University of Alberta

Iterative Interpolation of Daily Precipitation Data over Alberta

by

Qingfang Dai

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Applied Mathematics

Department of Mathematical and Statistical Sciences

Spring 2010
Edmonton, Alberta

Examining Committee

Peter Hooper, Mathematical and Statistical Sciences

Samuel Shen, Mathematical and Statistical Sciences

John Bowman, Mathematical and Statistical Sciences

Yanping Lin, Mathematical and Statistical Sciences

Gerhard Reuter, Earth and Atmosphere Sciences

James Ramsay, Psychology, McGill University

Dedication

This thesis is dedicated to my parents and my families.

Abstract

This thesis develops an optimal interpolation method that takes daily precipitation values collected from weather stations and produces precipitation estimates on a grid. The method, called Hybrid 2.0, combines EOF-based linear interpolation with the nearest-station method. Gridded monthly precipitation is first obtained via EOF, then distributed among days via nearest station. Hybrid 2.0 builds on an earlier method, called Hybrid 1.0, that applies an inverse-distance weighting method to obtain gridded monthly values. Hybrid 2.0 uses these monthly Hybrid 1.0 values as inputs when constructing EOF functions.

The data used in this thesis were obtained from the Meteorological Service of Canada. Few weather stations were located in the northern and mountain regions of Alberta prior to 1950. As a result, the Hybrid 1.0 gridded results underestimate precipitation in these regions for that period. The main contribution of Hybrid 2.0 is a substantial reduction in this bias, obtained by implicitly taking topographic elevation into account. Bias reduction is achieved by extracting EOFs from Hybrid 1.0 output for 1951-2002, when many more stations were present in the northern and mountain regions. Hybrid 2.0 is shown to be more accurate in interpolating both monthly and daily precipitation in Alberta, when compared with Hybrid 1.0 and other methods. The thesis also provides detailed analyses of precipitation trends and droughts using the gridded Hybrid 2.0 daily values. Optimality of the selected EOF modes and sensitivity to data error in the EOF-based linear reconstruction are also discussed in this thesis.

Agricultural uses of historical climate data have become extremely important. Applications include: enabling prompt, optimal decisions on market prices and disaster aid, designing future agricultural practices such as adaptation to climate and technology changes, and managing risks for agricultural producers and governments in areas such as drought monitoring. Many applications require a reliable interpolation technique to accurately reconstruct daily climate estimates onto grids of various resolutions. The gridded Hybrid 2.0 daily precipitation values produced by this thesis satisfy this requirement and can be used as inputs for many agricultural applications.

Acknowledgement

I am grateful to a number of people for their assistance, patience, and encouragement in completing this thesis. Without them, this work may not be possible to be completed by now.

First and foremost, I would like to express my deep and sincere gratitude to my supervisors, Professor Samuel S. P. Shen of the Department of Mathematics and Statistics at San Diego State University and Professor Peter Hooper of the Department of Mathematical and Statistical Sciences at University of Alberta. During my doctoral study, they always inspired, encouraged and financially supported me. Without their patience and generous support, as well as the useful suggestions during the time of research and writing of this thesis, this thesis would not have come as far as it did.

I also thank all support staffs of Department of Mathematical and Statistical Sciences at University of Alberta who have assisted me in many different ways during the past five and half years. Special thanks to my doctoral committee members for the valuable comments and recommendations eliminated many ambiguous points from the thesis.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AAFRD Alberta Agriculture, Food and Rural Development

CV Cross Validation

EDP Eco-District polygons

EOF Empirical orthogonal function

EPIC Erosion / Productivity Impact Calculator

IDW Inverse distance weighting

MAE Mean absolute error

MBE Mean biased error

MJJA from May to August

M-K Mann-Kendall

MSC Meteorological Service of Canada

MSE Mean Squared Error

PC principal component

PCA Principal Component Analysis

RMSE Root mean square error

SLC Soil Landscapes of Canada

SPI Standardized Precipitation Index

SST sea surface temperature

WEPP Water Erosion Prediction Project

Chapter 1

Introduction

1.1 Climate in Alberta

Alberta is a Canadian Province located between the 49th and 60th parallels of the northern latitude and between 110°W and 120°W in longitude. Such northern location results in the cold winter in Alberta. Although the winter is very cold, the summer can be warm and most days throughout the year are sunny. The arctic air masses in the winter produce extreme minimum temperatures as low as -54 °C in northern Alberta and -46°C in southern Alberta (Chetner et al., 2003). The continental air masses can produce maximum temperatures ranging from 32°C in the mountains to more than 40°C in the southern plains. These air masses can move quickly in the spring and fall, resulting in rapid seasonal changes. A famous climate characteristic in Alberta is its chinook winds, which sweep into southern Alberta several times each winter. These dry, warm winds can rapidly lift Alberta out of a deep freeze. The warming effect of the chinook winds near the mountains produces a west-to-east trend in winter temperatures. Precipitation is generally highest along the mountains and in west central Alberta due to the combined effects of frontal and topographic precipitation. Precipitation from May 1 to August 31 varies from slightly below 200 millimetres (mm) in the driest

prairie areas to more than 325 mm in the mountains. From September 1 to April 30, precipitation varies from less than 150 mm in the driest prairie region to more than 275 mm in the mountains (Chetner et al., 2003).

1.2 Measuring Precipitation

As the glossary from National Climate Data and Information Archive described by Environment Canada, Rain, drizzle, freezing rain, freezing drizzle and hail are usually measured by using the standard Canadian rain gauge, a cylindrical container 40 cm high and 11.3 cm in diameter. The precipitation is funneled into a plastic graduate which serves as the measuring device. Snowfall is the measured depth of the newly fallen snow, measured by using a snow ruler. Measurements are made at several points which appear representative of the immediate area, and then averaged. "Precipitation" is the water equivalent of all the above types of precipitation.

At most ordinary stations, the water equivalent of snowfall is computed by dividing the measured amount by ten. At principal stations, the water equivalent of snowfall is usually determined by melting the snow that falls into Nipher gauges. These are precipitation gauges designed to minimize turbulence around the orifice and to be high enough above the ground to prevent most blowing snow from entering. The amount of snow determined by this method normally provides a more accurate estimate of precipitation than that provided by using the "ten-to-one" rule. Even at ordinary climate stations, the normal

precipitation values will not always be equal to the rainfall plus one tenth of the snowfall. Missing observations are one cause of such discrepancies. Precipitation measurements are usually made four times per day at principal stations and usually once or twice per day ordinary stations.

1.3 Some of the Existing Interpolation Methods

Precipitation is measured only at limited number of locations, while the practical applications require precipitation data on a dense grid. Thus, the mathematical interpolation of the observed station data onto a grid is necessary. Many methods are available to interpolate the station data onto grid points, such as nearest-station assignment, inverse-distance weighting (Jones et al., 1986), kriging (Hudson and Wachkernagel, 1994; Cressie, 1993), thin plate smooting splines (Hutchinson, 1995, 1998 a, b), and empirical orthogonal function (EOF) method (Smith et al., 1998). To make the materials of this thesis self-contained, these commonly used interpolation methods are recapitulated below as was done in Shen at al. (2001).

(1). Nearest-station Assignment

The nearest-station assignment method is that each grid point is assigned the observed value of the nearest station that has data for a particular day. Since the nearest-station method uses only one station's data for a grid for a given day, the interpolated grid should adequately preserve the variance of a single point, although the nearest station with observed data available may change from day to

day. This method assigns the observed climate data directly from the nearest station to a grid point and should not yield a large bias when the observational stations are sufficiently dense. However, this method is by no means optimal since no computational optimization is implemented. When the observational stations are very sparse and the climate conditions are complex, this method will result in substantial errors for a climate parameter which varies over short length scales.

(2). Inverse-distance Weighting

The inverse distance weighting (IDW) method is based on the assumption that the predicted location should be influenced more by the nearby points and less by the more distant points. The general formula of the IDW is as follows:

$$u(x)=\frac{\sum_{k=1}^{N} w_k(x)\,u_k}{\sum_{k=1}^{N} w_k(x)}, \tag{1.1}$$

where

$$w_k(x)=\frac{1}{d(x,x_k)^p}, \tag{1.2}$$

x denotes an interpolated grid point, x_k is an known station, d is a given distance from the known station x_k to the unknown point x, N is the total number of known stations used in the interpolation, and p is a positive real number used as the power parameter. The IDW estimates the predicted values by weighted averaging of the values of the known data points in the vicinity of the predicted location. The weight assigned to each neighbor point decreases as

the distance from the predicted point to the neighbor point increases. With IDW, we can control the significance of the known points upon the interpolated values based on their distance from the predicted point. Usually, p is taken to be one. If a higher power is used, more emphasis will be placed on the nearest points, and the resulting surface will be less smooth. Using a lower power will put more emphasis on the farther points and result in a smoother surface. Another parameter that can control the characteristics of the interpolated surface is the search radius, which can be fixed or variable, and determines the number of known points that can be used for calculating the predicted value. Jones et al. (1986) used this method to interpolate the monthly mean temperature data onto the 5-by-10 degrees latitude and longitude grid points for the Northern Hemisphere over the period of 1851 – 1984. The obvious advantage of this method is that its computations are simple and efficient. For well-spaced data points, it is a good all-purpose interpolation method. However, for daily precipitation the resulting surface is too smooth.

(3). Kriging

Kriging is an interpolation method using geostatistics and based on statistical models that include autocorrelation – the statistical relationship among the known points. It has been accepted as a tool for interpolating many types of climate data including precipitation data (Daly et al., 1994). The kriging interpolation procedure consists of two steps: (1) calculating the semivariogram and modeling the spatial structure of the known station data, and (2) interpolating the values at the predicted points by using a fitted model for the spatial structure (Li et al.,

2006). Similar to the IDW method, kriging also weighs the contribution of the surrounding station data based on the distance between the surrounding station and the predicted location. In IDW, the weight depends solely on the distance to the predicted point. However, with the kriging method, the weights are based not only on the distance between the known points and the predicted location but also on ambient covariance properties of the interpolated parameter (Conolly and Lake, 2006). Hudson and Wachkernagel (1994) integrated the elevation information into the kriging to obtain an improved map of the January mean temperature in Scotland by comparing with the map interpolated by kriging based on the temperature data only. However, as Shen et al. (2001) pointed out, kriging's drawback is that it requires the studied field to be relatively stationary in time and homogeneous in space, so that the accuracy of the interpolation is questionable for daily precipitation data.

(4). Thin Plate Smoothing Splines

The original thin plate smoothing splines method was first described by Wahba (1979), who later provided a comprehensive introduction to its techniques, with various extensions in (1990). When bivariate thin plate splines are used, the climate surface is interpolated as a function of the latitude and longitude only, while with trivariate thin plate splines, the elevation information is included in the interpolation of the climate surface by a function with latitude, longitude and elevation as the independent variables. In the 1990s, Hutchinson extensively explored the meteorological applications of the method. Hutchinson (1995, 1998ab) provided a theoretical description of their application to annual mean

rainfall data and the applications to the daily rainfall values.

(5). Empirical Orthogonal Function (EOF) Method

The empirical orthogonal function (EOF) Analysis (or principal component analysis - PCA) is a method to determine the main spatial patterns of variability, as well as their variation in time, and to measure the relative importance of each pattern. The EOF has been applied in meteorological research since 1950s. However, Kutzbach (1967) provided detailed explanations and demonstrations of the EOF analysis, and following their papers, the EOF technique became popular in climatological research after the 1970s.

EOF analysis is conventionally used in climatology for examining data sets of space - time - distributed observations because this analysis reduces large amount of temporal-spatial data to a smaller number of spatial patterns, termed EOFs, and their temporal coefficients. The EOF application to the data interpolation is an expansion of the climate data into a finitely many EOFs and an estimation of the expansion coefficients. This method has been used for various climate parameter reconstructions. Smith et al. (1998) successfully used this method to interpolate the tropical Pacific's sea surface temperature (SST). Shen et al. (2004) used it to predict the SST anomalies from the land stations' SAT (surface air temperature) anomalies, and Smith et al. (2008) extended this method to predict oceanic precipitation from land stations.

The theory behind EOF computations is straightforward (see, for example, Lagerloef and Bernstein, 1988, for a particularly concise and lucid description). Basically, the eigenvectors of the data covariance matrix are calculated. Because

the covariance matrix is real and symmetric, the set of eigenvectors forms an orthogonal basis set. The norm of each eigenvector can be taken as any nonzero number and thus is usually taken as one so that the eigenvectors form an orthonormal basis for the data.

1.4 Thesis Objective: Hybrid 2.0 Interpolation

Shen et al. (2001) developed a hybrid method which is named as Hybrid 1.0 to interpolate daily precipitation data and produced the gridded climate data named ABClim 1.0. This method is a hybrid of inverse-distance-weight (IDW) interpolation for monthly data and a nearest-station assignment regression for redistributing the monthly total onto days. The procedure of Hybrid 1.0 is as follows. Firstly, we use IDW to interpolate the monthly data as

$$R_{grid}(m)=\frac{\sum_{k=1}^{N} R_k(m)/d_k}{\sum_{k=1}^{N} 1/d_k}, \qquad (1.3)$$

where $R_{grid}(m)$ is the estimated monthly precipitation at the grid point, $R_k(m)$ is the observed monthly data at the station k, d_k is the distance between the grid point and the station k, and N is the total number of the stations nearest to the grid point. N stations are selected according to the distance between the station and the grid point. That is, the ith station is the ith nearest station to the grid. No more than eight stations with $d_k \le 60km$ are chose in Hybrid 1.0 method. If the grid point is on the station k, $R_{grid}(m)=R_k(m)$. Secondly, the daily precipitation is computed by using the following nearest-station assignment method.

$$R_{grid}(d)=R_{grid}(m)\times\frac{R_{nearest}(d)}{R_{nearest}(m)}, \qquad (1.4)$$

where $R_{nearest}(d)$ is the precipitation of the station nearest to the grid point for the given day and $R_{nearest}(m)=\sum_{\text{all days in month}} R_{nearest}(d)$. When $R_{nearest}(m)$ was zero for a given month, $R_{grid}(d)$ is assigned the value zero. Thus (1.3) and (1.4) formulate the Hybrid 1.0 method. It is important to clarify a fact here that the nearest station method is different with the kernel_based smoothing method. The nearest station method searches the unknown grid's nearest station with available observed precipitation for each individual day of the month and then assign this observed data onto the predicted grid. There is no smoothing function used in the prediction and the predicted result displays the variance nature of the true data. The kernel_based smoothing method uses a smoothing kernel function to predict the unknown grid value and results in a smooth result. The only similarity between both methods is that the predicted grid's nearest station has most influence on the unknown grid.

In general, the Hybrid 1.0 method satisfies the mass conservation law

$$\sum_{\text{all days in month}} R_{grid}(d)=R_{grid}(m), \qquad (1.5)$$

that is, the monthly total precipitation for a grid is not changed after using the precipitation frequency formula.

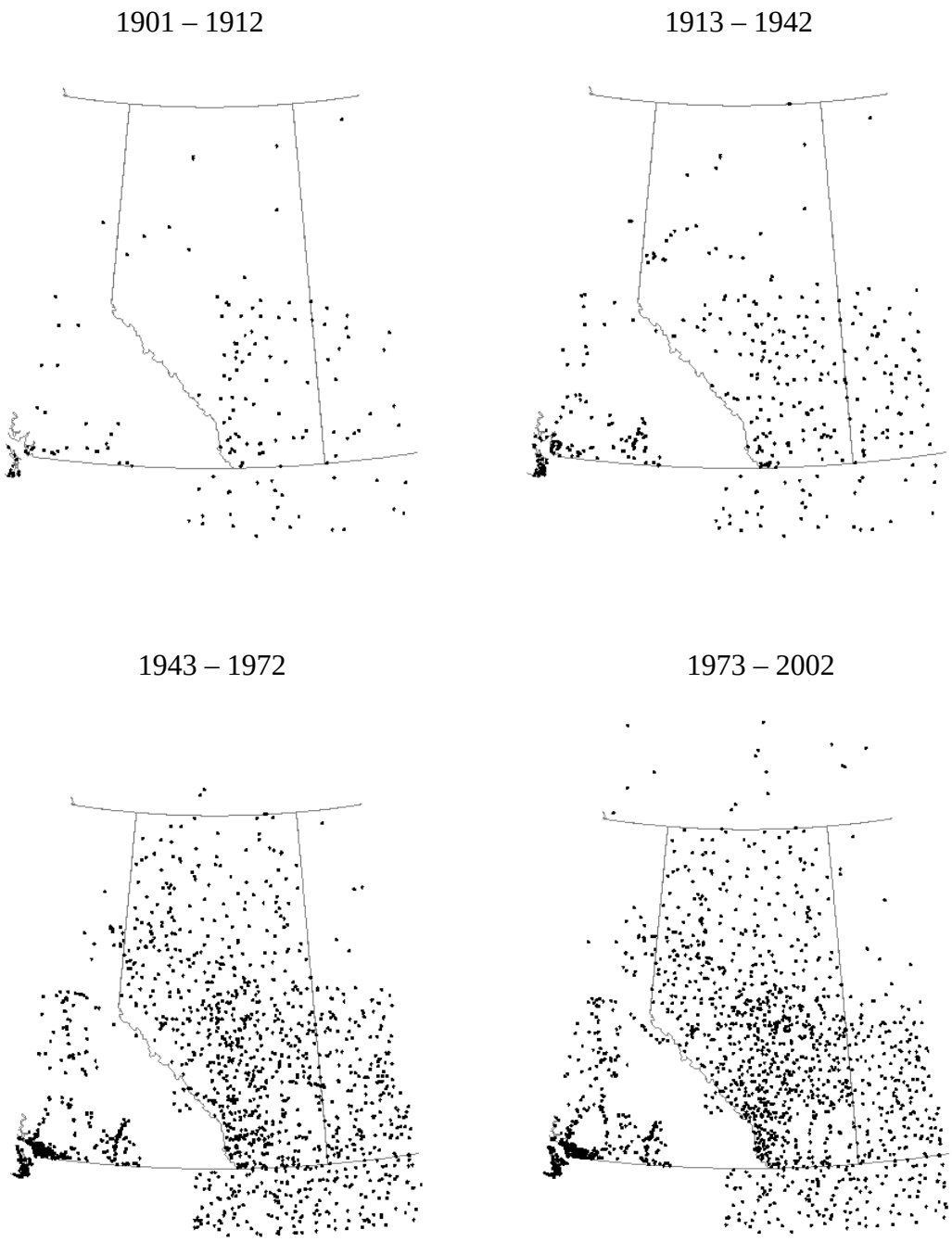


Figure 1.1. Distributions of the Alberta precipitation stations in the periods of 1901-1912, 1913 - 1942, 1943 - 1972, and 1973 - 2002.

Because of the sparse precipitation station distribution in high-elevation regions in the earlier period of the last century, the use of the IDW method can lead to serious underestimation of the precipitation over those regions in that period. Figure 1.1 shows the distribution of the precipitation stations in Alberta in four periods: 1901 - 1912, 1913 - 1942, 1943 - 1972, and 1973 - 2002. The figure indicates that only four stations (Fort Chipewyan, two stations in Fort Vermillion and Fort McMurray) were north of 56°N before 1912. Moreover, they had low elevations and, hence, did not measure the topographic precipitation over the mountain (i.e., Caribou Mountains, Buffalo Hills, Clear Hills, Birch Mountains) and the lake (i.e., Lake Athabasca and Lake Claire) regions. Thus, the results in the earlier part of the last century from the IDW interpolation method have a low precipitation bias in northern Alberta and the mountain area. To overcome this problem, an interpolation method needs to be developed to take topographic elevation and climatic variance into account. Hybrid 2.0, the second generation of Hybrid 1.0 is thus developed in this study to interpolate the daily precipitation over Alberta. The method is a hybrid of EOF-based linear reconstruction for monthly precipitation data and a nearest-station assignment regression for redistributing the monthly total onto each day of the month.

1.5 Thesis Outline

This thesis is organized as follows. Chapter 2 applies the EOF analysis to the 1961 – 2000 Hybrid 1.0 output and reconstructs the monthly precipitation data on the 0.25 by 0.5 degrees latitude and longitude grids by using the EOF-

based linear interpolation method for the province of Alberta. Chapter 3 uses the developed Hybrid 2.0 interpolation method to estimate the daily precipitation data based on the results of Chapter 2 and interpolates both the monthly precipitation data and daily precipitation data onto a finer grid. Chapter 4 introduces some applications and analysis of the Hybrid 2.0 data obtained in Chapter 3. Chapter 5 discusses the optimal selection of EOF modes in the EOF-based linear reconstruction and the sensitivity of the observational error used in the calculation of the optimal weights. Conclusions and suggestions for the future work are presented in Chapter 6.

Chapter 2

Spectral Method for Interpolating Monthly Precipitation Data over Alberta

2.1 Introduction

Alberta Agriculture, Food and Rural Development (AAFRD), a provincial government department, in partnership with the agriculture industry, has been developing a strategy for sustainable agriculture. AAFRD is committed to environmental sustainability and is working with researchers to develop quantitative measures of it. Climate data are particularly important to the sustainable agriculture and agricultural information technology (Changnon and Kunkel, 1999).

Soil quality is one of the initial indicators of environmental sustainability being developed. One aspect of sustainability is to ensure that land-management practices maintain or improve soil quality. Soil organic matter is one of the key soil attributes associated with soil quality. Soils with higher levels of organic matter are generally considered to be of better quality than other soils and tend to have (i) better nutrient-retention characteristics for good crop growth; (ii) better water infiltration rates, resulting in slower rates of water erosion of soil; and (iii) better structure, reducing susceptibility to wind erosion. AAFRD is using several models, such as EPIC (Erosion / Productivity Impact Calculator), to assess soil

quality in Alberta. Most models need a complete, continuous climate data set at a given resolution and with no missing data as their input. The models quantitatively estimate the effect on soil quality due to changes in land management practices under the climate conditions used in the models. Therefore actual climate data are needed to use the models to compare their results with carefully measured soil data.

Various applications of the climate data require improved accuracy of the gridded data. This Chapter will introduce the EOF interpolation method for interpolating the gridded monthly precipitation, which will be the Hybrid 2.0 gridded monthly precipitation, over Alberta province.

The EOFs are calculated as the normalized eigenvectors of the covariance matrix of the gridded Hybrid 1.0 monthly output. These projections are called the principal component (PC) time series or the expansion coefficients of the EOFs. Because the method finds a set of orthonormal basis vectors that maximizes the projection of the data onto the basis vectors, the EOF modes are uncorrelated over space, and the expansion coefficients are uncorrelated in time.

The discussion in this chapter takes elevation into account by using the EOF interpolation method to interpolate the station monthly precipitation data onto the grids in Alberta. The next chapter will use the nearest-station assignment method to redistribute the EOF-interpolated monthly data on each day. For this study, we developed hybrid 2.0, a hybrid of the EOF-interpolation method and the nearest-station assignment method, and also the next generation of the hybrid method.

This chapter is organized as follows. Section 2.2 describes the data used in this chapter. The detailed EOF- interpolation method is addressed in Section 2.3. Section 2.4 displays the reconstructed results of monthly data on the 0.25 by 0.5 degrees latitude and longitude grids, and the accuracy of the interpolated monthly precipitation is assessed in Section 2.5. Section 2.6 presents our conclusions and discussion.

2.2 Data

The dataset used in this study is the daily precipitation data at the 2645 climate stations within Alberta and those outside Alberta but near its borders: 4^o of longitude to the east and west, 4^o of latitude to the north, and 2^o of latitude to the south. The time period is between January 1, 1901 and December 31, 2002. The data is from the Meteorological Service of Canada (MSC). The website of the data is http://climate.weatheroffice.ec.gc.ca/prods_servs/documentation_ index_e.html#dly. The daily data are in the following format.

Monthly Record of Daily Data (DLY) - Length 233

| STN ID | YEAR |MO |ELEM |S| VALUE |F|

|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|

This 7-space data is repeated 31 times. Here, the legends are below.

Abbrv	Length	Field	Data Type
STN ID	7	Station Identification	alphanumeric
YEAR	4	Year, e.g., 1998	Numeric

MO	2	Month, i.e., 01 = Jan. etc.	Numeric
ELEM	3	Element Number	Numeric
S	1	Sign	'-' = negative '0' = positive
VALUE	5	Data Value	Numeric
F	1	Flag	alphanumeric

Standard record formats have been adopted for climatological data, which are archived at fifteen-minute, hourly, daily or monthly intervals. Each record consists of station identification, date, and element number followed by the data repeated for each time interval. The datum for each time interval is recorded as a five-digit integer plus a leading sign field and a following flag field. The units and decimal position are implied by the assigned element number.

In the initial data preparation pre-process, all the daily precipitation data with flags A and F are removed, where A stands for accumulated amount, and F stands for the accumulated and estimated amount. These data are deemed unreliable, and to correct these data is the data homogenization task which is not the focus of our research here. The stations used in the interpolation are shown in Fig. 2.1. Southern Alberta has higher station density than northern Alberta, where the station distribution is very sparse.

In this study, the station data are interpolated onto the 0.25-by-0.5 degrees latitude and longitude grid points over Alberta. Figure 2.2 shows the distribution of grid points in Alberta. The EOF-interpolation method is used to generate the

monthly precipitation onto the grids. Thirty years of daily precipitation data from January 1, 1961 to December 31, 1990 are used for computing the monthly precipitation mean (Figure 2.3).

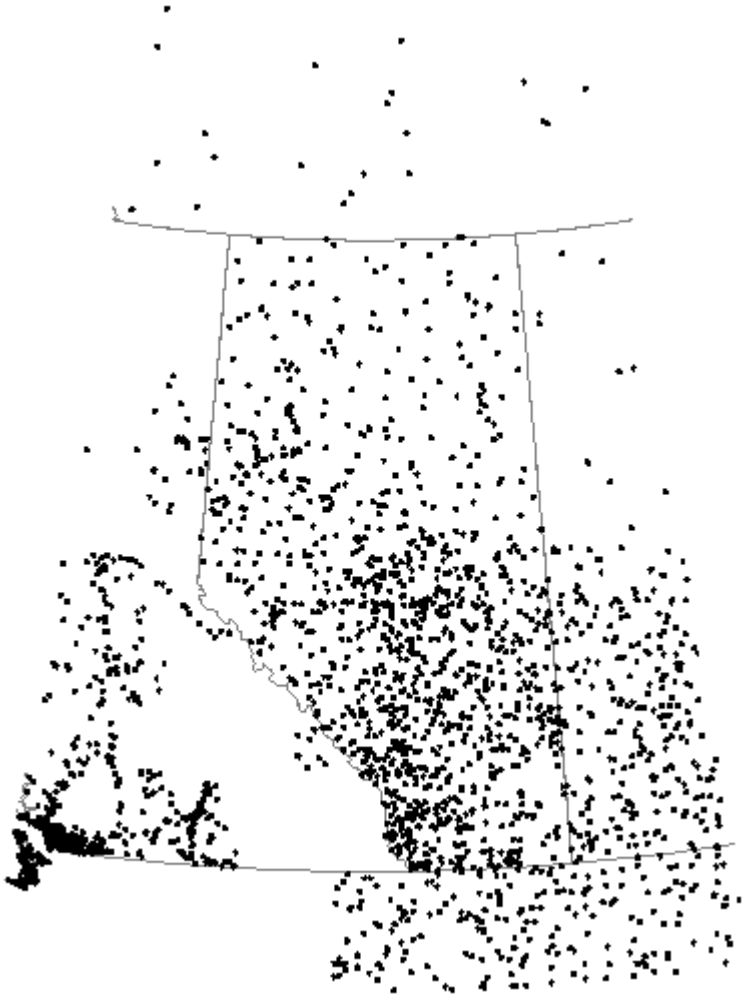


Figure 2.1. Precipitation station distributions in Alberta.

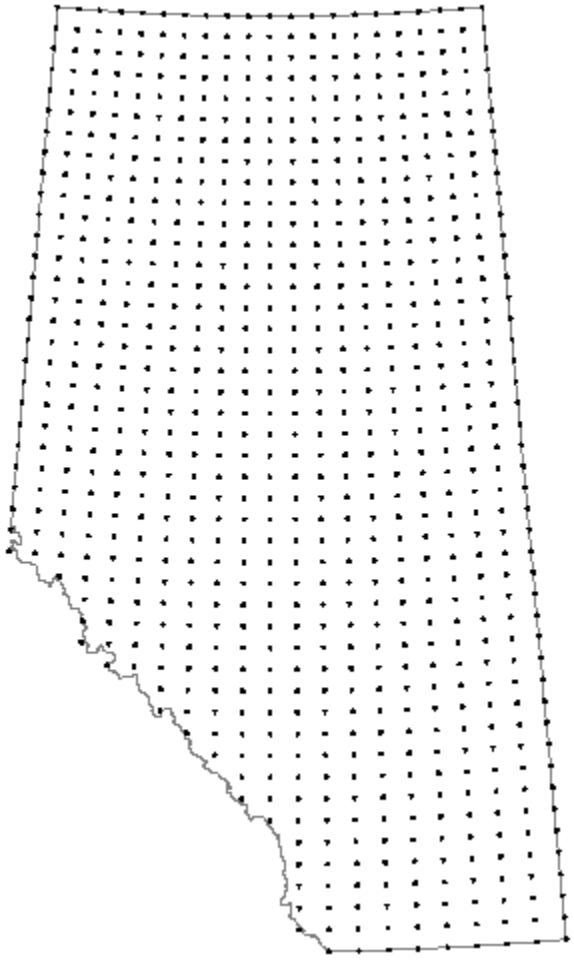
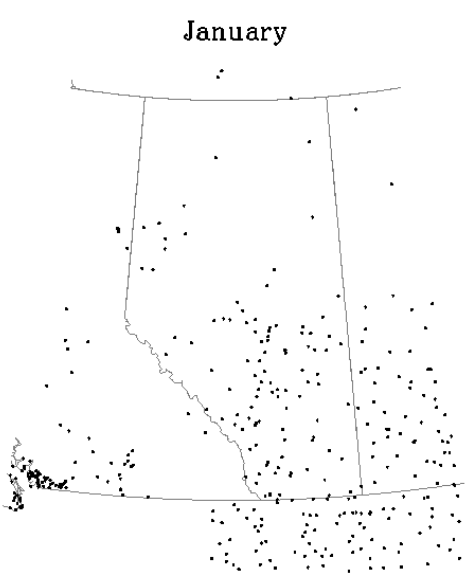
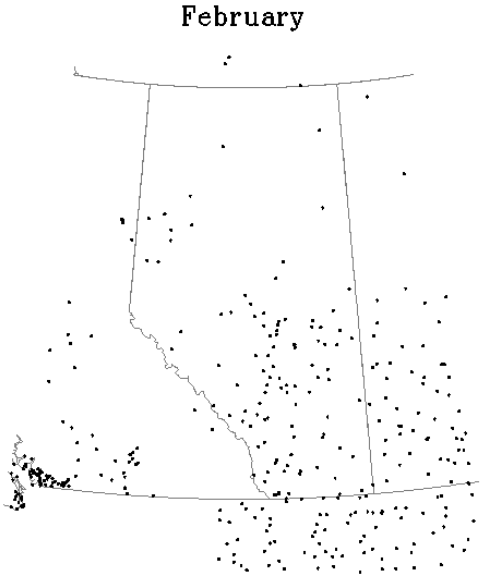
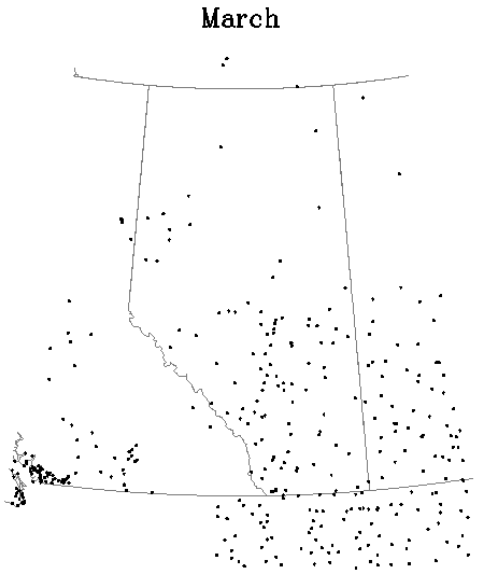
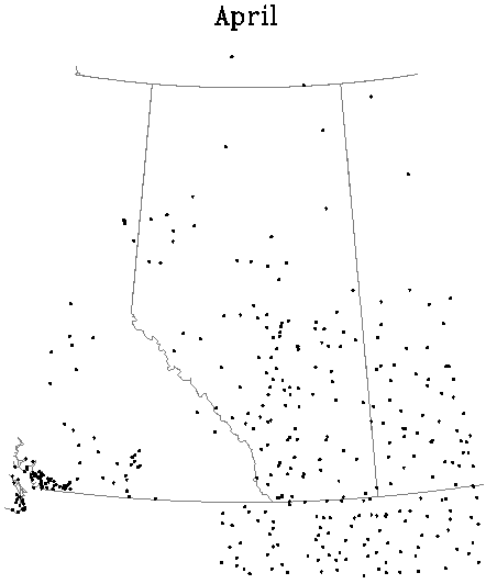


Figure 2.2. Distribution of the grid points (0.25-by-0.5 degrees latitude and longitude) over Alberta.

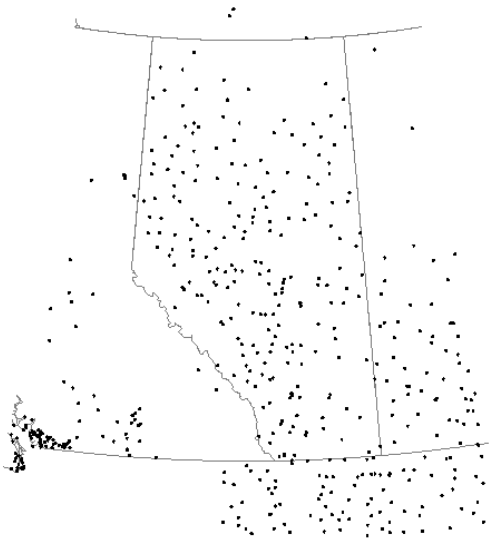

January


February

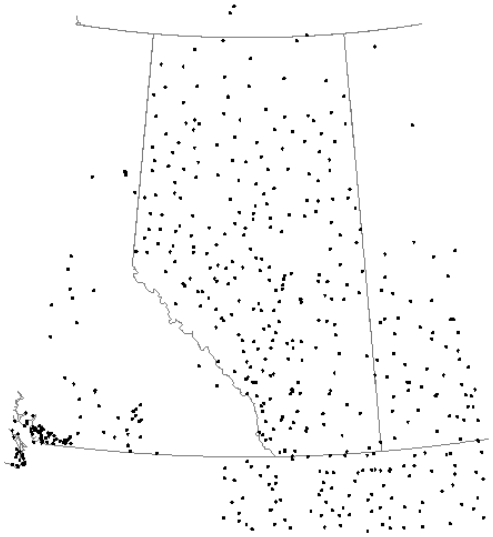

March


April

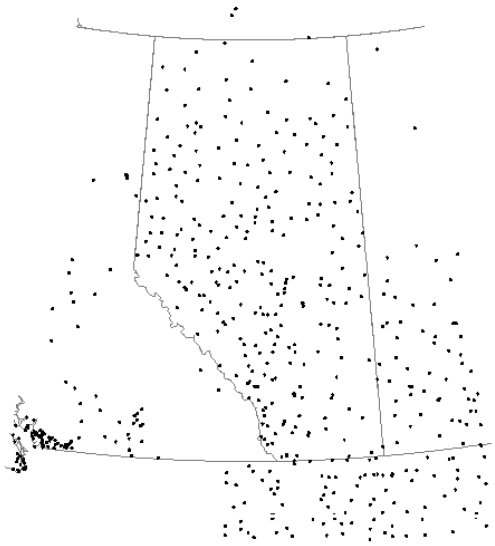
May



June



July



August

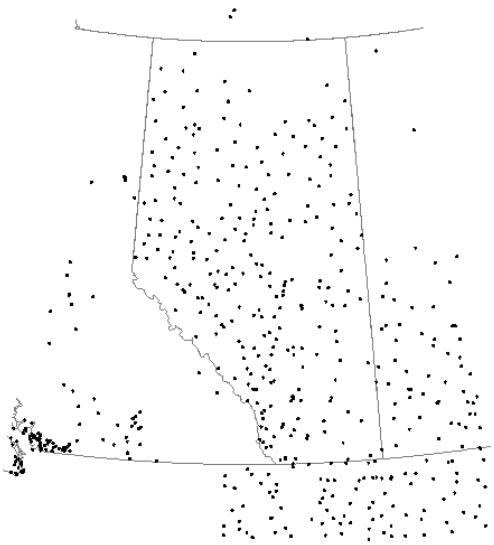


Figure 2.3. Distribution of the Alberta precipitation stations whose 1961-1990 monthly means were computed for each month: January to December.

Figures 2.4 and 2.5 show the daily and monthly series of the total number of precipitation stations in Alberta from 1901 to 2002, respectively. The seasonal fluctuations occurred because some stations were operating only in the growing season. Figure 2.6 shows the distribution of the precipitation stations within Alberta and those 4^{o} of longitude to the east and west, 4^{o} of latitude to the north, and 2^{o} of latitude to the south in the time interval between January 1, 1951 and December 31, 2002. Figures 2.4, 2.5 and 2.6 reveal that the precipitation stations have a good coverage over Alberta between January 1951 and December 2002, and that precipitation stations existed in the mountain and lake regions. Therefore, we can take the elevations into account by extracting the important patterns in the mountain and lake regions from Hybrid 1.0 gridded monthly precipitation in the time interval between January 1951 and December 2002. Thus, the EOF-interpolation method also can reliably interpolate the monthly precipitation in the mountain and lake regions.

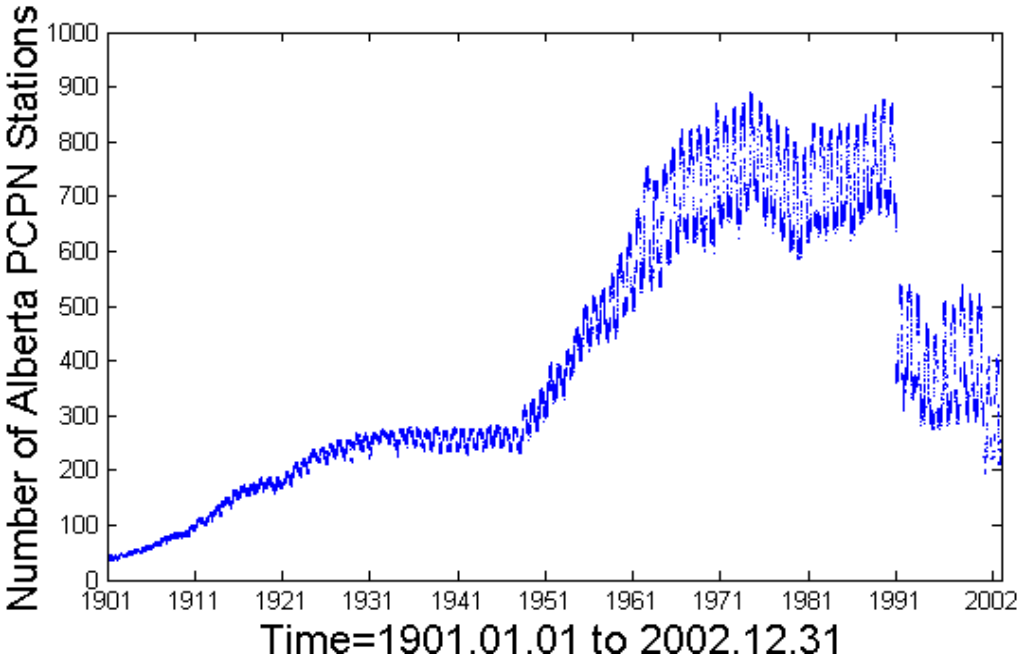


Figure 2.4. The daily number of Alberta stations used in the precipitation data interpolation.

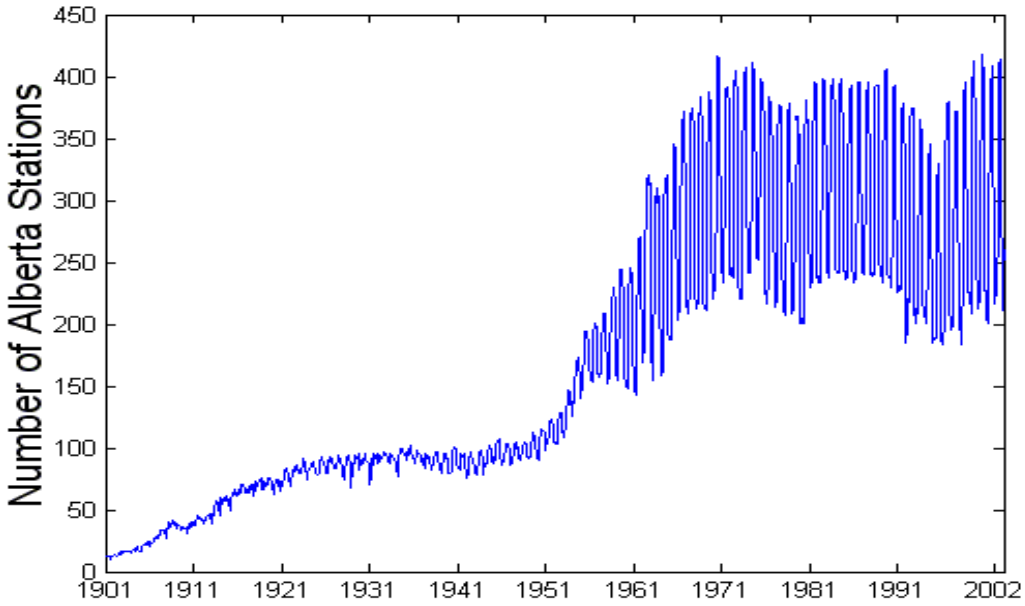


Figure 2.5. The monthly number of stations used in the precipitation data interpolation. The annual cycle indicate the seasonal variability of the number of stations: more in the summer and less in the winter.

Figure 2.6. Distribution of the precipitation stations for interpolation in the time interval between January 1, 1951 and December 31, 2002.

2.3 Interpolation Method

2.3.1 The Definition of EOF

The precipitation anomaly field, or departure from the mean, is defined at $(\mathbf{r},t)$ by

$$R(\mathbf{r},t) = P(\mathbf{r},t) - \mu(\mathbf{r}), \tag{2.1}$$

where $\mathbf{r}$ is any spatial location, and t is the time. $P(\mathbf{r},t)$ is the precipitation at $(\mathbf{r},t)$, and $\mu(\mathbf{r})$ is the precipitation mean at $\mathbf{r}$. The covariance function is defined as

$$C(\mathbf{r},\mathbf{r}') = \langle R(\mathbf{r},t)R(\mathbf{r}',t)\rangle \ , \tag{2.2}$$

where $\langle\bullet\rangle$ denotes the sample mean. The continuous EOFs $\psi_m(\mathbf{r})$ are the eigenfunctions of the covariance function (Shen, 1994):

$$\int_\Omega C(\mathbf{r},\mathbf{r}')\psi_m(\mathbf{r}')d\mathbf{r}' = \lambda_m\psi_m(\mathbf{r}) \ , \tag{2.3}$$

where λ_m is the m'th eigenvalue of $C(\mathbf{r},\mathbf{r}')$, and Ω is the region being studied. These EOFs are orthonormal and complete:

$$\int_\Omega \psi_m(\mathbf{r})\psi_n(\mathbf{r})d\mathbf{r} = \delta_{mn} = \begin{cases} 1, & m=n \\ 0, & m \neq n \end{cases} \ , \tag{2.4}$$

$$\sum_{m=1}^{\infty} \psi_m(\mathbf{r})\psi_m(\mathbf{r}') = \delta(\mathbf{r}-\mathbf{r}') \ , \tag{2.5}$$

where δ_{mn} is the Kronecker delta, and $\delta(x)$ is the Dirac delta , which is defined as

$$\int_{-\infty}^{\infty} \delta(x-x_0)f(x)dx = f(x_0) \tag{2.6}$$

for any sufficiently well-behaved function $f(x)$.

