100$	Winter	-0.8	5.92		
$2011-2040$	Spring	3.1	5.42		
$2041-2070$	Spring	4.8	5.31	0.44	5.06
$2071-2100$	Spring	6.6	5.40		
$2011-2040$	Summer	15.2	3.53		
$2041-2070$	Summer	16.9	3.55	12.54	3.08
$2071-2100$	Summer	18.7	3.59		
$2011-2040$	Fall	8.7	5.94		
$2041-2070$	Fall	10.5	5.83	6.10	5.58
$2071-2100$	Fall	12.3	6.19		

Table 4.18: For each period in each season for Tmin. The mean of the predicted future distribution and its standard deviation are shown. Also shown are the historical observed distribution mean and standard deviation. The units in this table are in degrees Celsius.

Period	Season	μ	σ	Historical μ	Historical σ
$2011-2040$	Winter	2.8	5.65		
$2041-2070$	Winter	4.2	5.67	0.67	5.47
$2071-2100$	Winter	5.7	5.34		
$2011-2040$	Spring	10.3	6.34		
$2041-2070$	Spring	11.7	6.06	8.16	6.07
$2071-2100$	Spring	13.2	6.38		
$2011-2040$	Summer	23.1	4.44		
$2041-2070$	Summer	24.5	4.35	20.97	3.99
$2071-2100$	Summer	25.9	4.37		
$2011-2040$	Fall	15.5	6.36		
$2041-2070$	Fall	16.9	6.28	13.34	5.91
$2071-2100$	Fall	18.3	6.55		

Table 4.19: For each period in each season for Tmax, the mean of the predicted future distribution and its standard deviation are shown. Also shown are the historical observed distribution mean and standard deviation. The units in this table are in degrees Celsius.

4.8 Future Projections Discussion

Two different ways to get future projections of Tmax and Tmin were presented in the previous two sections. It is worthwhile to briefly discuss the advantages and shortcomings of each. A GCM solves the fully non-linear equations in response to a prescribed forcing. This contains much more physics than the select predictors used in a statistical model. It is the subject of debate whether a linear statistical model (such as used in this thesis), is capable of representing the long term climate change signal. It is postulated that only a GCM is capable of doing this. As shown in Section 4.7, using the GCM to shift the mean instead of the statistical model (Section 4.6), yields different results. The GCM predicts close to two degrees Celsius more warming for Tmin and about 0.8 degrees Celsius more warming for Tmax compared to the main thesis method (regression).

The GCM includes the non-linear physics, which is a plus. However, the large grid spacing is problematic. The physical processes occurring at each location are probably misrepresented in the model because the large grid excluded local effects. The statistical model on the other hand does not contain all the physics, but hopefully has included some local effects because it was trained with local observations. It should also be noted that the predictors used in the statistical model evolve from the fully non-linear GCM. Using predictors like geopotential height (from the GCM) in the regression gives confidence

that the statistical model should be capable of capturing the climate signal (*Huth*, 1999). Even though the seasonal cycle was removed from the future predictors, the trend was not removed in the main method (discussed in Section 4.6). Recall the removed future seasonal cycle ($2001 - 2100$) included the mean removal. The means were then added back on to ensure the trend in the predictors was continuous.

It is the author's opinion that the CGCM3 overestimates the response at Shearwater, NS because the model does not consider local effects such as land/sea temperature contrasts and topography. It can not be determined with certainty whether the statistical model has all the necessary physics to represent the change properly. However, there are also many published studies using the statistical technique of using the regression to shift the mean (*Wilby et al.* (2002), *Huth et al.* (2002), *Cheng et al.* (2008), and *von Storch et al.* (1993)).

In support of the author's opinion is the fact that the main statistical downscaling thesis method produces similar results to the Canadian Regional Climate Model version 4 (CRCM4) dynamical downscaling results (*He et al.*,2010) for Shearwater. This suggests the regression contains the most important predictors for climate change signal. CRCM4 is based on CGCM3.7.1 (*Plummer et al.*, 2006), in which the dynamical downscaling is given its boundary conditions from the CGCM3. Note, the CGCM3 also produced the predictors used in the statistical downscaling in this thesis, which makes the comparison with the CRCM4 results more direct.

