5-2018

Analysis of Daily Precipitation Data from Selected Sites in the United States

Sahar Ahmed
Montclair State University

Abstract

Masters of Science

by

Sahar Ahmed

Global warming is a contentious topic since modern climate records only exist for the last 100 years in contrast to ice-core analysis that establishes ice ages tens of thousands of years ago. Nevertheless, patterns associated with events such as El Niño Southern Oscillation (ENSO), precipitation, tornadoes, and snowfall amounts over the last century can provide a useful and objective indicator of climate "change". This project focuses on daily precipitation totals for the state of New Jersey over the last 100 to 150 years from nineteen meteorological recording stations and involves large data sets with a million observations. This research utilizes time series analysis to present results and findings with a temporal emphasis. The project includes an extension to select states across the United States for a comparison of precipitation patterns.

Montclair State University

Analysis of Daily Precipitation Data from Selected Sites in the United States

by

Sahar Ahmed

A Master's Thesis Submitted to the Faculty of Montclair State University In Partial Fulfillment of the Requirements For the Degree of Masters of Science

May 2018

College of Science and Mathematics
College

Mathematical Sciences
Department

Thesis Committee:

[redacted]

Dr. Andrew J. McDougall
Thesis Sponsor

May 4, 2018
Date

[redacted]

Dr. Andrada Ivanescu
Committee Member

April 30, 2018
Date

[redacted]

Dr. Gregory Pope
Committee Member

April 30, 2018
Date

Analysis of Daily Precipitation Data from Selected Sites in the United States

Submitted in partial fulfillment of the requirements
For the degree of Masters of Science

by

Sahar Ahmed
Montclair State University
Montclair, NJ
May 2018

Acknowledgements

Firstly, I would like to express my deepest gratitude to my professor and research advisor, Dr. Andrew J. McDougall, for his support, guidance, and patience. I am truly honored to have been your research student for the last three years, this has been a remarkable experience.

I sincerely thank Dr. Andrada Ivanescu for help and support in this project and for her continued guidance during my undergraduate and graduate studies. I would also like to thank Dr. Gregory Pope for his invaluable input and support.

Finally, I would like to thank my friends and family for their support during the years that I have spent at Montclair State University.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

The Earth's climate is generated by complex interactions of solar energy, clouds, ocean currents, and atmospheric circulation. The atmosphere warms the planet by absorbing solar energy from the sun while at the same time, redirecting infrared radiation back into space to create a balance of energy on the planet (Seinfeld and Pandis, 2016). The atmosphere is comprised mainly of gases such as N_2(78%), O_2 (21%), and Ar (0.93%) along with small amounts of water vapor and trace gases (Seinfeld and Pandis, 2016). Trace gases make up less than 1% of the atmosphere but play a vital role in the Earth's radiative balance (Seinfeld and Pandis, 2016). Aside from gases in the atmosphere, clouds also play a role in regulating the climate. Some clouds cool the planet by reflecting solar radiation while others warm the Earth by trapping energy near the surface (Seinfeld and Pandis, 2016). On balance, clouds exert a cooling effect on the Earth although in some areas, heavy clouds warm the regional climate (Seinfeld and Pandis, 2016). In general, most of the solar radiation is absorbed by

the Earth near its equator with little solar energy reaching the polar regions. Over time, the energy absorbed, carried by winds and ocean currents, is spread out to colder regions of the globe thereby generating the climate as we know it (Seinfeld and Pandis, 2016). Until recently, the earth's climate was assumed to change on a gradual time scale far in the future yet rapidly warming temperatures and its effects, adverse as well as favorable, are being experienced globally. The Intergovernmental Panel on Climate Change (IPCC) reports that each of the last three decades has been successively warmer at the Earth's surface than any preceding decade with the most recent decade being the nation's warmest on record (IPCC: Core Writing Team and Reisinger, 2014).