From the properties (2.4) and (2.5), the covariance function can be expanded as

$$C(\mathbf{r},\mathbf{r}') = \sum_{m=1}^{\infty} \lambda_{\mathrm{m}} \psi_{\mathrm{m}}(\mathbf{r}) \psi_{\mathrm{m}}(\mathbf{r}') . \tag{2.7}$$

However, in practice the data are discrete, and their corresponding EOFs should be discrete. A discrete approximation should thus be used. The discretization of the above continuous EOFs is described below. We denote the value of any anomaly field at the discrete grid point r_j and time t_n by R_{jn} for j=1,$\cdots$,J and $n = 1,\cdots Y$. Then the anomaly field can be represented by the data matrix

$$\mathrm{R}=\begin{pmatrix} \mathrm{R}_{11} & \mathrm{R}_{12} & \ldots & \mathrm{R}_{1\mathrm{Y}} \\ \mathrm{R}_{21} & \mathrm{R}_{22} & \ldots & \mathrm{R}_{2\mathrm{Y}} \\ \ldots & \ldots & \ldots & \ldots \\ \mathrm{R}_{\mathrm{J1}} & \mathrm{R}_{\mathrm{J2}} & \ldots & \mathrm{R}_{\mathrm{JY}} \end{pmatrix}. \tag{2.7}$$

The aim of EOF analysis is to find the linear combination of all grid points to explain maximum variance, that is, to find a direction $\psi = (\psi(r_1),\cdots,\psi(r_J))'$ such that $R'\psi$ has maximum variability. Here the prime denotes the transpose of a matrix. We know that the anomaly field has a zero mean, so the variance of the time series $R'\psi$ is

$$Var(R'\psi) = \frac{1}{Y}\left\|R'\psi\right\|^2 = \frac{1}{Y}\left(R'\psi\right)' R'\psi = \psi' \frac{1}{Y} RR'\psi = \psi' C\psi, \tag{2.8}$$

where

$$C = \frac{1}{Y} RR' \tag{2.9}$$

is the covariance matrix. To make the above maximum variance problem bounded, we normally require the direction vector ψ to be unitary. The problem then becomes

$$\max_{\psi}(\psi' C\psi),\ s.t.\ \psi'\psi = 1. \tag{2.10}$$

The solution to (2.10) yields the first EOF, i.e., an eigenvector corresponding to the largest eigenvalue. The remaining EOFs can be motivated by a similar optimization problem with additional orthogonality constraints. The EOFs are thus eigenvectors obtained by solving

$$C\psi = \lambda\psi. \tag{2.11}$$

It is clear that the covariance matrix C is symmetrical and therefore diagonalizable. The m'th EOF (or mode) is the m'th eigenvector ψ_m of C if the eigenvalues and the corresponding eigenvectors have been sorted in decreasing order. As well the covariance matrix C is positive semidefinite, so all the eigenvalues are hence non-negative. The m'th eigenvalue (or variance) λ_m corresponding to the m'th EOF ψ_m gives a measure of the explained variance by ψ_m, $m = 1, \cdots, J$. The explained percentage variance is

$$\frac{\lambda_m}{\sum_{n=1}^{J}\lambda_n} 100\%. \tag{2.12}$$

2.3.2 EOF Calculation

In this study, the gridded Hybrid 1.0 output provided on a 0.25-by-0.5

degrees latitude and longitude grid are used to extract the EOF patterns. Clearly, the distribution of the Hybrid 1.0 output will be denser poleward. This non-uniform distribution can influence the structure of the computed EOFs. In order to avoid the effect of this geometrical artifact, we normally weight the data prior to analyzing them. The simplest and most useful way is to weight each data point by the local area of its location.

The details of our calculations are described as follows. The discrete form of (2.3) is written as (Li 2001; Shen et al., 1998)

$$\sum_{j=1}^{J} C_{ij}\psi_m(r_j)A_j=\lambda_m\psi_m(r_i), \quad i=1,2,\cdots,J;\ m=1,2,\cdots,J, \tag{2.13}$$

where

$$C_{ij}=C(r_i,r_j)=\frac{1}{y_2-y_1+1}\sum_{t=y_1}^{y_2} R(r_i,t)R(r_j,t) \tag{2.14}$$

is the covariance matrix, $R(r_i,t)$ is the Hybrid 1.0 precipitation anomaly at location r_i and time t,

$$A_j=R^2\left(\frac{\Delta\theta}{180}\pi\right)\left(\frac{\Delta\varphi}{180}\pi\right)\cos\varphi_j \tag{2.15}$$

is the area of the grid box **j**, $\Delta\theta\times\Delta\varphi$ is the resolution of the grid, φ_j is the latitude of center of the grid box **j**, R is the radius of Earth, approximately 6,376 km, J is the total number of grid boxes, and y_1 and y_2 are the years that the Hybrid 1.0 output in the time interval between them are used to extract EOF

patterns. It is easy to find that (2.13) is equivalent to the following symmetric form

$$\sum_{j=1}^{J}(\sqrt{A_i}C_{ij}\sqrt{A_j})(\psi_m(r_j)\sqrt{A_j})=\lambda_m(\psi_m(r_i)\sqrt{A_i}),\quad i=1,\,2,\,\cdots,\,J. \tag{2.16}$$

Therefore, we can solve the following new discrete eigenvalue problem

$$\sum_{j=1}^{J}\hat{C}_{ij}\upsilon_j^{(m)}=\lambda_m\upsilon_i^{(m)},\quad i=1,\,2,\,\cdots,\,J \tag{2.17}$$

to obtain the eigenvalues λ_m and the area-weighted eigenvectors

$$\hat{\upsilon}^{(m)}=(\upsilon_1^{(m)},\ \upsilon_2^{(m)},\ \cdots,\ \upsilon_J^{(m)})'$$

which are defined as the discrete EOFs, where

$$\hat{C}_{ij}=\sqrt{A_i}C_{ij}\sqrt{A_j} \tag{2.18}$$

is the area-weighted covariance matrix, and the area-weighted eigenvectors

$$\upsilon_j^{(m)}=\psi_m(r_j)\sqrt{A_j},\quad j=1,\,2,\,\cdots,\,J \tag{2.19}$$

satisfy the normalization condition

$$\sum_{j=1}^{J}(\upsilon_j^{(m)})^2=1. \tag{2.20}$$

Then we have

$$\int_{\Omega}(\psi_m(\mathbf{r}))^2d\mathbf{r}\approx\sum_{j=1}^{J}(\psi_m(\mathbf{r}_j))^2A_j=\sum_{j=1}^{J}(\psi_m(\mathbf{r}_j)\sqrt{A_j})^2=\sum_{j=1}^{J}(\upsilon_j^{(m)})^2=1. \tag{2.21}$$

The eigenvalues λ_m and the continuous EOFs

$$\psi_m(r_j)=\frac{\upsilon_j^{(m)}}{\sqrt{A_j}},\quad j=1,\,2,\,\cdots,\,J \tag{2.22}$$

are then obtained.

We denote the length of the observation data by Y, which is equal to y_2 minus y_1 plus 1. Since the covariance matrix computed from insufficient observations is not of full rank, the calculations of the EOFs follow different algorithms which depend on the values of Y and J. In this study, the length of the Hybrid 1.0 data we used, Y, is much less than the total number of grid boxes, J. Hence, the spatial covariance matrix is not of full rank and its determinant will vanish. In such a case, we can compute the spatial EOFs indirectly through space and time exchange. From the following algorithm, two benefits can be found for this indirect method: one is that the determinant of the temporal covariance will not vanish, and the computation will be reduced greatly. We know that different months have different physical patterns, so the EOFs have to be computed for each month. The detailed algorithm for calculating the EOFs is now described.

First, calculate the $Y \times Y$ matrix with the area factor by transposing the data matrix:

$$\hat{D} = \frac{1}{Y}\left(\sqrt{A}R\right)'\left(\sqrt{A}R\right), \tag{2.23}$$

where

$$\sqrt{A}R=[\sqrt{A_j}R(r_j,t)]_{J\times Y}, \quad j=1,2,\cdots,J, \quad \text{and} \quad t=y_1,\cdots,y_2. \tag{2.24}$$

Second, solve the eigenvalue problem

$$\sum_{k=1}^{Y}\hat{D}_{nk}u_k^{(m)}=\tilde{\lambda}_m u_n^{(m)}, \ n=1, 2, \cdots, Y \tag{2.25}$$

to get the eigenvalues $\tilde{\lambda}_m$ and the eigenvectors $\hat{u}^{(m)} = (u_1^{(m)}, u_2^{(m)}, \cdots, u_Y^{(m)})'$, which are normalized:

$$\sum_{k=1}^{Y} (u_k^{(m)})^2 = 1. \tag{2.26}$$

Third, multiply (2.25) by the matrix $\frac{\sqrt{A}R}{\sqrt{\tilde{\lambda}_m}\sqrt{Y}}$ on the left, and compare the resulting equation with the spatial eigenvalue problem (2.17) to obtain the following relationships:

$$\lambda_m = \tilde{\lambda}_m, \tag{2.27}$$

$$\hat{\upsilon}^{(m)} = \frac{\sqrt{A}R}{\sqrt{\tilde{\lambda}_m}\sqrt{Y}} \hat{u}^{(m)}, \ m=1, 2, \cdots, Y. \tag{2.28}$$

It is easy to find that $\frac{\sqrt{A}R}{\sqrt{\tilde{\lambda}_m}\sqrt{Y}} \hat{u}^{(m)}$ satisfies the normalization condition (2.20).

Finally, the continuous EOFs and the corresponding eigenvalues are obtained as

$$\lambda_m = \tilde{\lambda}_m, \tag{2.29}$$

$$\psi_m(r_j) = \frac{\upsilon_j^{(m)}}{\sqrt{A_j}}, \quad j=1,2,\cdots,J, \text{and} \quad m=1,2,\cdots,Y. \tag{2.30}$$

2.3.3 EOF Interpolation

The projection of the precipitation anomaly field $R(\mathbf{r},t)$ onto the m'th continuous EOF $\psi_m(\mathbf{r})$, i.e.,

$$R_m(t)=\int_\Omega R(\mathbf{r},t)\psi_m(\mathbf{r})d\mathbf{r} \tag{2.31}$$

is the m'th principal component (PC) or EOF coefficient. By the orthogonality property of the EOFs and the definition (2.31), it is clear that the EOF coefficients (PCs) are uncorrelated:

$$\begin{aligned}\langle R_m(t)R_n(t)\rangle &= \left\langle \int_\Omega R(\mathbf{r},t)\psi_m(\mathbf{r})d\mathbf{r}\int_\Omega R(\mathbf{r}',t)\psi_n(\mathbf{r}')d\mathbf{r}'\right\rangle \\ &= \int_\Omega\int_\Omega \langle R(\mathbf{r},t)R(\mathbf{r}',t)\rangle\psi_m(\mathbf{r})\psi_n(\mathbf{r}')d\mathbf{r}d\mathbf{r}' \\ &= \int_\Omega\int_\Omega C(\mathbf{r},\mathbf{r}')\psi_m(\mathbf{r})\psi_n(\mathbf{r}')d\mathbf{r}d\mathbf{r}' \\ &= \int_\Omega \lambda_m\psi_m(\mathbf{r}')\psi_n(\mathbf{r}')d\mathbf{r}' \\ &= \lambda_m\delta_{mn},\end{aligned} \tag{2.32}$$

where δ_{mn} is the Kroneker delta that is equal to 1 when $m=n$ and 0 otherwise, λ_m is the eigenvalue corresponding to the m'th EOF, and $\langle\bullet\rangle$ denotes the temporal mean. The completeness and orthogonality properties of the EOFs lead to the expansion of the precipitation anomaly field:

$$R(\mathbf{r},t)=\sum_{m=1}^{\infty} R_m(t)\psi_m(\mathbf{r})\,. \tag{2.33}$$

In this study, we use the observed station data in (2.31) to calculate the PCs $R_m(t)$. As we know, the station data are discrete and incomplete, so we have to numerically compute the integral (2.31) through the discrete form

$$\hat{R}_m(t)=\sum_{j=1}^{N}\tilde{R}(r_j,t)\psi_m(r_j)w_j^{(m)}(t)\,, \tag{2.34}$$

where $\tilde{R}(r_j,t)$ are the anomaly data on the station nearest to grid point j, N is the number of grid points assigned data from their nearest station, and $w_j^{(m)}(t)$ are the weights assigned to each grid corresponding to mode number m and month t, which satisfy the condition

$$\sum_{i=1}^{N} w_i^{(m)} = A \tag{2.35}$$

and can be computed by using the linear equations described in Shen et al. (2004), and A is the area of Alberta. The detailed derivation of the weights equations will be described in the next subsection. When Mc is the number of modes we retained in the calculations, we have the reconstructed anomalies as follows:

$$\hat{R}(r_j,t) = \sum_{m=1}^{Mc} \hat{R}_m(t)\psi_m(r_j),\ j=1,2,\cdots,J. \tag{2.37}$$

2.3.4 Optimal Weights Calculation

In order to solve the optimal weights in (2.34), Shen et al. (2004) derived a linear equation by minimizing the total mean squared error (MSE) between the reconstructed data and the true data. The detailed derivation is addressed as follows.

The total MSE can be written as

$$\begin{aligned} E^2 &= \int_\Omega \left\langle \left(R(\mathbf{r},t) - \hat{R}(\mathbf{r},t) \right)^2 \right\rangle d\mathbf{r} \\ &= \int_\Omega \left\langle \left(\sum_{m=1}^{\infty} R_m(t)\psi_m(\mathbf{r}) - \sum_{m=1}^{Mc} \hat{R}_m(t)\psi_m(\mathbf{r}) \right)^2 \right\rangle d\mathbf{r} \end{aligned}$$

$$= \int_{\Omega} \left\langle \left(\sum_{m=1}^{Mc} (R_m(t) - \hat{R}_m(t)) \psi_m(\mathbf{r}) + \sum_{m=Mc+1}^{\infty} R_m(t) \psi_m(\mathbf{r}) \right)^2 \right\rangle d\mathbf{r}$$

$$= \left\langle \int_{\Omega} \left(\sum_{m=1}^{Mc} (R_m(t) - \hat{R}_m(t)) \psi_m(\mathbf{r}) + \sum_{m=Mc+1}^{\infty} R_m(t) \psi_m(\mathbf{r}) \right)^2 d\mathbf{r} \right\rangle$$

$$= \left\langle \left(\sum_{m=1}^{Mc} (R_m(t) - \hat{R}_m(t))^2 + \sum_{m=Mc+1}^{\infty} R_m^2(t) \right) \right\rangle$$

$$= \sum_{m=1}^{Mc} \left\langle (R_m(t) - \hat{R}_m(t))^2 \right\rangle + \sum_{m=Mc+1}^{\infty} \left\langle R_m^2(t) \right\rangle \qquad (2.37)$$

$$= \sum_{m=1}^{Mc} \varepsilon_{(m)}^2 + \sum_{m=Mc+1}^{\infty} \left\langle R_m^2(t) \right\rangle \ ,$$

where $\varepsilon_{(m)}^2 = \left\langle (R_m(t) - \hat{R}_m(t))^2 \right\rangle$.

We also have

$$\left\langle R_m(t) \hat{R}_m(t) \right\rangle = \left\langle \int_{\Omega} R(\mathbf{r},t) \psi_m(\mathbf{r}) d\mathbf{r} \sum_{i=1}^{N} \tilde{R}(r_i,t) \psi_m(r_i) w_i^{(m)} \right\rangle$$

$$= \left\langle \sum_{i=1}^{N} \int_{\Omega} R(\mathbf{r},t) \tilde{R}(r_i,t) \psi_m(\mathbf{r}) \psi_m(r_i) w_i^{(m)} d\mathbf{r} \right\rangle$$

$$= \left\langle \sum_{i=1}^{N} \int_{\Omega} R(\mathbf{r},t) \left(R(r_i,t) + E_i \right) \psi_m(\mathbf{r}) \psi_m(r_i) w_i^{(m)} d\mathbf{r} \right\rangle$$

$$= \sum_{i=1}^{N} \int_{\Omega} \left(\left\langle R(\mathbf{r},t) R(r_i,t) \right\rangle + \left\langle R(\mathbf{r},t) E_i \right\rangle \right) \psi_m(\mathbf{r}) \psi_m(r_i) w_i^{(m)} d\mathbf{r}$$

$$= \sum_{i=1}^{N} \int_{\Omega} C(\mathbf{r},r_i) \psi_m(\mathbf{r}) \psi_m(r_i) w_i^{(m)} d\mathbf{r}$$

$$= \sum_{i=1}^{N} \int_{\Omega} \left(\sum_{n=1}^{\infty} \lambda_n \psi_n(\mathbf{r}) \psi_n(r_i) \right) \psi_m(\mathbf{r}) \psi_m(r_i) w_i^{(m)} d\mathbf{r}$$

$$= \sum_{i=1}^{N} \sum_{n=1}^{\infty} \lambda_n \psi_n(\mathbf{r}_i) \psi_m(\mathbf{r}_i) w_i^{(m)} \int_{\Omega} \psi_m(\mathbf{r}) \psi_n(\mathbf{r}) d\mathbf{r}$$

$$= \sum_{i=1}^{N} \sum_{n=1}^{\infty} \lambda_n \psi_n(\mathbf{r}_i) \psi_m(\mathbf{r}_i) w_i^{(m)} \delta_{mn}, \tag{2.38}$$

and

$$\left\langle \hat{R}_m^2(t) \right\rangle = \left\langle \left(\sum_{j=1}^{N} \tilde{R}(\mathbf{r}_j, t) \psi_m(\mathbf{r}_j) w_j^{(m)} \right)^2 \right\rangle$$

$$= \left\langle \left(\sum_{j=1}^{N} \left(R(\mathbf{r}_j, t) + E_j \right) \psi_m(\mathbf{r}_j) w_j^{(m)} \right)^2 \right\rangle$$

$$= \left\langle \sum_{i=1}^{N} \left(R(\mathbf{r}_i, t) + E_i \right) \psi_m(\mathbf{r}_i) w_i^{(m)} \sum_{j=1}^{N} \left(R(\mathbf{r}_j, t) + E_j \right) \psi_m(\mathbf{r}_j) w_j^{(m)} \right\rangle$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\left\langle R(\mathbf{r}_i, t) R(\mathbf{r}_j, t) \right\rangle + \left\langle R(\mathbf{r}_i, t) E_j \right\rangle + \left\langle R(\mathbf{r}_j, t) E_i \right\rangle + \left\langle E_i E_j \right\rangle \right) \psi_m(\mathbf{r}_j) w_j^{(m)} \psi_m(\mathbf{r}_i) w_i^{(m)}$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{N} \left\langle R(\mathbf{r}_i, t) R(\mathbf{r}_j, t) \right\rangle \psi_m(\mathbf{r}_j) w_j^{(m)} \psi_m(\mathbf{r}_i) w_i^{(m)} + \sum_{i=1}^{N} \left\langle E_i^2 \right\rangle \left(\psi_m(\mathbf{r}_i) w_i^{(m)} \right)^2$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{N} C_{ij} \psi_m(\mathbf{r}_i) \psi_m(\mathbf{r}_j) w_i^{(m)} w_j^{(m)} + \sum_{i=1}^{N} \left\langle E_i^2 \right\rangle \left(\psi_m(\mathbf{r}_i) w_i^{(m)} \right)^2$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\sum_{n=1}^{\infty} \lambda_n \psi_n(\mathbf{r}_j) \psi_n(\mathbf{r}_j) \right) \psi_m(\mathbf{r}_i) \psi_m(\mathbf{r}_j) w_i^{(m)} w_j^{(m)} + \sum_{i=1}^{N} \left\langle E_i^2 \right\rangle \left(\psi_m(\mathbf{r}_i) w_i^{(m)} \right)^2, \tag{2.39}$$

where E_j is the random observational error at the location $\mathbf{r}_j$ and time t. It is impossible to compute the exact error E_j, but some error statistics, such as the error variance, can be estimated. The systematic errors are assumed to have been removed from the observed data, and hence the remaining random error E_j has

the following properties (Shen 2004):

$$\left\langle R(\mathbf{r},t)E_j \right\rangle = 0 \text{ and } \left\langle E_i E_j \right\rangle = 0, \ \forall\ i \neq j. \tag{2.40}$$

By combining (2.38) and (2.39), we have

$$\begin{aligned}
\varepsilon_{(m)}^2 &= \left\langle (R_m(t) - \hat{R}_m(t))^2 \right\rangle \\
&= \sum_{n=1}^{\infty} \lambda_n \delta_{mn} \delta_{mn} - 2\sum_{i=1}^{N}\sum_{n=1}^{\infty} \lambda_n \psi_n(r_i)\psi_m(r_i) w_i^{(m)} \delta_{mn} \\
&\quad + \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\sum_{n=1}^{\infty} \lambda_n \psi_n(r_i)\psi_n(r_j) \right) \psi_m(r_i)\psi_m(r_j) w_i^{(m)} w_j^{(m)} + \sum_{i=1}^{N} \left\langle E_i^2 \right\rangle \left(\psi_m(r_i) w_i^{(m)} \right)^2 \\
&= \sum_{n=1}^{\infty} \lambda_n \left[\delta_{mn}\delta_{mn} - 2\sum_{i=1}^{N} \psi_n(r_i)\psi_m(r_i) w_i^{(m)} \delta_{mn} \right. \\
&\quad \left. + \sum_{i=1}^{N}\sum_{j=1}^{N} \psi_n(r_i)\psi_m(r_i) w_i^{(m)} \psi_n(r_j)\psi_m(r_j) w_j^{(m)} \right] \\
&\quad + \sum_{i=1}^{N} \left\langle E_i^2 \right\rangle \left(\psi_m(r_i) w_i^{(m)} \right)^2 \\
&= \sum_{n=1}^{\infty} \lambda_n \left[\delta_{mn} - \sum_{i=1}^{N} \psi_n(r_i)\psi_m(r_i) w_i^{(m)} \right]^2 + \sum_{i=1}^{N} \left\langle E_i^2 \right\rangle \left(\psi_m(r_i) w_i^{(m)} \right)^2 .
\end{aligned} \tag{2.41}$$

Combining (2.32) and (2.37) leads to

$$E^2 = \sum_{m=1}^{Mc} \varepsilon_{(m)}^2 + \sum_{m=Mc+1}^{\infty} \lambda_m, \tag{2.42}$$

which means that minimizing $\varepsilon_{(m)}^2$ for each m is equivalent to minimizing the total MSE E^2.

In order to minimize each MSE $\varepsilon_{(m)}^2$ with constraint (2.35) for the weights, a Lagrange function is defined as

$$J_m = \varepsilon_{(m)}^2 + 2\Lambda_m \left(\sum_{i=1}^{N} w_i^{(m)} - A\right) , \tag{2.43}$$

where Λ_m is the Lagrange multiplier, whose unit is the square of the precipitation unit (mm). The critical point for the Lagrange function is determined by

$$\frac{\partial J_m}{\partial w_j^{(m)}} = 0 \quad \text{and} \quad \frac{\partial J_m}{\partial \Lambda_m} = 0 . \tag{2.44}$$

$\frac{\partial J_m}{\partial w_j^{(m)}} = 0$ (j=1,2,...,N) gives us

$$\frac{\partial \varepsilon_{(m)}^2}{\partial w_j^{(m)}} + \frac{\partial}{\partial w_j^{(m)}} \left[2\Lambda_m \left(\sum_{j=1}^{N} w_i^{(m)} - A\right) \right] = 0;$$

that is,

$$\frac{\partial}{\partial w_j^{(m)}} \left\{ \sum_{n=1}^{\infty} \lambda_n \left[\delta_{mn} - \sum_{i=1}^{N} \psi_n(r_i)\psi_m(r_i) w_i^{(m)} \right]^2 + \sum_{i=1}^{N} \left\langle E_i^2 \right\rangle \left(\psi_m(r_i) w_i^{(m)} \right)^2 \right\} + 2\Lambda_m$$

$$= 2\sum_{n=1}^{\infty} \lambda_n \left[\delta_{mn} - \sum_{i=1}^{N} \psi_n(r_i)\psi_m(r_i) w_i^{(m)} \right] \left[-\psi_n(r_j)\psi_m(r_j) \right] + 2\left\langle E_j^2 \right\rangle \psi_m^2(r_j) w_j^{(m)} + 2\Lambda_m$$

$$= -\sum_{n=1}^{\infty} \lambda_n \left[\delta_{mn} - \sum_{i=1}^{N} \psi_n(r_i)\psi_m(r_i) w_i^{(m)} \right] \psi_n(r_j)\psi_m(r_j) + \left\langle E_j^2 \right\rangle \psi_m^2(r_j) w_j^{(m)} + \Lambda_m$$

$$= -\lambda_m \psi_m^2(r_j) + \sum_{i=1}^{N} \left(\sum_{n=1}^{\infty} \lambda_n \psi_n(r_i)\psi_n(r_j) \right) \psi_m(r_i)\psi_m(r_j) w_i^{(m)} + \left\langle E_j^2 \right\rangle \psi_m^2(r_j) w_j^{(m)} + \Lambda_m$$

$$= -\lambda_m \psi_m^2(r_j) + \sum_{i=1}^{N} C_{ij} \psi_m(r_i)\psi_m(r_j) w_i^{(m)} + \left\langle E_j^2 \right\rangle \psi_m^2(r_j) w_j^{(m)} + \Lambda_m$$

$$= 0,$$

which is equivalent to

$$\sum_{i=1}^{N} C_{ij}\psi_m(r_i)\psi_m(r_j)w_i^{(m)} + \left\langle E_j^2 \right\rangle \psi_m^2(r_j)w_j^{(m)} + \Lambda_m = \lambda_m \psi_m^2(r_j). \quad (2.45)$$

As well $\frac{\partial J_m}{\partial \Lambda_m} = 0$ yields

$$\sum_{i=1}^{N} w_i^{(m)} = A, \quad (2.46)$$

where A is the area of the region being studied as mentioned before. The equations (2.45) and (2.46) are used to solve the optimal weights in the EOF interpolation procedure (2.34). As we said before, the random observational error E_i cannot be determined exactly, so the error variance $\left\langle E_i^2 \right\rangle$ in the equation (2.45) must be estimated in order to compute the optimal weights.

2.4 Interpolation Results

In this study, 52 years of Hybrid 1.0 output are used to generate the EOF patterns, so at most 52 non-zero eigenvalues are obtained for each month. In Figure 2.7, the percentage variance explained of the spatial EOF modes as a function of the mode number is shown. In January, the first, second, third, and fourth mode explain 47.56%, 8.72%, 7.81%, and 5.00% of the monthly precipitation variance, respectively, and together explain a total of 69.08% of the monthly precipitation variance. In July, the first, second, third, and fourth mode explain 25.88%, 20.49%, 10.13%, and 6.71% of the monthly precipitation variance, respectively, and together explain a total of 63.21% of the monthly precipitation variance. Finally, the first ten modes together explain a total of

84.27% and 81.62% of the monthly precipitation variance in January and July, respectively. This finding indicates that the first ten modes can reasonably reflect the anomaly field's physical patterns over Alberta without losing much variance. The computations will hence be reduced greatly.

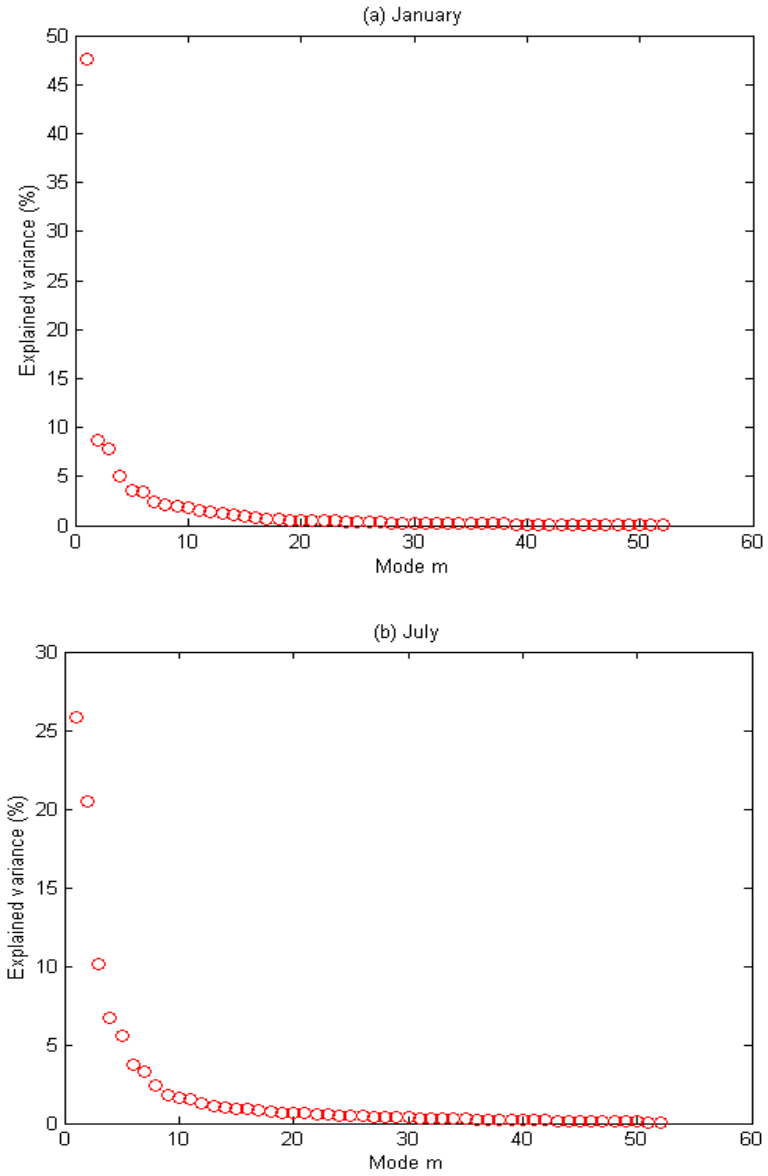


Figure 2.7. The percentage variance explained by the spatial EOF modes in (a) January, and (b) July, respectively.

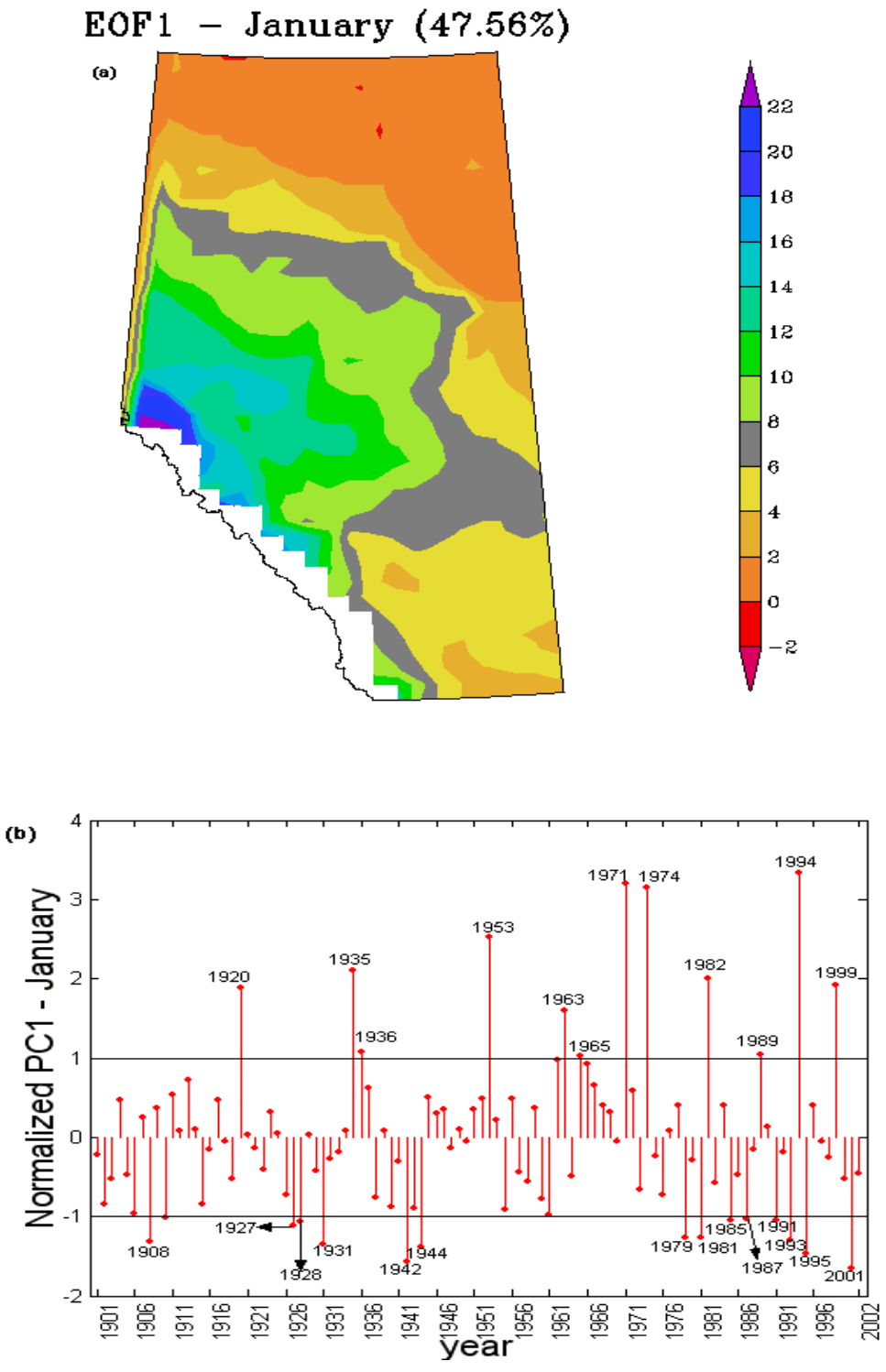
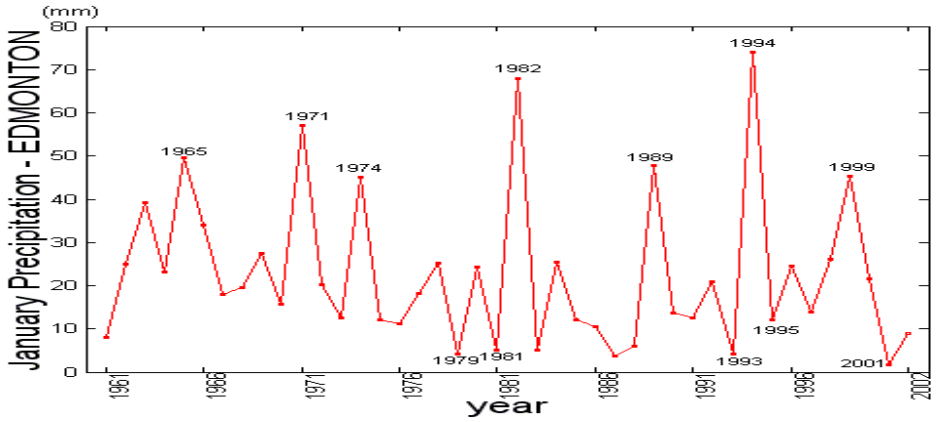
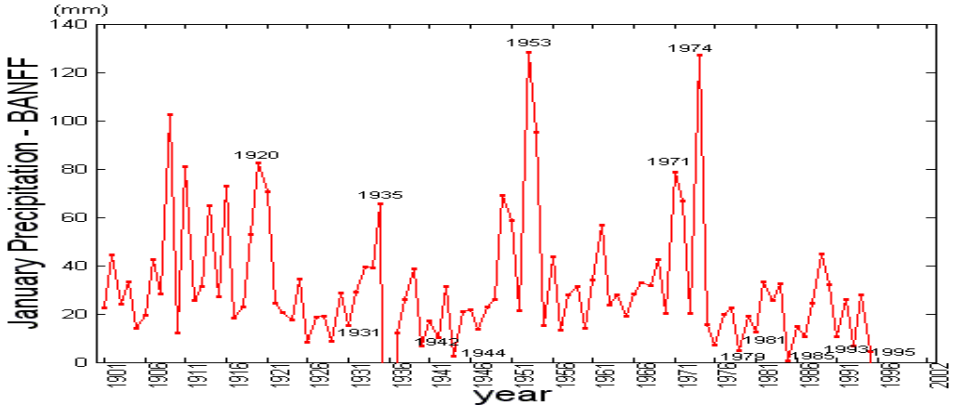


Figure 2.8. (a) The first mode of spatial EOF in January, and (b) its corresponding standardized PC, which is non-dimensional.

(a)



(b)



(c)

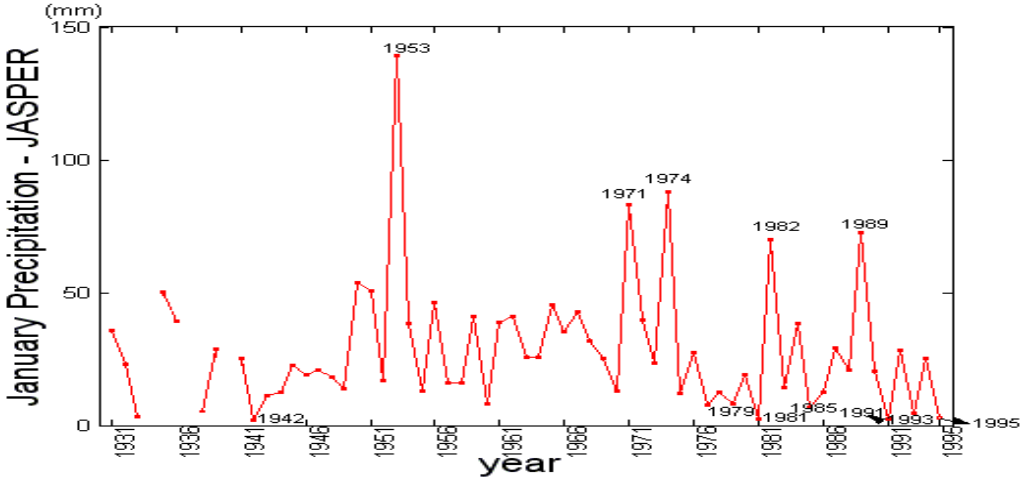


Figure 2.9. The observed January precipitation time series in (a) Edmonton, (b) Banff, and (c) Jasper.

Figure 2.8 displays the first spatial EOF, which explains 47.56% of the monthly precipitation variance and its corresponding PC1 in January. Here, PC1 is standardized by divided $\sqrt{\lambda_1}$ to have zero mean and unit variance. The PC1 reflects the time variation of the spatial pattern. The values associated with the first mode are almost positive everywhere in Alberta, so after multiplying the first mode by the PC1, the contribution of the first mode to the monthly precipitation without mean is positive (negative) when the PC1 value is positive (negative) and extreme when the PC1 values cross the horizontal lines ±1. For example, Edmonton, Banff, and Jasper have positive EOF1 values. The contribution of the first mode to the January drought in Edmonton, Banff, and Jasper is extreme when the PC1 values are greater than -1. Figure 2.9 shows the observed January precipitation time series in Edmonton, Banff, and Jasper. No observed precipitation data are available before 1961 in Edmonton and after 1995 in Banff. Continuous time series are available only in the time interval of 1938-1995 in Jasper. Figure 2.9 clearly shows that the PC1 captures the January drought event of 1979, 1981, 1993, 1995 and 2001 in the period of 1961-2002 in Edmonton, of 1931, 1942, 1944, 1979, 1981, 1985, 1985, 1993 and 1995 in the period of 1901-1995 in Banff, and of 1942, 1979, 1981, 1985, 1991, 1993 and 1995 in the time interval of 1938-1995 in Jasper.

The second spatial EOF, which explains 8.72% of the monthly precipitation variance, and the dimensionless PC2 in January are displayed in Figure 2.10. The values associated with the second mode are negative in the region along the Rocky Mountains and northern Alberta and positive in the other regions. Thus,

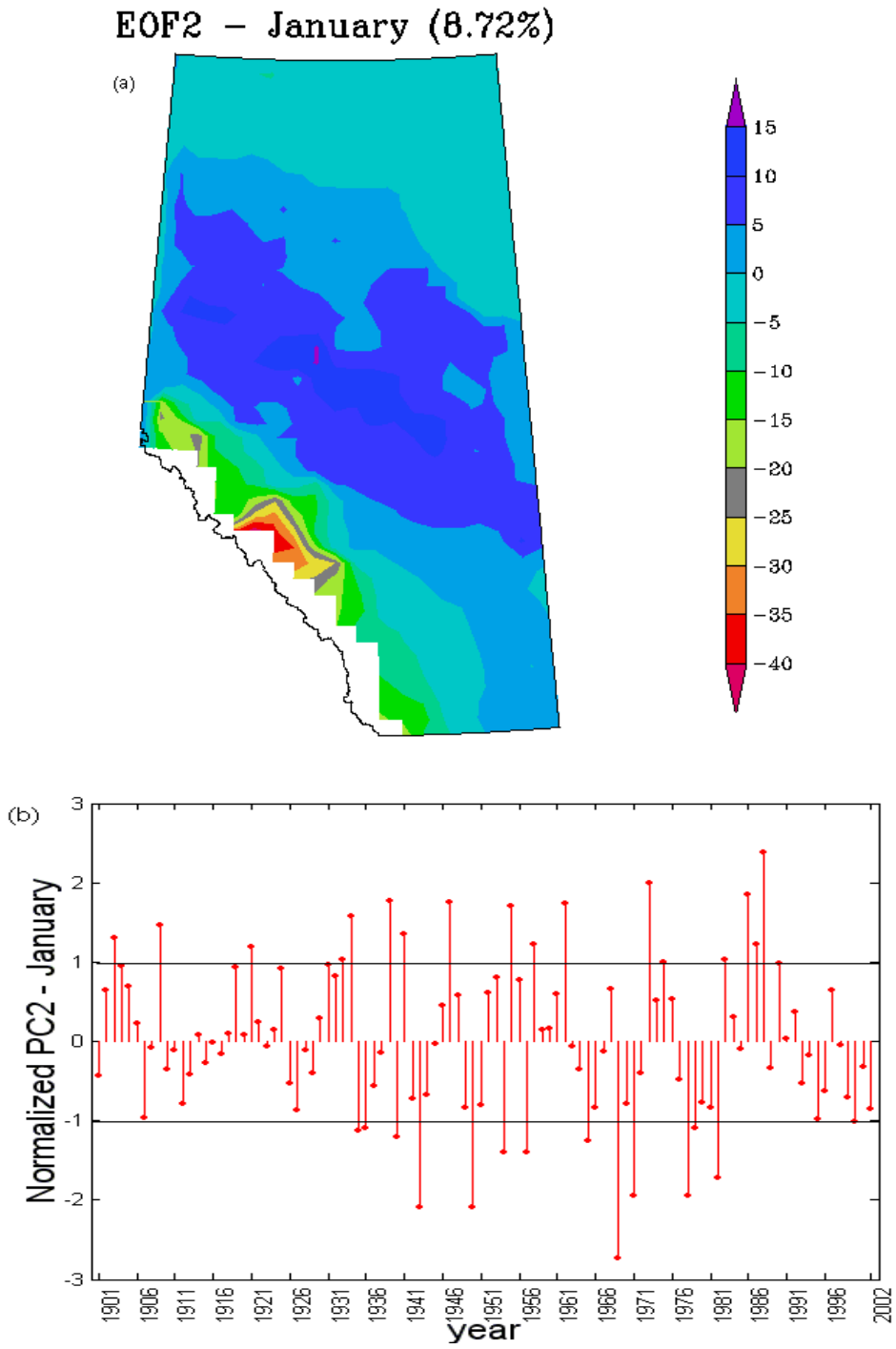


Figure 2.10. (a) The second mode of spatial EOF in January and (b) its corresponding standardized PC, which is non-dimensional.

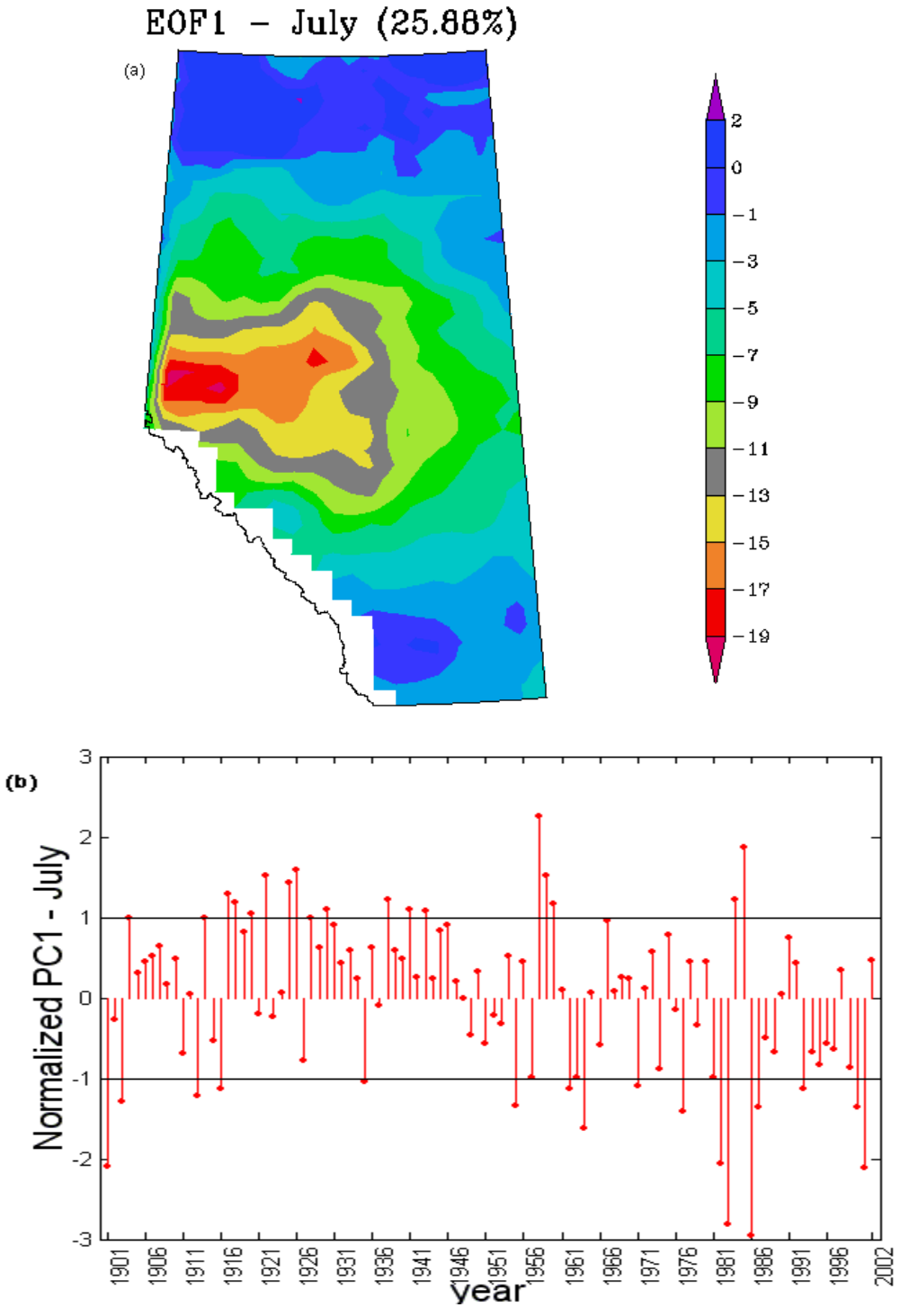


Figure 2.11. (a) The first mode of spatial EOF in July and (b) its corresponding standardized PC, which is non-dimensional.