Table 4.20 compares future projections of annual means, for Tmax and Tmin from the CRCM4, with the two future projection methods described in the thesis (Section 4.6, 4.7). Recall that the main method (Section 4.6) shifts the mean by applying the developed regression to future predictors, then adding back on the historical observed seasonal cycle. The alternate future projections (Section 4.7) are found via the trend from the CGCM3 temperature in the grid box containing Shearwater, which is used to shift the mean. The change in Table 4.20, is the difference in mean between the historical period $1961 - 2000$ and the $2071 - 2100$ period. The annual means (from Section 4.6) from the $2071 - 2100$ distribution are calculated from Tables 4.16 through 4.18, by taking the average of the four seasonal means of $2071 - 2100$.

Table 4.20 demonstrates that the downscaling method described in this thesis is better at capturing the change in mean for both Tmax and Tmin compared to the alternate method. Of course this statement assumes that the CRCM4 projected change is representative of

Seasonal	Dynamical (CRCM4)	Thesis method	Alternate method
$\delta Tmin$	5	4.1	6.2
$\delta Tmax$	4.3	4.2	5

Table 4.20: The annual change (δ) in projected means (Tmax and Tmin) for the $2071-2100$ distribution compared to the historical $1961-2000$ means. This comparison includes dynamical downscaling from the CRCM4, statistical downscaling via the main thesis method, and finally the mean change projected by the alternative method is shown. Units are in degrees Celsius

the truth which is not known. This is not proof that the main thesis method used to shift the mean is better than the alternate method (GCM trend shifts the mean), but it certainly adds weight to the previous argument that it is.

In theory, dynamical downscaling is currently the best tool for climate projections because it can produce higher resolution climate projections with the benefit of solving the fully non-linear governing equations for the forcing. However, in the past regional climate model results were not available due to their complexity to produce. The main problem right now is the relatively small amount of dynamical downscaling completed compared to GCM's available to run dynamical downscaling. It is well known that different GCM's using the same forcing produce different future results. It takes a large amount of computer power to dynamically downscale one GCM for one forcing. Given this problem, statistical downscaling can be a computationally quick tool to provide local projections for a location using various GCMs and forcings. In any climate study, it is best to consider a range of GCMs and forcing scenarios to get a range for climate change at a given location. Using one GCM with one emission scenario can be very misleading.

CHAPTER 5

CONCLUSIONS

5.1 Summary of the Downscaling Process

The methodology in this thesis can be described in two main parts (Figures 5.1 and 5.2). The first flow chart below is concerned with creating independent predictors to be used in the regression. The second flow chart in this section, is concerned with the development of the regression, and using it to make projections in the future.

The process begins with writing the predictors as Z-scores (*Predictor Z-scores*, see Section 3.5). Two sets of predictors were used, the first is the NCEP predictors, and the second is the CGCM3 predictors. The first step is to take the NCEP/CGCM3 predictors and remove their seasonal cycles to get the daily anomalies (*Seasonal anomalies*, see Section 3.2 and 3.5). Next, a predictor selection process was applied to the NCEP predictors, to get the best set of predictors (*Predictor subset*, see Section 4.1). The final step is to transform the final predictor set into its principal components to be used in the regression (*PC analysis*, see Section 2.3 and 4.2).

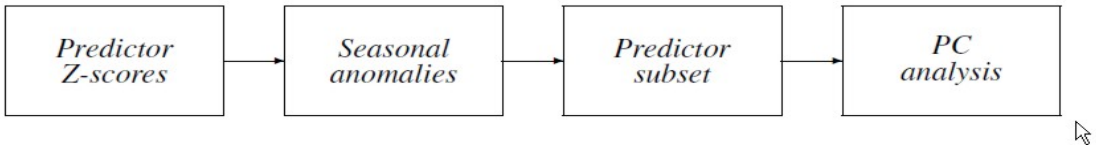

Figure 5.1: Flow chart overview of the first four steps in the statistical downscaling method described in the thesis.

The second chart is concerned with the regression development, and future projections. Once the principal components are developed (from NCEP) the regression can be

trained. The regression training is done between the NCEP principal components and both observed Tmax and Tmin in the historical period (*Regress Tmax/Tmin with PC's*, see Section 3.1, 3.2 and 4.2). The regression is validated via cross validation (*Validate regression*, see Section 4.4). Once it was determined that the regression had predictive skill, the CGCM3 predictors can be used via projecting them onto the NCEP derived eigenvectors to give the respective CGCM3 PC's (*Use CGCM3 predictors*, see Section 4.5). Finally, the CGCM3 PC's were used to make future projections (*Future projections, see Section 4.6 and 4.7*).