In the United States, the average temperatures have increased from 1.3°F to 1.9°F since record keeping began in 1895 with most of the of the increase having occurred after 1970 (Melillo et al., 2014). The observed warming, evident from the increase in global average air and ocean temperatures, rising sea levels, and melting ice caps, is due to increased emissions of trace gases such as carbon dioxide (CO_2), methane (CH_4), and nitrous oxides (NO_x) such as carbon monoxide (CO) and sulfur dioxide (SO_2) into the atmosphere (IPCC: Core Writing Team and Reisinger, 2014). These trace gases or greenhouse gases (GHG) act as atmospheric thermal insulators by absorbing infrared radiation from the Earth's surface and re-emitting a portion back to it (Seinfeld and Pandis, 2016). Changes in the atmospheric concentrations of GHGs, aerosols, and solar radiation alter the energy balance of the climate system causing changes such as the global warming we are currently experiencing (IPCC: Core Writing Team and Reisinger, 2007). Global increases in GHG emissions have been linked to human activities and have grown by 70% since pre-Industrial times (IPCC: Core Writing Team and Reisinger, 2007). Carbon Dioxide is by far the most prevalent anthropogenic GHG since

it is the most abundantly emitted and has increased by 80% from 1970 to 2004 (IPCC: Core Writing Team and Reisinger, 2007). It is anticipated that continued emissions of GHG's will cause further warming and long lasting changes in all components of the climate system and environmental phenomena including precipitation (IPCC: Core Writing Team and Reisinger, 2014).

Precipitation is a critical component of the Earth's ecosystem and essential for life on Earth. Over the continental land masses, it is the primary source of all freshwater used for general water consumption and agriculture. It not only impacts all of humanity, but also the natural environment around us by contributing to the maintenance of soil moisture and replenishing natural underground water reservoirs (Kidd et al., 2017; Michaelides et al., 2009; IPCC: Core Writing Team and Reisinger, 2014). As temperatures rise and the Earth's surface warms, characteristics of precipitation such as amount, duration, and frequency are directly influenced and altered (Trenberth et al., 2003). In the last century, surface air temperatures and precipitation over land have increased and are expected to continue to rise over the 21st century under the continued current levels of GHGs emissions. It is very likely that extreme precipitation events will become more intense and occur more frequently (IPCC: Core Writing Team and Reisinger, 2014). IPCC models predict changes in the hydrological cycle such as increases in evaporation and precipitation which can lead to an intensified water cycle and result in increased severe natural disasters (Huntington, 2006; Trenberth et al., 2003).

Higher temperatures increase the atmosphere's water holding capacity and in turn allows for more moisture absorption which favors stronger rainfall and snowfall events usually resulting in storms (Trenberth et al., 2003). Global warming also leads to an increase in evapotranspiration which is the combined process of water evaporation from the Earth's surface

and from vegetation (IPCC: Core Writing Team and Reisinger, 2007). In general, evaporation occurs to cool the surface of the planet and would be expected in increasing temperatures (Trenberth et al., 2003). In areas that have less precipitation, this can have devastating effects such as drought which can lead to further issues like increased risk of wildfires and heatwaves (Trenberth et al., 2003). On the other hand, areas that are predominantly moist are more inclined to experience episodes of intense precipitation which results in flooding (Trenberth et al., 2003). A warmer climate therefore increases the risks of both droughts in areas that are usually dry and floods in areas that are already wet since the characteristics of precipitation are more prone to change (Trenberth, 2012). Both of these precipitation extremes, as a response to climate change, have been the subject of extensive study with models having found an intensification of precipitation extremes with important regional variations (O'Gorman, 2015).

Changes in the major ocean currents caused by global warming also affect precipitation patterns. It is known that the variability of precipitation in the tropics is highly influenced by the cyclic variations of sea surface temperatures (SST) across the equatorial Pacific caused by the El Nino Southern Oscillation (ENSO) every two to seven years (Strangeways, 2006). Since 1975, there has been a shift in temperatures to warmer conditions which have caused the ENSO events to have become more intense, frequent, and persistent (Dore, 2005). Such changing patterns call for renewed efforts for adaptation to climate change since changing precipitation patterns affect regional availability of food and supply (Dore, 2005). The North Atlantic Oscillation (NAO) is another SST anomaly that affects terrestrial precipitation through ocean-atmosphere interactions (Chang et al., 2017), and is the main cause of much of the winter precipitation variability in the North Atlantic region (Strangeways, 2006).