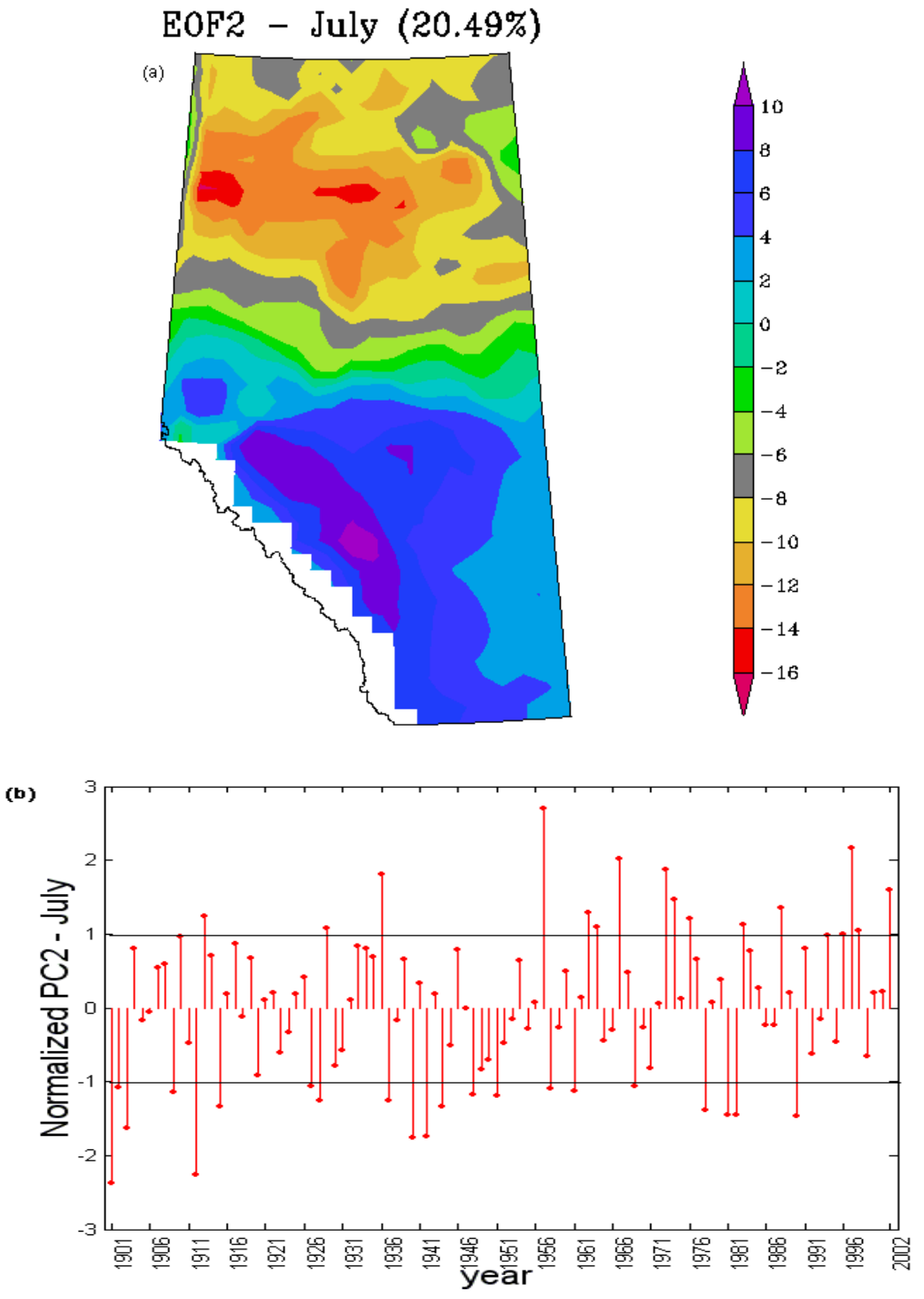


Figure 2.12. (a) The second mode of spatial EOF in July and (b) its corresponding standardized PC.

the contribution of the second mode to the January precipitation will be extreme when the PC2 values exceed 1 or -1, but opposite in the above two regions.

Figures 2.11 and 2.12 show the first two EOFs and their corresponding PCs in July. The PCs are also standardized to have zero mean and unit variance. The regions with larger absolute values of EOF1 and EOF2 are very sensitive to the variations of the PC1 values and PC2 values, respectively. The central Alberta has negative EOF1 values with large absolute values. Thus, the contribution of the first mode to the July drought event (or storm event) is relatively high when the PC1 values exceed +1 (or -1) in this region. Northern Alberta has negative EOF2 values with large absolute values, and the region along the Rocky Mountains has large positive EOF2 values, which indicate that the second mode contributes highly to the July drought event (or storm event) when the PC2 values cross the horizontal lines +1 (or -1) in northern Alberta, and has the opposite contribution in the Rocky Mountain region.

The geographical variability is relatively large across Alberta, so the Alberta climate varies considerably from region to region: Precipitation is generally highest along the mountains and in west central Alberta due to the combined effects of frontal and topographic precipitation, and generally lowest in the southeastern Alberta. The climate change over time also differs from region to region in Alberta. The difference between the average annual precipitation (mm) in the time intervals of 1913-1942 and 1973-2002 is displayed in Figure 2.13. The values in the figure were obtained by the 1973-2002 average annual precipitation minus the 1913-1942 average annual precipitation. The figure shows that, over 60

years, the increase of annual precipitation over northern Alberta and central Alberta is higher than that over southern Alberta. For example, an increase of 60-70 mm of annual precipitation in Edmonton, 20-30 mm in Calgary, 10-20 mm in Medicine Hat, and even over 100 mm in some areas in northern Alberta occurred over 60 years.

The weights are involved in the EOF-based linear interpolation and they have a key impact on the accuracy of the interpolation as shown in equation (2.34): weights with large absolute values would increase the effect of the corresponding observed station data regardless of the negative weights or positive weights, and vice versa. The relatively large (small) change of weight values would result in the relatively large change (small) of reconstructed precipitation values. That is, the accuracy of EOF-based interpolation rests with the values of weights. The needs of optimal weights thus arise, and the equations (2.45) and (2.46) are derived to compute the weights. There is an undetermined value, the value of data error variance $\left\langle E_j^2 \right\rangle$, in calculating the optimal weights. In practice, we have to choose a reasonable value of $\left\langle E_j^2 \right\rangle$ in order to get an optimal weight then the accurate interpolation results. However, in a specific range of $\left\langle E_j^2 \right\rangle$ values such as from 0.001 to 10, the optimal weight values are almost unchanged with changing $\left\langle E_j^2 \right\rangle$ values (the detailed analysis of this result is in Chapter 5.). So both the optimal weights and the interpolation results are insensitive to the choice of $\left\langle E_j^2 \right\rangle$ in this range. In our calculation, the values of 0.5, 1.0, and 0.9 are assigned to

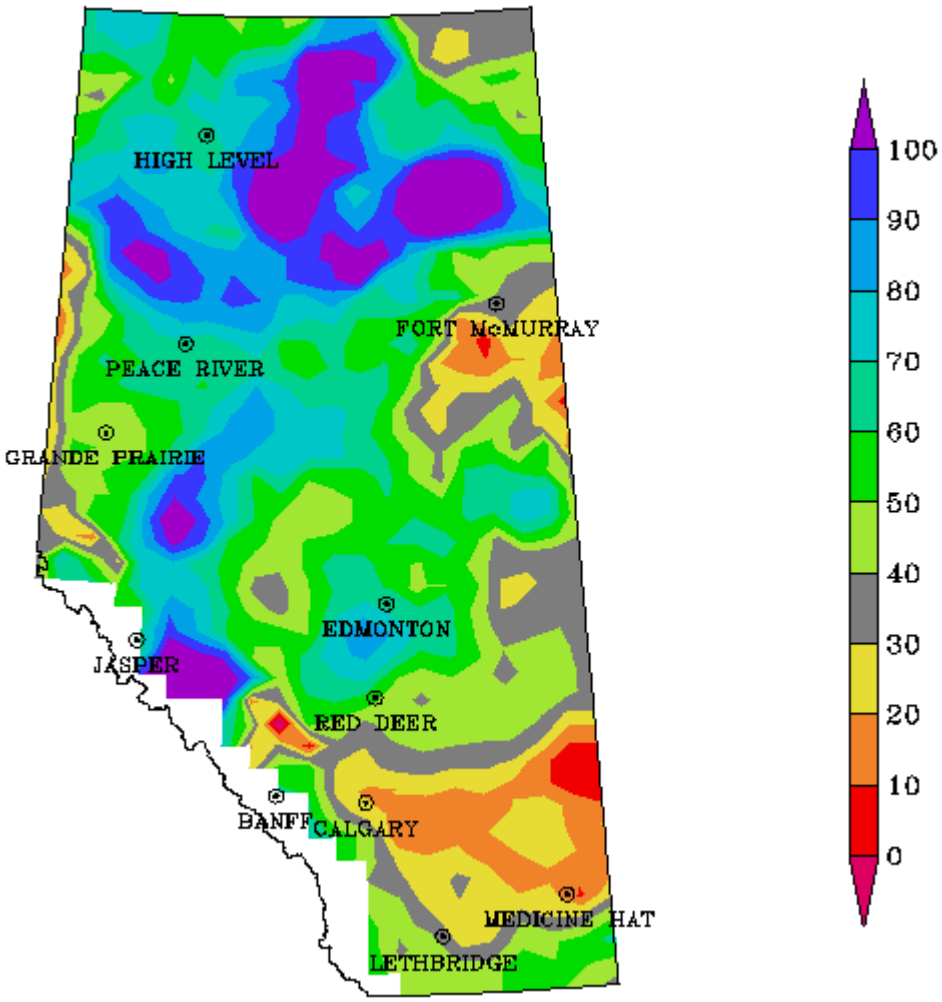


Figure 2.13. The difference between 1973-2002 mean and 1913-1942 mean for annual precipitation (mm).

$\left\langle E_j^2 \right\rangle$ and the interpolation results reveal the insensitivity to $\left\langle E_j^2 \right\rangle$ (detailed analysis, examples and figures are in Chapter 5).

2.5 The Accuracy of the Interpolated Monthly Data: Cross-Validation (CV)

The most effective method now commonly used to assess the error of climate data estimation is cross-validation (Cressie, 1993). The procedure compares the estimated data for a point with the observed station data at that point. Of course, the station data are withheld from the estimation. The data from other stations are interpolated into the station location. The statistics for the difference, or errors, between the true data and the interpolated data are used to evaluate the accuracy of the interpolation scheme.

To evaluate the interpolation accuracy of the spectral method, three types of errors were computed:

(1). Root mean square errors (RMSE)

$$\mathrm{RMSE}=\left[\frac{1}{K}\sum_{t=1}^{K}(X_{true}(t)-X_{estimate}(t))^2\right]^{1/2}, \qquad (2.47)$$

(2). Mean absolute errors (MAE)

$$\mathrm{MAE}=\frac{1}{K}\sum_{t=1}^{K}\left|X_{true}(t)-X_{estimate}(t)\right|, \qquad (2.48)$$

(3). Mean biased errors (MBE)

$$\mathrm{MBE}=\frac{1}{K}\sum_{t=1}^{K}\left[X_{true}(t)-X_{estimate}(t)\right], \qquad (2.49)$$

where K is the number of months used for cross-validation studies, t is the time with a unit of one month, and X denotes a climate parameter (precipitation in this study) at a cross-validation location.

The RMSE and MAE measure the variation of the interpolated results away from the real data and are a kind of measure of the closeness of the fit. The MBE indicates whether the interpolation is biased toward or against one side. Thus, these three measures are routinely used in checking the quality of interpolated results in statistics.

We also consider the CV errors of the IDW method and compare them with those of the EOF - interpolation. Cross-validation stations are selected based on the principles of (i) long-term data stream and (ii) even distribution around the province, particularly in the north-south orientation. Eight stations are chosen here for Alberta. The eight cross-validation sites are listed in Table 2.1 in the order from south to north.

Table 2.2 shows the cross-validation results of the IDW method and the EOF interpolation method for the period of 1901-2002. Comparing the RMSE, MAE and MBE values of the two methods indicates that the EOF-interpolated monthly data have a smaller error than that of the IDW method, especially in the high elevation regions and northern Alberta region such as Banff, Jasper, and High level A. The reason is that Hybrid 2.0 takes the elevation into account by extracting the EOF patterns from 1951-2002 Hybrid 1.0 gridded output when many stations available in the mountain and high elevation regions.

In Section 2.1, Figure 2.1 indicates only a small number of stations in

Table 2.1. Eight Alberta cross-validation sites

Station Name	Station ID	Latitude	Longitude	Elevation [m]
Medicine Hat A	3034480	50°01′	110°43′	716.90
Calgary Int'l A	3031093	51°07′	114°01′	1084.10
Banff	3050520	51°11′	115°34′	1383.70
Jasper	3053520	52°53′	118°04′	1062.20
Edmonton Int'l A	3012205	53°19′	113°35′	723.30
Beaverlodge CDA	3070560	55°12′	119°24′	744.90
Peace River A	3075040	56°14′	117°27′	570.90
High Level A	3073146	58°37′	117°10′	338.30

Table 2.2. Monthly precipitation estimation errors accessed by cross-validation for eight long-term stations in Alberta from south to north (units: millimeters). The values in the round brackets are the cross validation errors for the IDW method, and the preceding values for the EOF method.

Station Name	RMSE	MAE	MBE
Medicine Hat A	12.84 (13.52)	8.04 (8.49)	0.97 (-0.32)
Calgary Int'l A	14.94 (15.61)	9.37 (9.94)	-0.93 (-1.66)
Banff	18.93 (22.65)	13.68 (15.86)	-5.71 (-7.10)
Jasper	19.68 (32.78)	14.20 (21.43)	-9.40 (-13.49)
Edmonton Int'l A	9.00 (9.45)	6.15 (6.13)	-1.01 (-1.51)
Beaverlodge CDA	12.49 (14.45)	8.73 (9.44)	-1.23 (-1.34)
Peace River A	8.25 (12.84)	5.84 (9.06)	-1.16 (-4.18)
High Level A	12.65 (18.57)	8.75 (12.69)	-5.97 (-6.78)

northern Alberta and west central Alberta in the earlier period of 1901-1912. This factor would cause the IDW method to obtain inaccurate results over those regions and to seriously underestimate the precipitation over the high-elevation regions. In order to show the improvement of Hybrid 2.0, Figure 2.14 shows the difference between the July precipitation interpolated by using the EOF-interpolation method and the IDW method in the periods of 1901-1912 and 1973-2002, respectively. The values in the figure were obtained by using the EOF-interpolated data minus the IDW-interpolated data. Fig 2.16 (a) indicates a difference of 10-50 mm in northern Alberta and west central Alberta because the IDW method underestimates the precipitation in those regions because of the sparsity of precipitation stations in the period of 1901-1912, and the EOF-interpolation method overcomes this problem. Fig 2.16 (a) also indicates a very small difference in southern Alberta in the period of 1901-1912 because the stations were relatively dense in this region. Fig 2.16 (b) indicates a difference of near zero precipitation in most part of Alberta since the dense stations in the period of 1973-2002 enabled both methods to obtain more accurate interpolation results.

2.6 Conclusion

This chapter used the EOF-interpolation method to reconstruct the monthly precipitation on the grids over Alberta, and our Hybrid 2.0 method, not the IDW method, to reconstruct the monthly precipitation from the Hybrid 1.0 method. The

purpose was to overcome the problem of the IDW method, which underestimates the precipitation in the high-elevation regions and northern Alberta because of scarcity of precipitation stations in those areas in the earlier period of the last century. The comparison of the results and the error analysis for both methods reveals that the EOF-interpolation method for reconstructing the monthly precipitation overcame the underestimation problem of the IDW method to a certain extent and obtained a more accurate result than that of the IDW method. As discussed above, the iterative nature of our EOF approach is using the Hybrid 1.0 data - ABCclim 1.0 to build the basis functions (EOFs) for the Hybrid 2.0 gridding. This is also why we named our developed method as Hybrid 2.0. The 1951 – 2002 Hybrid 1.0 data was selected to calculate the EOFs as there were so dense precipitation station distribution in this time interval that the gridded Hybrid 1.0 data is accurate and is reliable to represent the precipitation patterns over Alberta. The above comparison was thus obtained.

However, some problems still need to be discussed, such as the problem of determining the number of modes to be retained in the EOF analysis, and the problem of estimating the error variance $\left\langle E_i^2 \right\rangle$ in the linear equation (2.47) for computing the optimal weights. These problems will be discussed in Chapter 5.

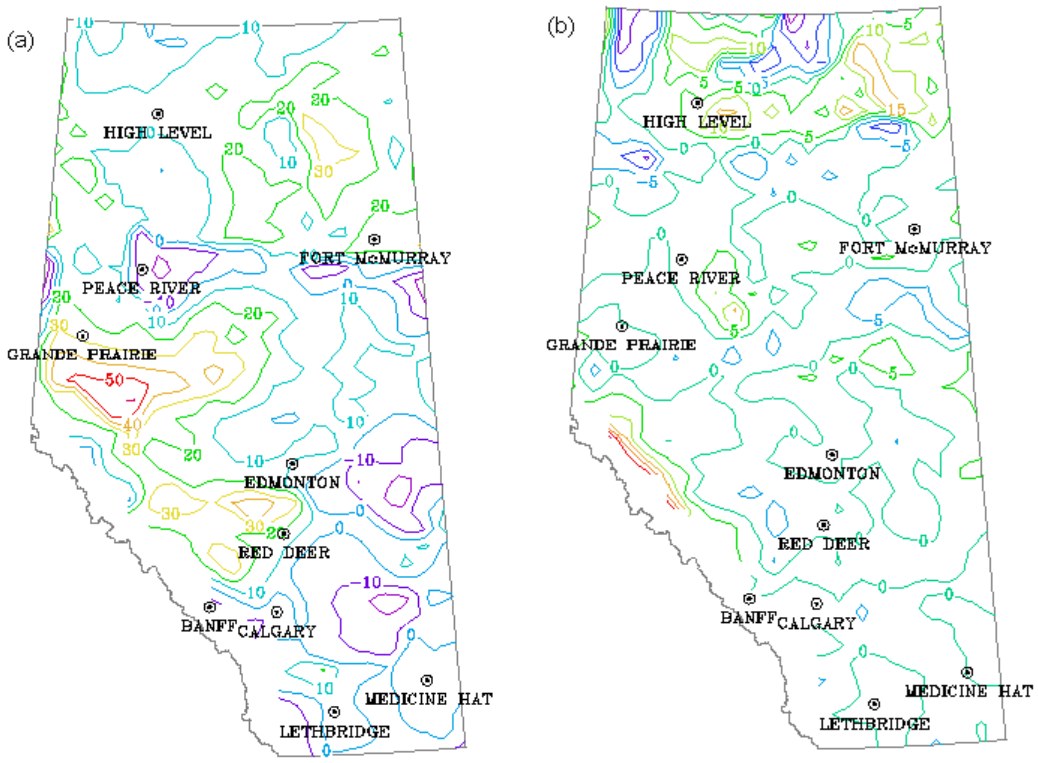


Figure 2.14 The difference the interpolated July precipitation (mm) by EOF method minus the IDW method in the periods of (a) 1901-1912 and (b) 1973-2002.

Chapter 3

Hybrid 2.0 Interpolation for Daily Precipitation Data

3.1 Introduction

In the previous Chapter, we reviewed the Hybrid 1.0 method in detail and discussed its problem. Then the EOF-based linear interpolation method was developed in order to overcome the shortcoming of Hybrid 1.0. Hybrid 2.0, the next generation of Hybrid 1.0 method, is continued to be completed by using a nearest-station assignment regression for redistributing the monthly total onto days. In Chapter 2, we interpolated the monthly data onto the 0.25 by 0.5 degrees latitude and longitude grids over Alberta by using the EOF interpolation method, This chapter will use nearest-station assignment regression to distribute the monthly total obtained in Chapter 2 onto days to obtain the Hybrid 2.0 daily data over Alberta with applications of Hybrid 2.0 method on a more fine resolution grid. Various applications of these gridded data will be explored in Chapter 4, including the assessment of agroclimate changes, input climate data for soil quality modeling, and evaluation of drought indices. These data will be useful to Alberta Agriculture, Food and Rural Development, Agriculture and Agri-Food Canada, and other governmental and private agencies. The monthly data

interpolated onto the grids of 0.25 by 0.5 degrees latitude and longitude from January 1901 to December 2002 used in this chapter are calculated from the EOF interpolation method discussed in Chapter 2. The daily station data from January 1, 1901 to December 31, 2002 used in this chapter are the same as the data used in Chapter 2.

The organization of this chapter is as follows. Section 3.2 describes the nearest-station assignment method for interpolating daily precipitation data. Section 3.3 displays the results of the Hybrid 2.0 interpolation onto the 0.25 by 0.5 degrees latitude and longitude grids and assesses its error. Section 3.4 conducts the Hybrid 2.0 interpolation method for the 0.1 by 0.2 degrees latitude and longitude grids. Section 3.5 compares the error with that of other interpolation methods. Section 3.6 presents our conclusions and discussion.

3.2 The Method of Nearest-Station Assignment

The interpolation method used in this study is Hybrid 2.0, the next generation of Hybrid 1.0 (Shen et al., 2001). This method consists of two procedures: interpolating the monthly total first and then redistributing the monthly total to each individual day of the month. The monthly total was obtained by using EOF interpolation in Chapter 2, and here the redistribution will be realized by using the nearest-station assignment method discussed in Chapter 1 (see the formula (1.4)).

As one of characteristic of daily precipitation data, precipitation frequency is very important for the decision makers in Agriculture. The precipitation

frequency is calculated by the average of the number of days with precipitation in a month in the discussed period. So the accuracy of the estimated precipitation frequency is an important factor to evaluate the interpolation accuracy and will be calculated for Hybrid 2.0 in this chapter.

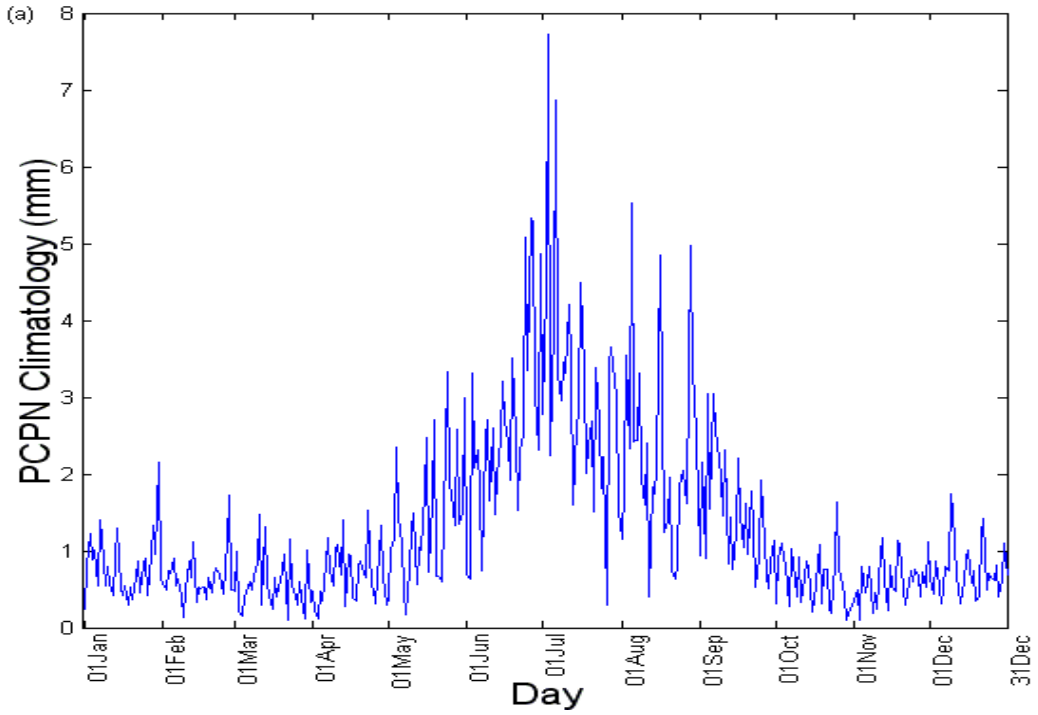
3.3 Hybrid 2.0 Results and Error Analysis

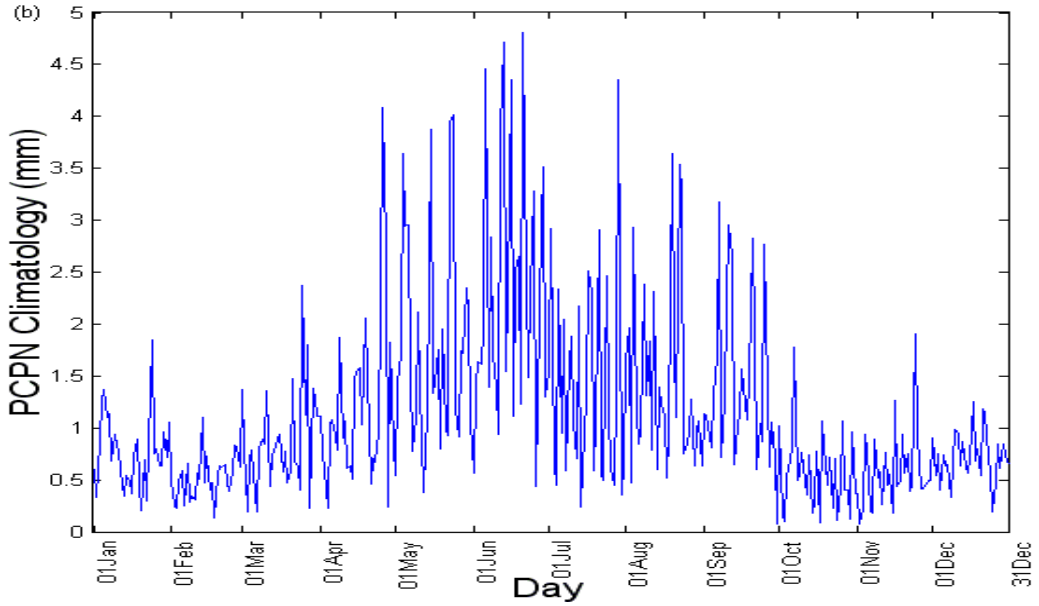
3.3.1 Hybrid 2.0 daily data interpolated onto the grids of 0.25 by 0.5 degrees latitude and longitude

The daily data are interpolated onto a total of 808 grid points with latitude interval 0.25° and longitude interval 0.5° over Alberta. The interpolation results for the daily precipitation are stored in one file with the format (I4, 2I3, I8, F6.1):

1901 1 1 10001 0.0
1901 1 1 10002 0.0
……
1901 1 1 10808 0.0
1901 1 2 10001 0.0
……
1901 1 2 10808 1.4
……
……
2002 12 31 10001 0.0
……
2002 12 31 10808 1.1

where the first column is the year number, the second one is the month number, the third one is the date number, the fourth one is the grid ID, and the last one is the precipitation amount on the day.


(a)
PCPN Climatology (mm)
0
1
2
3
4
5
6
7
8
01Jan
01Feb
01Mar
01Apr
01May
01Jun
01Jul
01Aug
01Sep
01Oct
01Nov
01Dec
31Dec
Day


(b)
PCPN Climatology (mm)
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
01Jan
01Feb
01Mar
01Apr
01May
01Jun
01Jul
01Aug
01Sep
01Oct
01Nov
01Dec
31Dec
Day

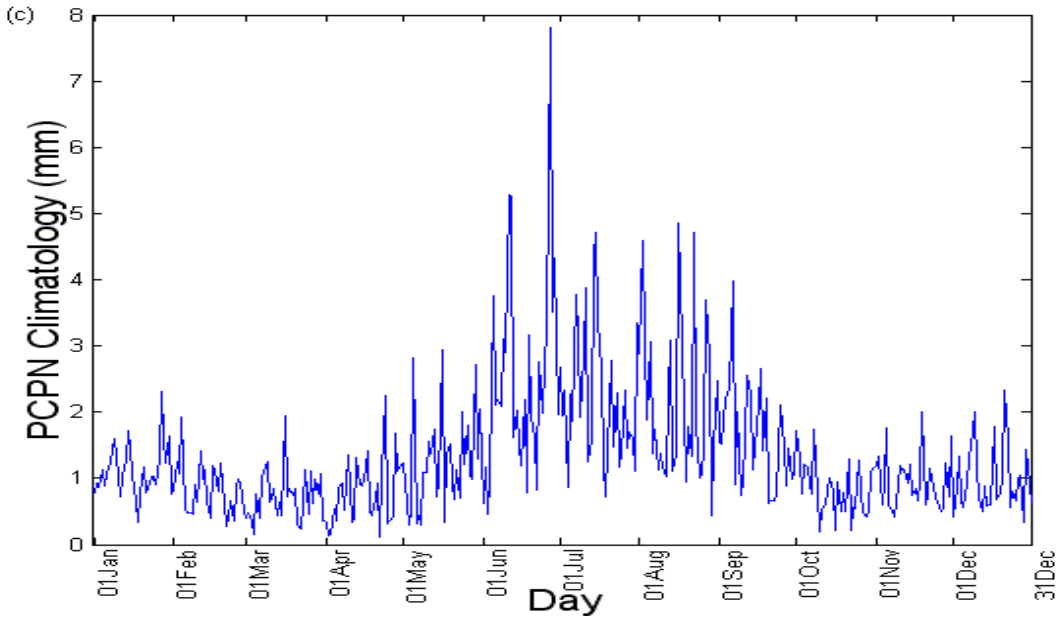


Figure 3.1 The 30-year (1961-1990) mean for daily precipitation at (a) Edmonton, (b) Lethbridge and (c) Grande Prairie.

Figure 3.1 shows the 1961-1990 daily precipitation mean in Edmonton, Lethbridge and Grande Prairie, respectively. The figure indicates the seasonal fluctuation in the daily precipitation for these three cities. The highest daily precipitation in Edmonton and Grande Prairie is higher than that in Lethbridge since southern Alberta is generally drier than central Alberta. Figure 3.2 displays the annual total precipitation normal in the time interval between 1961 and 1990. The highest annual precipitation amount of more than 600 mm occurs in the Rocky Mountain region, and the lowest annual precipitation amount of less than 350 mm occurs in the south east Alberta region. As well, northern Alberta is drier

than central Alberta. Among the major cities in Alberta, Edmonton has an annual total precipitation of 500 – 550 mm, a little higher than Calgary. Lethbridge is a little drier than Calgary. Medicine Hat is the driest city with less than 350 mm annual precipitation. Figure 3.3 shows the May 1 to August 31 total precipitation normal from 1961 to 1990 with less than 200 mm in the driest areas and more than 325 mm over the high-elevation regions. In general, this period has about 60 percent of the total annual precipitation over Alberta. Figure 3.2 indicates that this period's total precipitation amount distribution over Alberta from lowest to highest is similar to that of the annual total precipitation amount distribution; i.e., the highest precipitation amount is in the central Alberta area and the mountain regions, and the driest area is south east Alberta and north Alberta. Figure 3.4 displays the eight-month period total precipitation normal from September 1 to April 30 in the time interval between 1961 and 1990. The precipitation in this period varies from less than 150 mm in the driest areas to more than 275 mm in the high-elevation regions. More than 275 mm precipitation occurs along the Rocky Mountain region and 150-175 mm and even less than 150 mm precipitation occur in the south east Alberta during this period. Figure 3.3 and Figure 3.4 indicate that less precipitation occurs during the September to April period than in the May to August period.

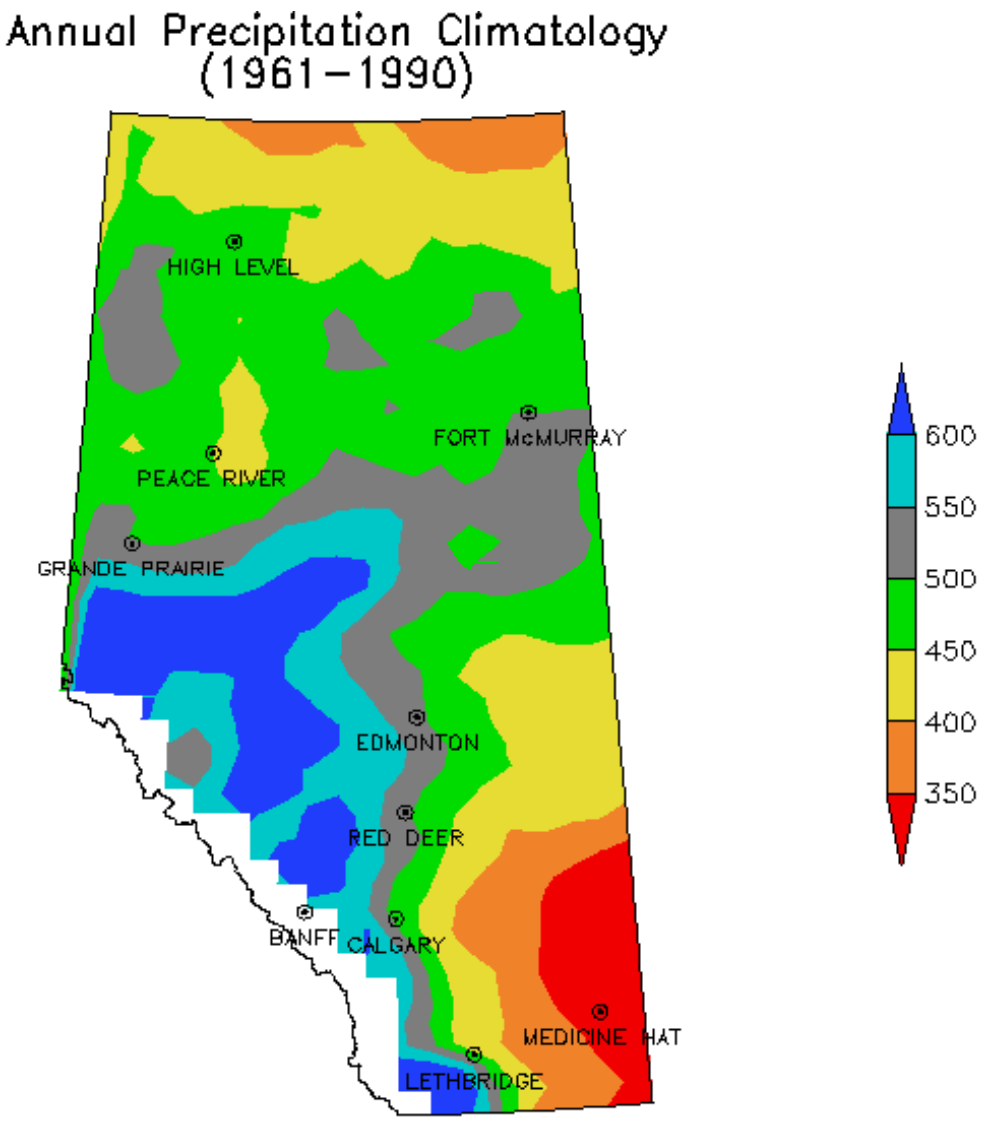


Figure 3.2. The annual total precipitation normals (mm) in the period of 1961 – 1990 onto the grids of 0.25-by-0.5 degrees lattitude and longitude over Alberta.

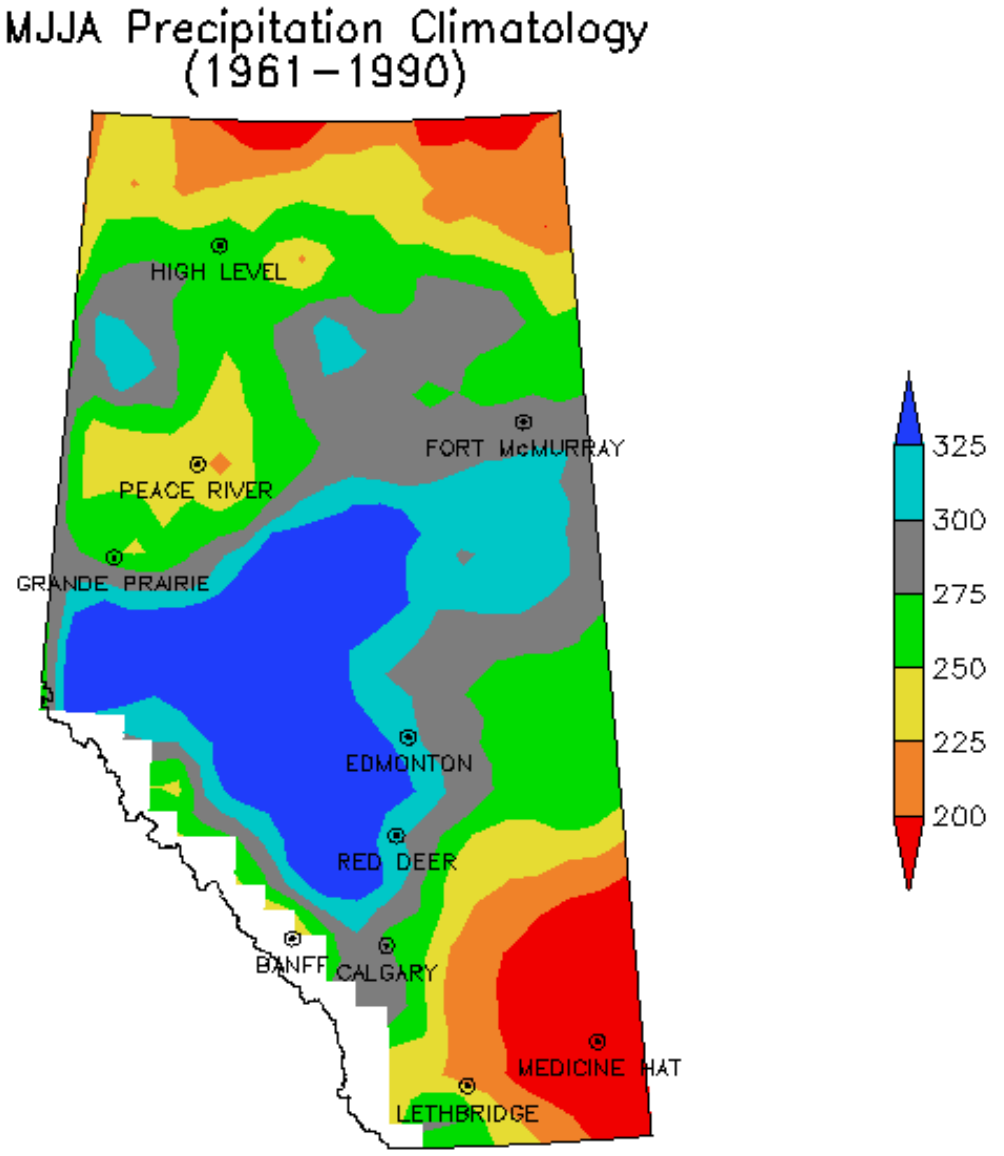


Figure 3.3. May 1 to August 31 total precipitation normals (mm) in the period of 1961 – 1990 onto the grids of 0.25-by-0.5 degrees lattitude and longitude over Alberta.

.

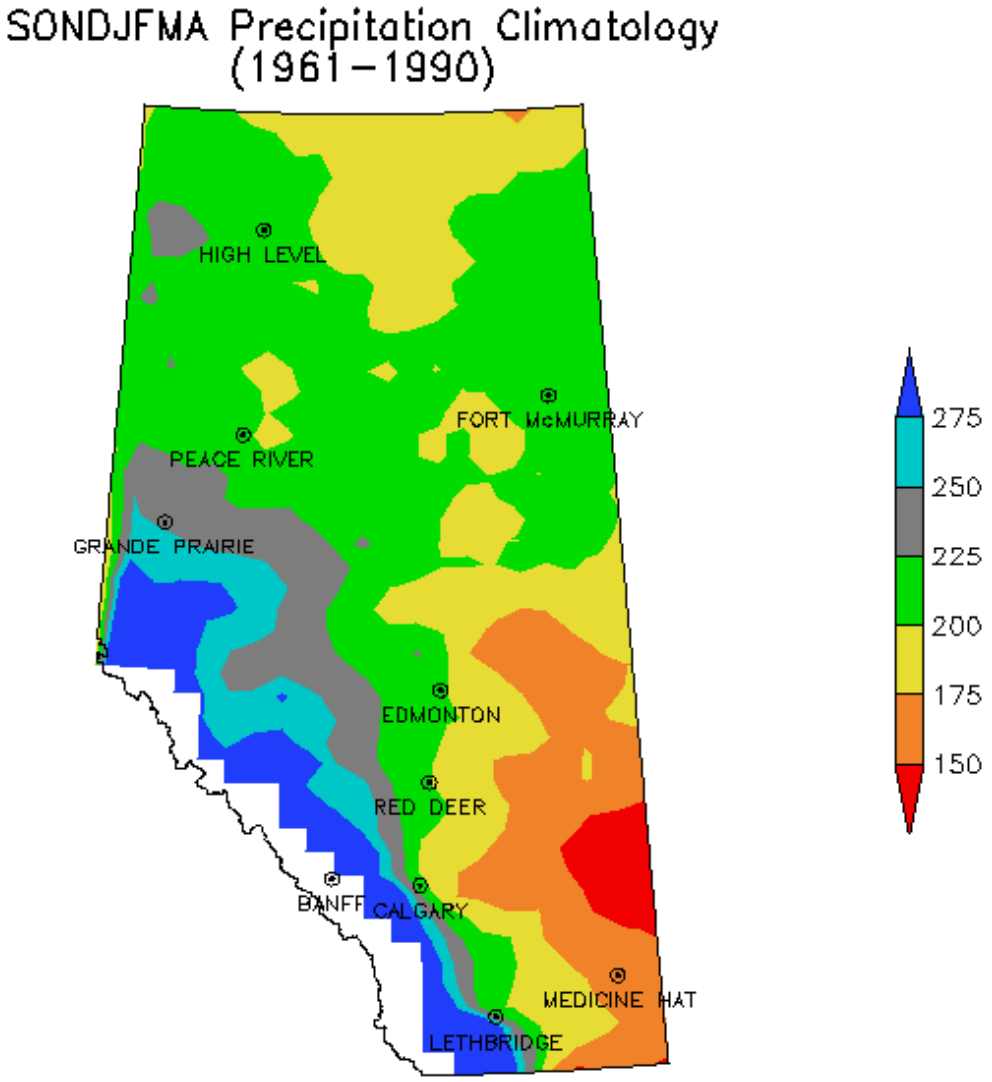


Figure 3.4. September 1 to April 31 total precipitation normals (mm) in the period of 1961 – 1990 onto the grids of 0.25-by-0.5 degrees lattitude and longitude over Alberta.

.

3.3.2 The accuracy of Hybrid 2.0 daily data: cross-validation (CV)

The previous CV error formulas in Chapter 2 are used here at the daily scale to assess the error of both Hybrid 2.0 and Hybrid 1.0 daily data, and the eight stations listed in Table 2.1 are selected as cross-validation stations. Table 3.1 shows the cross-validation results of the Hybrid 2.0 method and the Hybrid 1.0 method for the period of 1901-2002. Each target station is excluded from the interpolation. Comparing the RMSE, MAE and MBE values of the two methods indicates that the Hybrid 2.0 daily data have a smaller error than those of Hybrid 1.0. The RMSE, MAE and MBE values in the high-elevation regions, such as the Calgary Int'l A, Jasper and Banff stations, are higher than in the low-elevation areas. One reason for this difference is that the high-elevation regions receive more precipitation than the low-elevation regions; another reason is that fewer precipitation stations were located in the high-elevation areas than in the low-elevation areas in the earlier period, or the period of 1901 – 1912 (Figure 2.1).

Figure 3.5 shows the difference between the arithmetic average of the station daily precipitation and the areal average of the Hybrid 2.0 gridded daily data over Alberta in the period of January 1, 1901 – December 31, 2002. The different data are obtained by using the station data minus the gridded data. The figure indicates that most different data are close to zero; that is, the areal average of the H2.0 gridded daily data nearly agrees with the arithmetic average of the station daily precipitation. Because more precipitation stations were used in the Hybrid 2.0 interpolation, the difference in the period of 1901 – 2002 is smaller than in the earlier period.

Table 3.1. Daily precipitation estimation errors accessed by cross-validation for eight long-term stations in Alberta from south to north (units: millimeters). The values in the round brackets are the cross validation errors for the Hybrid 1.0 method, and the preceding values for the Hybrid 2.0 method.