It should be noted that the predictands (Tmax, Tmin) were never transformed into Z-scores. The predictands did have their average seasonal cycles removed to generate the daily anomalies. The regression (using NCEP or CGCM3 predictors) predicts the daily anomaly. The variance of the predicted daily anomaly is then inflated via the inflation factor. Finally, the average seasonal cycle removed from observations is added back on to the inflated daily anomaly to give the total predicted Tmax or Tmin.

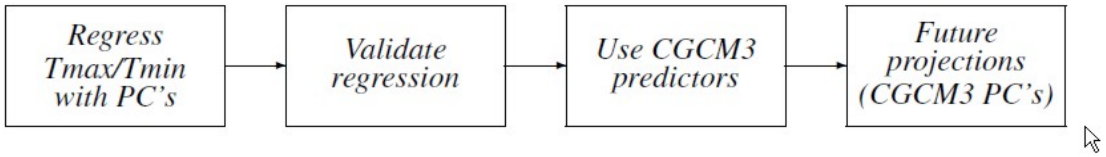

Figure 5.2: Flow chart overview of the last four steps in the statistical downscaling method described in the thesis.

5.2 Conclusions

An important question to be addressed is how credible are the future projections? Of course this is a difficult question to answer with certainty. The future climate is unknown and depends on many interrelated complicated variables. Since there are no observations to compare the future projections with, we focus on the downscaling method to comment on the results.

This thesis took as its starting point, the standard Statistical Downscaling Method (SDSM) used previously (Wilby et al, 2002), and its known problems. The SDSM takes the actual predictand data (Tmax or Tmin) and predictor data (NCEP) to create a regression in the historical period. The standard SDSM does have the capability to create a

regression by season as done in this thesis, but it is not capable of using different predictors in each season. Once the predictors are chosen in the predictor selection process, the same predictors are used in all seasons. Furthermore, the predictors used in SDSM still contain the seasonal cycle which strongly influences the regression development. Standard SDSM also has no capability to compare the NCEP predictors to the GCM predictors, a missing essential step. For temperature, SDSM often has an explained variance of above ninety percent. This is misleading because the method is predicting the seasonal cycle as well as the day to day variability and it is the seasonal cycle that accounts for a large part of the variance in the data.

To address these problems the new method developed in the thesis (hereafter the thesis method) and was investigated. The main steps in the thesis method include: removal of the seasonal cycle, comparing the NCEP and the CGCM3 predictors in a careful predictor selection process, and principal component regression by season using different predictors in each season.

To gauge the improvement of the thesis method compared to SDSM, the predictors (NCEP), and the predictand (Tmax) used in this study were used in the standard SDSM to develop the regression in the historical period. SDSM was given the same suite of predictors from NCEP used in the main thesis method, to produce the historical NCEP regression. SDSM determined the set of predictors to be used, then produced a regression in all four seasons, using the same predictors in each season. It should be noted that these predictors included their seasonal cycles. SDSM picked different predictors in the regression than the main thesis method. SDSM typically chooses predictors with a strong seasonal cycle, since the predictand (Tmax) has a strong seasonal cycle.

Once the SDSM historical prediction was completed, a three harmonic seasonal cycle (same as discussed in Chapter 3) was fitted to the prediction and then removed, leaving the SDSM daily anomaly. The thesis method predicts the daily anomaly (thesis method anomaly), directly as discussed in the thesis. Finally, in each season the correlations between the observed Tmax daily anomaly from $1961-2000$ and the SDSM Tmax daily anomaly for the same period is shown in Table 5.1. Table 5.1 also shows the correlation between the $1961-2000$ observed Tmax daily anomaly and the Tmax daily anomaly predicted via the thesis method for each season.

It is noteworthy to mention that it makes no difference in creating the SDSM anomaly

Season	Thesis method vs Observed	SDSM versus Observed
$Winter$	0.86	0.57
$Spring$	0.62	0.45
$Summer$	0.68	0.32
$Fall$	0.81	0.58

Table 5.1: Correlations between the $1961 - 2000$ SDSM predcition Tmax, the thesis method prediction Tmax and observed Tmax daily anomalies.

whether the observed seasonal cycle was removed or the seasonal cycle fitted to the SDSM prediction was removed. This is because SDSM has skill in predicting the seasonal cycle. However it is clear from Table 5.1 that SDSM is poor at predicting the daily anomaly compared to observations. By comparison, the thesis method produces a more credible result for predicting the observed Tmax anomaly variance, as compared to SDSM. The improvement in correlation depends on season and ranges from roughly 0.2 to 0.3. One thing to keep in mind, is in spring and summer, the thesis method does produce a better regression than SDSM. However the explained variance is still low and requires a significant inflation factor to be able to account for the observed variance. Therefore, spring and summer are not as trustworthy in terms of projections compared to fall and winter.