The NAO consists of two pressure centers with one located at an area of high pressure over the Azores and the other located at an area of low pressure near Iceland (Jones et al., 1997). Fluctuations in these pressure centers causes the NAO to take on a positive or negative phase. In the positive phase, there is a stronger than usual Azores high pressure and a deeper than usual Icelandic low which causes intense winter depressions to cross the Atlantic (Strangeways, 2006). The increase in the pressure gradient of the positive phase also alters the orientation of the jet stream and ultimately affects temperature and precipitation especially on the east coast of the United States causing wetter winters with stronger storms (Hurrell et al., 1995; Strangeways, 2006). In the negative phase, both pressure centers are weakened causing fewer depressions to cross the Atlantic and drier winters (Strangeways, 2006). In the last thirty years, for unknown reasons, the NAO has leaned towards a more positive phase which is unusual compared to the past century (Hurrell et al., 1995). This has become the topic of many debates since changes in these naturally occurring patterns of atmospheric and oceanic variability, such as the ENSO and the NAO, coordinate large variations in weather over much of the globe (Hurrell and Deser, 2010). Concerns over such phenomena has generated the desire for a deeper understanding of how climate change has altered precipitation patterns so far and how it will continue to affect them in the future.

Although the subject of climate change is immense, the topic of changing precipitation patterns around the world requires urgent and systematic attention (Dore, 2005). To determine the extent of precipitation patterns being affected, accurate rain trend analysis is required to establish climate model evaluation (Strangeways, 2006). A study of rainfall trends of past years, when compared to those of recent years, can potentially provide a good indication of any change that has been brought by climate change. Fluctuations and deviations

from the norm in the past may give a signal to possible future predictions in association to climate change. Analysis can be used to perform climate-scale comparisons of individual areas.

Climate change is a global issue, therefore analysis on rainfall trends can be applied on a global scale as well as local. However, accurately analyzing precipitation on global scale can have several limitations which makes it a difficult task for the following reasons:

- Climate research requires an integrated climatologically sound global data set which unfortunately does not exist (Michaelides et al., 2009; Tapiador et al., 2017).
- There is a disparity between the number of precipitation records on land versus the records on sea with more records existing for land as opposed to sea (Kidd et al., 2017).
- Data available from different countries or regions depend on the organization in command of acquiring it, and often there are several different organizations in charge of measuring rainfall and usually are not consistent with each other (Kidd et al., 2017).
- Accurate long term precipitation records spanning more than a century without breaks in between do not generally exist (Strangeways, 2006).

In the past, precipitation was not recorded for climate research, rather it was collected mainly for weather forecasting purposes or for water resource assessments (Strangeways, 2006). Now that we want to understand the effects of climate change, complete records without large periods of missing measurements are crucial. This is also a major drawback for analysis of precipitation at a local scale which is the central focus of this project. In this study, analysis of observed daily rainfall records over the past hundred years is performed

on several selected sites and states in the United States to determine the impact of climate change on precipitation patterns and to discern some general patterns at major regional and national levels. This study performs several analysis methods such as time series analysis to search for patterns and spectral analysis for cyclic patterns.