Station Name	RMSE	MAE	MBE
Medicine Hat A	3.07 (3.14)	0.93 (0.95)	0.04 (-0.01)
Calgary Int'l A	3.54 (3.60)	1.13 (1.13)	-0.03 (-0.05)
Banff	3.77 (3.88)	1.46 (1.48)	-0.19 (-0.23)
Jasper	3.61 (4.04)	1.47 (1.59)	-0.31 (-0.45)
Edmonton Int'l A	2.98 (3.03)	1.00 (1.01)	-0.04 (-0.05)
Beaverlodge CDA	2.89 (3.96)	1.05 (1.37)	-0.04 (-1.04)
Peace River A	2.99 (3.15)	1.06 (1.13)	-0.03 (-0.13)
High Level A	2.99 (3.16)	1.13 (1.16)	-0.19 (-0.22)

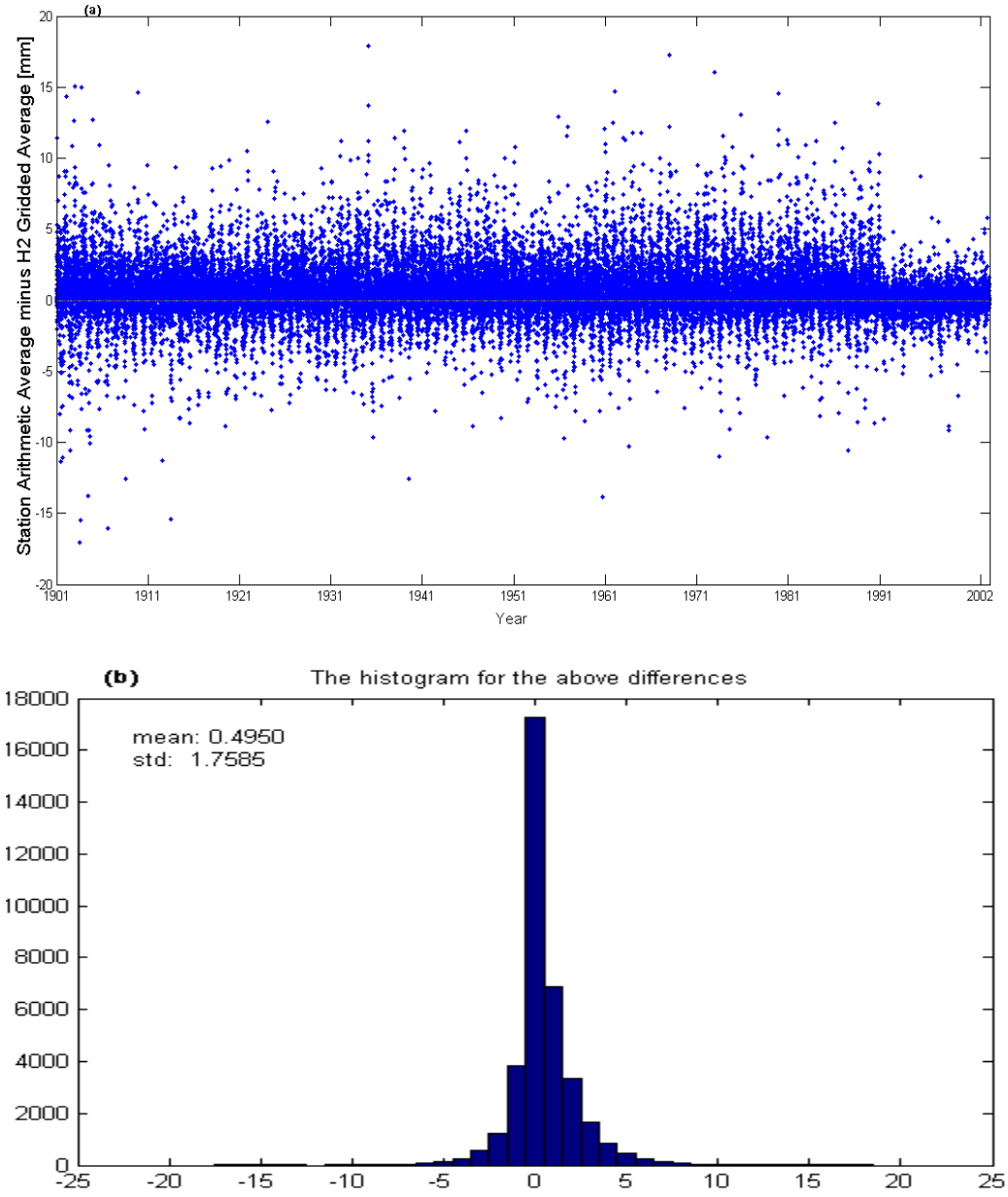


Figure 3.5. (a) The difference (mm) between the arithmetic average of station daily precipitation and the areal average of Hybrid 2.0 gridded daily data over Alberta in the period of January 1, 1901 – December 31, 2002, (b) Histogram of the differences.

3.4 The Hybrid 2.0 Results on a $0.1^\circ \times 0.2^\circ$ Grid

Many agriculture applications of historical climate data require the data to be available on the daily time scale and a spatial scale as fine as a township or a 10 km spatial grid. In this section, Hybrid 2.0 daily data will be computed on a $0.1^\circ \times 0.2^\circ$ grid (approximately a 10 km by 10 km grid) over Alberta. The base point of the of the grid is the northwest corner of Alberta: $(60^\circ N, 120^\circ W)$.

3.4.1 Data

Like the data used in Chapter 2, the daily station data from January 1, 1901 to December 31, 2002 obtained from MSC are used in this section. By using our Hybrid 2.0 method, these observed daily station data are to be interpolated onto the grid of 0.1 by 0.2 degrees latitude and longitude over Alberta. Figure 3.6 shows the distribution of the grid points over Alberta.

3.4.2 Interpolated monthly dataset and its accuracy

In this study, 52 years of Hybrid 1.0 data from 1951 to 2002 are used to extract the EOF patterns, so we can obtain at most 52 non-zero eigenvalues for each month. Figure 3.7 shows the explained variance of the spatial EOF modes as a function of the mode number. In January, the first, second, third, and fourth mode explain 47.84%, 9.30%, 7.61%, and 4.83% of the monthly precipitation variance, respectively, and together explain a total of 69.57% of the monthly

precipitation variance. In July, the first, second, third, and fourth modes explain 26.29%,

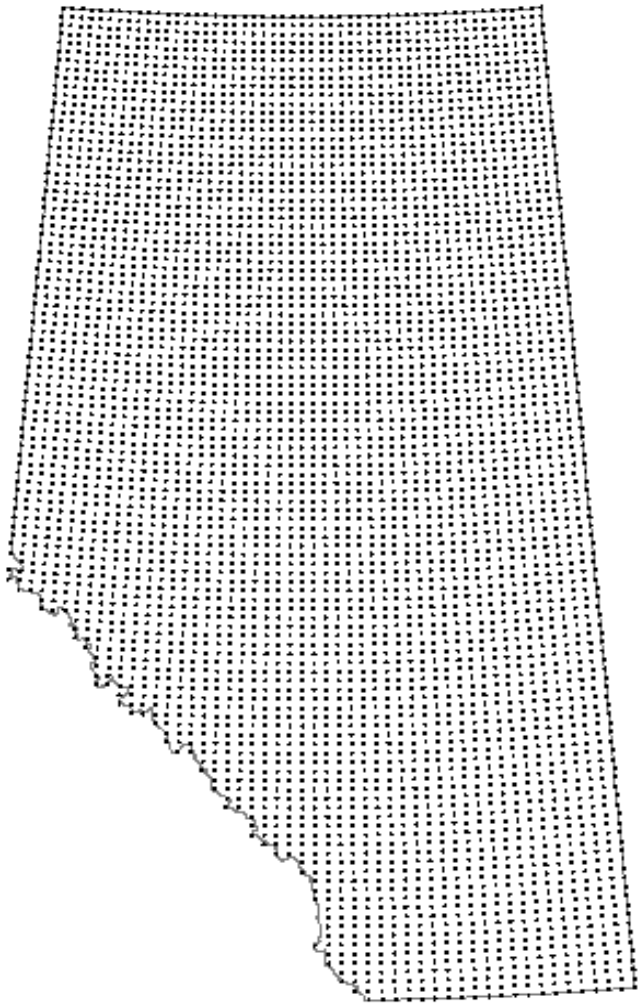


Figure 3.6. Distribution of the grid points (0.1-by-0.2degrees latitude and longitude) over Alberta.

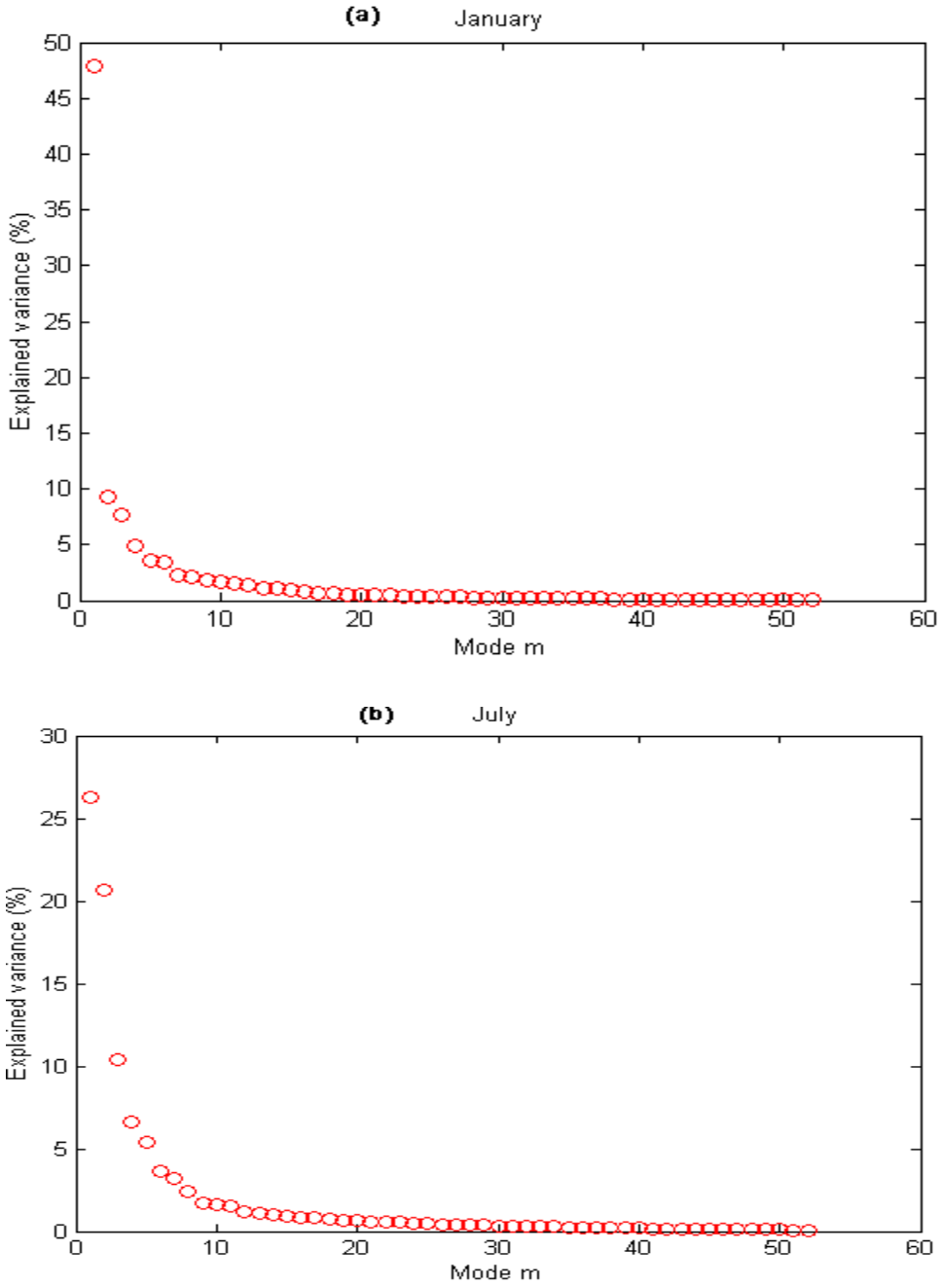


Figure 3.7. Percentage variance explained by the spatial EOF analysis in (a) January and (b) July, expressed as percentage of total variance.

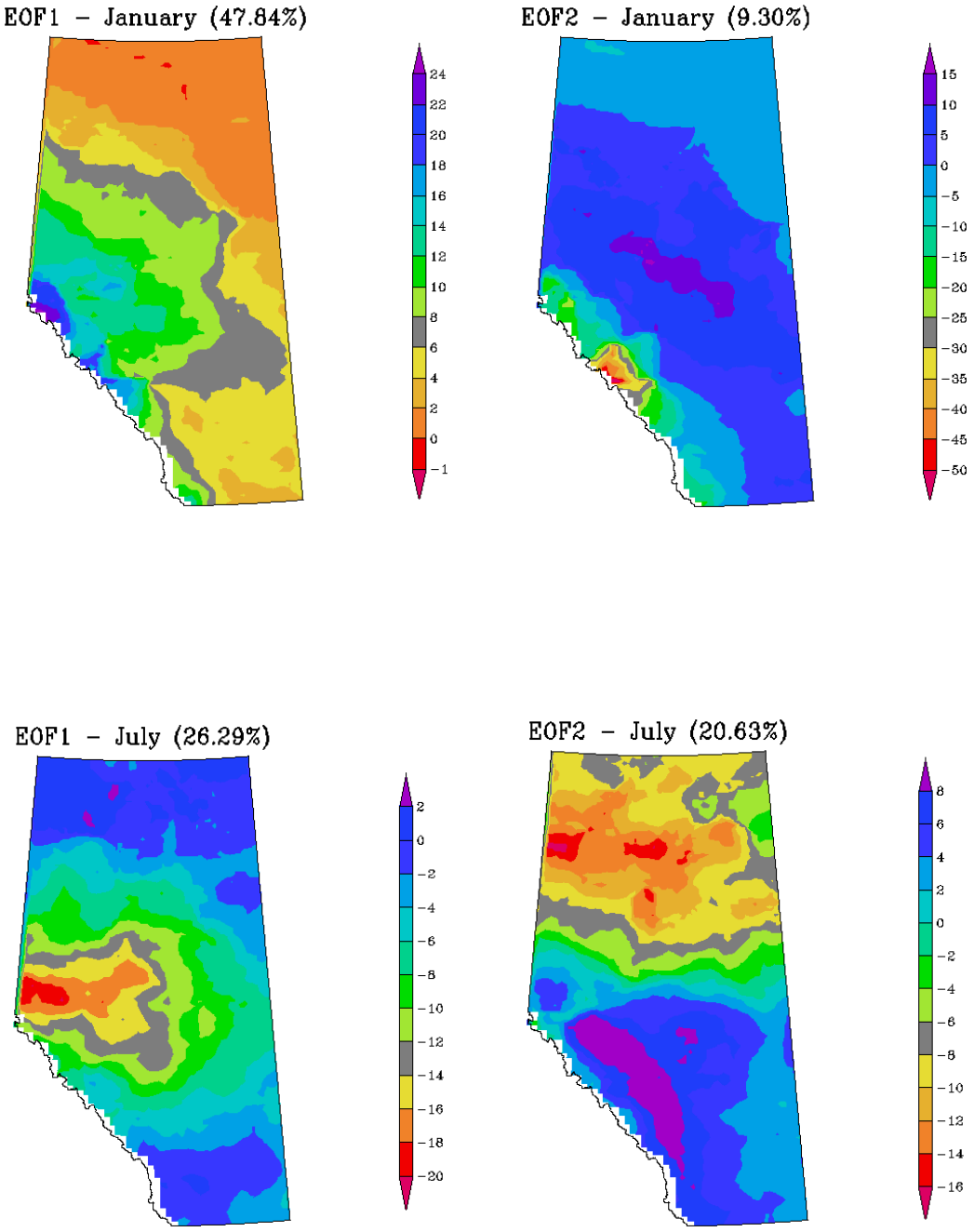


Figure 3.8. The first two modes of the spatial EOF in January and July.

20.63%, 10.38%, and 6.62% of the monthly precipitation variance, respectively, and together explain a total of 63.91% of the monthly precipitation variance. This figure indicates that the first eigenvalue seems separated from the rest in January, but for July and that the first three eigenvalues seem to be separated from the rest. In both months, the other eigenvalues appear to be generally continuous, and hence there is no clear natural cut for taking the EOF truncation.

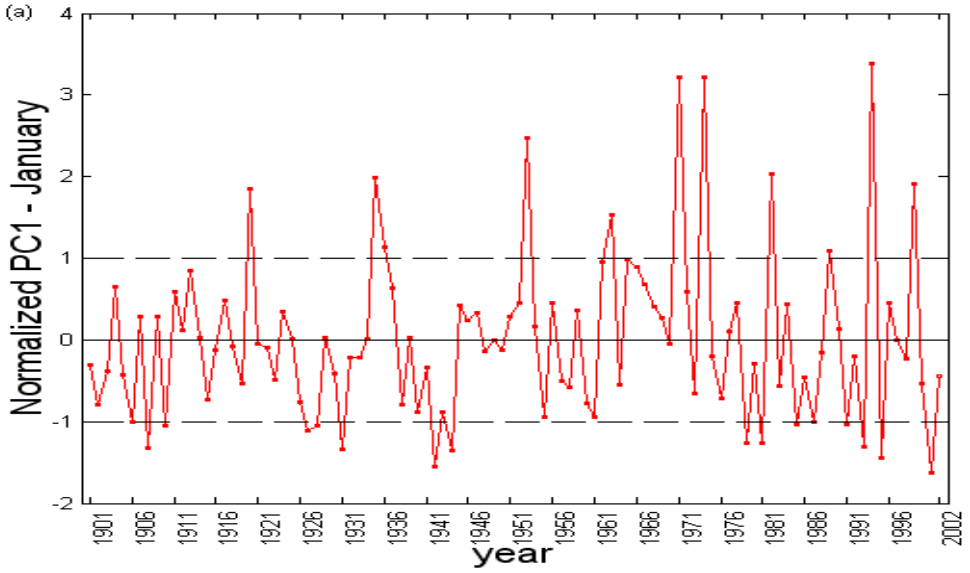
Figure 3.8 shows the first two EOF modes in January and July. Figure 3.9 shows the first two principal components (PC) of January and July. The PCs are standardized to have zero mean and unit variance. The PCs reflect the time variation of the spatial pattern. In January, the values associated with the first mode are almost positive everywhere, with the lower values in north Alberta and south Alberta, and the higher values in the Rocky Mountain regions. After multiplying the first mode by its PC, the contribution of the first mode to the monthly precipitation without the mean is positive (negative) when the PC value is positive (negative) and extreme when the PC values cross the horizontal lines ±1 as the PCs have zero mean and unit variance. The values associated the second mode are negative in the region along the Rocky Mountains and positive in the other regions, so in the same year, the second mode has an opposite contribution in these two regions. In July, the values associated with the first mode are negative with high amplitude in central Alberta. The values associated with the second mode are negative in northern Alberta and nearly positive in the southern regions, so the contribution of the first mode is particularly high in the central

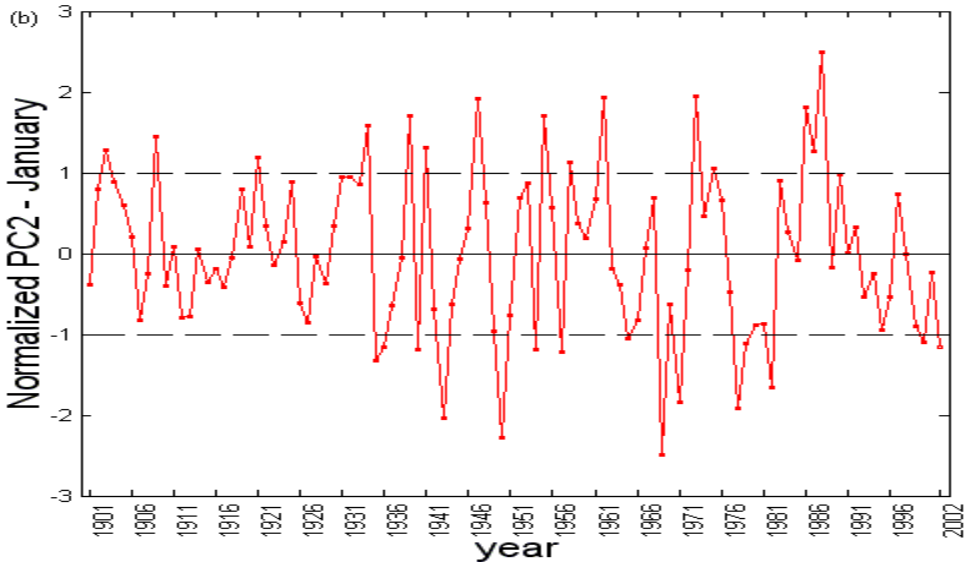
regions when the PC values cross the horizontal lines ±1. The second mode has an opposite contribution in the northern regions and southern areas.

The difference between the observed monthly precipitation data and the interpolated monthly precipitation data at the stations in Jasper, Banff, Peace River A, and Beaverlodge CDA and the corresponding histograms are shown in Figures 3.10 – 3.13. The figures indicate that the difference between the observed monthly data and the EOF-interpolation monthly data is closer to zero than the difference between the observed monthly data and the IDW-interpolated data since 1951 at each cross validation station. That is, the error of the EOF-interpolated data is smaller than that of IDW-interpolated data at these four stations. Both sets of interpolated data have smaller errors at the Peace River A and Beaverlodge CDA stations than at the Jasper and Banff stations because of the high precipitation amount and sparse station distribution in the high-elevation regions.

The nearest grid point used to replace cross-validation station in the error evaluation and the distance between them are presented in Table 3.2. Table 3.3 displays the monthly precipitation estimation errors accessed by cross-validation for eight long-term stations in Alberta from south to north. Each target station is excluded from the interpolation. Comparing this table with Table 2.4 indicates the EOF-interpolated monthly data have smaller RMSE, MAE and MBE values for the grids of 0.1 by 0.2 degrees latitude and longitude than for the grids of 0.25 degrees latitude and 0.5 degrees longitude. That is, the EOF-interpolation method

provides more accurate estimations for the higher-resolution grids, even compared with to the IDW-interpolation method.





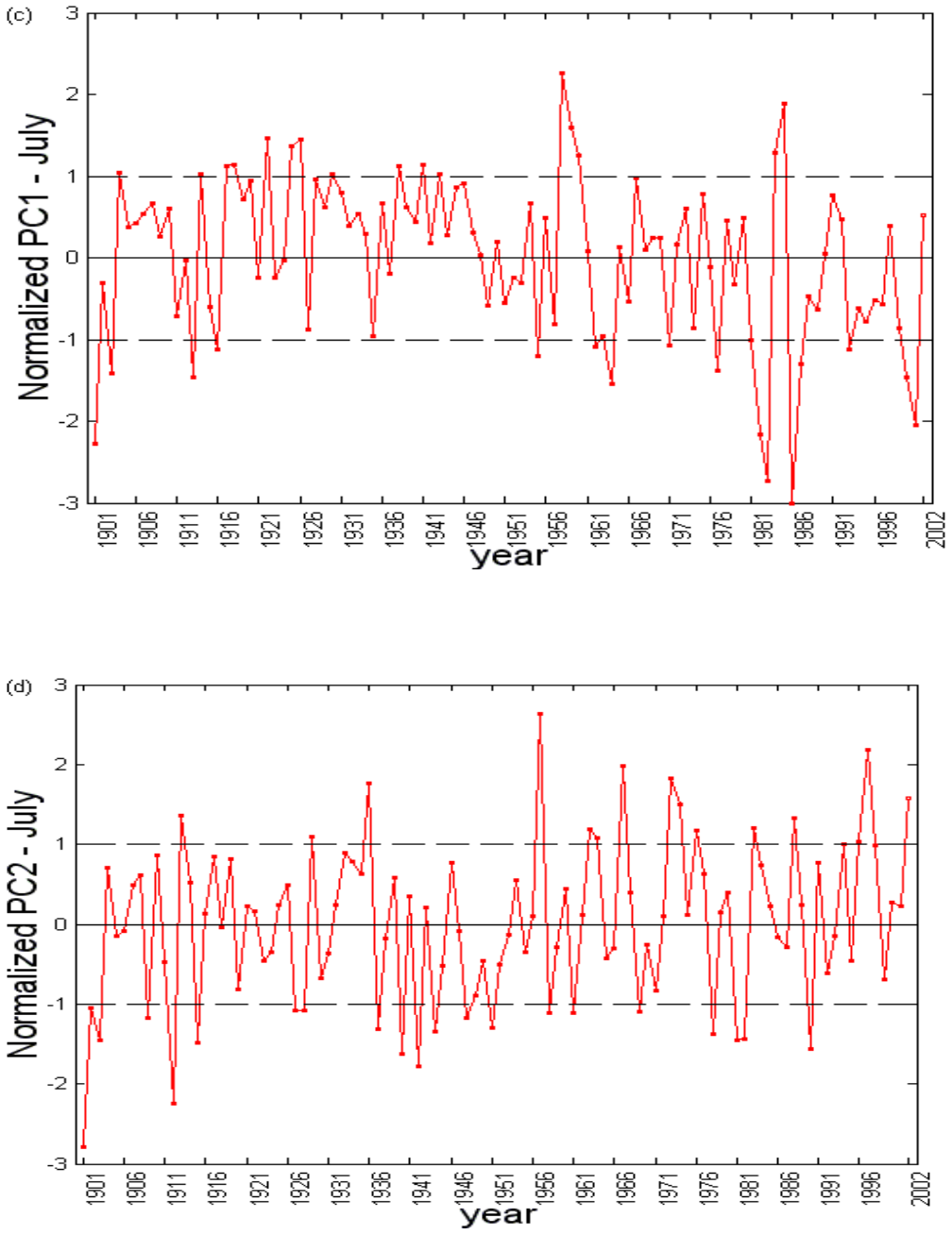


Figure 3.9. Principal components in the EOF analysis: (a) first mode in January, (b) second mode in January, (c) first mode in July, and (d) second mode in July. The components are nondimensional.

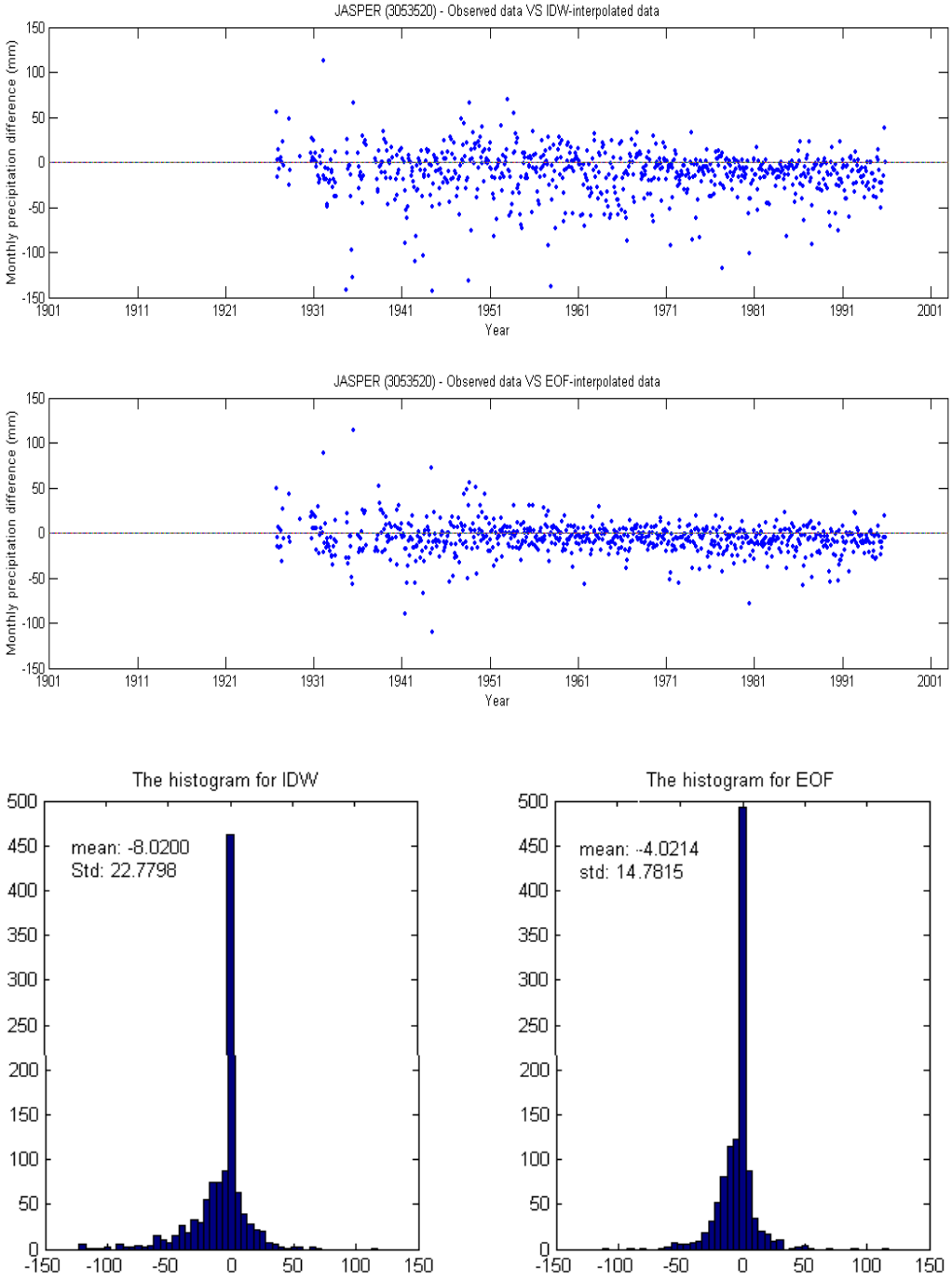


Figure 3.10. The differences between the observed monthly data and the interpolated monthly data at Jasper station (blank period indicates no observed data during that time period), and the corresponding histograms.

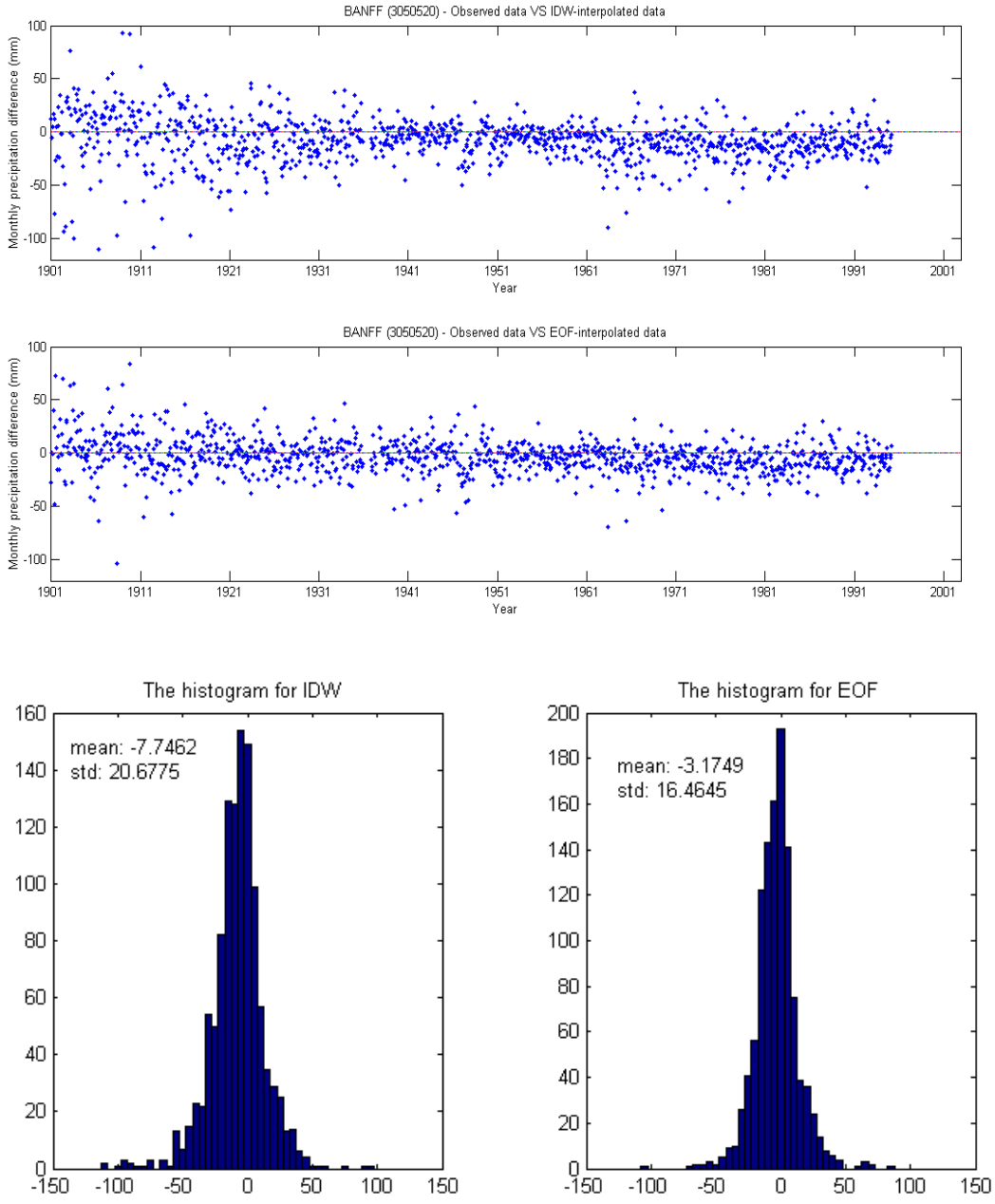


Figure 3.11. The differences between the observed monthly data and the interpolated monthly data at Banff station and the corresponding histograms.

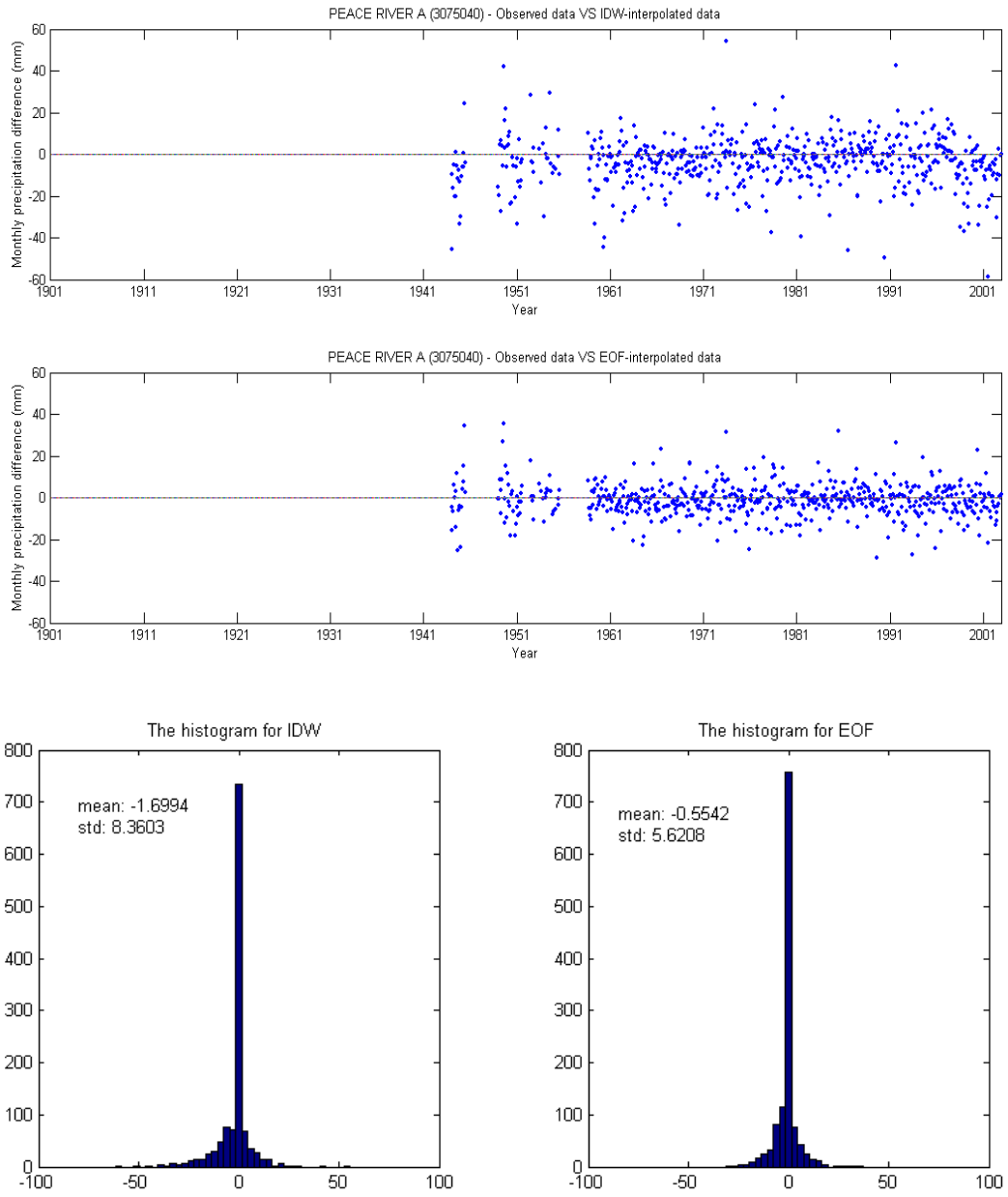


Figure 3.12. The differences between the observed monthly data and the interpolated monthly data at Peace River A station (blank period indicates no observed data during that time period), and the corresponding histograms.

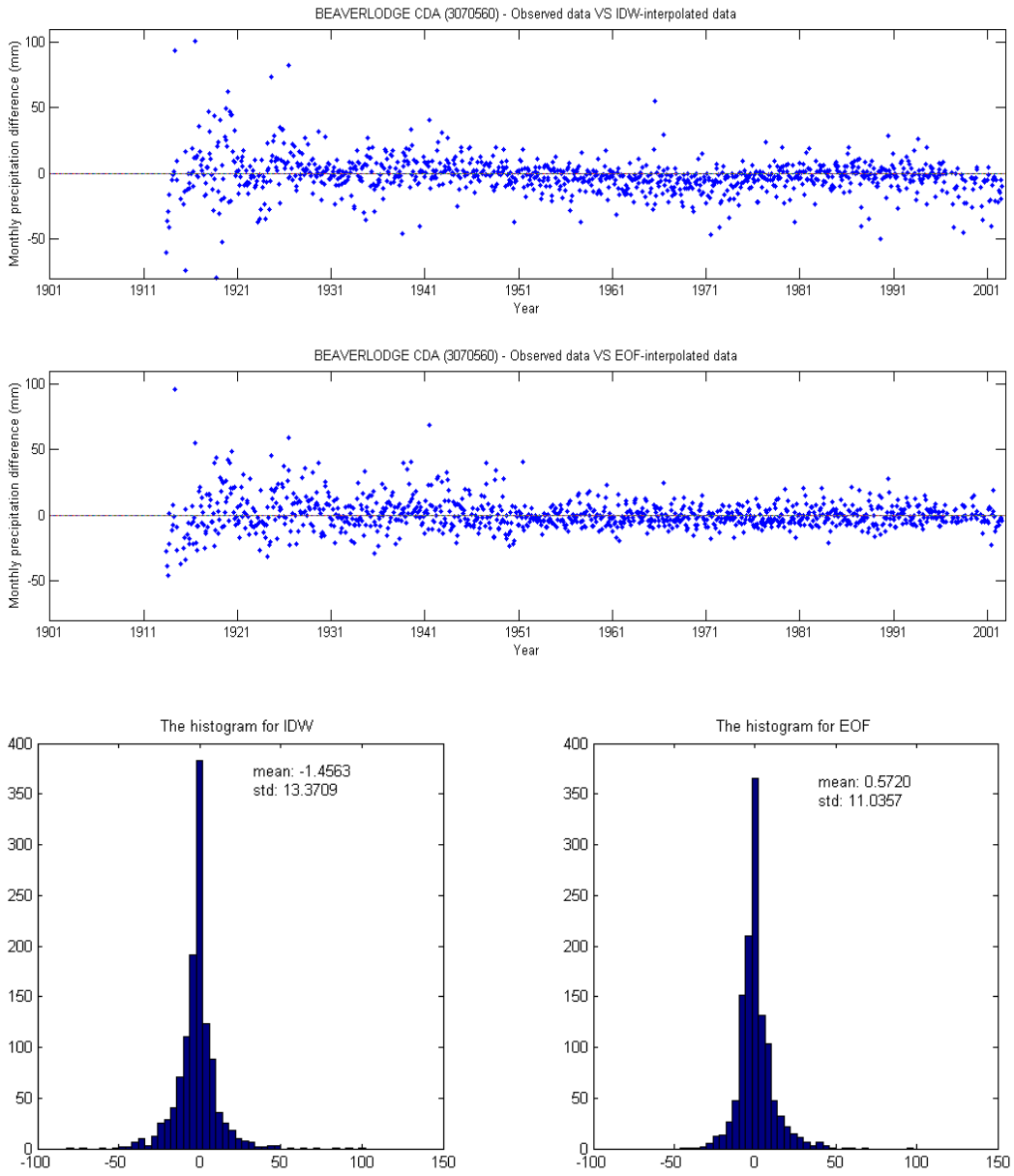


Figure 3.13. The differences between the observed monthly data and the interpolated monthly data at Beaverlodge CDA station (lank period indicates no observed data during that time period), and the corresponding histograms.

Table 3.2. The nearest grid point of the station & the distance between them for 0.25-by-0.5 degrees and 0.1-by-0.2 degrees, respectively. The values in the round brackets is for the 0.25-by-0.5 degrees grid, and the preceding values for the 0.1-by-0.2 degrees grid.

Station Name	Nearest grid point	Distance (km)
Medicine Hat A (50.02°, -110.72°)	14624: 50.00°, -110.80° (10768: 50.00°, -110.50°)	6.134 (15.88)
Calgary Int'l A (51.12°, -114.02°)	14310: 51.10°, -114.00° (10706: 51.25°, -114.00°)	2.626 (14.52)
Banff (51.18°, -115.57°)	14270: 51.20°, -115.60° (10703: 51.25°, -115.50°)	3.052 (9.185)
Jasper (52.88°, -118.07°)	13606: 52.90°, -118.00° (10606: 52.75°, -118.00°)	5.196 (15.2)
Edmonton Int'l A (53.32°, -113.58°)	13445: 53.30°, -113.60° (10579: 53.25°, -113.50°)	2.591 (9.427)
Beaverlodge CDA (55.20°, -119.40°)	12452: 55.20°, -119.40° (10401: 55.25°, -119.50°)	0 (8.434)
Peace River A (56.23°, -117.43°)	11952: 56.20°, -117.40° (10321: 56.25°, -117.50°)	3.817 (4.864)
High Level A (58.62°, -117.17°)	10729: 58.60°, -117.20° (10133: 58.50°, -117.00°)	2.822 (16.59)

Table 3.3. Monthly precipitation estimation errors accessed by cross-validation for eight long-term stations in Alberta from south to north (units: millimeters). The values in the round brackets are the cross validation errors for the IDW method, and the preceding values for the EOF method.

Station Name	RMSE	MAE	MBE
Medicine Hat A	11.96 (13.45)	7.56 (8.53)	0.58 (0.01)
Calgary Int'l A	14.23 (15.85)	8.68 (10.30)	1.88 (-1.13)
Banff	16.89 (22.24)	12.20 (16.05)	3.22 (-7.86)
Jasper	18.73 (29.52)	12.98 (20.17)	6.01(-11.98)
Edmonton Int'l A	7.71 (9.76)	5.28(6.18)	0.51(-1.36)
Beaverlodge CDA	11.98 (14.58)	8.24 (9.51)	0.67(-1.71)
Peace River A	8.07 (12.62)	5.72 (8.87)	1.13 (-3.83)
High Level A	11.06 (20.84)	7.89 (13.69)	-5.82 (-7.77)

3.4.3 Interpolated daily dataset and its accuracy

As Figures 3.2 - 3.4, Figures 3.14 – 3.16 also display the annual total precipitation normal, the May 1 to August 31 total precipitation normal and the eight-month period total precipitation normal from September 1 to April 30 in the time interval between 1961 and 1990, respectively. These two sets of figures display the almost identical precipitation distribution over Alberta except in a few

major cities. For example, the annual total precipitation normal in the Edmonton region is 500 – 550 mm in Figure 3.2 but 450 – 500 in Figure 3.14. The May 1 to August 31 total precipitation normal in Edmonton is 300 -325 mm in Figure 3.3 but 275 – 300 mm in Figure 3.15. The September 1 to April 30 total precipitation normal in Edmonton is 200 -225 mm in Figure 3.4 but 175 – 200 mm in Figure 3.16. As we discussed in Section 3.4.2, the EOF-interpolated data are more accurate on the grids of 0.1 by 0.2 degrees latitude and longitude than on the grids of 0.25 by 0.5 degrees. Thus, Figures 3.14 – 3.16 display more reliable maps than Figures 3.2 - 3.4.

The comparison of the RMSE, MAE and MBE values for the Hybrid 1.0 daily precipitation data and Hybrid 2.0 daily data is displayed in Table 3.4. Each target station is excluded from the interpolation. Table 3.4 shows the Hybrid 2.0 daily precipitation data have smaller errors than those of the Hybrid 1.0 daily data.