It is important to note that if one wants to have confidence in the future projections, it is essential to have the best regression model possible. The method described in the thesis does produce a trustworthy regression and is done so in a scientifically defensible way. The thesis method addresses known problems with SDSM which is a step in the right direction. It cannot be said with certainty that the thesis method will produce an improvement for other locations. However, it is concluded that it produces a more credible result in this case. The most important improvement is the method itself. In the end, if you want to have confidence in the projections, it is essential to have a defendable method.

5.3 Future Work

The main issue for future work is to determine if the thesis method is valid for other locations. It is possible that Halifax has a certain type of climate that allows the thesis method to do a good job. Locations with different climates may not be such good candidates for this technique. Future work should include downscaling other sites in Atlantic Canada.

A possible first candidate would be Fredericton, NB. Fredericton has a more continental type climate and experiences more extremes, since it is further away from the marine influence.

Another major problem is the limited number of predictors available from the GCM. This limit on predictor variables comes from the inability of the GCM to represent predictors in a realistic manner. In seasons when synoptic variability is less important, such as spring and summer, predictors like ocean temperature would be useful. This is actually a climate modeling problem. There is not much the statistical downscaling process can do to fix this problem. As GCM's improve, so should statistical downscaling. This is probably the main point on which statistical downscaling can be improved. More realistic GCM predictors and a predictor selection process that considers higher moments (skewness, kurtosis) of the predictor distributions are essential. This will allow the regression to capture the realistic shape of the observed predictand distribution.

It would be also worthwhile to make use of other reanalysis data. Some research shows downscaling is very sensitive to the choice of reanalysis data. *Koukidis and Berg* (2009) utilizes ERA-40 and NCEP reanalysis data independently to downscale temperature at a southern Ontario site. The result was that there are statistically different results in downscaled temperature. It would be very instructive to use ERA-40 data instead of NCEP and repeat the work in this thesis.

Although GCM's and dynamical downscaling are likely to continue to evolve, statistical downscaling is likely to continue to be a useful technique for a researcher who requires high resolution climate projections. This thesis addresses some of the known problems with statistical downscaling and focuses on the development of the regression to achieve trustworthy projections. Projections of temperature for Shearwater could be used to plan for health care in Halifax. The impacts of higher temperatures on human health was highlighted in recent years in the European heatwave. Even building codes require credible temperature projections, to ensure structures do not fail and can withstand thermal expansion. Climate change projections are essential to plan for the future. Credible projections give governments confidence to invest funds in adaptation strategies to help mitigate the negative effects of climate change.

BIBLIOGRAPHY

Cheng, C., G. Li, Q. Li, and H. Auld, Statistical downscaling of hourly and daily climate scenarios for various meteorological variables in South-central Canada, *Theoretical and Applied Climatology*, *91*, 129–147, 2008.

Dibike, Y., P. Gachon, A. St-Hilaire, T. Ouarda, and V. Nguyen, Uncertainty analysis of statistically downscaled temperature and precipitation regimes in Northern Canada, *Theoretical and Applied Climatology*, *91*, 149–170, 2008.

Flato, G., and G. Boer, Warming asymmetry in climate change simulations., *Geophysical Research Letters*, *28*, 195–198, 2001.

Gachon, P., and Y. Dibike, Temperature change signals in northern Canada: convergence of statistical downscaling results using two driving GCMs, *International Journal of Climatology*, *27*, 1623–1641, 2007.

Gachon, P., A. Harding, and M. Radojevic, Predictor datasets derived from the CGCM3.1 T47 and NCEP/NCAR reanalysis, *Tech. rep.*, 2008.

He, Y., A. Monahan, C. Jones, A. Dai, S. Biner, D. Caya, and K. Winger, Probability distributions of land surface wind speeds over North America, *J. Geophys. Res*, *115*, 2010.

Holton, J., *An Introduction to Dynamic Meteorology*, Academic Press, 2004.

Houghton, J., Y. Ding, D. Griggs, M. Noguer, P. van der Linden, X. Dai, and K. Maskell, *Climate Change 2001: The Scientific Basis*, Elsevier, 2001.

Huth, R., Statistical downscaling in central Europe: Evaluation of methods and potential predictors, *Climate Research*, *13*, 91–101, 1999.