1.2 Aims and Objectives

In the United States, the average precipitation amount has increased since 1900, with some areas showing an increase greater than the national average and others a decrease (Melillo et al., 2014). In the last three to five decades, heavy downpours have increased nationally (Melillo et al., 2014), and as global temperatures continue to increase, it seems reasonable to want to take a closer look at individual states and see how precipitation patterns have changed. The primary objective of this project is to examine how precipitation patterns have been changing in the last century or over time in association with climate change of selected areas in the United States. In this study, there are nineteen sites selected from the state of New Jersey that will be studied. This study aims to:

- Examine the rainfall data to determine whether precipitation patterns have changed
- Explore possible time series models to represent the data
- Analyze rain intensity and investigate whether precipitation has become more intense over recent years
- Extend analysis to select states in the United States and to compare results across states to examine change on a national level

There are several factors that could be expected during analysis, the following questions will be used as a guideline for the execution of the analysis:

1. Is there an evident change in precipitation amounts in recent years?
2. Is there a noticeable trend in precipitation patterns and if so, is it periodic?
3. If there is a change, is it being experienced locally or statewide?

1.3 Organization of Thesis

Chapter two contains a brief explanation of how precipitation is measured and the errors often associated with obtaining these measurements. Presented in this chapter is the study area and information about the database which provided the data used in this study.

Chapter 3 focuses on the importance of data cleaning and identifies some of the cleaning methods used on the rainfall data. Presented in this chapter are two techniques to display the rainfall data used throughout the thesis. The first applies various forms of time series graphs while the second utilizes functional data sets to construct functional data plots where each observation is a function and each curve represents a year for all available recorded years.

Chapter 4 focuses on the results obtained from running analysis on the NJ data. Possible models to represent the data are explored and rain intensity is examined. Chapter 5 is an extension of the study in which analysis of rainfall patterns is extended to selected states across the United States. Chapter 6 contains the conclusion and final comments.

Chapter 2

Data Collection

2.1 Measuring Precipitation

Precipitation is of interest to a number of different scientific communities such as the atmospheric and environmental communities and its monitoring and proper measurements has great economic and scientific value. For climate studies, accuracy of measurements and the homogeneity of the data records are crucial for properly and accurately assessing the change in environmental factors brought by climate change (Kidd et al., 2017). However, accurately measuring precipitation is extremely difficult because of its highly variable properties and so consequently is one of the most poorly monitored environmental parameter especially on a global scale (Kidd et al., 2017).

Precipitation is generally measured by a device known as a rain gauge which collects rainfall through a circular funnel called the orifice and measures the amount of water captured (Strangeways, 2006). The total amount of precipitation is the sum of the collected liquid

(Tapiador et al., 2017). The concept of the gauge is simple, but its importance and practical use has led to a large number of different types of rain gauges (Kidd et al., 2017). There are currently more than fifty different types of rain gauges in use around the world with different designs and each with its own associated mechanical errors (Tapiador et al., 2017). As a result, this has led to much variation among precipitation records over time across the globe.

2.1.1 Errors in Measurements

Rain gauges cannot simply be put outside and expected to give accurate measurements and are usually susceptible to many errors in their readings. Most gauges have difficulty accurately measuring precipitation due to several factors or a combination of factors. One natural cause that leads to inaccurate measurement of rainfall is wind flow. Turbulence caused by wind and gauge interaction leads to an undercapture of precipitation especially in low intensities of rainfall and higher wind speeds (Tapiador et al., 2017). Similar errors may also arise from physical problems such as blockages of the gauge orifice by surrounding trees or other factors, consequently causing less rainfall to be collected and measured (Strangeways, 2006). Precipitation is also susceptible to low readings by the rain gauge when collected water in the gauge is lost through the process of evaporation. Evaporative loss is a source of error prevalent in precipitation records with the proportion of evaporative loss greater in light showers than during occasional heavy rains (Strangeways, 2006). During heavy downpours, precipitation is often lost through outspash when the orifice of the gauge fills up. Several rain gauge models of rain gauges overcome this issue through use of deep steeply sloping funnels which virtually eliminates outsplash, however the increased area that can be “wet” leads to higher evaporative loss (Strangeways, 2006). To minimize some of these mechanical errors, rain

gauges must be strategically positioned in places representative of the surrounding area with a balanced amount of sheltering since some can help decrease the adverse effects of wind, while too much can increase blockages (Strangeways, 2006). Despite the errors, rain gauges remain the most accurate instrument to measure rainfall at the surface (Tapiador et al., 2017).