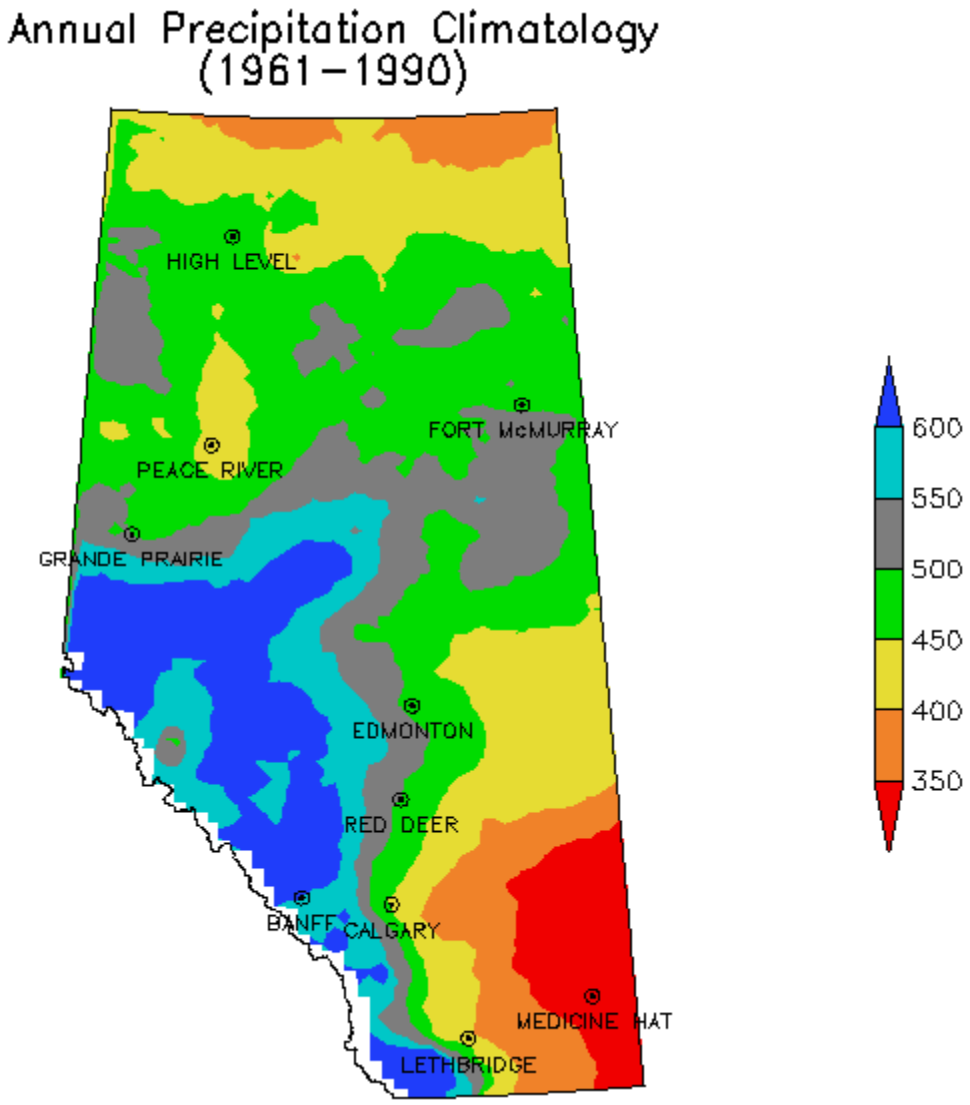


Figure 3.14. The annual total precipitation normals (mm) in the period of 1961 – 1990 onto the grids of 0.1-by-0.2 degrees lattitude and longitude over Alberta.

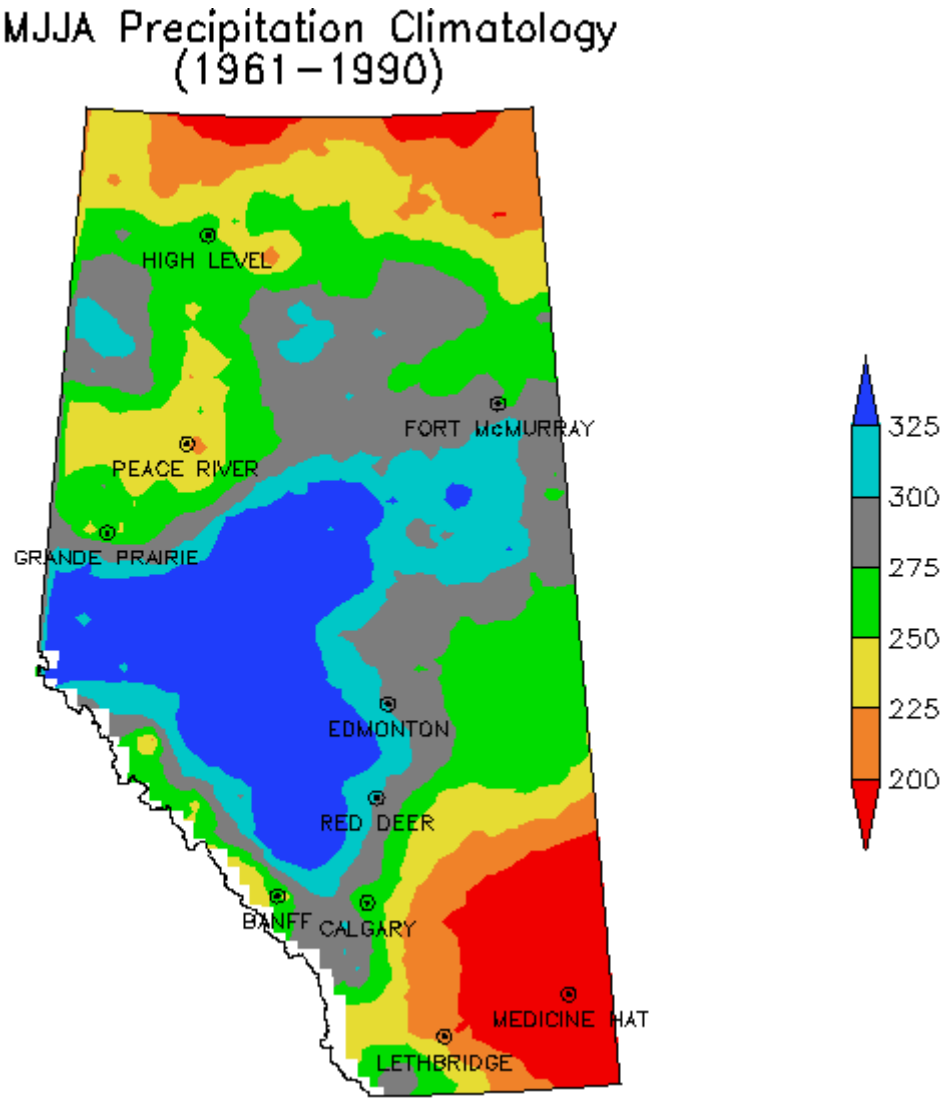


Figure 3.15. May 1 to August 31 total precipitation normals (mm) in the period of 1961 – 1990 onto the grids of 0.1-by-0.2 degrees lattitude and longitude over Alberta.

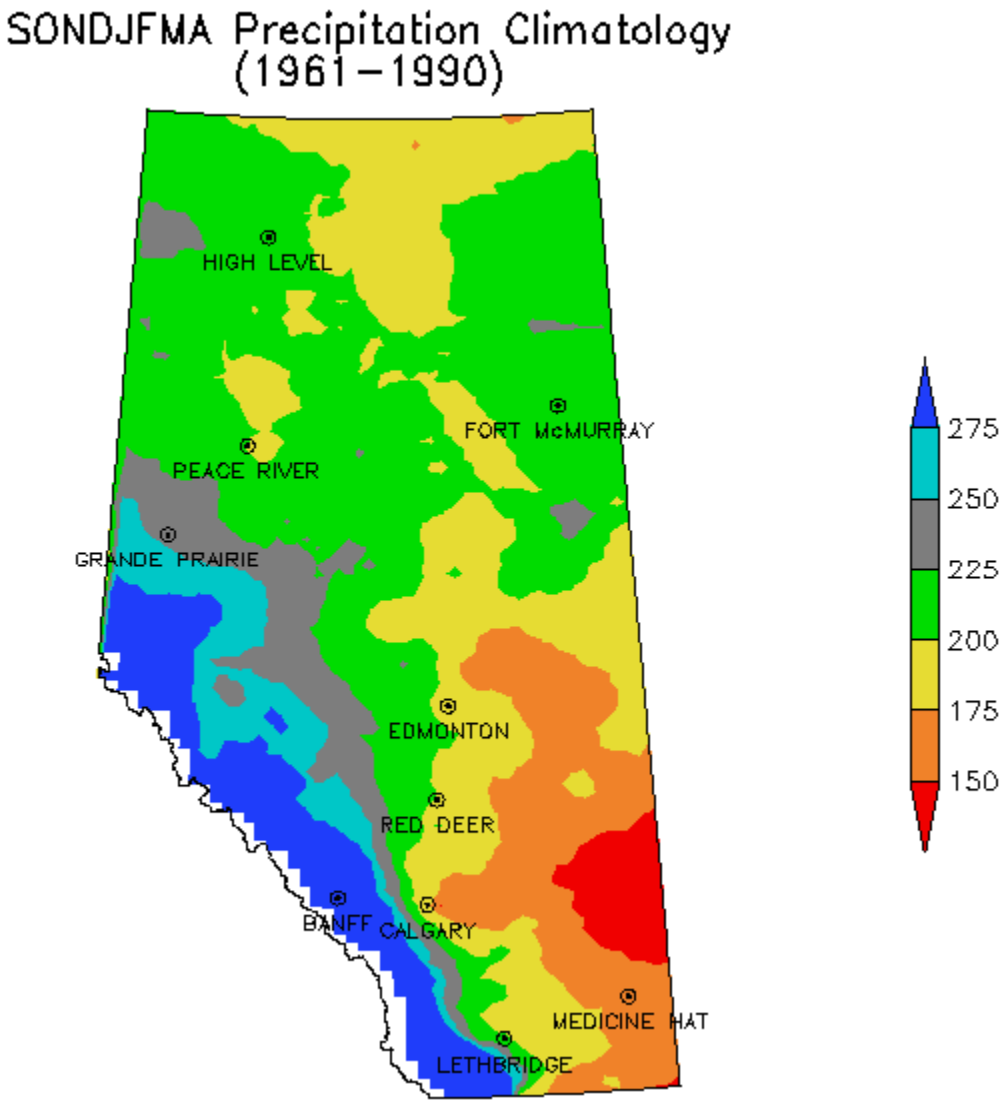


Figure 3.16. September 1 to April 30 total precipitation normals (mm) in the period of 1961 – 1990 onto the grids of 0.1-by-0.2 degrees lattitude and longitude over Alberta.

Table 3.4. Daily precipitation estimation errors accessed by cross-validation for eight long-term stations in Alberta from south to north (units: millimeters). The values in the round brackets are the cross validation errors for the Hybrid 1.0 method, and the preceding values for the Hybrid 2.0 method.

Station Name	RMSE	MAE	MBE
Medicine Hat A	3.07 (3.11)	0.92 (0.94)	0.02 (0.0003)
Calgary Int'l A	3.40 (3.51)	1.06 (1.11)	0.06 (-0.04)
Banff	3.60 (3.88)	1.39 (1.48)	-0.10 (-0.26)
Jasper	3.51 (3.95)	1.41 (1.56)	-0.19 (-0.39)
Edmonton Int'l A	3.08 (3.15)	1.02 (1.04)	-0.02 (-0.04)
Beaverlodge CDA	2.80 (4.01)	1.02 (1.38)	0.03 (-1.06)
Peace River A	3.03 (3.16)	1.08 (1.14)	-0.03 (-0.12)
High Level A	2.99 (3.31)	1.12 (1.18)	-0.19 (-0.25)

3.5 Comparison with Other Interpolation Results

ANUSPLIN is a multivariate non-parametric surface fitting approach to develop spatially continuous climate models and it makes use of thin plate-smoothing splines. Thin plate smoothing splines method has been further developed and made operational as a climate mapping tool by Professor Michael Hutchinson at the ANU over the last 20 years. In this section, we compare our

interpolation results with those of other studies. The Hybrid 2.0, Hybrid 1.0, and ANUSPLIN methods are used to grid the monthly and daily precipitation data over Alberta at a spatial resolution of 300 arc seconds of latitude and longitude, for the period 1961-1990. We compare the gridded monthly precipitation, gridded daily precipitation, and daily precipitation occurrence at seven stations based on a cross validation procedure that withholds these seven stations' data from the analysis. The period 1961 to 1990 is selected for the current comparison. The number of precipitation stations with available data in any year during this period ranged from 2000 to 3000 (Fig. 3.17). As an example, Figure 3.18 shows the stations used on March 31, 1977. According to the withheld station selection rule, these seven withheld stations have nearly complete daily observation data over the studied period of 1961 – 1990. The locations of the withheld stations are shown in Figure 3.19. The summaries of the comparison of the withheld data errors for the interpolated monthly total precipitation amount are presented in Table 3.5, which includes the RMSE, MAE and MBE of the monthly precipitation amounts. Comparison of the RMSE, MAE and MBE values for these three methods shows that the Hybrid 2.0 has much smaller errors than Hybrid 1.0 and ANUSPLIN in interpolating the monthly total precipitation at the northern stations. And as the station changes from north to south, the differences between Hybrid 2.0 and other two methods changes from large to small, but Hybrid 2.0 always has a more accuracy result. Compared with ANUSPLIN, Hybrid 2.0 has 2.10mm – 5.73mm smaller values in RMSE and 1.17mm – 3.17mm smaller values in MAE. Compared with Hybrid 1.0, Hybrid 2.0 has 1.49mm – 9.41mm

smaller values in RMSE and 0.85mm – 6.97mm in MAE. Table 3.6 displays the compared summaries of the average precipitation frequency (days) in each month at the withheld stations. The comparison shows that Hybrid 2.0 has the comparable accurate results in estimating the precipitation frequency as Hybrid 1.0 method. Both of them perform averagely better than ANUSPLIN when estimating the precipitation frequency. In general, ANUSPLIN overestimates the precipitation frequency and both hybrid methods underestimate the precipitation frequency.

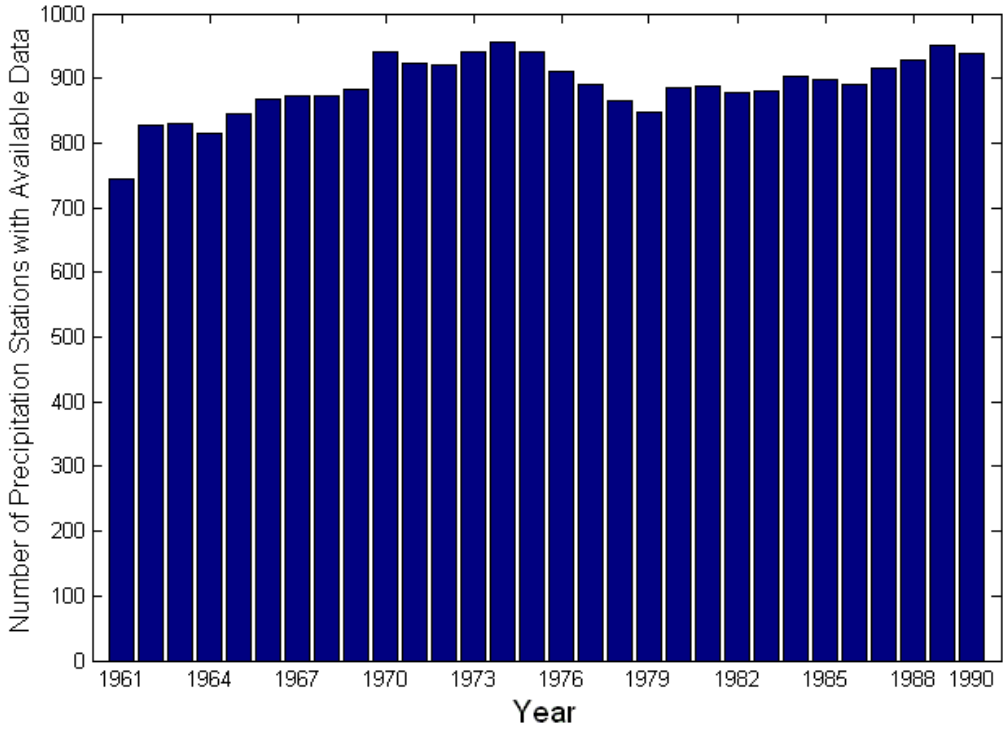


Figure 3.17. The number of precipitation stations with available observed data including the vicinity stations outside Alberta in each year over the period of 1961 – 1990.

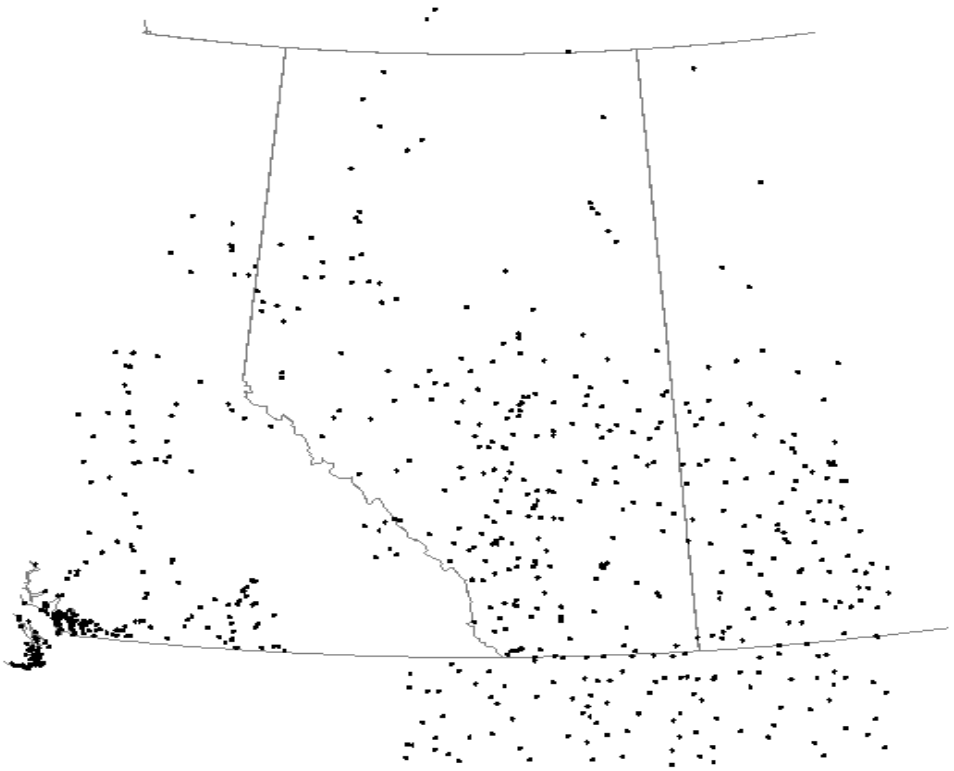


Figure 3.18. The locations of 678 precipitation stations with available observed data on March 31, 1977 as an example in Alberta.

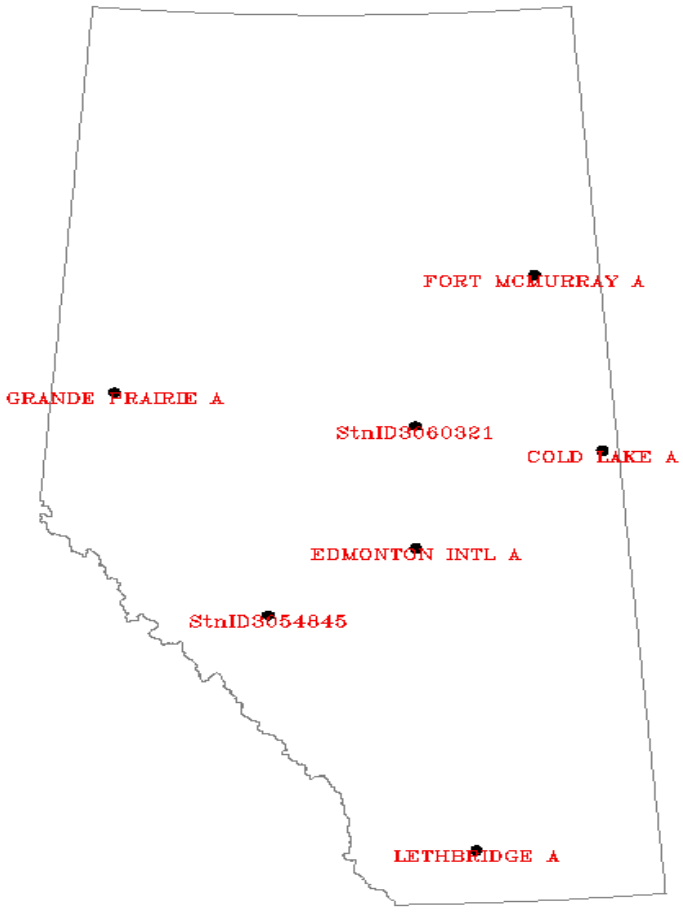


Figure 3.19. The locations of the seven precipitation stations withheld in the comparion of Hybrid 2.0 and other interpolated methods.

Table 3.5 Compared summaries of withheld data errors (mm) for the monthly total precipitation amount at 6 selected withheld stations across Alberta.

Station	Hybrid 1.0			Hybrid 2.0			ANUSPLIN		
	RMSE	MAE	MBE	RMSE	MAE	MBE	RMSE	MAE	MBE
Fort McMurray A	16.28	12.26	-4.37	6.87	5.29	-1.88	12.60	8.39	1.91
Grande Prairie A	13.30	8.83	-6.24	5.82	4.37	-1.67	10.96	7.30	-2.28
Athabasca 2	8.89	6.07	1.43	5.32	3.64	0.06	9.08	6.19	1.65
Cold Lake A	14.37	9.19	0.93	5.90	4.32	0.95	12.35	7.49	1.63
Edmonton Int'l A	9.76	6.07	-1.01	5.10	3.58	0.06	9.99	6.34	0.51
Lethbridge A	7.81	5.22	-0.55	6.32	4.37	-0.88	8.42	5.54	0.30

Table 3.6 Compared summaries of the average precipitation frequency (days) in each month at 7 withheld stations across Alberta: (a). Fort McMurray A (ID: 3062693), (b). Grande Prairie A (ID: 3072920), (c). Athabasca 2 (ID: 3060321), (d). Cold Lake A (ID: 3081680), (e). Edmonton Int'l A (ID: 3012205), (f). Nordegg RS (ID: 3054845), and (g). Lethbridge A (ID: 3033880).

(a). Fort McMurray A (ID: 3062693)				
Month	Observation	ANUSPLIN	Hybrid 1.0	Hybrid 2.0
JAN	12	15.1	8.3	8.1
FEB	9.4	11.3	6.1	6.0
MAR	9.7	10.8	7.0	6.5
APR	7.3	7.8	6.2	6.2
MAY	9.8	12.8	10.1	10.0
JUN	12.5	15.2	13.0	12.6
JUL	14.4	17.1	15.0	14.7
AUG	12.5	15.2	12.6	12.5
SEP	11.2	14.0	11.4	10.9
OCT	10.3	11.5	8.4	8.6
NOV	11.9	14.1	8.2	8.4
DEC	12.1	15.9	8.4	8.4

(b). Grande Prairie A (ID: 3072920)				
Month	Observation	ANUSPLIN	Hybrid 1.0	Hybrid 2.0
JAN	12.8	12.5	11.4	11.5
FEB	9.3	9.4	8.2	8.2
MAR	9.4	10.2	8.9	8.4
APR	6.9	7.2	5.9	5.9
MAY	8.8	10.4	8.8	8.3
JUN	11.3	13.2	10.6	10.6
JUL	12.0	13.7	10.8	10.8
AUG	11.4	12.9	10.7	10.4
SEP	11.3	13.6	11.6	11.5
OCT	8.4	8.6	7.9	7.4
NOV	10.8	10.0	8.7	8.7
DEC	11.1	12	9.9	9.9

(c). Athabasca 2 (ID: 3060321)				
Month	Observation	ANUSPLIN	Hybrid 1.0	Hybrid 2.0
JAN	9.5	12.4	9.3	9.3
FEB	7.7	9.8	7.2	7.2
MAR	8.0	9.9	6.0	6.2
APR	6.3	7.2	4.9	4.9
MAY	9.3	10.9	8.0	8.0
JUN	12.8	14.4	11.0	10.8
JUL	14.1	16.1	13.2	13.2
AUG	12.4	13.8	10.5	10.5
SEP	10.3	12.7	9.0	9.0
OCT	7.1	8.4	4.9	4.9
NOV	7.6	9.7	6.8	6.8
DEC	9.6	12.3	9.3	9.2

(d). Cold Lake A (ID: 3081680)				
Month	Observation	ANUSPLIN	Hybrid 1.0	Hybrid 2.0
JAN	10.5	10.3	7.2	7.0
FEB	8.2	8.3	5.6	5.7
MAR	7.7	8.0	5.8	5.6
APR	6.7	7.1	5.7	5.8
MAY	9.3	10.9	9.1	8.8
JUN	12.6	14.3	13.0	12.8
JUL	13.8	15.8	14.7	14.5
AUG	11.4	13.4	11.7	11.5
SEP	10.1	11.8	9.8	9.5
OCT	6.9	7.7	5.6	5.6
NOV	9.1	9.6	5.6	5.5
DEC	10.7	10.5	6.8	6.8

(e). Edmonton Int'l A (ID: 3012205)				
Month	Observation	ANUSPLIN	Hybrid 1.0	Hybrid 2.0
JAN	10.8	13.1	10.3	10.4
FEB	7.8	9.8	7.9	7.7
MAR	8.7	9.3	7.5	7.3
APR	6.7	7.8	6.3	6.3
MAY	9.5	11.1	9.4	9.1
JUN	12.7	15.1	13.0	13.1
JUL	13.2	14.7	13.8	14.0
AUG	11.6	13.8	12.8	12.8
SEP	9.6	11.6	9.8	9.8
OCT	5.6	7.1	5.0	4.9
NOV	8.0	9.8	7.5	7.3
DEC	9.7	10.9	9.3	9.2

(f). Nordegg RS (ID: 3054845)				
Month	Observation	ANUSPLIN	Hybrid 1.0	Hybrid 2.0
JAN	6.1	9.5	7.4	7.5
FEB	4.6	7.3	5.6	5.5
MAR	6.5	8.4	6.6	6.6
APR	7.6	8.7	7.4	7.5
MAY	10.5	12.5	10.9	10.9
JUN	12.8	15.3	13.8	13.9
JUL	13.5	16.5	13.7	13.8
AUG	12.2	14.8	12.3	12.3
SEP	11.2	12.9	10.3	10.3
OCT	6.7	7.4	6.8	6.8
NOV	5.4	8.1	5.9	5.7
DEC	6.0	8.7	5.8	5.8

(g). Lethbridge A (ID: 3033880)				
Month	Observation	ANUSPLIN	Hybrid 1.0	Hybrid 2.0
JAN	10.1	9.7	9.4	9.0
FEB	7.1	7.2	6.8	6.5
MAR	8.9	8.5	8.2	8.2
APR	7.7	8.3	7.9	7.8
MAY	9.6	10.5	9.9	9.9
JUN	9.1	10.0	9.4	9.2
JUL	7.7	8.2	7.7	7.6
AUG	7.4	8.2	7.7	7.7
SEP	7.4	8.0	7.4	7.4
OCT	5.4	5.5	5.1	5.0
NOV	6.4	6.5	6.2	6.0
DEC	9.3	9.3	9.0	9.1

3.6 Conclusion

We have used the Hybrid 2.0 method to grid the daily precipitation data over Alberta from January 1, 1901 to December 31, 2002. The resolutions of the two grids were 0.25° latitude by 0.5° longitude and 0.1° latitude by 0.2° longitude, respectively. The interpolated data were named ABClim2.0. The cross-validations were used to do the error analysis of the interpolated monthly total data and daily data for two grid resolutions, respectively. The RMSE, MAE and MBE values were calculated for both the monthly data and the daily data on the two grids to compare Hybrid 2.0 and Hybrid 1.0. The comparison showed that the accuracy of the Hybrid 2.0 interpolation was higher than that of Hybrid 1.0 interpolation, especially in the high-elevation regions and the regions with sparse station distribution in the earlier period of the last century. The interpolation accuracy of Hybrid 2.0, Hybrid 1.0 and ANUSPLIN was also compared. The results revealed that Hybrid 2.0 had comparable accuracy as Hybrid 1.0 and both of them are a little bit more accurate than ANUSPLIN in estimating the precipitation frequency in each month. Hybrid 2.0 also had more accurate results than ANUSPLIN and Hybrid 1.0 in interpolating the monthly total precipitation amount. Further improvements in the current methodology may be aimed at taking the elevation into account explicitly when redistributing the monthly total to each individual day of the month. The monthly total precipitation values of this study can be used with some confidence across Alberta.

Chapter 4

Applications and Analysis of ABClim2.0 Data

4.1 Introduction

The gridded daily data -- ABClim 2.0 -- have numerous applications, ranging from soil quality models to crop models and to drought monitoring. Three applications will be discussed here: trend analysis of ABClim 2.0 data, historical flood monitoring by maps of the maximum number of consecutive wet days, and historical drought monitoring by maps of the maximum number of consecutive dry days and by the Standardized Precipitation Index (SPI) (Yin, 2005; Shen, 2005 (a); Shen, 2005 (b)).

Alberta is one of major agriculture provinces in Canada. Therefore, information on its precipitation trend is very important. In practice, some statistical methods such as linear regression, moving average, and the Mann-Kendal test are usually used to analyze the trend of the climate time series. We will focus these three methods on the historic trend analysis of ABClim 2.0 data here. The linear regression analysis provides us an overall trend which is a clear picture of decreasing or increasing or unchanged line in the whole analysis period However, the moving average will show us the detailed trend, a clear picture of curve which reveals the trend in each shorter period not the overall trend in the whole analysis period. The non-parametric Mann-Kendall (M-K) test was used to

test randomness against trend. The advantage of this non-parametric test is that it does not assume any special form for the distribution function of the data while this test's power is nearly as high as that of parametric tests. Therefore, this test was found to be an excellent tool for trend detection in different applications, such as detecting the trend in climatological time series (Goosens and Berger, 1986).

Besides the information on the precipitation trend, the information on the historic flood monitoring and drought monitoring is also important for agriculture in Alberta. We will apply the maps of the maximum number of consecutive wet days and consecutive dry days and the SPI method to the ABClim 2.0 data here. A drought index value is typically a single number, far more useful than raw data for monitoring drought. Various indicators and climatic indices have been developed by water resource and climatological professionals for use in drought planning. McKee et al. (1993) developed the Standardized Precipitation Index (SPI) for the purpose of defining and monitoring drought. The nature of the SPI allows an analyst to determine the rarity of a drought or an anomalously wet event at a particular time scale for any location in the world that has a precipitation record. The SPI is better than other indexes, such as Standardized Anomaly Index (SAI), and Principal Component Index (PCI), in reflecting the intensity and duration of drought. The SPI can reflect different aspects of the condition of the water resources at different time scales (Yin, 2005).

The organization of this chapter is as follows. Section 4.2 uses the methods of linear regression, moving average, and the Mann-Kendal test to analyze the trend of ABClim 2.0 data in the period of 1901 – 2002 for some

stations in Alberta. The maps of consecutive dry days and of consecutive wet days in each decade from 1901 – 2002 are created from the ABClim 2.0 data for Alberta in Section 4.3 and Section 4.4, respectively. Section 4.5 uses the SPI as the drought index to detect the historic extreme events like drought and flood for some stations in Alberta from the ABClim 2.0 data. The conclusions and discussion are presented in section 4.6.

4.2 Trend Analysis of Precipitation Data

The time series of the Hybrid 2.0 summer total (from May to August, i.e., MJJA) precipitation data has a trend. Three statistics methods are used to detect the trends. The methods are as follows.

(1). Linear Regression Analysis

Let R_i denote the precipitation time series which has the sample size of n, and t_i denote the corresponding time. The simplest linear regression equation between R_i and t_i is constructed as

$$\hat{R}_i = a + bt_i + \varepsilon_i, \text{ i=1,2,}\cdots\text{,n,} \qquad (4.1)$$

where a and b are the regression parameters for the above straight line model, ε_i are the residuals. This simple linear regression model is based on the assumption that the residuals are independent and normally distributed. The null hypothesis of this linear regression based trend test is $H_0: b = 0$ which states that variable R_i is independent of t_i and there is no significant linear relationship

between the variable them. To test whether the slope b significantly apart from 0, t-test is used and the test statistic is defined as

$$t = \frac{\hat{b}}{SE\left(\hat{b}\right)},$$

where $\hat{b}$ is the estimated slope of the regression line and $SE\left(\hat{b}\right)$ is the standard error of $\hat{b}$. According to the value of above test statistic, we can reject or accept the null hypothesis at some significance level like 0.01, 0.05, and 0.10.

(2). K-year Central Moving Average

Moving average is a basic method of the trend analysis technique, which is equivalent to a low-pass filter. The trend is revealed by the smoothness of the time series data. k-year central moving average is used to detect the trend in the precipitation time series here and is expressed as follows.

$$MA_i = \frac{1}{k} \sum_{j=-(k-1)/2}^{(k-1)/2} R_{i+j} \text{ , } i = (k+1)/2, \text{ ..., n-(k-1)/2,} \qquad (4.2)$$

where k is the moving length, R_i denotes the precipitation time series, and n is the total number of years. Here we use 11-year central moving average to analyze the detailed trend of the ABClim 2.0 data in the period of 1901-2002.

(3). Mann-Kendall Trend Test

Considering the time series x with sample size n, we define a statistic as

$$s_k = \sum_{i=1}^{k} r_i,\ k = 2,3,\cdots,n, \qquad (4.3)$$

where r_i is the cumulative number of $x_i > x_j$ $(j = 1,2,\cdots,i-1)$:

$$y_j = \begin{cases} 1, \text{ if } x_i > x_j, \\ 0, \text{ otherwise,} \end{cases}, j = 1,2,\cdots,i-1,$$

$$r_1 = 0 \text{ and } r_i = \sum_{j=1}^{i-1} y_j,\ i = 2,3,\cdots,n.$$

Under the null hypothesis of no trend, the statistics s_k is distributed as a normal distribution with expected value $E(s_k)$ and variance $Var(s_k)$ as follows:

$$\begin{cases} E(s_k) = \dfrac{k(k-1)}{4}, \\ Var(s_k) = \dfrac{k(k-1)(2k+5)}{72}. \end{cases} \qquad (4.4)$$

The statistics s_k is then standardized as

$$UF_k = \frac{s_k - E(s_k)}{\sqrt{Var(s_k)}},\ k = 1,2,3,\cdots,n, \qquad (4.5)$$

where $UF_1 = 0,$ so UF_k has a standard normal distribution. The null hypothesis of no trend will be rejected at a significance level of α if the absolute value of UF_n is greater than $z_{\alpha/2}$. That is, a statistically significant trend exists in the time series. For example, $z_{\alpha/2} = 1.96$ when $\alpha = 0.05,$ and $z_{\alpha/2} = 2.58$ when $\alpha = 0.01$.

In order to localize the beginning of the trend, the same procedure applied for the UF_k statistic is also applied to the retrograde series $x_n, x_{n-1}, \cdots, x_1$ to get

UB_k . The graphical representations of the series UF_k and UB_k are denoted as UF and UB , respectively. In the case of a significant trend, the intersection of these two curves localizes the time when the trend starts (Feidas, 2007).

Figure 4.1 displays the positive and negative MJJA precipitation trends at four stations. The linear trend lines show a non-significant decreasing trend in Edmonton (0.3%) and Calgary (4%) but a significant increasing trend in Peace Rive (38%) and High Level (28%). The 11-year moving average lines give more details about the MJJA precipitation changing trend. For Edmonton, it is characterized by two obvious decreasing periods (the early 1910's - the early 1920's, and the early 1950's to the early 1960's) and two obvious increasing periods (from the early 1920's to the end of the 1940's, and from the early 1960's to the end of the 1970's). Calgary has a significant decreasing trend from 1906 to the early 1920's and a significant increasing trend from the end of the 1930's to the early 1950's, and from the end of the 1980's to the middle of 1990's. One obvious decreasing period (from 1906 to the end of the 1920's) and one increasing period (from the early 1940's to the end of the 1990's) are characterized for Peace River. A significant increasing trend is located in the period of the early 1940's to the end of the 1950's, and a decreasing trend is located in the 1960's for High Level.

In order to identify the intersection of the curves and thus enable the detection of the beginning of the trend or the change, graphical analysis is applied to the Mann-Kendall statistics time series UF_k and UB_k to create the curves UF and UB (Feidas, 2007). The plots for four stations in Alberta are given in Fig. 4.2.

The analysis of the full range of figures shows the decreasing or increasing trend per station, as well as the approximate year when an abrupt MJJA precipitation change occurred. In Fig. 4.2(a), UF and UB intersect in 1976, and after 1976, the curve UF increases to positive, indicating that the period of 1976 – 2000 was a relatively wet period for Edmonton. However, $|UF_{102}| = 0.51 < 1.96$. This result means that no significant trend occurred in MJJA precipitation from 1901 to 2002 for Edmonton at the 0.05 significance level. Fig. 4.2(b) locates the intersection of UF and UB in 1987, and then UF increases to positive. Figure 4.2 (b) shows that the period of 1987-1999 was a relatively wet period for Calgary. As well, $|UF_{102}| = 0.38 < 1.96$ shows that no significant trend occurred in MJJA precipitation from 1901 to 2002 for Calgary at the 0.05 significance level. In Fig. 4.2(c), $UF_{102} = 1.97 > 1.96$ reflects a statistically significant increasing trend in MJJA precipitation from 1901 to 2002 for Peace River at the 0.05 significance level. The intersection of UF and UB, located between the interval (-1.96, 1.96), reveals that this increasing trend was an abrupt change which started in 1977. In Fig. 4.2(d), $UF_{102} = 2.12 > 1.96$ reveals a statistically significant increasing trend in MJJA precipitation from 1901 to 2002 for High Level at the 0.05 significance level. The location of the intersection of UF and UB identifies this increasing trend as an abrupt change which started in 1949. The M-K test's detection of the trend in MJJA precipitation is summarized in Table 4.1 for the stations represented in Fig. 4.3. Table 4.1 displays whether a statistically significant trend at the 0.05 significance level occurred in the period of 1901 – 2002 as well as the

first year of the abrupt change for this trend.

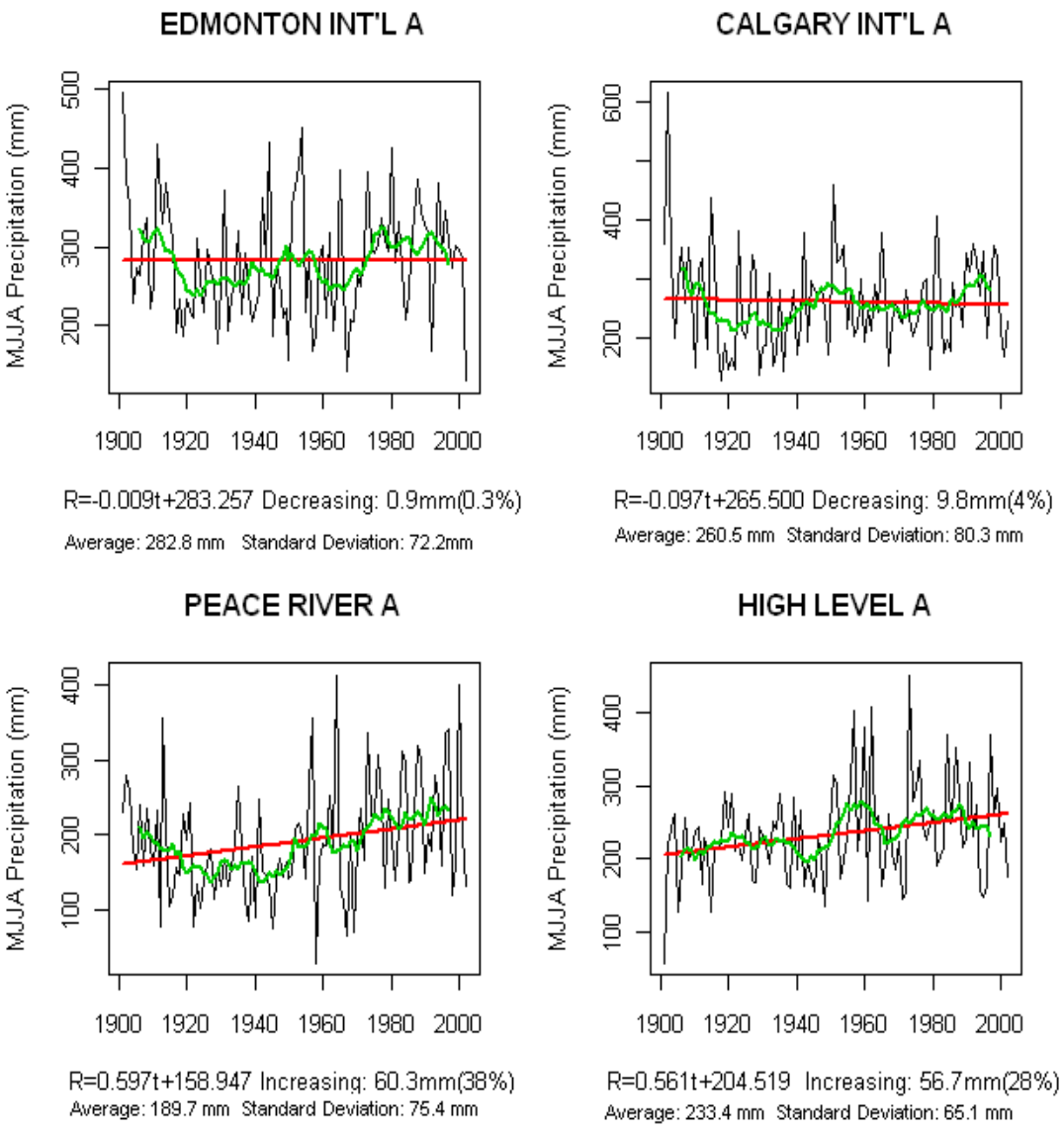


Figure 4.1. Observed values, trend line and 11-year moving average of MJJA precipitation in some stations for the period 1901–2002 (black line: observed values, red line: linear trend, green line: 11-year moving average).

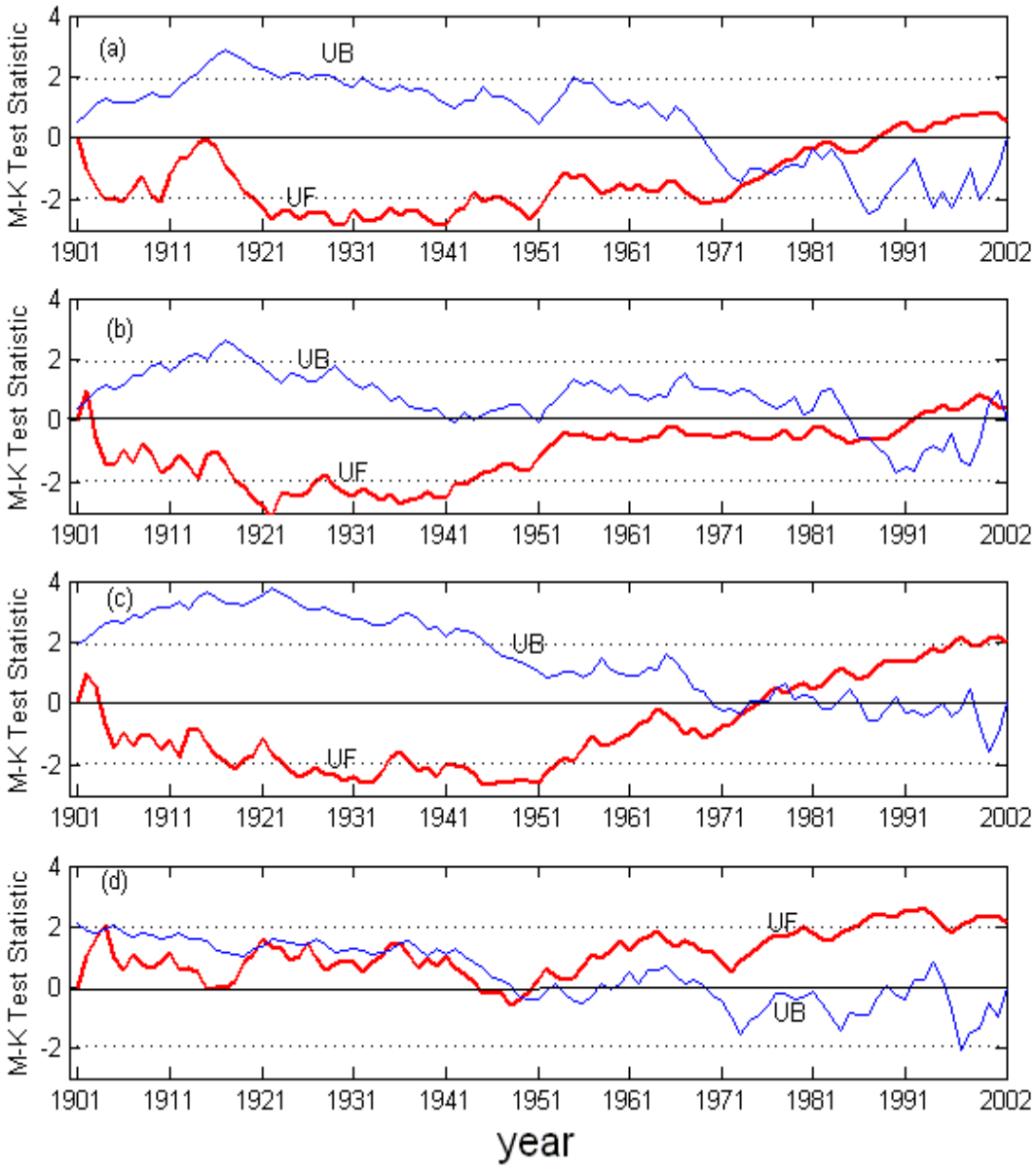


Figure 4.2. Mann-Kendall test of MJJA precipitation at four stations in Alberta for the period 1901–2002: (a) Edmonton Int'l A, (b) Calgary Int'l A, (c) Peace River A, (d) High Level A (red line: UF, blue line: UB).