Huth, R., Statistical downscaling of daily temperature in central Europe, *Journal of Climate*, *15*, 1731–1742, 2002.

Johnson, R., and D. Wichern, *Applied Multivariate Statistical Analysis*, Prentice Hall Upper Saddle River, NJ, 2002.

Karl, T., W. Wang, M. Schlesinger, R. Knight, and D. Portman, A method of relating general circulation model simulated climate to the observed local climate. Part I: Seasonal statistics, *Journal of Climate*, *3*, 1053–1079, 1990.

Kistler, R., and E. Kalnay, The NCEP/NCAR 50-year reanalysis, *Bulletin of the American Meteorological Society*, *82*, 247–268, 2001.

Koukidis, E., and A. Berg, Sensitivity of the statistical downscaling model (SDSM) to reanalysis products, *Atmosphere-Ocean*, *47*, 1–18, 2009.

Laprise, R., D. Caya, A. Frigon, and D. Paquin, Current and perturbed climate as simulated by the second-generation Canadian Regional Climate Model (CRCM-II) over northwestern North America, *Climate Dynamics*, *21*, 405–421, 2003.

Levine, D., P. Ramsey, R. Smidt, P. Ramsey, and R. Smidt, *Applied Statistics for Engineers and Scientists: using Microsoft Excel and Minitab*, Prentice Hall, 2001.

Lines, G., M. Pancura, and C. Lander, Building climate change scenarios of temperature and precipitation in Atlantic Canada using the Statistical Downscaling Model (SDSM), *Tech. rep.*, 2005.

McFarlane, N., J. Scinocca, M. Lazare, R. Harvey, D. Verseghy, and J. Li, The CCCma third generation atmospheric general circulation model (AGCM3), *Tech. rep.*, Internal Report, Canadian Centre for Climate Modelling and Analysis, 2006.

Nakicenovic, N., and R. Swart, Emissions scenarios 2000–special report of the intergovernmental panel on climate change, *Cambridge University Press, Cambridge, UK*, *96*, 98, 2000.

Pancura, M., and G. Lines, Variability and extremes in statistically downscaled climate change projections at Greenwood Nova Scotia, *Tech. rep.*, 2005.

Plummer, D., D. Caya, A. Frigon, H. Côté, M. Giguère, D. Paquin, S. Biner, R. Harvey, and R. De Elia, Climate and climate change over North America as simulated by the Canadian RCM, *Journal of Climate*, *19*, 3112–3132, 2006.

Schoof, J., and S. Pryor, Downscaling temperature and precipitation: A comparison of regression-based methods and artificial neural networks, *International Journal of Climatology*, *21*, 773–790, 2001.

Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K. Averyt, M. Tignor, and H. Miller, The physical science basis, contribution of working group 1 to the fourth assessment report of the intergovernmental panel on climate change, 2007.

Swansburg, E., N. El-Jabi, and D. Caissie, Climate change in New Brunswick (Canada): statistical downscaling of local temperature, precipitation, and river discharge, *Tech. rep.*, 2005.

Thompson, K., and J. Sheng, Subtidal circulation on the Scotian Shelf: Assessing the hindcast skill of a linear, barotropic model, *Journal of Geophysical Research*, *102*, 24,987–25,004, 1997.

Vincent, L., X. Zhang, B. Bonsal, and W. Hogg, Homogenization of Daily Temperatures over Canada, *Journal of Climate*, *15*, 1322–1334, 2002.

von Storch, H., E. Zorita, and U. Cubasch, Downscaling of global climate change estimates to regional scales: an application to Iberian rainfall in wintertime, *Journal of Climate*, *6*, 1161–1171, 1993.

Wilby, R., and C. Dawson, Using SDSM Version 3.1–A Decision support tool for the assessment of regional climate change impacts, *User manual*, 2004.

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

Wilby, R., C. Dawson, and E. Barrow, Sdsm: A decision support tool for the assessment of regional climate change impacts, *Environmental Modelling and Software*, *17*, 145–157, 2002.

Wilby, R., S. Charles, E. Zorita, B. Timbal, P. Whetton, and L. Mearns, Guidelines for use of climate scenarios developed from statistical downscaling methods, *IPCC Task Group on Data and Scenario Support for Impacts and Climate Analysis*, 2004.

Wunsch, C., The interpretation of short climate records, with comments on the North Atlantic and Southern Oscillations, *Bulletin of the American Meteorological Society*, *80*, 245–255, 1999.