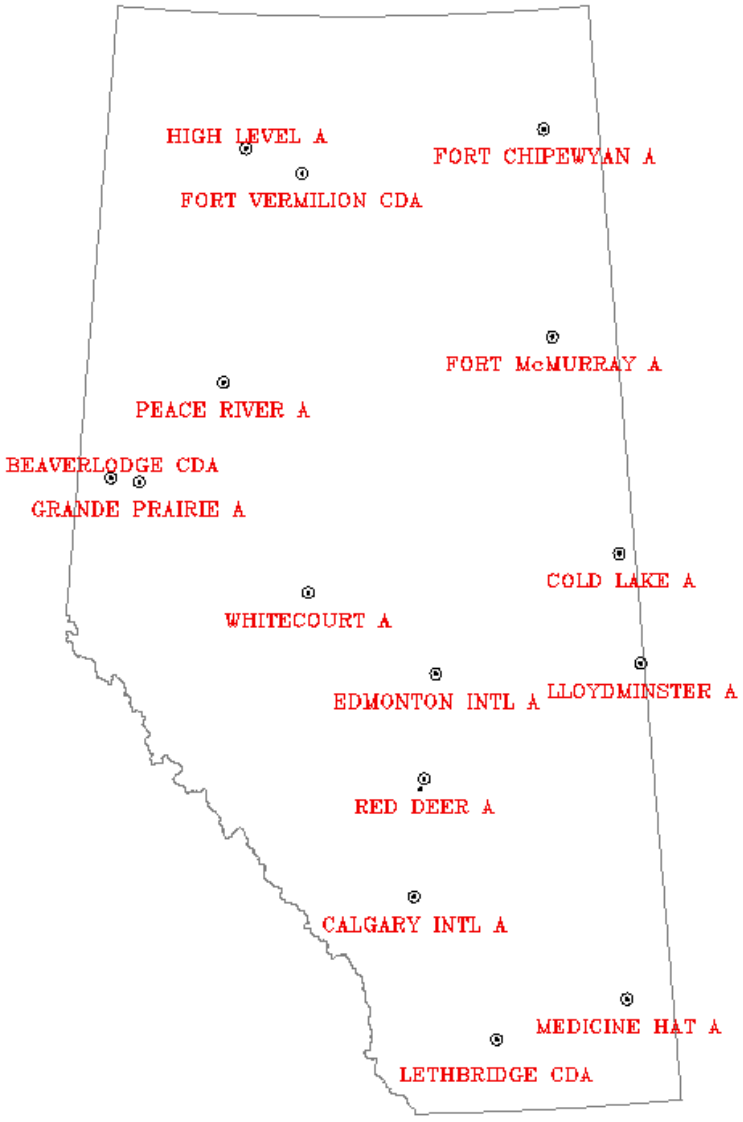


Figure 4.3. The stations to which the Mann-Kendall test was applied to detect the MJJA precipitation trend for the period 1901–2002.

Table 4.1. Significant trend analysis at the 95% significance level by the M-K test for 1901-2002 MJJA precipitation.

Stations	MJJA precipitation trend
HIGH LEVEL A	1949 +
FORT CHIPEWYAN A	1949 +
FORT VERMILION CDA	1982 +
FORT MCMURRAY A	1959 +
PEACE RIVER A	1977 +
BEAVERLODGE CDA	o
GRANDE PRAIRIE A	o
COLD LAKE A	o
WHITECOURT A	o
LLOYDMINSTER A	o
EDMONTON INT'L	o
RED DEER A	o
CALGARY INT'L A	o
MEDICINE HAT A	o
LETHBRIDGE CDA	o

Note: 1). 'o' means no significant trend.
2). '+' means increasing trend which is an abrupt change.
3). The year means the beginning year of the increasing trend.
4). The data of each station used here is the ABClim 2.0 data of the grid point nearest to this station.

4.3 Historical Drought Monitoring: Maps of the Maximum Number of Consecutive Dry Days in 1951-2000

To calculate the maximum number of consecutive dry days in a certain time period, a decade is chosen here (Shen et al., 2005 (a)). A severe drought event is very likely to happen if the number is very large. The maximum number of wet and dry days depends on the chosen criteria. In this study, the criteria of 0.5 mm is used; that is, if the daily precipitation amount is less than 0.5 mm, that day will be considered a dry day for the reason that the small amount of precipitation like 0.5 mm does not relieve the drought situations during the continuous dry days (Shen et al., 2005 (a)). This thesis does not use a rigorous estimate. We calculate the maximum number of dry days for every grid point in each decade period; that is, each grid point has a number for each decade. Five maps are generated for the ten decades: 1951-1960, 1961-1970, …, 1991-2000 (Figs. 4.4 – 4.8). These maps help reveal the locations of historical severe drought events. More than 100 days without effective precipitation would certainly have a significant impact on the local agriculture and everything else related to the water supply. Figure 4.4 reveals that severe drought events happened mainly in a small region of southern Alberta. By comparing the corresponding maps of Saskatchewan and Manitoba in (Shen et al., 2005 (a)), Manitoba had more severe drought events than Alberta and Saskatchewan during 1951 – 2000.

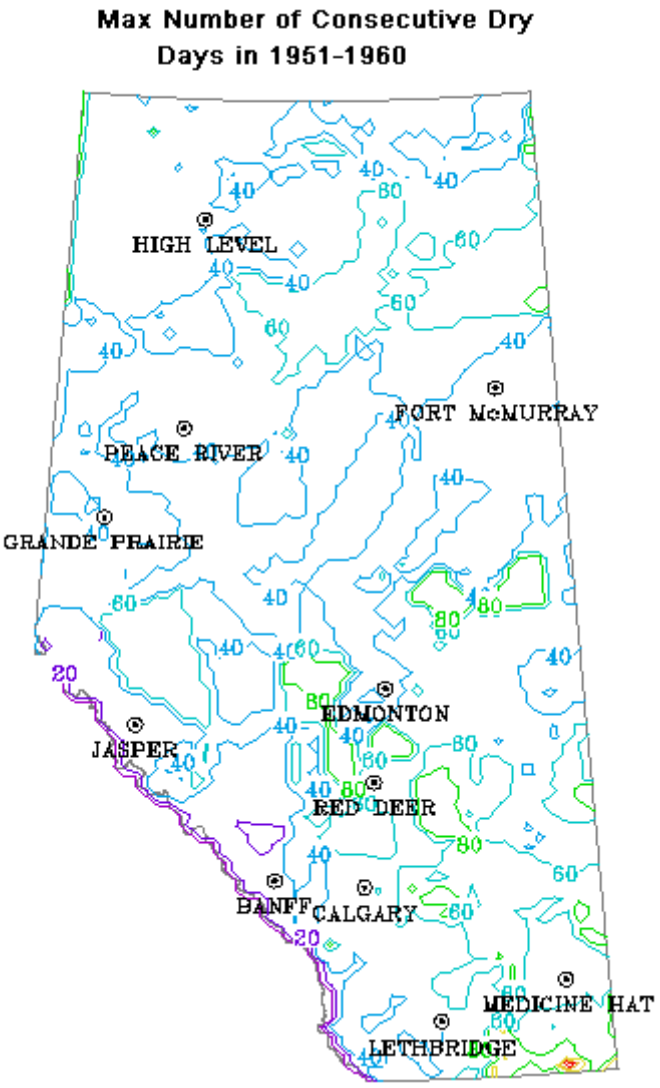


Figure 4.4. Maximum number of consecutive dry days over Alberta during 1951-1960.

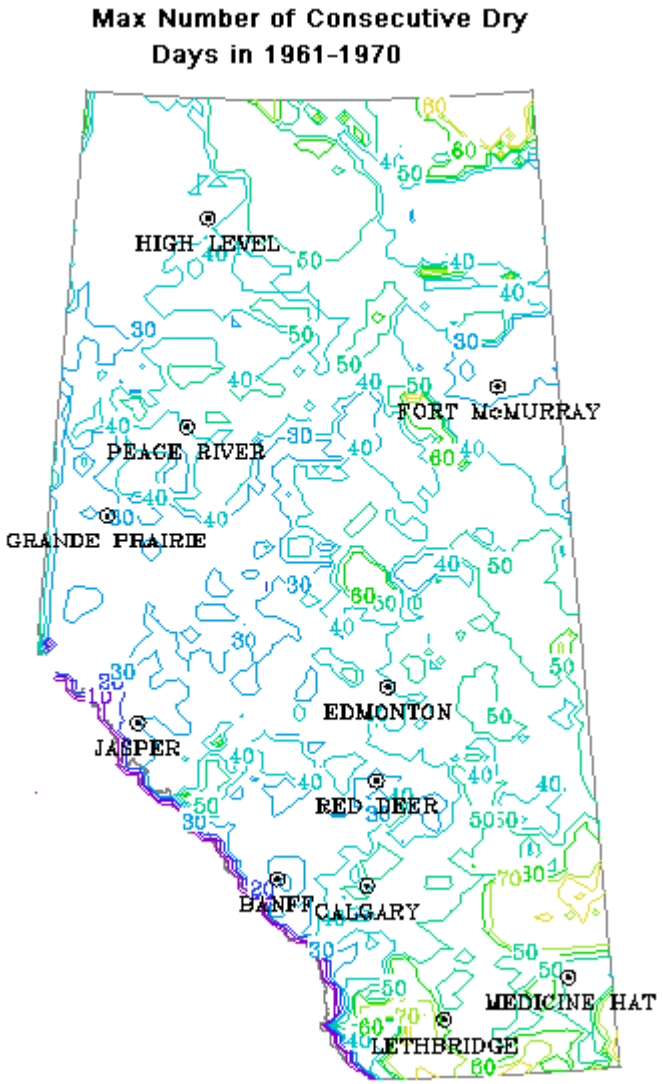


Figure 4.5 Maximum number of consecutive dry days over Alberta during 1961-1970.

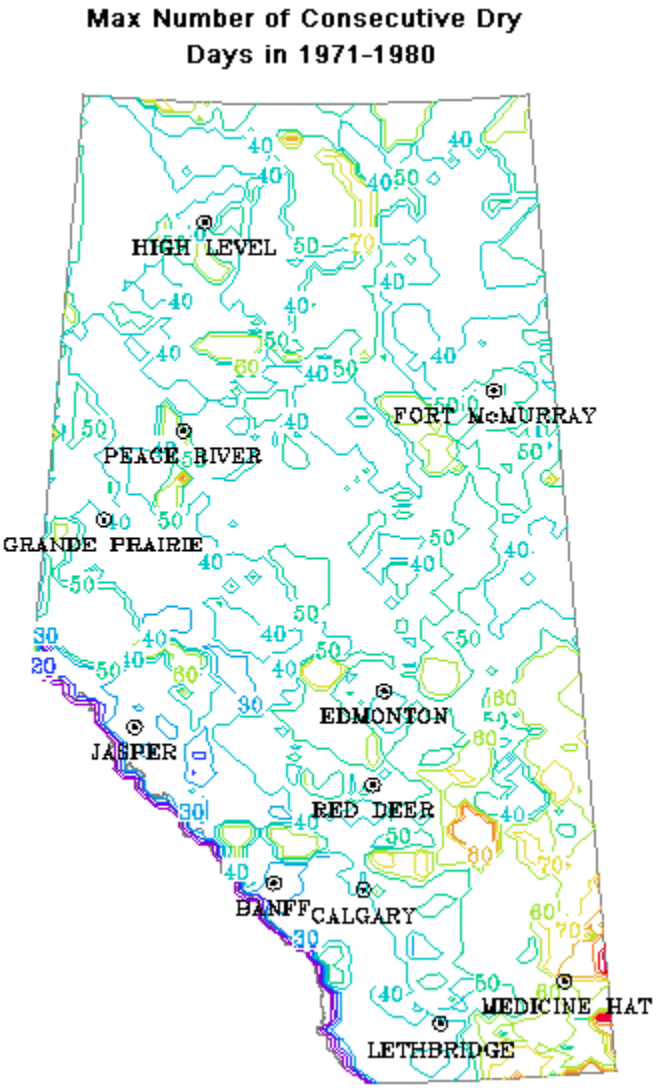


Figure 4.6. Maximum number of consecutive dry days over Alberta during 1971-1980.

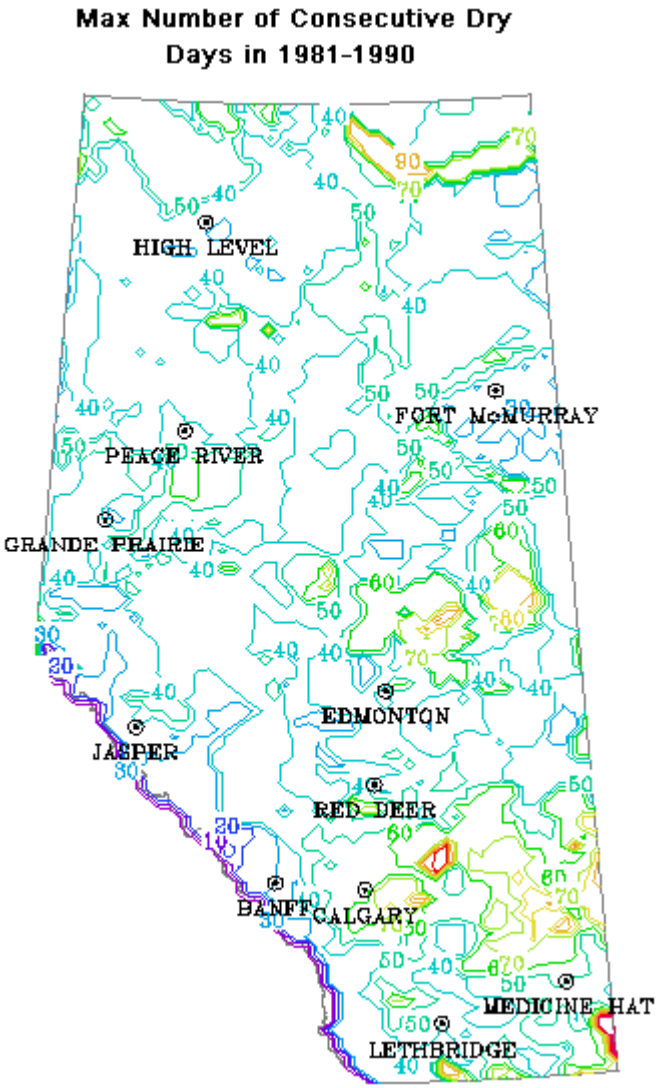


Figure 4.7. Maximum number of consecutive dry days over Alberta during 1981-1990.

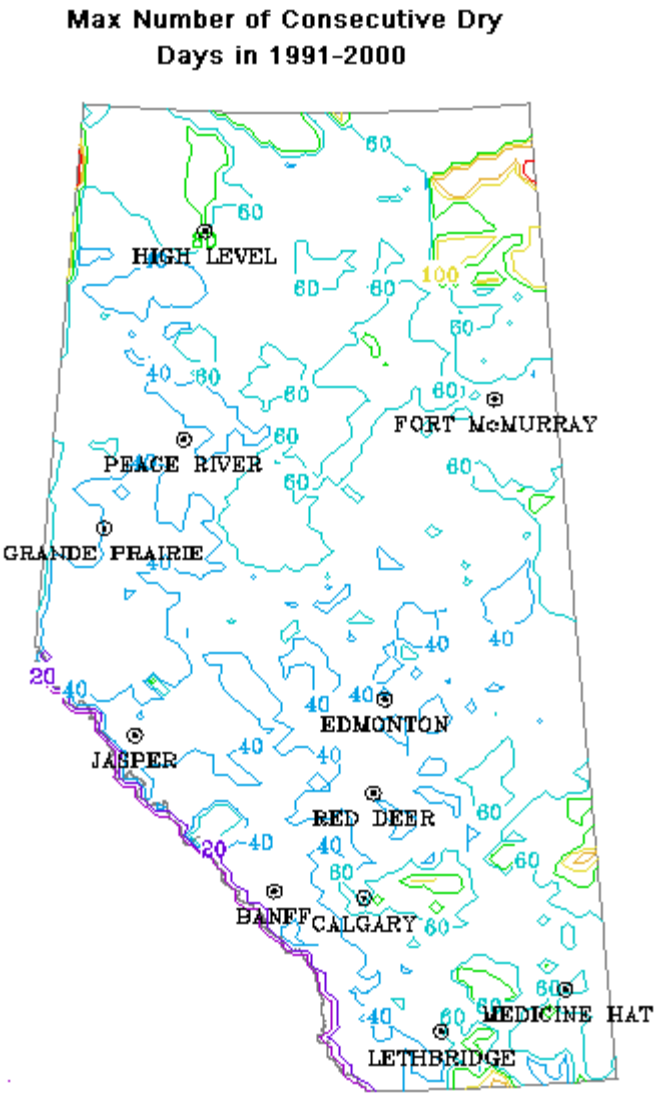


Figure 4.8. Maximum number of consecutive dry days over Alberta during 1991-2000.

4.4 Historical Flood Monitoring: Maps of the Maximum Number of Consecutive Wet Days in 1951-2000

This section studies the maximum number of wet days in 1951 - 2002. As the station instrument's common resolution is 0.2 mm a day, 0.2 mm is used as the wet-day criterion for the calculation here (Shen et al., 2005 (a)). We calculate the maximum number of days with a daily precipitation greater or equal to 0.2 mm for each decade time window and for every grid point. Each grid point thus has a number for every decade. Ten maps are generated for the ten decades: 1951-1960, 1961-1970, …, 1991-2000 (Figs. 4.9 – 4.13). Comparing Figs. 4.9 – 4.13 and Figs. 4.4 – 4.8, we find that the maximum numbers of consecutive wet days are much smaller than those of dry days. Figures 4.4 – 4.8 demonstrate that the maximum number of consecutive wet days is relatively larger in the area along the Rocky Mountains and in part of central Alberta than that in other areas like southern Alberta and northern Alberta. The comparison to the corresponding maps of Saskatchewan and Manitoba in (Shen et al., 2005 (a)) shows that Alberta had larger numbers of consecutive wet days than Saskatchewan and Manitoba during the period of 1951 -2000.

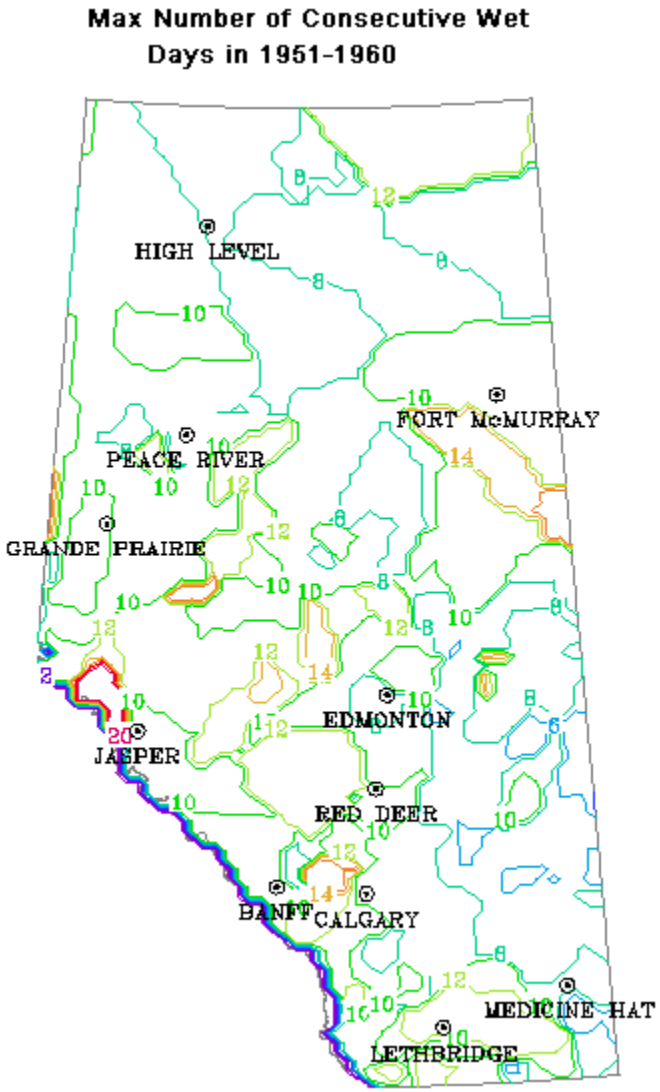


Figure 4.9. Maximum number of consecutive wet days over Alberta during 1951-1960.

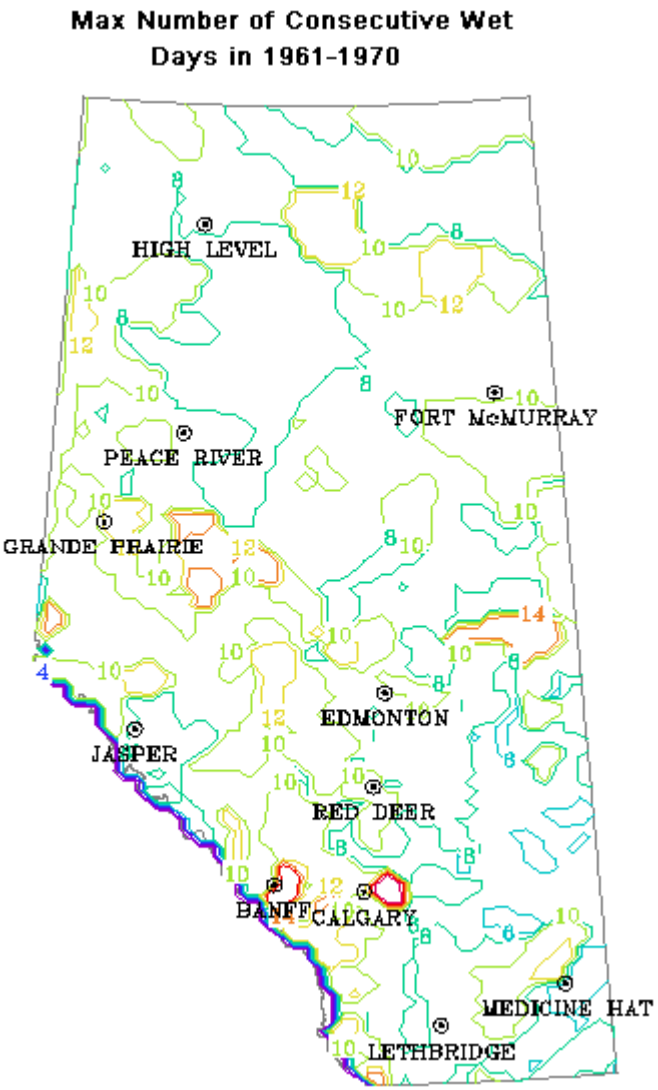


Figure 4.10. Maximum number of consecutive wet days over Alberta during 1961-1970.

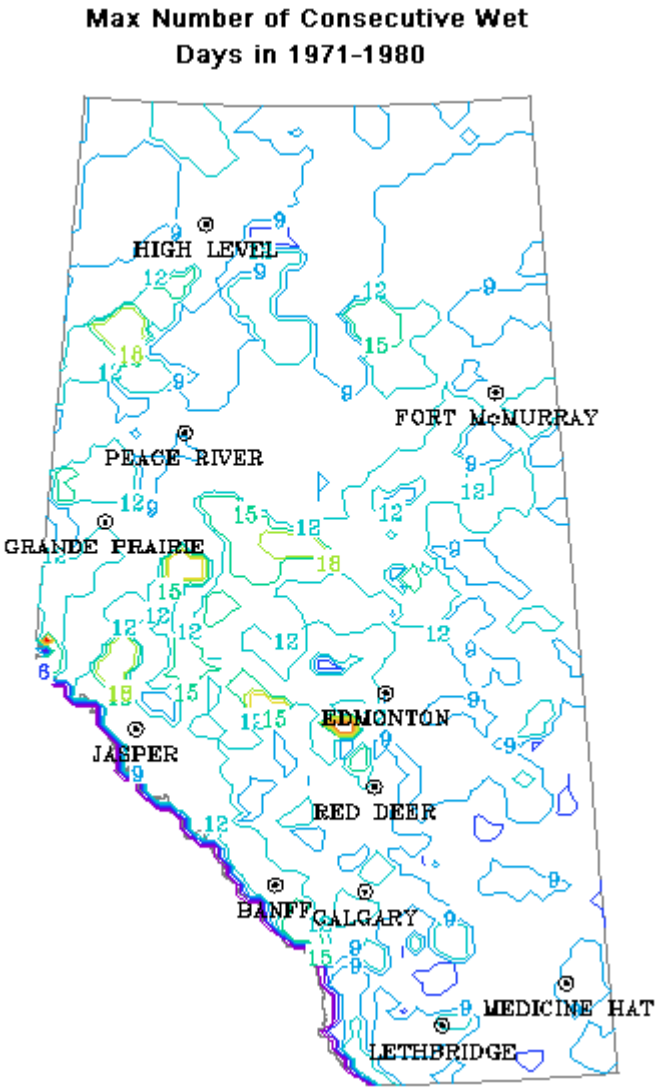


Figure 4.11. Maximum number of consecutive wet days over Alberta during 1971-1980.

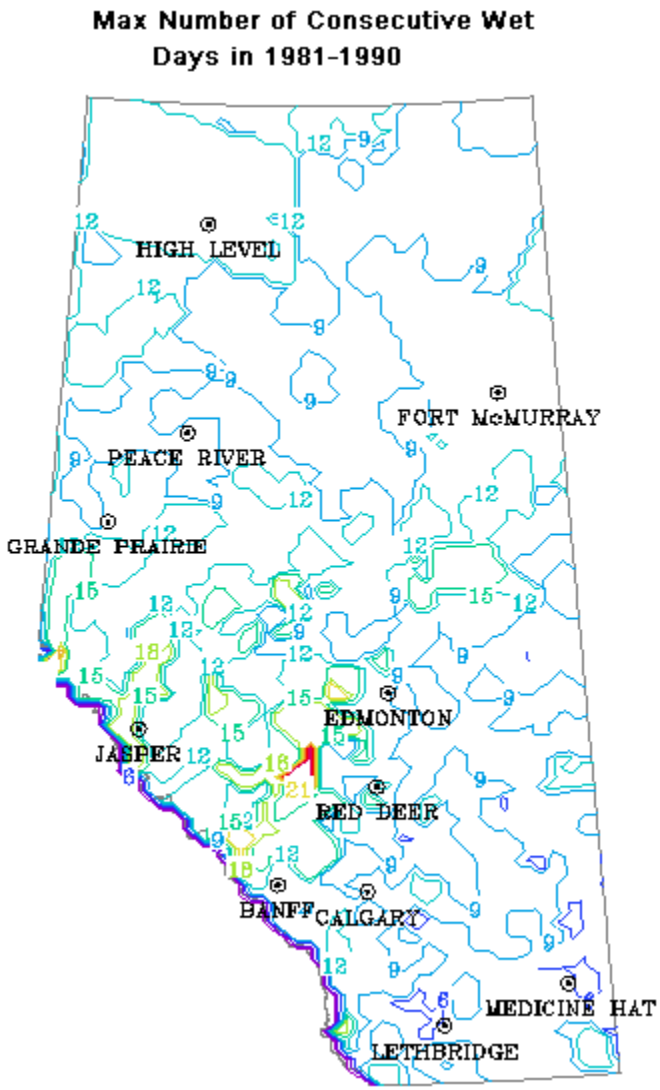


Figure 4.12. Maximum number of consecutive wet days over Alberta during 1981-1990.

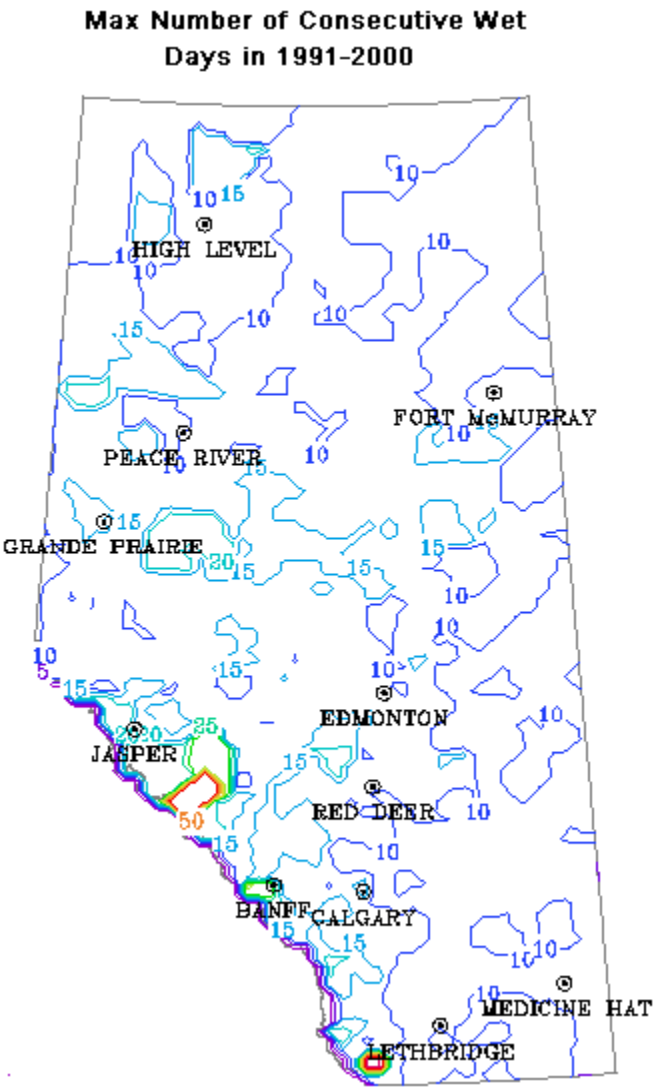


Figure 4.13. Maximum number of consecutive wet days over Alberta during 1991-2000.

4.5 Drought Monitoring: Standardized Precipitation Index Analysis

Precipitation distribution is usually a skewed distribution, not a normal distribution, particularly in a short time scale. Thom (1966) found that the gamma distribution fits climatological precipitation time series well. The gamma distribution is used to construct a "Standardized Precipitation Index" (SPI); i.e., a transformation of precipitation amounts that approximates a standard normal distribution. The gamma probability density function is

$$f(x \mid \alpha, \beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \ (x > 0), \tag{4.6}$$

where $\alpha > 0$ is the shape parameter, $\beta > 0$ is the scale parameter, x is the precipitation amount, and $\Gamma(\alpha)$ is the gamma function which has the following form

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1} e^{-y} dy. \tag{4.7}$$

In the computation of SPI, the parameters α and β in (4.6) need to be estimated for each station, each time scale of interest, and each month of the year. The maximum likelihood estimate of α and β can be obtained by an iterative algorithm. We employ the following simple approximation of the maximum likelihood estimator proposed by Thom (1966):

$$\hat{\alpha} = \frac{1 + \sqrt{1 + 4A/3}}{4A}, \tag{4.8}$$

$$\hat{\beta} = \frac{\overline{x}}{\hat{\alpha}}, \tag{4.9}$$

where

$$A = \ln(\overline{x}) - \sum_{i=1}^{n} \ln(x_i) \Big/ n, \tag{4.10}$$

and n is the length of the precipitation time series. Thus, the cumulative probability of an observed precipitation event for the given month and time scale for the station can be calculated as follows:

$$F(x) = \begin{cases} \int_0^x f(u \mid \alpha, \beta) du = \dfrac{1}{\hat{\beta}^{\hat{\alpha}} \Gamma(\hat{\alpha})} \int_0^x u^{\hat{\alpha}-1} e^{-u/\hat{\beta}} du, & \text{for } x > 0, \\ 0, \quad \text{for } x \le 0. \end{cases} \tag{4.11}$$

Let X be a random variable denoting the amount of precipitation at a specified location over a specified period of time. There is a probability q > 0 of no precipitation. The Gamma cumulative distribution function F is used to model the conditional distribution of X given that X is positive. Let Φ be the cumulative distribution function for the standard normal distribution. A simple calculation shows that the conditional distribution $\Phi^{-1}(F(X))$ given $X > 0$ is standard normal. This result assumes only that F is a continuous cumulative distribution function. The SPI is obtained using a similar transformation for the unconditional distribution of X ; i.e.,

$$H(x) = \begin{cases} q + (1-q)F(x), & \text{for } x \ge 0, \\ 0, & \text{for } x < 0. \end{cases} \tag{4.13}$$

The unconditional distribution has a discontinuity at $x = 0$, so it is not possible to

obtain a transformation yielding exact normality. Thom (1966) introduced a transformation that approximates a standard normal. The detailed transformation and approximation was provided by Abramowitz and Stegun (1965) on Page 933 as following:

$$SPI = \begin{cases} -\left(t - \frac{c_0 + c_1 t + c_2 t^2}{1 + d_1 t + d_2 t^2 + d_3 t^3}\right), & \text{when } 0 < H(x) \le 0.5, \\ \left(t - \frac{c_0 + c_1 t + c_2 t^2}{1 + d_1 t + d_2 t^2 + d_3 t^3}\right), & \text{when } 0.5 < H(x) < 1, \end{cases} \tag{4.14}$$

where

$$t = \begin{cases} \sqrt{\ln\left(1/H(x)^2\right)}, & \text{when } 0 < H(x) \le 0.5, \\ \sqrt{\ln\left(1/(1-H(x))^2\right)}, & \text{when } 0.5 < H(x) < 1, \end{cases} \tag{4.15}$$

$$\begin{aligned} c_0 &= 2.515517, \\ c_1 &= 0.802853, \\ c_2 &= 0.010328, \\ d_1 &= 1.432788, \\ d_2 &= 0.189269, \\ d_3 &= 0.001308. \end{aligned}$$

McKee et al. (1993) defined the criteria for a "drought event", which, occurs any time the SPI value is continuously negative and reaches -1.0 or less. The event ends when the SPI value becomes positive. Each drought event, therefore, has a duration defined by its beginning and end, and an intensity for each month that the event continues. The drought and flood levels determined by the SPI value are displayed in Table 4.2. The ABClim2.0 data on the nearest grid point of each station (Edmonton Int'l A, Calgary Int'l A, High Level A, and Peace River A) are used here to compute the SPI values and create the figures 4.14 – 4.21.

Figures 4.14, 4.16, 4.18, and 4.20 show that the SPI values for the time scale of 3 months and 6 months have larger fluctuations of positive or negative but smaller fluctuations for 12-month and 24-month time scales. The SPI value for different time scales has different sensitivity to the precipitation amount. As the time scale becomes smaller, the SPI value has a larger change, even in the fluctuation of positive or negative for the changes in one month's precipitation amount. In contrast, for the larger time scales, the SPI value responds more slowly to one month's precipitation variation. Only many periods of continuous precipitation can make the SPI value fluctuate. Thus, there are fewer but longer periods with negative and positive SPI value. The SPI characteristics for the different time scales are highlighted by the results presented in Figures 4.14, 4.16, 4.18, and 4.20. These results agree with the observations reported by McKee et al. (1993). Thus, it is reasonable to use SPI for the larger time scales for monitoring the long-term water status (McKee et al., 1993 and Seiler et al., 2002).

Figures 4.15, 4.17, 4.19 and 4.21 demonstrate the comparison of the SPI values with the precipitation for the period of May to August in Edmonton, Calgary, Peace River and High Level, respectively. For Edmonton, the SPI can capture 19 drought events, which are in 1917, 1919, 1922, 1929, 1931, 1939, 1945, 1948, 1950, 1957, 1958, 1961, 1963, 1967, 1968, 1969, 1984, 1992 and 2002. As well, the drought events of 1950, 1967 and 2002 are extreme dry, as the SPI values for these years are less than -2.0. SPI can capture two extremely wet events as well, which were in 1901 and 1954. There were 494 mm and 451.5 mm precipitation in total for the period of May to August in those two years,

respectively (Fig 4.15).

Table 4.2. Classification scales for the SPI values.

SPI Values	**Drought and Flood Level**
2.0+	extremely wet
1.5 to 1.99	very wet
1.0 to 1.49	moderately wet
-.99 to .99	near normal
-1.0 to -1.49	moderately dry
-1.5 to -1.99	severely dry
-2.0 and less	extremely dry

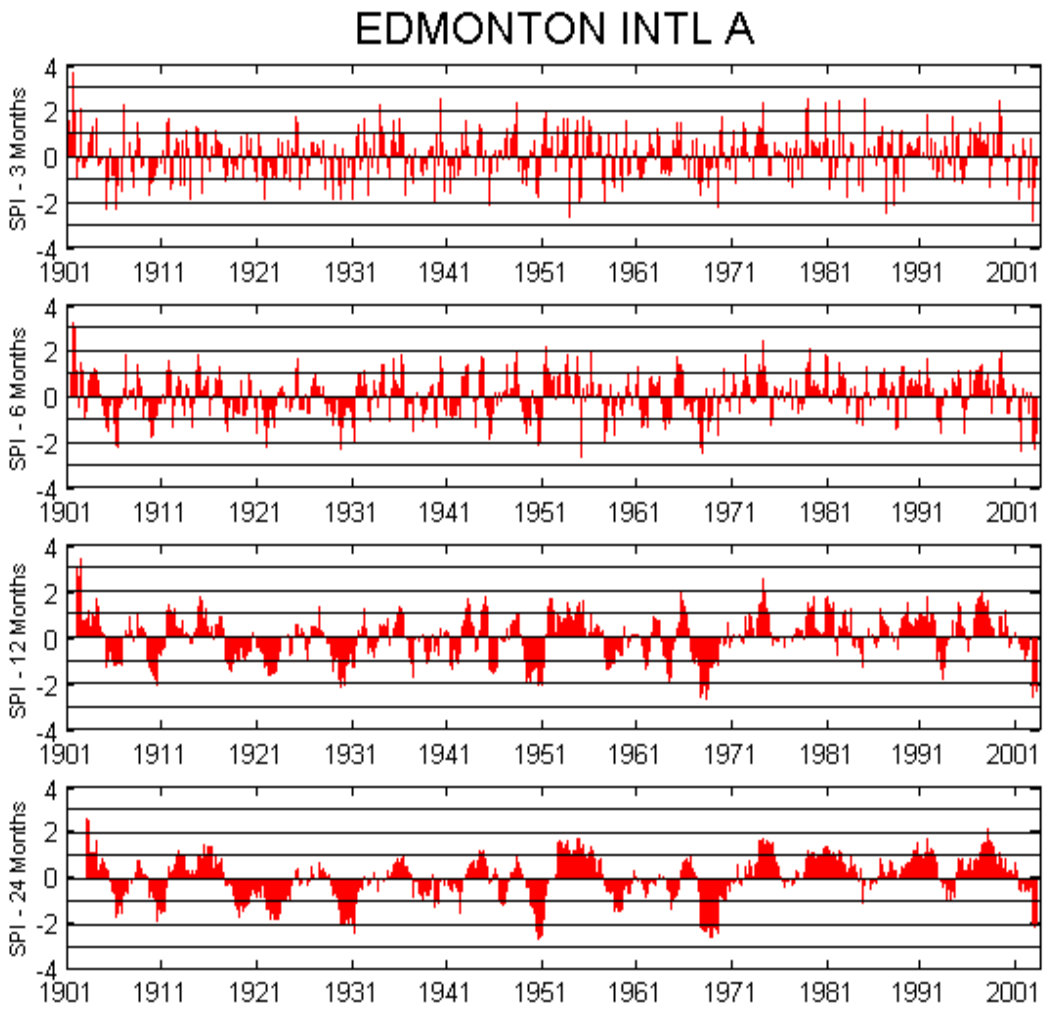


Figure 4.14. SPI values calculated from the 1901 – 2002 precipitation time series for different time scales at Edmonton Int'l A station.

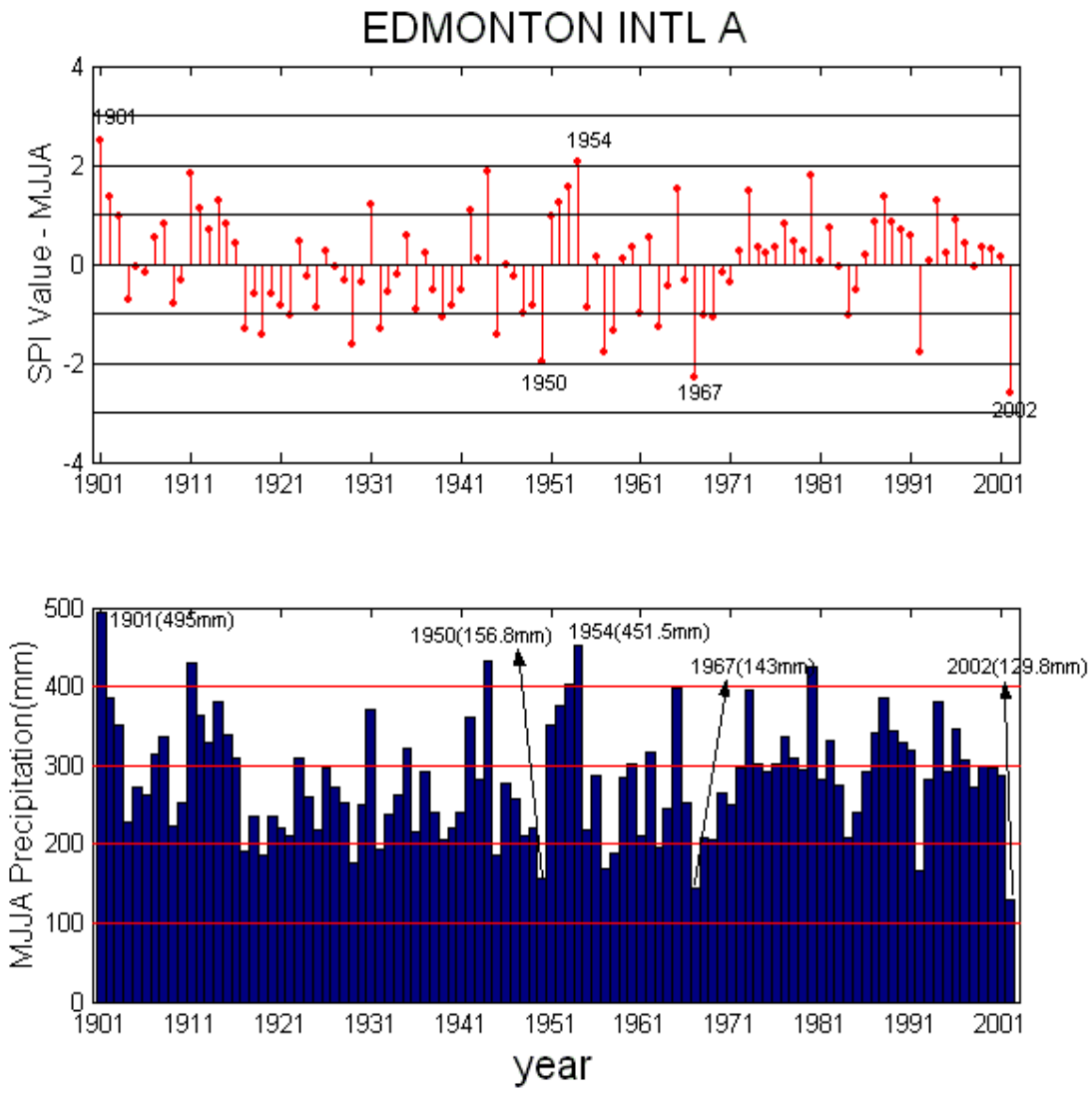


Figure 4.15. SPI values calculated from the 1901 – 2002 MJJA precipitation time series VS MJJA precipitation at Edmonton Int'l A station.

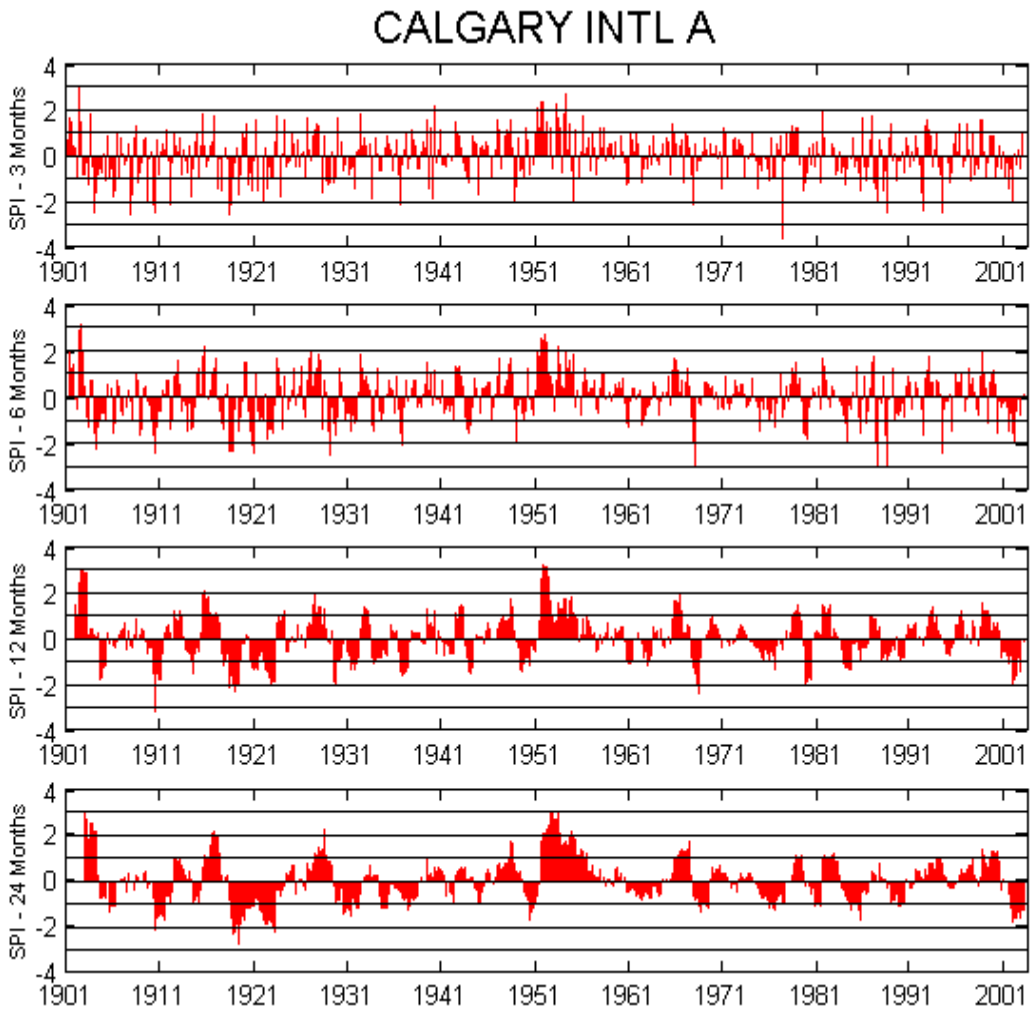


Figure 4.16. SPI values calculated from the 1901 – 2002 precipitation time series for different time scales at Calgary Int'l A station.

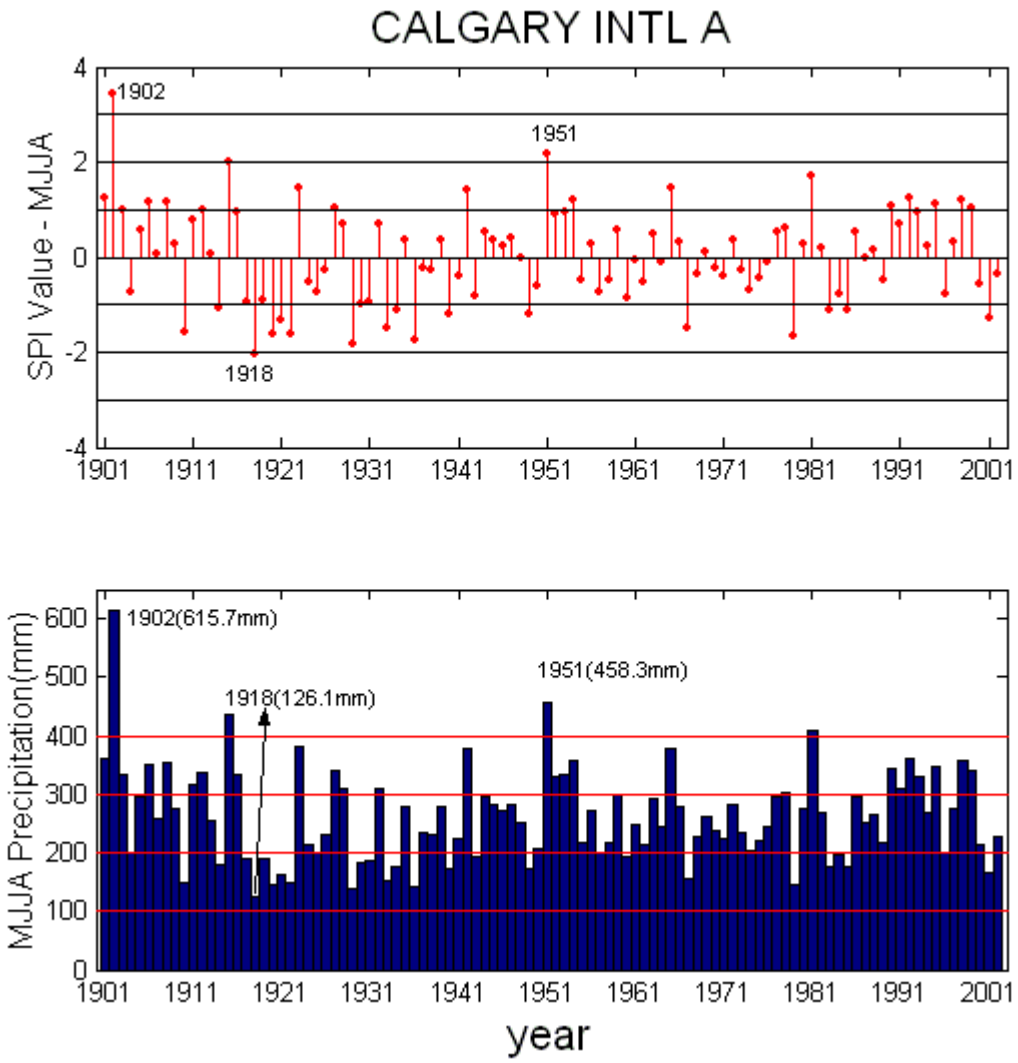


Figure 4.17. SPI values calculated from the 1901 – 2002 MJJA precipitation time series VS MJJA precipitation at Calgary Int'l A station.

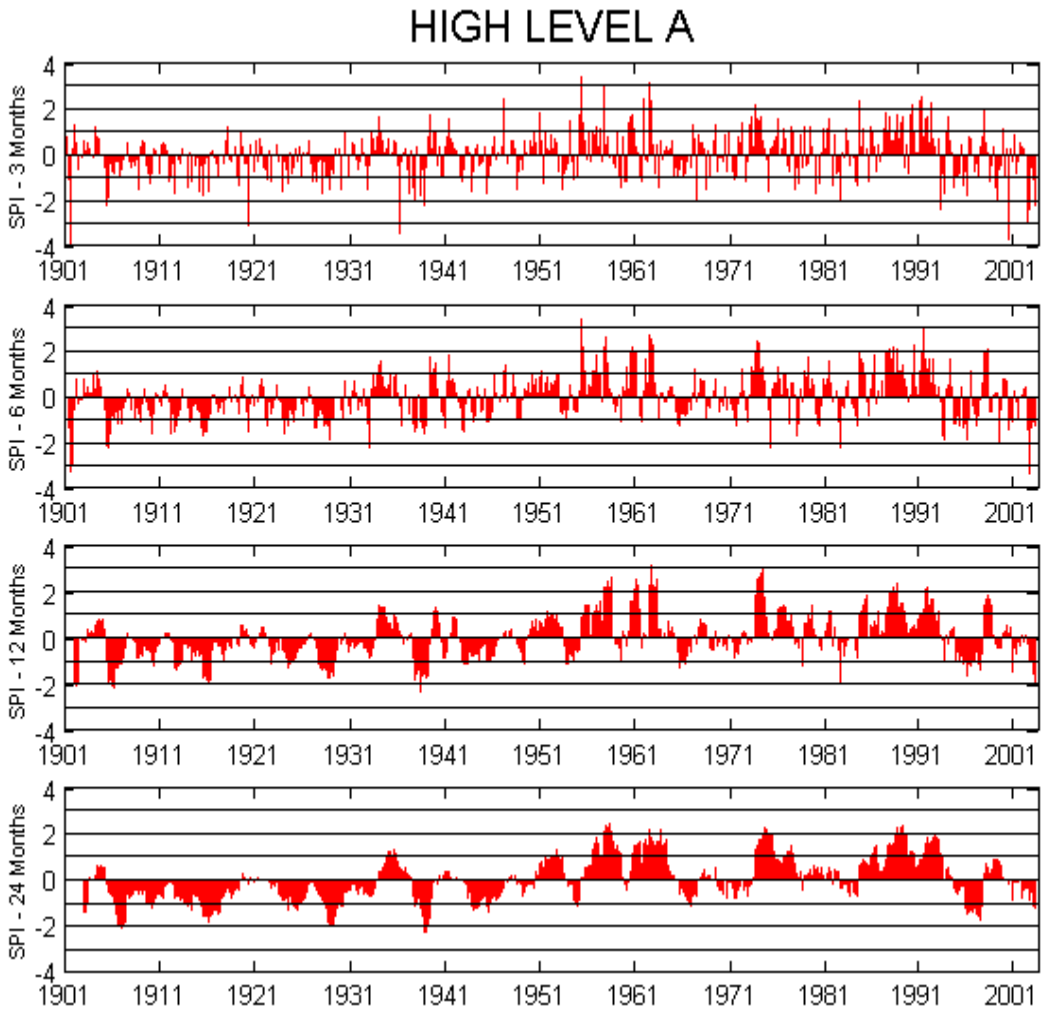


Figure 4.18. SPI values calculated from the 1901 – 2002 precipitation time series for different time scales at High Level A station.

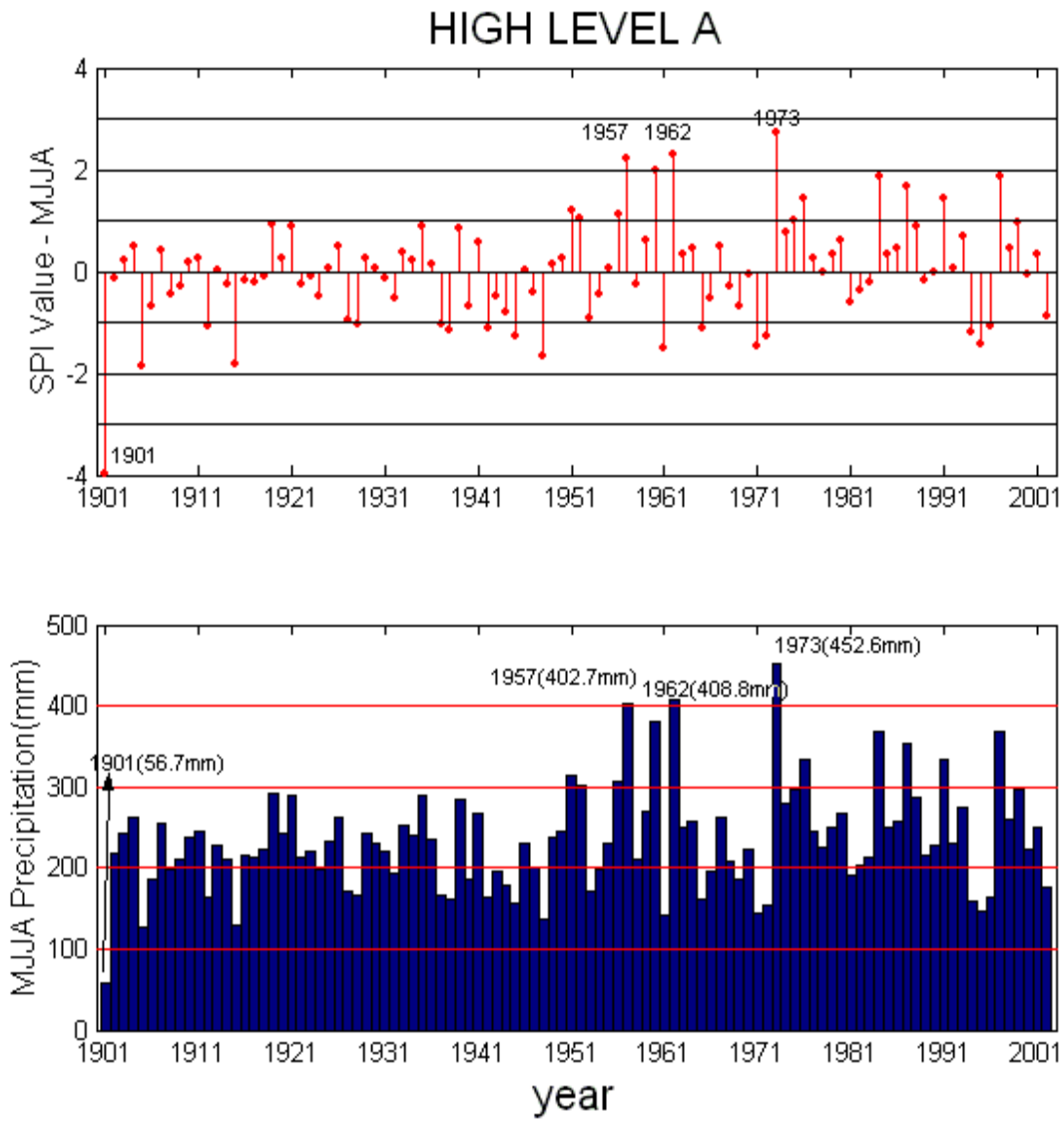


Figure 4.19. SPI values calculated from the 1901 – 2002 MJJA precipitation time series VS MJJA precipitation at High Level A station.

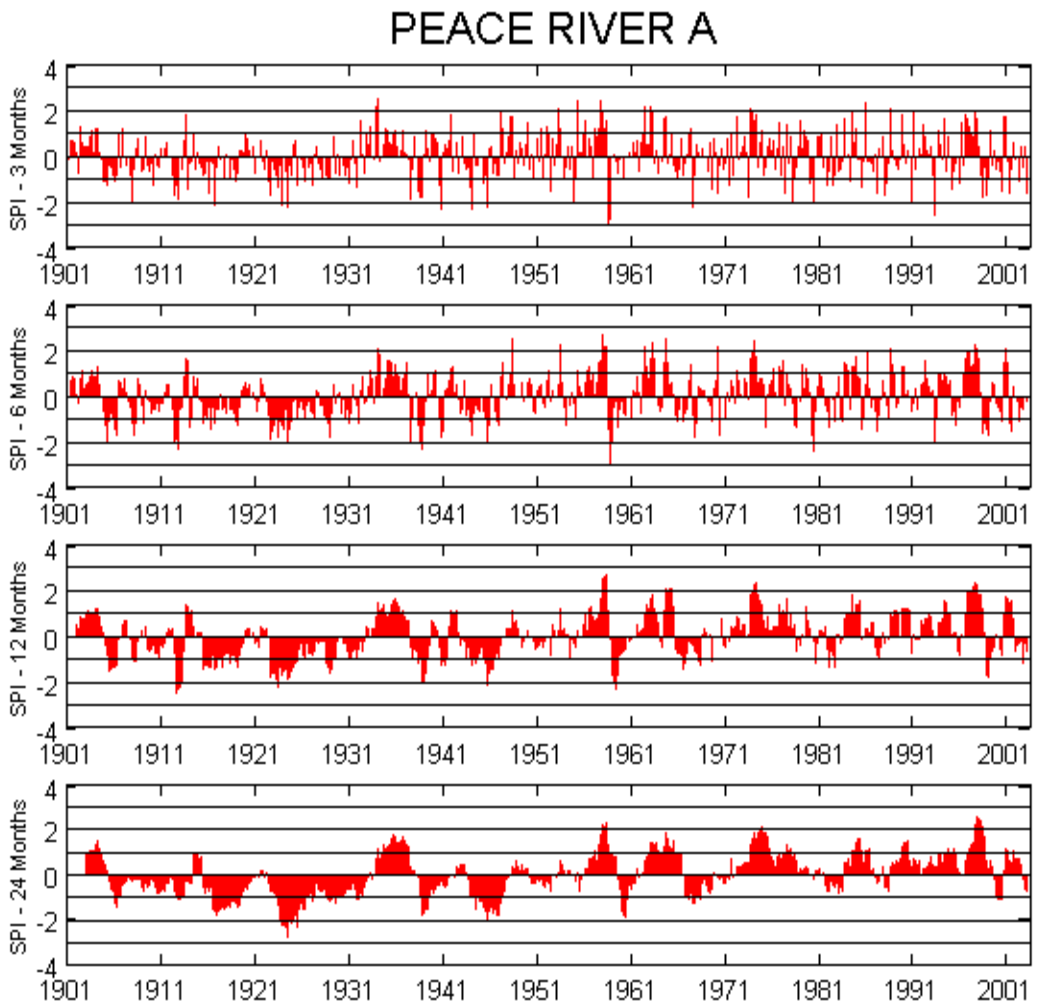


Figure 4.20. SPI values calculated from the 1901 – 2002 precipitation time series for different time scales at Peace River A station.

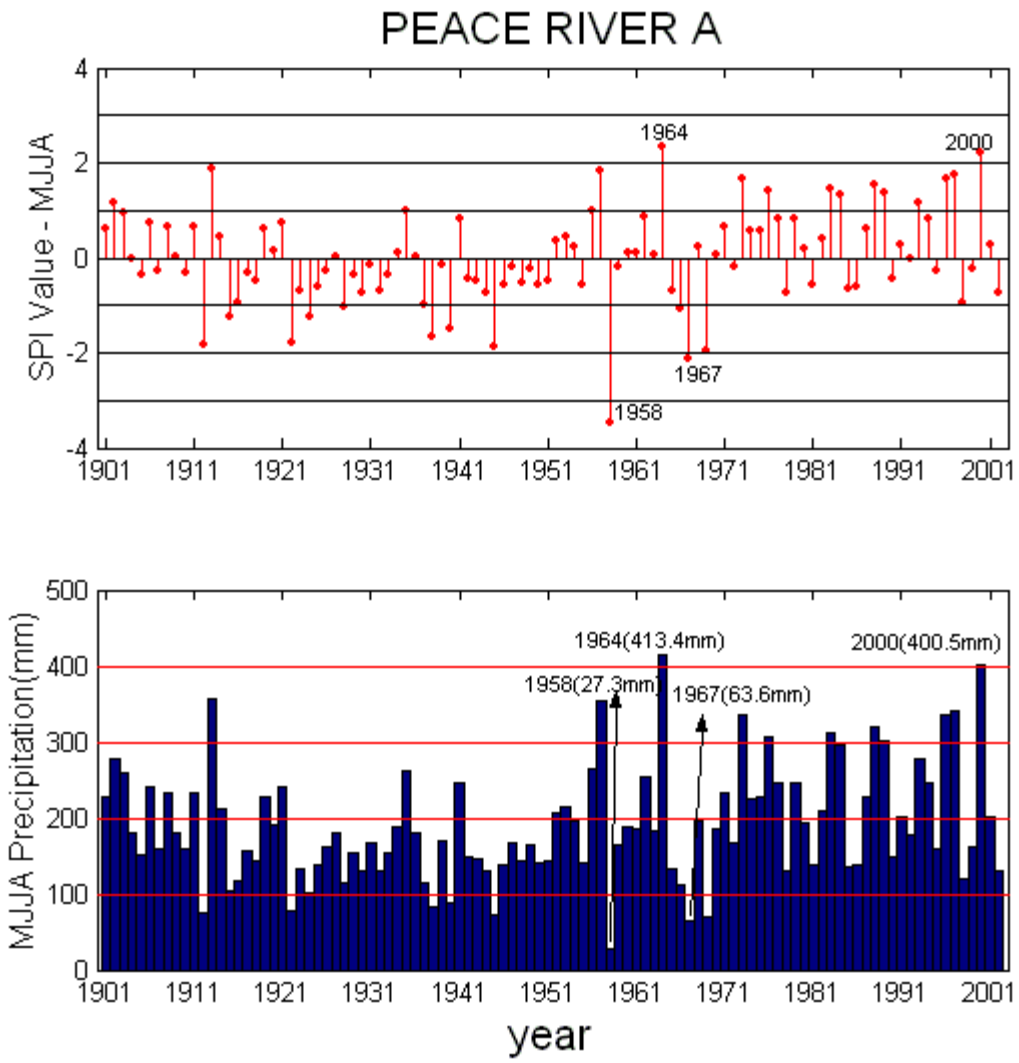


Figure 4.21. SPI values calculated from the 1901 – 2002 MJJA precipitation time series VS MJJA precipitation at Peace River A station.

At Calgary, 15 droughts are captured by SPI and the droughts of 1910, 1918, 1920, 1922, 1929, 1933, 1936, 1967 and 1979 are severe events, as the corresponding SPI values are less than -1.50. The SPI also captures some of the extremely wet events of 1902 and 1951, with MJJA precipitation of 615.7 mm and 458.3 mm, respectively (Fig 4.17). Figure 4.19 represents some of the severe droughts of 1901, 1905, 1915, 1948 and 1961 ($SPI \leq -1.50$) and the very wet events of 1957, 1960, 1962, 1973, 1984, 1987 and 1997 ($SPI \geq 1.50$) at High Level. The drought event of 1901 was extremely dry, as its MJJA precipitation was only 56.7 mm, and the corresponding SPI value is -3.97. The wet events of 1957, 1962 and 1973 are extremely wet, as their MJJA precipitations were over 400 mm, and the corresponding SPI values are greater than 2.0. For Peace River, the SPI captures some of the historic severe drought events of 1912, 1922, 1938, 1940, 1945, 1958, 1967 and 1969 and the historic wet events of 1913, 1957, 1964, 1973, 1988, 1996, 1997 and 2000, among which there are extremely droughts in 1958 (the MJJA precipitation was 27.3 mm and its corresponding SPI is even less than -3.0) and 1967 (the MJJA precipitation was 63.6 mm and its corresponding SPI is less than -2.0), as well as extremely wet events in 1964 and 2000 (the MJJA precipitation was over 400 mm, and the corresponding SPI values are larger than 2.0 in these two years).

4.6 Conclusions

A trend analysis of the MJJA total precipitation time series for the period of 1901 to 2002 at some stations in Alberta like Edmonton and Calgary was been

carried out by using the linear regression and moving average methods. In this study, the linear regression analysis showed that a slight decreasing trend for Edmonton and Calgary and an obvious increasing trend for Peace River and High Level in the MJJA total precipitation time series from 1901 to 2002. The detailed increasing or decreasing trend in each shorter time period was given by the 11-year moving average. The graph of the moving average also showed the relatively dry and relatively wet periods, respectively. In this study, we also detected the trend in the MJJA total precipitation for the whole period from 1901 to 2002 at some stations in Alberta by using the Mann-Kendall test at the 5% significance level. Significant upward trends were found at High Level and Peace River, and no significant trends were found at Edmonton and Calgary. The results of the Mann-Kendall test were consistent with those of the linear regression. However, the Mann-Kendall test is still just a statistical method for processing data although it has been applied in many cases and been found to be an excellent tool for trend detection. Theoretical research work is needed in order to learn the physical meaning of the temporal and spatial patterns of the climatic data.

The maps of the maximum consecutive dry days for each decade were generated over Alberta to give the spatial distribution of the severe droughts in every decade. Southern Alberta had the largest number of maximum consecutive dry days and the smallest number of maximum consecutive wet days. Both results are consistent with each other and show that southern Alberta is drier than other areas in Alberta.

In this chapter, we also used SPI to assess the Hybrid 2.0 data and then monitor the past drought conditions at some stations in Alberta. Most of the severely dry or wet events were captured by the SPI assessment. The advantage of using SPI is that its calculation is simple and it is easy to obtain the information needed in the calculation. One superior characteristic of SPI is that it can be computed at multi-time scales and, thereby, meet the needs of different water resource assessments and drought monitoring services. Thus, SPI provides a uniform drought indicator for various departments and areas like water-resources assessment and different time scales for drought-monitoring services.

Chapter 5

Discussions on EOF Modes and the Sensitivity of Data Errors

5.1 Introduction

Generally, a reduced set of EOFs is able to describe most of the variance in a data set, in particular those sets with the largest eigenvalues. The EOF method will fail to represent important characteristics of the low-frequency variability if too few EOFs are selected. As well, if too many EOFs are retained, the statistical significance will be compromised due to the noise. The ability to represent a physical signal by a reduced set of EOFs, and to preserve the signal's physical significance is the main advantage of using EOF analysis in practice, where such a reduction can significantly diminish the size of the algebraic calculations involved.

Due to the fundamental characteristic of EOF analysis, the problem in using EOFs is how to distinguish the EOFs which represent physically significant modes from those purely noisy EOFs and then to retain the significant EOFs only. In general, three groups of methods are used for selecting the significant EOFs depending on whether they focus on the eigenvalues, the principal components or the EOFs. The first group contains the methods based on the amount of the data

variance explained, or on the size of the eigenvalues. North's rule of thumb (North et al. 1982) belongs to this group. The second group includes the methods based on statistical hypotheses for separating signal and noise eigenvalues. The most often used of these methods is "Rule N" (Preisendorfer and Barnett, 1977; Overland and Preisendorfer, 1982; Preisendorfer, 1988). However, due to its statistical nature, this selection rule could fail when the sample size is small. The third group includes the spatial map methods based on the examination of the EOF patterns that should resemble some predefined or true signal mode patterns.

In our EOF-based linear interpolation, the random measurement error E_j for a given location $\mathbf{r}_j$ and time t are included. The error variance $\left\langle E_j^2 \right\rangle$ is required in Eq. (2.48) for computing the optimal weights. Because the exact value of the random measurement error E_j is not available, the estimated error variance is used in our study. It is thus necessary to test if the optimal weights are sensitive to the exact size of the error variance.

In this chapter, the Rule N and North's rule of thumb are tested for selecting the significant EOFs, and the method for choosing the optimal set of EOF based on the EOF reconstruction is discussed. We also test the sensitivity of the EOF-based linear reconstruction to the size of the error variance. The organization of this chapter is as follows. In Section 5.2, we discuss the rules for selecting the significant EOFs and apply them to the monthly total precipitation data in the period of 1961-2000. In Section 5.3, the optimal set of EOFs in the EOF reconstruction is selected for the same data set and time period. In Section 5.4, we

test the sensitivity of the EOF-based linear reconstruction to the size of the error variance. Our conclusions and discussion are presented in Section 5.5.

5.2 Significance Test of EOF Modes

As we discussed above, several methods can be used to identify the significant EOFs. The Rule N and the North's rule of thumb will be discussed here.

5.2.1 Rule N Test

The Rule N technique for EOF mode selection is a dominant-variance rule which selects eigenvalues for which the geophysical signal is above the level of noise. The technique is based on a Monte Carlo experiment which simulated the sampling data from the standard normal distribution. The procedure of this significance test is as follows. First, we generate a floating-point, pseudo-random $N \times P$ data matrix R with a standard normal distribution by a random generator. Second, the covariance matrix $C = R^T R/N$ and its eigenvalues $d^{'}s$ are computed and sorted in descending order. Third, the above two steps are repeated one hundred times. Fourth, we let d_j^r denote the jth $(j = 1, 2, \cdots, P)$ eigenvalue in the rth $(r = 1, 2, \cdots, 100)$ Monte Carlo experiment. We compute the normalized eigenvalues produced by the rth Monte Carlo experiment as

$$U_j^r = d_j^r \Big/ \sum_{i=1}^{P} d_i^r \,,\ j = 1, 2, \cdots, P, \tag{5.1}$$

which are sorted in ascending order; that is, for fixed j, we have $U_j^1 \le U_j^2 \le \cdots \le U_j^{100}$. Fifth, we compute the eigenvalues λ_j $(j = 1, 2, \cdots, P)$ of the covariance matrix of the real data and sort them in descending order: $\lambda_1 > \lambda_2 > \cdots > \lambda_P$. Also, the normalized eigenvalues are computed as

$$T_j = \lambda_j \Big/ \sum_{i=1}^{P} \lambda_i, \; j = 1, 2, \cdots, P. \tag{5.2}$$

Finally, for each $j = 1, 2, \cdots, P$, the value T_j is compared with the value U_j^{95}. This latter value provides an estimate of the 95th percentile of the distribution of the jth eigenvalue when the data are white noise (no signal). The Rule N gives us a number M_c which is the largest integer m such that $T_m > U_m^{95}$. It means that the first M_c eigenvalues are selected for which the geophysical signal is above the level of noise, and the corresponding M_c EOF modes may contain physical meaning which can be interpreted. We thus retain only the eigenvalues for which the ratio $T_j / U_j^{95} > 1$.

Overland and Preisendorfer (1982) presented a table of U_j^{95} values for different values of N and P which showed that the values of U_j^{95} become smaller as the values of N and P become larger. Thus, if the length of the time series and the number of stations are smaller, the values of U_j^{95} will be larger so that fewer eigenvalues will be retained. In this case, some eigenvalues may not be retained for which the ratio T_j / U_j^{95} is not large enough to have physical meaning,

but they do contribute to the significant percentage of the total variance. The data record should therefore be reasonably long in an EOF analysis.

Next, we apply the Rule N test to the monthly precipitation over Alberta for the period of 1961 - 2000. The Hybrid 1.0 data on the 0.25-by-0.5 degrees latitude and longitude grid and the 0.1-by-0.2 degrees latitude and longitude grid are used to this application, respectively. That is, we have 40 years' data on 808 and 4862 grid points, respectively. Tables 5.1 and 5.2 contain the normalized eigenvalues of the above two data sets. Rule N is satisfied with values of $T_j / U_j^{95} > 1.0$.

Therefore, for the resolution of the 0.25-by-0.5 degrees latitude and longitude grids, only the first 6 modes in January (representing 76% of the total variance), March (69%), April (72%), May (76%), June (77%), July (72%), August (75%), October (72%) and November (77%); the first 5 modes in February (72%) and September (77%); and the first 7 modes in December (76%) can be interpreted as signals. For the resolution of the 0.1-by-0.2 degrees latitude and longitude grids, only the first 8 modes in January (82%) and April (78%); the first 7 modes in March (73%), May (79%), August (78%), October (76%) and December (76%); and the first 6 modes in February (76%), June (77%), July (73%), September (81%) and November (77%) can be interpreted as signals. We can see that in this EOF analysis, the number of EOF modes which pass the Rule N test satisfy the hoc rule of thumb, that is, explain over 70% of the total variance for the analyzed data set.

In applying Rule N to EOF analysis, some eigenvalues may not pass the Rule N test but do represent a significant percentage of the total variance. Another

Table 5.1. Normalized EOF eigenvalues for 1961-2000 Hybrid 1.0 data set on the 0.25-by-0.5 degrees latitude and longitude grids: N=40, P=808 (Here T_j and U_j^{95} are multiplied by 100.). The boldface values in the table indicate the smallest values of $T_j / U_j^{95} > 1.0$.

Month		j=1	j=2	j=3	j=4	j=5	j=6	j=7	j=8	j=9
JAN	T_j	40.84	14.43	7.68	5.33	4.00	3.69	3.21	2.83	2.08
	U_j^{95}	3.83	3.67	3.57	3.47	3.40	3.31	3.24	3.18	3.11
	T_j / U_j^{95}	10.66	3.93	2.15	1.54	1.18	**1.11**	0.99	0.89	0.67
FEB	T_j	43.31	5.08	5.63	4.63	3.81	3.08	2.69	2.26	1.95
	U_j^{95}	3.83	3.67	3.57	3.47	3.40	3.31	3.24	3.18	3.11
	T_j / U_j^{95}	11.31	4.11	1.58	1.33	**1.12**	0.93	0.83	0.71	0.63
MAR	T_j	30.91	16.87	6.86	5.59	5.04	4.16	3.17	2.81	2.61
	U_j^{95}	3.83	3.67	3.57	3.47	3.40	3.31	3.24	3.18	3.11
	T_j / U_j^{95}	8.07	4.60	1.92	1.61	1.48	**1.26**	0.98	0.88	0.84
APR	T_j	26.93	17.30	9.67	8.79	5.58	3.68	3.09	2.85	2.35
	U_j^{95}	3.83	3.67	3.57	3.47	3.40	3.31	3.24	3.18	3.11
	T_j / U_j^{95}	7.03	4.71	2.71	2.53	1.64	**1.11**	0.95	0.90	0.76
MAY	T_j	34.80	20.20	6.95	5.55	4.91	3.54	2.83	2.45	2.01
	U_j^{95}	3.83	3.67	3.57	3.47	3.40	3.31	3.24	3.18	3.11
	T_j / U_j^{95}	9.09	5.50	1.95	1.60	1.44	**1.07**	0.87	0.77	0.65
JUN	T_j	29.47	18.28	13.46	6.89	5.14	3.64	2.75	2.51	1.78
	U_j^{95}	3.83	3.67	3.57	3.47	3.40	3.31	3.24	3.18	3.11

	T_j/U_j^{95}	7.69	4.98	3.77	1.99	1.51	**1.10**	0.85	0.79	0.57
JUL	T_j	22.29	19.91	12.85	7.69	5.06	4.64	2.67	2.53	2.41
	U_j^{95}	3.83	3.67	3.57	3.47	3.40	3.31	3.24	3.18	3.11
	T_j/U_j^{95}	5.82	5.43	3.60	2.21	1.49	**1.40**	0.82	0.80	0.77
AUG	T_j	37.47	14.59	8.78	6.03	4.30	3.44	2.95	2.61	1.97
	U_j^{95}	3.83	3.67	3.57	3.47	3.40	3.31	3.24	3.18	3.11
	T_j/U_j^{95}	9.78	3.98	2.46	1.74	1.26	**1.04**	0.91	0.82	0.63
SEP	T_j	41.42	17.48	7.13	6.18	4.73	3.31	2.65	2.33	1.77
	U_j^{95}	3.83	3.67	3.57	3.47	3.40	3.31	3.24	3.18	3.11
	T_j/U_j^{95}	10.81	4.76	2.0	1.78	**1.39**	1.0	0.82	0.73	0.60
OCT	T_j	32.36	15.33	8.33	6.86	5.59	4.01	2.90	2.66	2.23
	U_j^{95}	3.83	3.67	3.57	3.47	3.40	3.31	3.24	3.18	3.11
	T_j/U_j^{95}	8.45	4.18	2.33	1.98	1.64	**1.21**	0.90	0.84	0.72
NOV	T_j	41.29	15.55	7.26	4.60	4.31	3.60	2.77	2.02	1.90
	U_j^{95}	3.83	3.67	3.57	3.47	3.40	3.31	3.24	3.18	3.11
	T_j/U_j^{95}	10.78	4.24	2.03	1.33	1.27	**1.09**	0.85	0.64	0.61
DEC	T_j	32.45	17.75	7.03	5.91	5.26	3.77	3.57	2.68	2.29
	U_j^{95}	3.83	3.67	3.57	3.47	3.40	3.31	3.24	3.18	3.11
	T_j/U_j^{95}	8.47	4.84	1.97	1.70	1.55	1.14	**1.10**	0.84	0.74

Table 5.2. Normalized EOF eigenvalues for 1961-2000 Hybrid 1.0 data set on the 0.1-by-0.2 degrees latitude and longitude grids: N=40, P=4862 (Here T_j and U_j^{95} are multiplied by 100.). The boldface values in the table indicate the smallest values of $T_j / U_j^{95} > 1.0$.

Month		j=1	j=2	j=3	j=4	j=5	j=6	j=7	j=8	j=9
JAN	T_j	41.31	14.30	7.71	5.26	4.13	3.64	3.12	2.82	2.05
	U_j^{95}	3.04	2.99	2.95	2.92	2.88	2.86	2.83	2.81	2.79
	T_j / U_j^{95}	13.59	4.78	2.61	1.80	1.43	1.27	1.10	**1.01**	0.73
FEB	T_j	43.67	15.17	5.59	4.56	3.77	3.10	2.72	2.21	1.90
	U_j^{95}	3.04	2.99	2.95	2.92	2.88	2.86	2.83	2.81	2.79
	T_j / U_j^{95}	14.37	5.07	1.89	1.56	1.31	**1.08**	0.96	0.79	0.68
MAR	T_j	31.27	16.92	6.86	5.61	4.97	4.13	3.20	2.77	2.63
	U_j^{95}	3.04	2.99	2.95	2.92	2.88	2.86	2.83	2.81	2.79
	T_j / U_j^{95}	10.29	5.66	2.33	1.92	1.73	1.44	**1.13**	0.99	0.94
APR	T_j	27.28	17.19	9.56	8.81	5.56	3.63	3.12	2.84	2.29
	U_j^{95}	3.04	2.99	2.95	2.92	2.88	2.86	2.83	2.81	2.79
	T_j / U_j^{95}	8.97	5.75	3.24	3.02	1.93	1.27	1.10	**1.01**	0.82
MAY	T_j	35.13	20.40	6.88	5.51	4.95	3.49	2.89	2.38	2.04
	U_j^{95}	3.04	2.99	2.95	2.92	2.88	2.86	2.83	2.81	2.79
	T_j / U_j^{95}	11.56	6.82	2.33	1.89	1.72	1.22	**1.02**	0.85	0.73
JUN	T_j	29.84	18.25	13.44	6.84	5.16	3.60	2.71	2.53	1.75
	U_j^{95}	3.04	2.99	2.95	2.92	2.88	2.86	2.83	2.81	2.79
	T_j / U_j^{95}	9.82	6.10	4.56	2.34	1.79	**1.26**	0.96	0.90	0.63

JUL	T_j	22.60	20.19	13.06	7.54	4.98	4.59	2.64	2.52	2.35
	U_j^{95}	3.04	2.99	2.95	2.92	2.88	2.86	2.83	2.81	2.79
	T_j/U_j^{95}	7.43	6.75	4.42	2.58	1.73	**1.60**	0.93	0.90	0.84
AUG	T_j	37.83	14.84	8.67	6.05	4.29	3.46	2.84	2.51	1.92
	U_j^{95}	3.04	2.99	2.95	2.92	2.88	2.86	2.83	2.81	2.79
	T_j/U_j^{95}	12.44	4.96	2.94	2.07	1.49	1.21	**1.01**	0.89	0.69
SEP	T_j	41.81	17.62	7.13	6.06	4.65	3.28	2.59	2.33	1.73
	U_j^{95}	3.04	2.99	2.95	2.92	2.88	2.86	2.83	2.81	2.79
	T_j/U_j^{95}	13.75	5.89	2.42	2.08	1.61	**1.15**	0.92	0.83	0.62
OCT	T_j	32.55	15.41	8.28	6.90	5.57	3.94	2.87	2.66	2.26
	U_j^{95}	3.04	2.99	2.95	2.92	2.88	2.86	2.83	2.81	2.79
	T_j/U_j^{95}	10.71	5.15	2.81	2.36	1.93	1.38	**1.01**	0.95	0.81
NOV	T_j	41.53	15.58	7.17	4.71	4.18	3.56	2.78	2.03	1.87
	U_j^{95}	3.04	2.99	2.95	2.92	2.88	2.86	2.83	2.81	2.79
	T_j/U_j^{95}	13.67	5.21	2.43	1.61	1.45	**1.24**	0.98	0.72	0.67
DEC	T_j	32.94	17.57	7.09	5.91	5.25	3.77	3.66	2.68	2.24
	U_j^{95}	3.04	2.99	2.95	2.92	2.88	2.86	2.83	2.81	2.79
	T_j/U_j^{95}	10.84	5.88	2.40	2.02	1.82	1.32	**1.29**	0.95	0.80

point worth noting is that some EOF modes may pass the Rule N test but represent a very low percentage of the total variance. In the latter case, these modes do not need to be analyzed, for the signal is very weak.

5.2.2 North's Rule of Thumb

North et al. (1982) presented a "rule of thumb" by showing that the standard error in the estimation of the eigenvalues λ_k can be computed as

$$\delta\lambda_k \approx \lambda_k \sqrt{\frac{2}{N}}, \tag{5.3}$$

and the standard error of the corresponding EOF ψ_k is then computed as

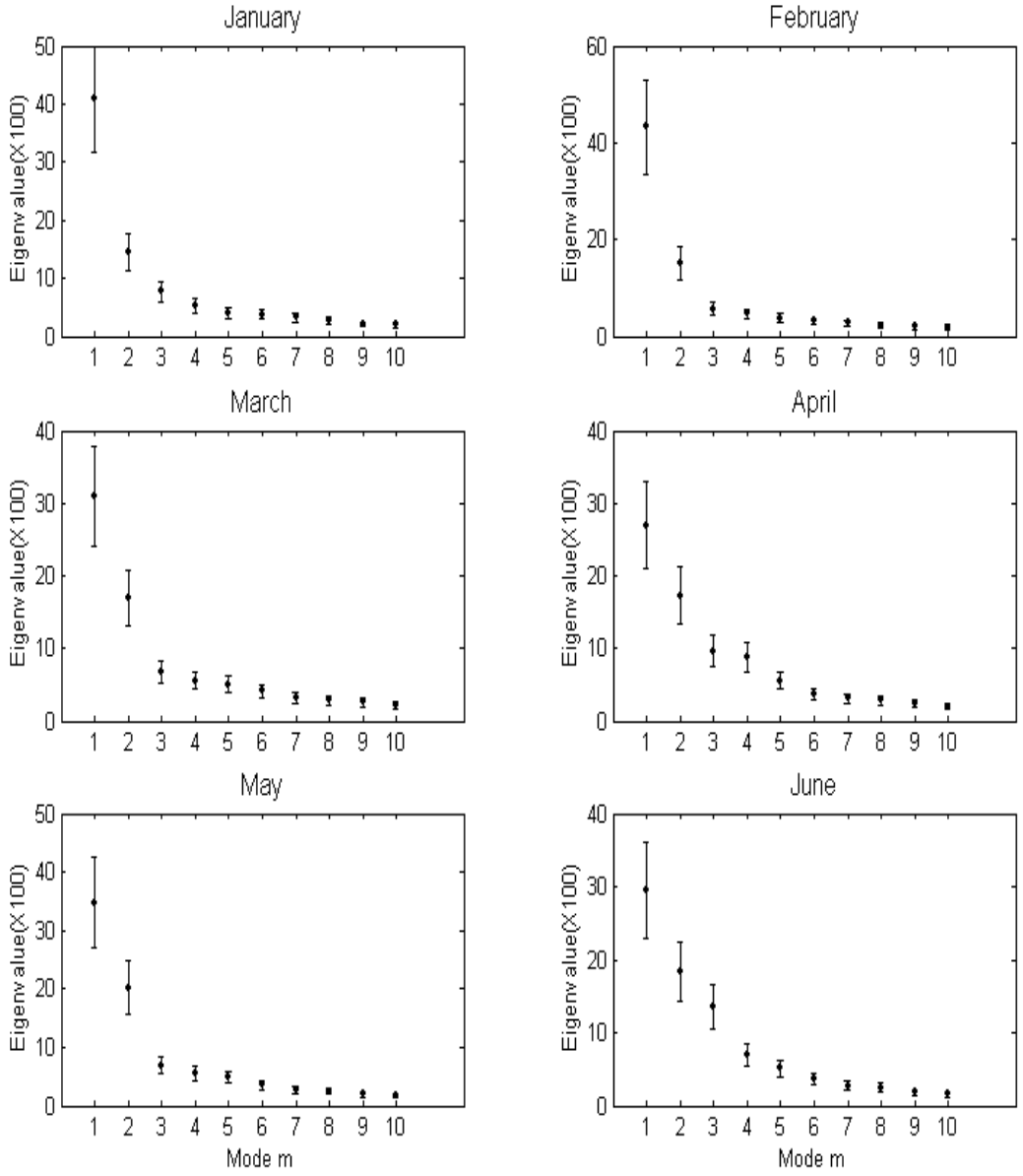
$$\delta\psi_k \approx \frac{\delta\lambda_k}{\lambda_j - \lambda_k}\psi_j \tag{5.4}$$

if the space between λ_k and its closest neighbor is small and those between λ_k and other eigenvalues are very large, where N denotes the number of independent samples. North's rule of thumb is very useful to decide whether an EOF is likely to be subject to large sampling fluctuations and to determine the maximum number of EOFs that are well separated from each other and are to be retained in an EOF analysis. North's rule of thumb states that if the standard error $\delta\lambda_k$ of λ_k is comparable or larger than the space between the eigenvalue λ_k and its closest neighbor, then the estimation error $\delta\psi_k$ of the corresponding EOF will be comparable to the size of the neighboring EOF. That is, only the EOF for which the standard error of the corresponding eigenvalue is less than the

difference between the eigenvalue and its closest eigenvalue will be retained in the EOF analysis.

Here, we use the same data sets as in Section (5.2.1) to test North's rule of thumb for deciding on the number of EOFs to keep. Fig. 5.1 presents a schematic diagram of the first ten eigenvalues derived from the Hybrid 1.0 data set with the resolution of the 0.25-by-0.5 degrees latitude and longitude grids. If no overlap is found between the confidence intervals $\lambda_k \pm \delta\lambda_k$ of successive eigenvalues, then the corresponding EOF modes are not degenerated (Anderson, 1963). An overlap with successive mode is found from the third mode for January, February, March, April, May, August, September, October, November and December (JFMAMASOND), from the second mode for June and from the first mode for July. Therefore, the two leading patterns are well separated and potentially meaningful to keep for JFMAMASOND, which explain roughly 55%, 58%, 48%, 44%, 55%, 52%, 59%, 48%, 57%, and 50% of the total variance, respectively, and the first leading patterns are significant for June, which represents 29% of the total variance. All modes are degenerated for July by North's rule of thumb. Fig. 5.2 shows a schematic diagram of the first ten eigenvalues derived from the Hybrid 1.0 data set with the resolution of the 0.1-by-0.2 degrees latitude and longitude grids. Figure 5.2 reveals that the same number of leading patterns are well separated and therefore significant, JFMAMASOND, as in Fig. 5.1, and that these patterns represent roughly 55%, 58%, 48%, 44%, 56%, 53%, 59%, 48%, 57%, and 51% of the total variance, respectively. As well, the first leading patterns are significant for June, which represents 30% of the total variance, and

all EOF modes are non-significant for July. As no significant modes will be retained for July by this rule, we will not use North's rule of thumb as the method for selecting the significant modes in our EOF based linear reconstruction. However, the Rule N test and North's rule of thumb will be combined in the following section to define the lower threshold for the truncation level in our EOF-based linear reconstruction.



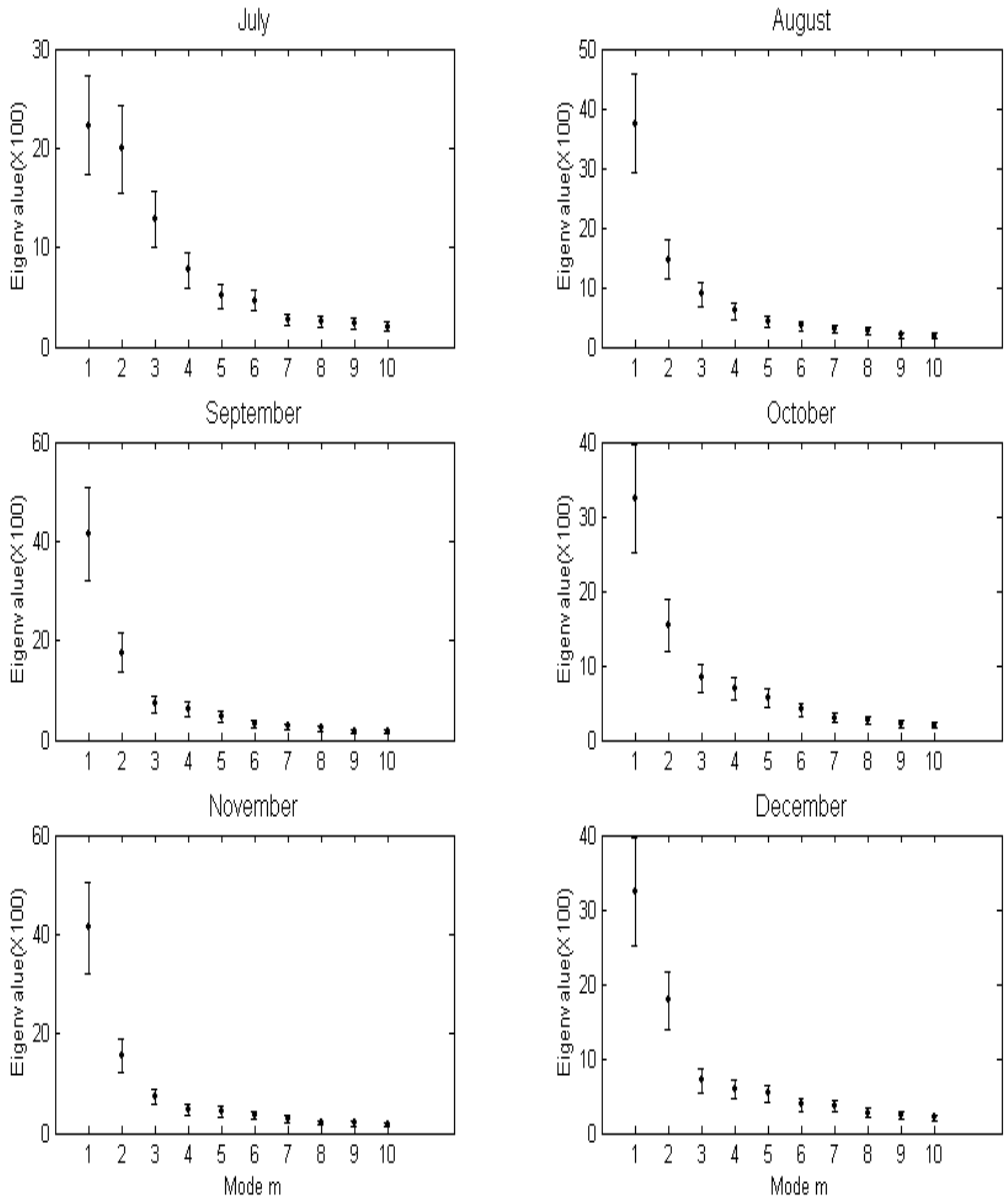
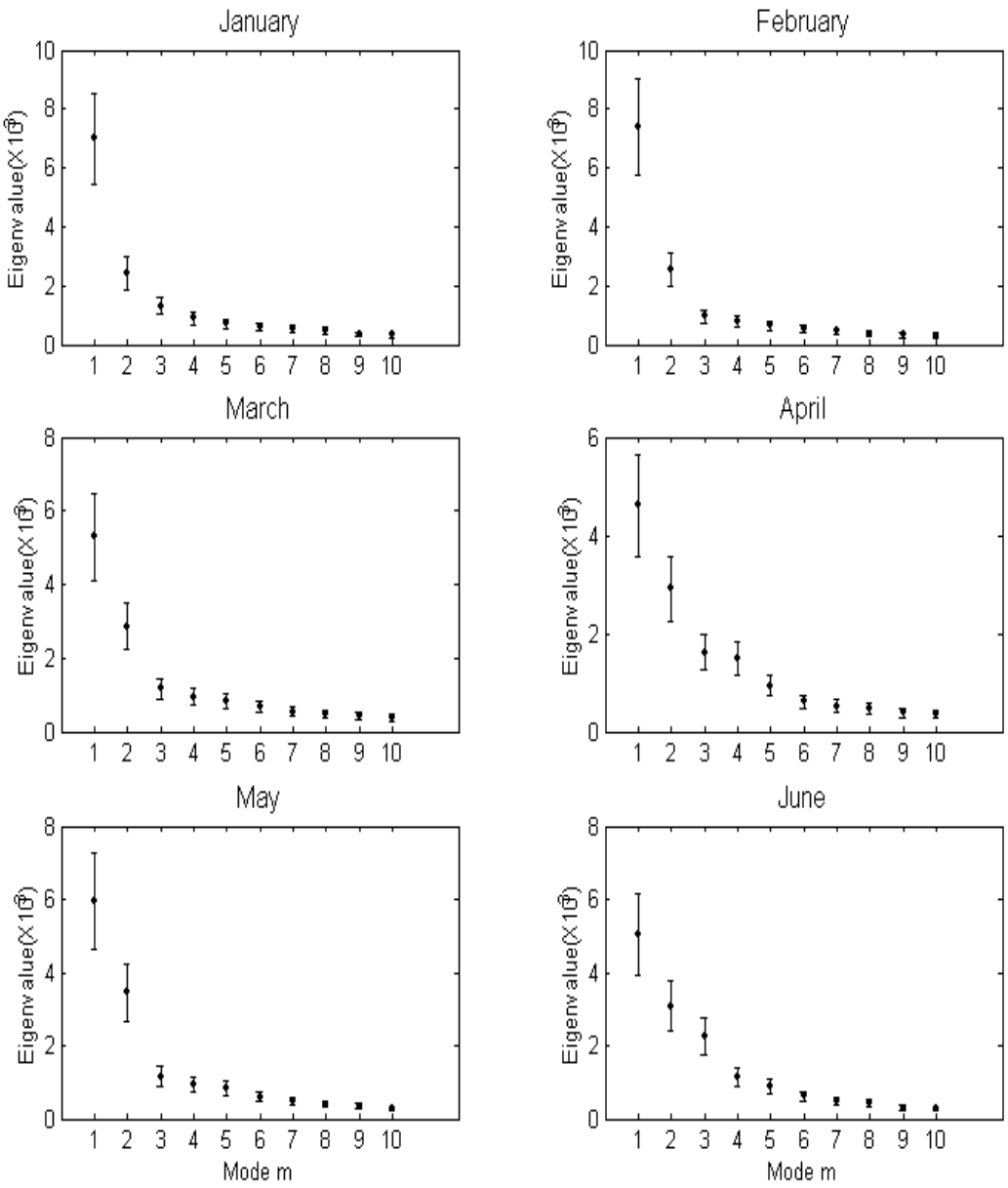


Figure 5.1. Schematic diagram of the first ten eigenvalues derived from the Hybrid 1.0 data set on the 0.25-by-0.5 degrees latitude and longitude grids: N=40, P=808.


January
February
March
April
May
June
Eigenvalue(X10^3)
Mode m
1 2 3 4 5 6 7 8 9 10

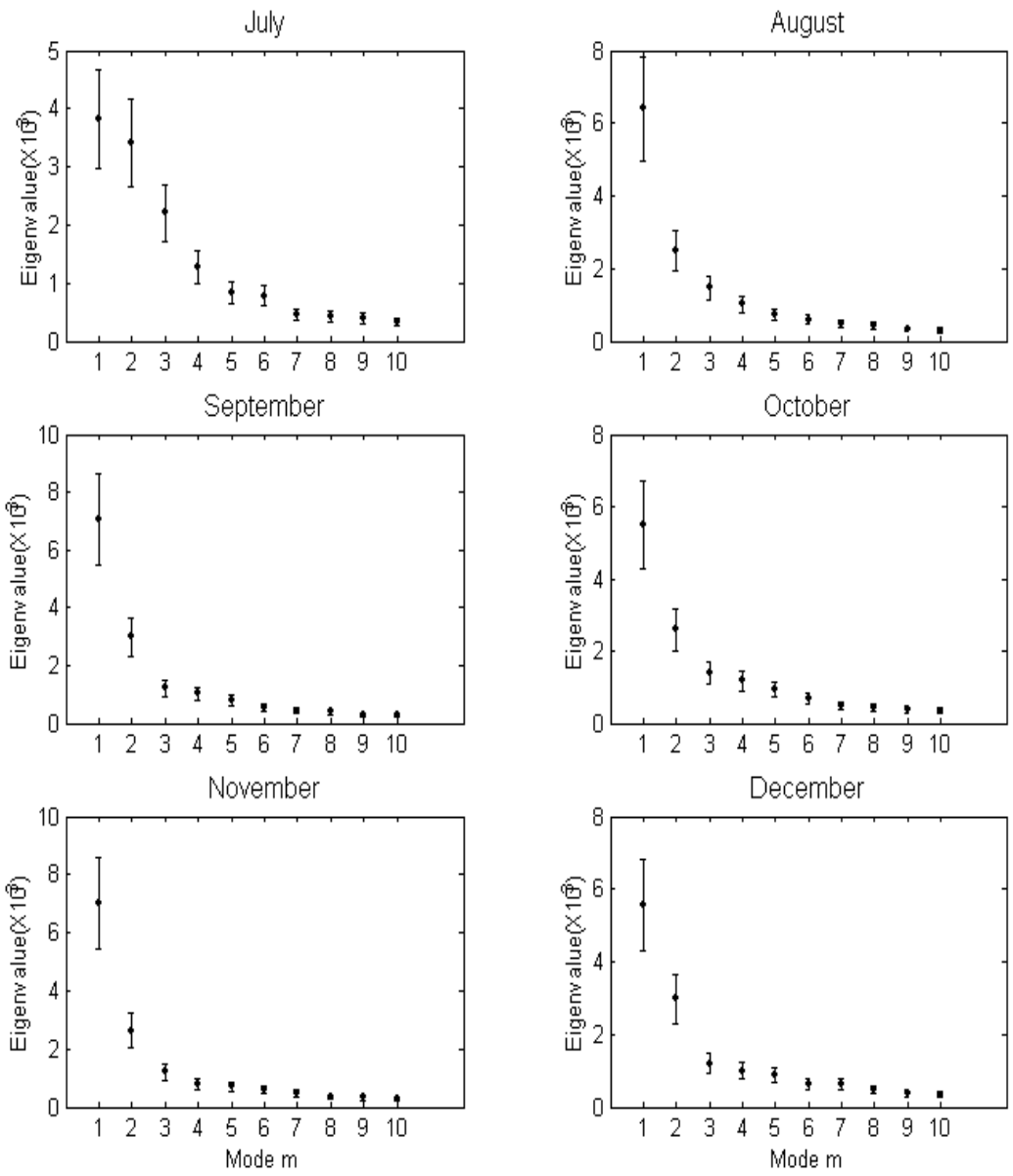


Figure 5.2. Schematic diagram of the first ten eigenvalues derived from the Hybrid 1.0 data set on the 0.1-by-0.2 degrees latitude and longitude grids: N=40, P=4862.

5.3 Selection of the Optimal Set of EOFs

In the EOF-based linear reconstruction, the reduced number (M) of EOFs is retained in order that the difference between the reconstructed data and the observed data is minimized and that the computation is reduced. In this section, our purpose is to find the optimal value of M so that the signal of the lower orders $(\leq M)$ can be used in the reconstruction while the noise of the higher orders $(> M)$ can be ignored. A criterion based on the analysis of the size of the eigenvalues and the interpolation errors (RMSE and MAE) between the observed data and the data reconstructed by the reduced set of EOFs will be used to decide whether a reduced set of EOFs is optimal, that is, to find the optimal set of EOFs. First, the North et al. (1982) error is combined with the Monte Carlo test to decide on the lower threshold for the truncation level, say M'. Next, we choose a set of withheld stations that have complete daily observation data over the studied period and have been reasonably sampled in the studied area. Then we discard the selected stations from the data set and interpolate these values by using a reduced set of EOF modes. Lastly, we compare the interpolated values to the data we put aside and calculate the average interpolation error for these stations. Of course, the found optimal number of EOFs is the one that minimizes the interpolation errors, which are the RMSE and MAE between the interpolated fields at these stations and the observed data there. We first apply the EOF-based linear interpolation method with the M' EOFs retained and calculate the average of both the RMSE and MAE values for the withheld stations. Second, (M'+1)th EOF is

now taken into account in the interpolation method and used to interpolate the values on these stations. Third, we calculate the average RMSE and MAE values again and compare them with the last ones. If both the RMSE and MAE values start to increase at the first time, the procedure stops; otherwise, we continue with more and more EOFs and repeat the above three steps. With this cross-validation technique, we can thus find the optimal number, say M , of EOFs and an RMSE and MAE estimation of the interpolation procedure. Note that the number of modes in our reconstruction is varied from month to month.

In this study, the EOFs are calculated from the 1961 – 2000 dataset, and the interpolated period is also 1961 -1990. Twenty-two withheld stations across Alberta are selected to calculate the average RMSE values, and their locations are presented in Figure 5.3. Figure 5.4 shows the percentage of variance explained by the 20 leading EOF modes derived from the 1961 – 2000 Hybrid 1.0 dataset on the 0.25-by-0.5 degrees latitude and longitude grids. In Figure 5.4, the EOF modes error bars were estimated by using North et al. (1982)'s rule of thumb test, and the thick line represents the significance level obtained by using the Monte Carlo test (Li et al. 2000). Both tests have 95% confidence. The Monte Carlo test allows us to distinguish the explained variance obtained from the EOF analysis by testing it against a white noise null hypothesis. The significances of the EOF modes estimated by using North's rule of thumb are based on comparing the separation among the neighboring eigenvalues with the sampling error (Taschetto and Wainer, 2008). By considering the EOF modes for which the North's rule of thumb error bars are above the Monte Carlo test lines in Figure 5.4, we determine

the lower threshold for the truncation level, say M', for each month (Table 5.3). By comparing the average RMSE and MAE values of the interpolation and retaining one more EOF mode for each step starting from the M' EOF modes, we fixed the optimal number of EOFs in the interpolation for each month as shown in Table 5.3, which reveals a different truncation level for each month: more EOF modes are retained for the EOF-based linear reconstruction for the summer and autumn months than for the winter and spring months. Also, the fixed M EOF modes in Table 5.3 explain enough variances (over 90%) for each month.

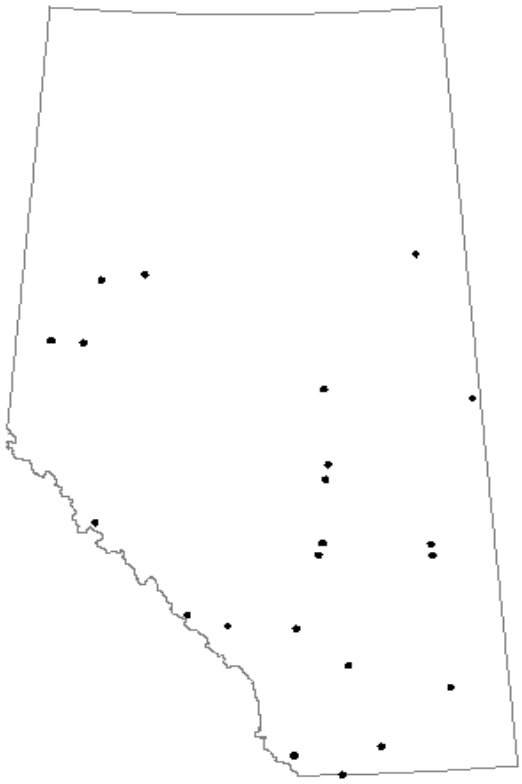
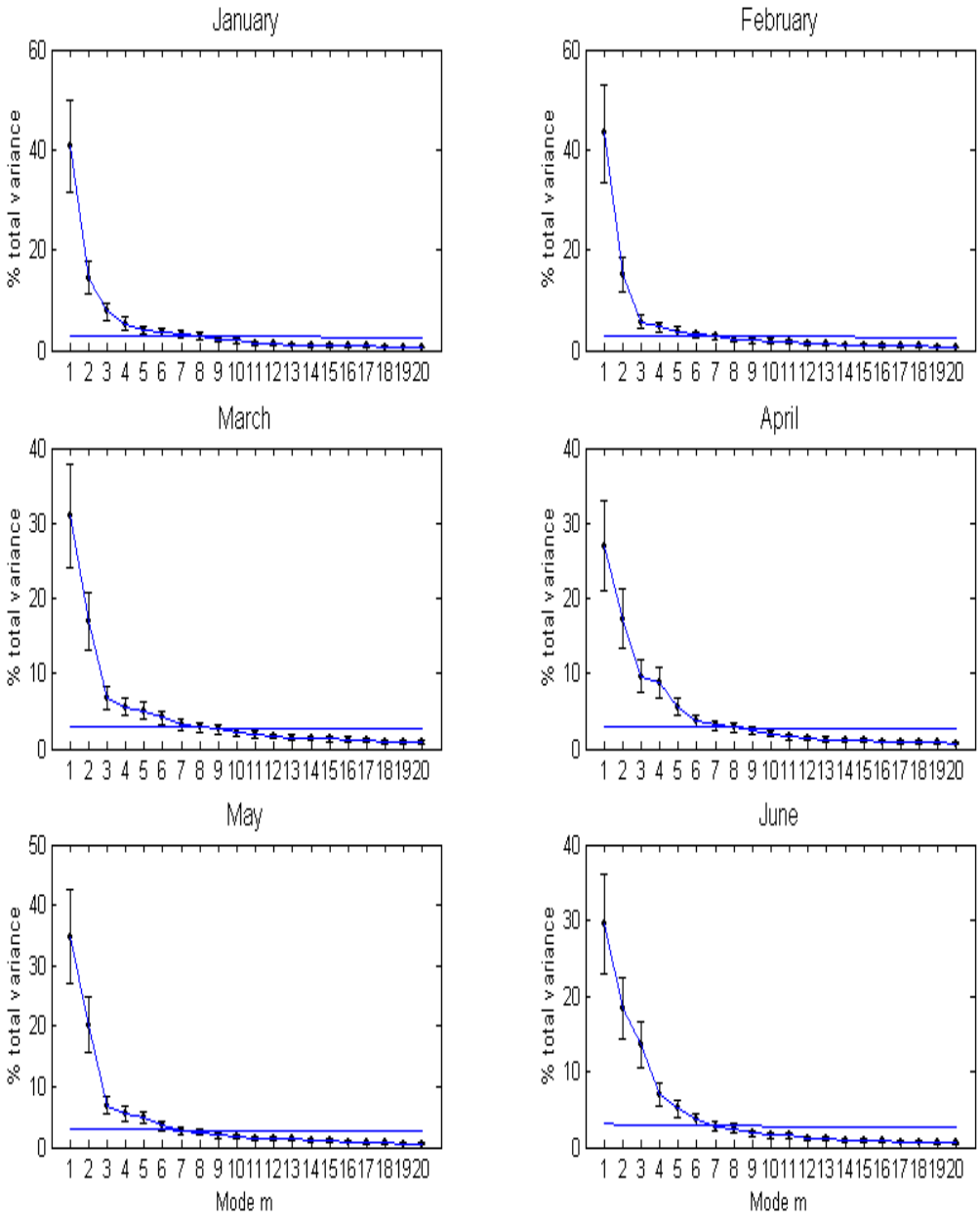


Figure 5.3. The locations of the 22 stations withheld across Alberta for this current study.


January
February
March
April
May
June
% total variance
Mode m

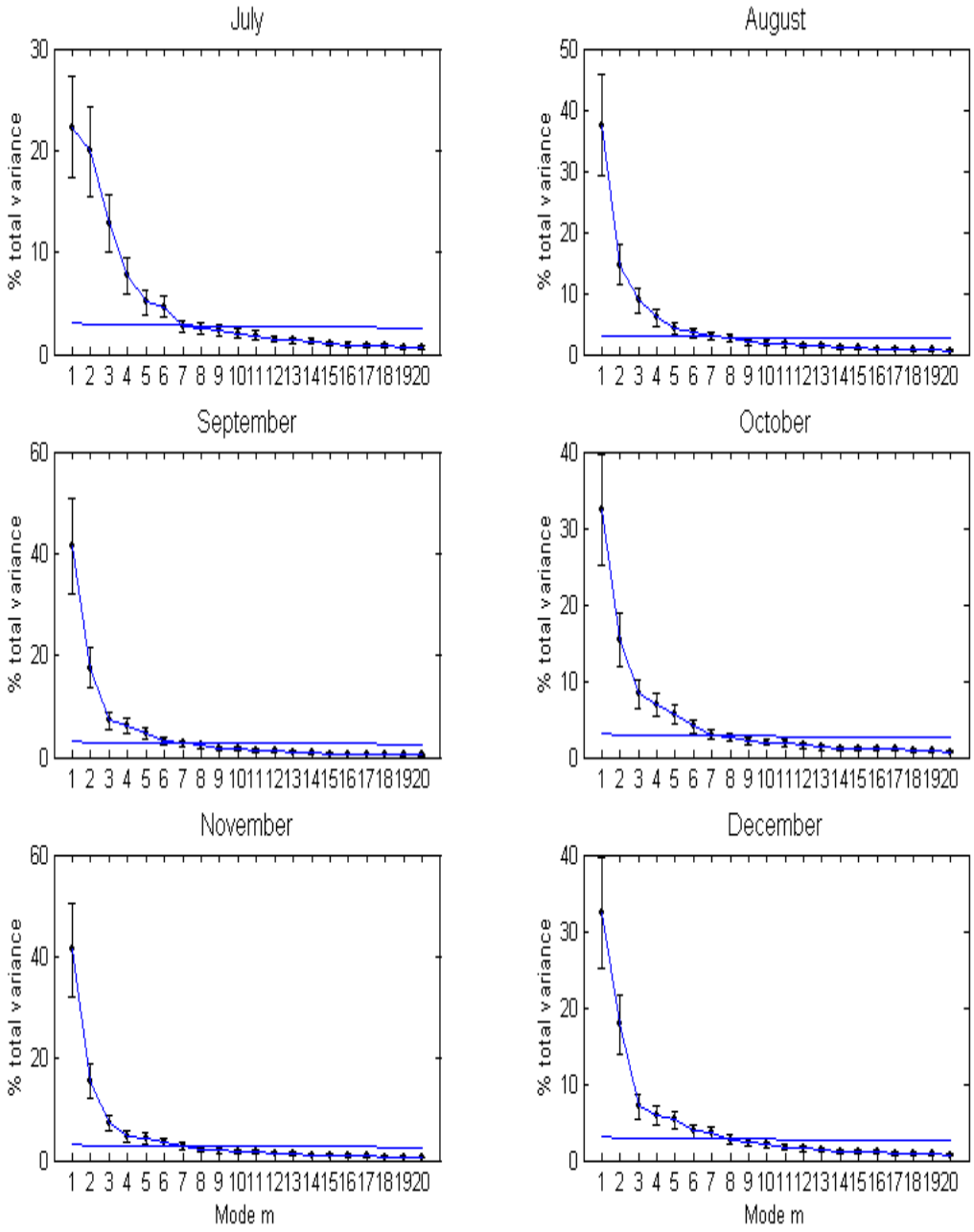


Figure. 5.4. Percentage of explained variance of the EOF modes derived from the Hybrid 1.0 data set on the 0.25-by-0.5 degrees latitude and longitude grids (N=40, P=808). Error bars estimated by using North et al.'s (1982) rule of thumb. The thick line represents the significance level obtained by using the Monte Carlo test.

Table 5.3. The determined lower threshold for the truncation level (M') and the optimal number (M) of EOFs for each month ($M \geq M'$).

Month	M'	M
JAN	5	15
FEB	5	17
MAR	6	18
APR	6	18
MAY	5	27
JUN	6	31
JUL	6	32
AUG	5	28
SEP	5	31
OCT	6	26
NOV	5	26
DEC	5	15

5.4 Test of Sensitivity to Data Errors

In this study, the random measurement error E_j for a given location $\mathbf{r}_j$ and time t are included in our EOF-based linear interpolation. The error variance $\left\langle E_j^2 \right\rangle$ is required in Eq. (2.48) to compute the optimal weights. Because the exact

value of random measurement error E_j is not available, the estimated error variance is used in our study. It is thus necessary to test if the optimal weights are sensitive to the exact size of the error variance. It is obvious that Eqn. (2.48) and (2.49) are a linear system which can be re-written as

$$\mathbf{A}_m \mathbf{X}_m = \mathbf{b}_m \tag{5.5}$$

where $m = 1, 2, \cdots, M_C$ is the modes number,

$$\mathbf{A}_m = \begin{bmatrix} \left(C_{11} + \left\langle E_1^2 \right\rangle\right)\psi_m^2(r_1) & \cdots & C_{N1}\psi_m(r_N)\psi_m(r_1) & 1 \\ \cdots & \cdots & \cdots & \cdots \\ C_{1N}\psi_m(r_1)\psi_m(r_N) & \cdots & \left(C_{NN} + \left\langle E_N^2 \right\rangle\right)\psi_m^2(r_N) & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \tag{5.6}$$

is a symmetric matrix,

$\mathbf{X}_m = \left[w_1^{(m)}, \cdots, w_N^{(m)}, \Lambda \right]'$, and $\mathbf{b}_m = \left[\lambda_m \psi_m^2(r_1), \cdots, \lambda_m \psi_m^2(r_N), A \right]'$, the prime denotes the transpose of a vector, and A is the area of Alberta in this study.

First, we test the sensitivity of the optimal weights to the exclusion and inclusion of the random measurement error E_j. Without loss of generality, in the computation of the optimal weights for July 2000 as an example (Eqn. (2.48) and (2.49)), we exclude the error variance, that is, let $\left\langle E_j^2 \right\rangle = 0$ in the matrix (5.6). Then for $m = 1$ as an example without loss of generality, the condition number of matrix A_m, $k(A_m) = \|A_m\| \|A_m^{-1}\|$, is very large, and its determinant is close to zero; that is, the linear system (5.5) used to compute the optimal weights is ill-conditioned when we exclude the random measurement error E_j. The optimal

weights are thus highly sensitive to the values of the coefficient matrix A_m: when we make a slight change in the coefficient matrix A_m, for example, when we let $\left\langle E_j^2 \right\rangle = 0.001$, the change in the optimal weights is quite significant as follows. Let X_m be the solution of the original system of $A_m X_m = b_m$, and consider the coefficient matrix A_m is changed to $A_m + \Delta A_m$, and the solution changes from X_m to $X_m + \Delta X_m$. Meyer (2000) proved that with a small change ΔA_m of A_m, the relative change in X_m and the relative change in A_m has the following relationship:

$$\frac{\|\Delta X_m\|}{\|X_m\|} \le k(A_m) \frac{\|\Delta A_m\|}{\|A_m\|} \tag{5.7}$$

Thus, the optimal weights are quite sensitive to the change of the coefficient matrix A_m when the linear system (5.5) is ill-conditioned. When we choose the value of the error variance $\left\langle E_j^2 \right\rangle$ in the range of a small number, like 0.001, to a reasonable size, like 10, the linear system (5.5) becomes well conditioned, and its optimal weight solution becomes non-sensitive to the small change of the coefficient matrix A_m; that is, it is non-sensitive to the exact size of the error variance $\left\langle E_j^2 \right\rangle$. However, when the value of $\left\langle E_j^2 \right\rangle$ goes to infinity, the linear system (5.5) goes to a form like

$$\begin{bmatrix} \infty & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \infty & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ \cdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} 0 \\ \cdots \\ 0 \\ 0 \end{bmatrix}, \tag{5.8}$$

whose solution obviously goes to zero. The detailed example for July 2000 and the mode number $m=1$ is shown in Figure 5.5. This figure shows the changes of the reconstructed anomaly data with changing the value of error variance $\left\langle E_j^2 \right\rangle$ in the weights equations. The changes of anomaly data rest with that of weight values, that is, big (little) changes of anomaly data come from big (little) changes of weight values. By comparing the graphs of $\left\langle E_j^2 \right\rangle = 0$ and $\left\langle E_j^2 \right\rangle = 0.001$, it is clear to see there is a big change of reconstructed anomaly data, that is, a big change of weight values, although a very small change of error variance $\left\langle E_j^2 \right\rangle$. However, the second figure panel reveals the totally reversed results when the value of error variance $\left\langle E_j^2 \right\rangle$ changes from 0.001 to 10: a big change of error variance value results very little change of the reconstructed anomaly data or the weight solution. The third figure panel shows the reconstructed anomaly data or the weight solution goes to zero when the value of the error variance is very large. All above results of this figure are as expected to coincide with our previous analysis in this section.

5.5 Conclusions

In this chapter, we tested the North's rule of thumb and Rule N on the explained variance of each mode. By combining the Monte Carlo test and North's rule of thumb error bars, we identified the lower threshold for the truncation level in EOF analysis. The criterion based on the analysis of the size of the interpolation errors between the observed data and the data reconstructed by the

reduced set of EOFs on the stations we put aside in the interpolation was used to determine the optimal set of EOFs for each month. The result showed that more EOFs were taken into the EOF-based linear interpolation for summer and autumn than for winter and spring. However, the mode selection method varies from study to study.

Without loss of generality, July 2000 and the first mode were used as an example to test the sensitivity of the weight solution to the exact value of the error variance $\left\langle E_j^2 \right\rangle$ in equations (2.48) and (2.49). The result showed that the weight solution was non-sensitive to the value of error variance $\left\langle E_j^2 \right\rangle$ in the range of 0.001 to 1.0, but went to zero as $\left\langle E_j^2 \right\rangle$ went to infinity.

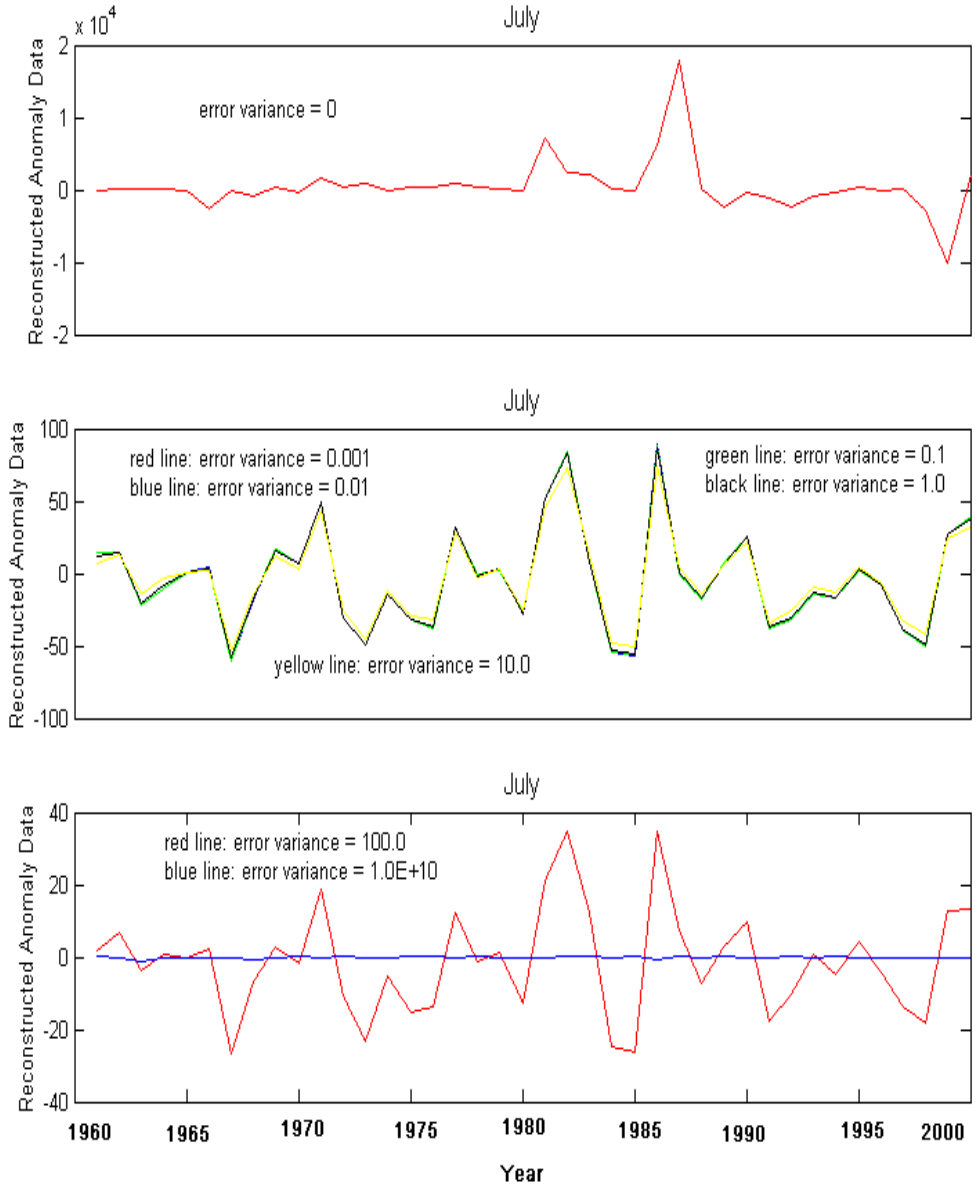


Figure. 5.5. The changes of the reconstructed anomaly data with changing the value of error variance $\left\langle E_j^2 \right\rangle$ in the equations (2.48) and (2.49).

Chapter 6

Conclusions

In order to overcome the problem of the IDW method which underestimates the precipitation in Alberta's high-elevation regions and northern Alberta because of the sparse precipitation station distribution in the lower elevation regions in the earlier period of last century, the EOF-interpolation method in our Hybrid 2.0 method was used to reconstruct the monthly precipitation on the regular grids over Alberta. The comparison of the results of the EOF-based linear interpolation and the IDW method was compared by withholding eight stations for cross-validation purposes. The criterion to select the cross-validation stations is that each station has nearly complete observations in the studied period and some stations are located in high-elevation area which aims to evaluate whether EOF-based linear interpolation overcome the particular problem of Hybrid 1.0 as we stated before. It is happy that the cross-validation results at the high-elevation stations showed that the EOF-based linear interpolation method improved the accuracy of reconstruction of monthly total precipitation in the high-elevation regions greatly over the period of 1901 – 2002 for Alberta.

We have developed the Hybrid 2.0 method based on the Hybrid 1.0 method developed by Shen et al. (2001) to grid the daily precipitation data over Alberta from January 1, 1901 to December 31, 2002. The grid resolution was 0.25°

latitude by 0.5° longitude and 0.1° latitude by 0.2° longitude, respectively. The interpolated data were named ABClim 2.0. The interpolation accuracy of the Hybrid 2.0, Hybrid 1.0, and ANUSPLIN methods in the precipitation frequency and monthly total precipitation was also compared over the period of 1961 – 1990. We selected seven withheld stations that had nearly complete observed data in the studied period and that were evenly distributed over Alberta. The predicted errors between the interpolated values and the observed data for these stations were compared for the monthly total precipitation amount and daily precipitation frequency. The results showed that the both Hybrid 1.0 and Hybrid 2.0 obtained the comparable average number of precipitation days for each month but more accurate than ANUSPLIN at some withheld stations for some month and less accurate than ANUSPLIN at other withheld locations for some month in estimating the precipitation frequency by comparing with the observed precipitation frequency. Hybrid 2.0 also had more accurate results than both ANUSPLIN and Hybrid 1.0 in interpolating the monthly total precipitation amount.

The trend analysis of the MJJA total precipitation time series for the period of 1901 - 2002 at some stations like Edmonton and Calgary, carried out by using linear regression, the moving average method, and the Mann-Kendall test, showed that the significant upward trends were found at High Level and Peace River and that no significant trends were found at Edmonton and Calgary. SPI was also used to assess the Hybrid 2.0 data and then monitor the past drought conditions at some stations in Alberta.

North's rule of thumb and Rule N were tested on the explained variance of each mode. A criterion based on the analysis of the explained variance of each mode and the size of the interpolation errors between the observed data and the data reconstructed by the reduced set of EOFs on the withheld stations was used to determine the optimal set of EOFs for each month. The result showed that more EOFs were retained in the EOF-based linear interpolation for the summer and autumn months than for the winter and spring months. The test of the sensitivity of the weight solution to the exact value of the error variance $\left\langle E_j^2 \right\rangle$ in the equations (2.48) and (2.49) showed that the weight solution was non-sensitive to the value of error variance $\left\langle E_j^2 \right\rangle$ in the range of 0.001 to 1.0, but the weight solution went to zero as $\left\langle E_j^2 \right\rangle$ went to infinity.

The monthly total precipitation values obtained in this study can also be used with some confidence across Alberta. Based the facts that EOFs can be rotated without perturbing the total fit and the rotated EOFs often are in better agreement with physical patterns than the orthogonal EOFs, the rotated EOFs will be used to represent the true patterns when we analyze the precipitation patterns in the regional area. Future improvement in our current methodology could be aimed at taking the elevation into account when redistributing the monthly total to each individual day of the month.

Bibliography

Abramowitz, M., and I. Stegun, 1965. Handbook of mathematical functions. Dover, New York, 1046pp.

Akinremi, O., S. McGinn, and H. Cutforth, 1999. Precipitation trends on the Canadian prairies. J. Climate, 12: 2996–3003.

Anderson, T., 1963. Asymptotic theory for principal component analysis. The Annals of Mathematical Statistics, 34: 122–148.

Beckers, J., and M. Rixen, 2003. EOF calculations and data filling from incomplete oceanographic datasets. J. Atmos. Ocean Technol., 20: 1839–1856.

Changnon, S., and K. Kunkel, 1999. Rapidly expanding uses of climate data and information in agriculture and water resources: causes and characteristics of new applications. Bull. Amer. Meteor. Soc., 80: 821–830.

Chetner, et al., 2003. Agroclimatic Atlas of Alberta 1971 – 2000. AAFRD, Edmonton, Alberta, Canada, 97 pp.

Conolly, J., and M. Lake, 2006. Geographical information systems in archaeology. Cambridge University Press, Cambridge, UK, 326pp.

Cressie, N., 1993. Statistics for spatial data. John Wiley and Sons, 900 pp.

Daly, C., R. Neilson, and D. Phillips, 1994. A statistical-topographic model for mapping climatological prediction over mountainous terrain. J. Appl. Meteor., 33: 140–158.

Feidas, H., C. Noulopoulou, T. Makrogiannis, and E. Bora-Senta, 2007. Trend analysis of precipitation time series in Greece and their relationship with circulation using surface and satellite data: 1955–2001. Theor. Appl. Climatol., 87: 155-177.

Griffith, D., 2001.Processing of daily agroclimatic data. MSc Thesis. Univ. Of Alberta, Edmonton, Alberta, Canada.104 pp.

Gallaudet, T., and J. Simpson, 1994. An empirical orthogonal function analysis of remotely sensed sea surface temperature variability and its relation to interior oceanic processes off Baja California. Remote. Sens. Environ., 47**:** 375–389.

Goosens, C., A. Berger, 1986. Annual and seasonal climatic variations over the northern hemisphere and Europe during the last century. Ann. Geophys., B 4: 385–399.

Hudson, G., and H. Wackernagel, 1994. Mapping temperature using kriging with external drift: theory and example from Scotland. Int. J. Climatol., 14: 77–91.

Hutchinson M., 1995. Interpolating mean rainfall using thin plate smoothing splines. International Journal of GIS, 9: 385-403.

Hutchinson, M., 1998a. Interpolation of rainfall data with thin plate smoothing splines. Part I: Two-dimensional smoothing of data with short range correlation. J. Geogr. Inf. Decis. Anal., 2: 152–167.

Hutchinson, M., 1998b. Interpolation of rainfall data with thin plate smoothing splines. Part II: Analysis of topographic dependence. J. Geogr. Inf. Decis. Anal., 2: 168–185.

Jones, P., S. Raper, R. Bradley, H. Diaz, P. Kelly, and T. Wigley, 1986. Northern Hemisphere surface air temperature variations: 1851–1984. J. Climate Appl. Meteor., 25: 161–179.

Karl, T., and R. Knight, 1998. Secular trends of precipitation amount, frequency, and intensity in the United States. Bull. Amer. Meteor. Soc., 79: 231–241.

Kutzbach, J., 1967. Empirical eigenvectors of sea-level pressure, surface temperature and precipitation complexes over North America. J Appl. Met., 6: 791-801.

Lagerloef, G., and R. Bernstein, 1988. EOF analysis of AVHRR surface temperature patterns in Santa Barbara Channel. J. Geophys. Res. 93, C6: 6863-6873.

Li, G., 2001. Accurate estimation of standard errors in the global averaging of surface temperatures. MSc Thesis. Univ. Of Alberta, Edmonton, Alberta, Canada. 85 pp.

McKee, T., N. Doesken, and J. Kleist. 1993. The relationship of drought frequency and duration to time scales. In Proceedings of the 8th Conference of Applied Climatology, 17-22 January, Anaheim, CA. American Meterological Society: 179-184.

McKee, T., N. Doesken, and J. Kleist, 1995. Drought monitoring with multiple time scales. Preprints, 9th Conference on Applied Climatology, 15-20 January, Dallas, Texas, American Meteorological Society: 233-236.

Meyer, C., 2000. Matrix analysis and applied linear algebra. SIAM, Philadelphia, PA.

Mitchell, J., B. Dzerdzeevski, H. Flohn, and W. Hofmery, 1966. Climatic change. WMO Tech Note 79. WMO No. 195. TP. 100, Geneva, 79 pp.

North, G., T. Bell, R. Cahalan, and F. Moeng, 1982. Sampling errors in the estimation of empirical orthogonal functions. Mon. Wea. Rev., 110: 699–706.

North, G., 1984. Empirical orthogonal functions and normal modes. J. The Atmos. Sci., 41: 879–887.

Overland, J., and R. Preisendorfer, 1982. A significance test for principal components applied to a cyclone climatology. Mon. Weather Rev., 110: 1-4.

Preisendorfer, R., 1988. Principal component analysis in meteorology and oceanography. Elsevier, 425 pp.

Preisendorfer, R., and T. Barnett, 1977. Significance tests for empirical orthogonal functions. Preprints, Fifth Conf. on Probability and Statistics in Atmospheric Sciences, Las Vegas, NV, Amer. Meteor. Soc., 169–172.

Richman, M., 1986. Rotation of principal components. J. Climatol., 6: 293-335.

Seiler, R., M. Hayes, and L. Bressanl, 2002. Using the standardized precipitation index for flood risk monitoring. Int. J. Clim., 22: 1365-1376.

Sharpley, A., and J. Williams, Eds.1990. EPIC: Erosion/Productivity Impact Calculator. Part 1. Model documentation. U.S. Department of Agriculture Technical Bulletin Rep. 1768, 235pp.

Shen, S., T. Smith, C. Ropeleeski, and R. Livezey, 1998. An optimal regional averaging method with error estimates and a test using tropical pacific SST data. J. Climate, 11: 2340-2350.

Shen, S., P. Dzikowski, G. Li, and D. Griffith, 2001. Interpolation of 1961–97 daily temperature and precipitation data onto Alberta polygons of ecodistrict and soil landscapes of Canada. Journal of Applied Meteorology, 40: 2162–2177.

Shen, S., W. Lau, K. Kim, and G. Li, 2001. Error estimation of an ensemble statistical seasonal precipitation prediction model. NASA Technical Memorandum.

Shen, S., A. Basist, G. Li, C. Williams, and T. Karl, 2004. Prediction of sea surface temperature from the global historical climatology network data. Environmetrics, 15: 233 – 249.

Shen, S., B. Shen, N. Newland, T. Yang, H. Hill, and A. Howard, 2005 (a). Interpolating daily precipitation and temperature data on a regular grid in Manitoba and Saskatchewan. National Agroclimate Information Service. 89 pp.

Shen, S., H. Yin, K. Cannon, A. Howard, S. Chetner, and T. Karl, 2005. Temporal and spatial changes of the agroclimate in Alberta, Canada, from 1901 to 2002. J. Appl. Meteor., 44: 1090-1105.

Shen, S., G. North, and K. Kim, 1994. Spectral approach to optimal estimation of the global average temperature, J. Climate, 7: 1999-2007.

Shields, J., C. Tarnocai, K. Valentine, and K. MacDonald, 1991. Soil landscapes of Canada: Procedures manual and user's handbook. Land Resource Research Centre, Research Branch, Agriculture Canada Publ. 1868/E, 74 pp.

Smith, T., R. Reynolds, R. Livezey, and D. Stokes, 1996. Reconstruction of sea

surface temperature using empirical orthogonal functions. J. Climate, 9: 1403–1420.

Smith, T., R. Livezey, and S. Shen, 1998. An improved method for interpolating sparse and irregularly distributed data onto a regular grid. J. Climate, 11: 2340–2350.

Smith, T., and R. Reynolds, 2004. Improved extended reconstruction of SST (1854–1997). J. Climate, 17: 2466–2477.

Taschetto, A., and I. Wainer. 2008. Reproducibility of south American precipitation due to subtropical south Atlantic SSTs. J. Climate, 21: 2835-2851.

Turgay, P., and E. Kahya, 2006. Trend analysis in Turkish precipitation data. Hydrological Processes, 20: 2011–2026.

Thom, H., 1966. Some methods of climatological analysis. WMO Tech. Note 81, 53 pp.

Thornton, P., S. Running, and M. White, 1997. Generating surfaces of daily meteorological variables over large regions of complex terrain. Journal of Hydrology, 190: 241-251.

Wahba, G., 1990. Spline models for observational data. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia, PA, 169 pp.

Wahba, G., 1979. How to smooth curves and surfaces with splines and cross-validation. 24th Conference on the Design of Experiments, U.S. Army Research Office: 167-192.

Yin, H., 2005. Statistical data analysis of Alberta agroclimate and global surface

air temperature. PhD Thesis. Univ. Of Alberta, Edmonton, Alberta, Canada. 165 pp.