The errors mentioned are associated primarily with older, mechanical rain gauges. Modern electric rain gauges, such as the optical rain gauge, are able to minimize mechanical errors by directly measuring precipitation with modern sensors as it falls instead of measuring collected water (Strangeways, 2006). While mechanical recording rain gauges rely on paper charts to record pen traces of rainfall that need to be manually read and then converted to tables before any of the data can be downloaded for use, newer electrically recording gauges use data loggers which use memory chips of large capacities to record measurements that can automatically be downloaded into computers, reducing both the time it takes to compile the data and human error (Strangeways, 2006). Together, the cost of mechanical rain gauges and the invention of modern electronic recording rain gauges have also led to the development of automatic (quasi) rain gauges that can measure, record, and report rainfall in near real time (Kidd et al., 2017). The availability of gauge measurements in real time greatly enhances its practical use in terms of accuracy and efficiency in rain trend analysis useful for meteorological and hydrological purposes (Kidd et al., 2017). Physical rain gauges also verify new methods of remote precipitation estimates, taken by land-based or satellite-based radars.

2.2 GHCN-Daily Database

The data used in this study was provided by the Global Historical Climatology Network (GHCN)-Daily from the National Oceanic and Atmospheric Administration (NOAA) and can be acquired through the website: data.nodc.noaa.gov. The GHCN-Daily database provides daily gridded precipitation records from meteorological stations worldwide including in the United States. The GHCN-Daily archive is updated daily and is composed of data records from over forty thousand stations distributed across the continents with several station records extending back to the 19th century (Menne and Houston, 2012). GHCN-Daily collects climate records from several sources and merges them for quality assurance and then provides a single format data set for each site ideal for analysis and usage.

2.3 Description of Study Area

The state of New Jersey (NJ), located in north eastern United States, experiences a wide range of temperatures throughout the year with warm summers during the months from June to September and cold wet winters during the months of December to February. Typically, rainfall is evenly distributed throughout most of the year with the greatest amounts in July and August. To examine precipitation patterns for the past several decades representative of all of New Jersey, precipitation records were collected from more than one area in the state. Cities with major international (intl), regional (rgnl) airports, or meteorological stations were selected for sites since both retain long detailed records of precipitation observations. A total of 19 sites across the state were selected. The sites were selected from North Jersey, South Jersey, Central Jersey, and a few from the coast to examine coastal precipitation behavior.

Figure 2.1 displays a plot of NJ generated in R (R Core Team, 2017) using the `maps` library with the locations of the selected 19 sites marked.

2.3.1 Study Sites

For this project, data for the selected station sites was requested from the NOAA website. A summary of the requested New Jersey data is provided in Table 2.1 with a complete list of the selected stations, their corresponding longitudinal coordinates, and available record dates for each site. A majority of the sites have available precipitation records beginning from the late 19^{th} century and a few starting from the early 20^{th} century. The total number of data records (N) available for each site is also listed in table 2.1 with the total observations calculated in the last row.

Site	Start Year	End Year	Latitude	Longitude	N
1. Atlantic City Marina	01/01/1948	03/09/2016	39.449	-74.567	24906
2. Belleplain State Forest	03/01/1922	10/31/2006	39.248	-74.843	30404
3. Belvidere Bridge	01/01/1893	03/12/2016	40.829	-75.083	44741
4. Charlotteburg Reservoir	04/01/1893	03/12/2016	41.034	-74.423	44578
5. Flemington	01/01/1898	03/11/2016	40.508	-74.815	41177
6. Hightstown	01/01/1893	03/12/2016	40.265	-74.564	44966
7. Indian Mills	05/01/1901	03/11/2016	39.814	-74.788	41910
8. Lambertville	01/01/1893	01/31/2016	40.366	-74.947	42902
9. Little Falls	09/01/1903	03/12/2016	40.878	-74.220	38855
10. Long Branch	11/01/1907	03/11/2016	40.279	-74.004	36305
11. New Brunswick	01/01/1893	03/12/2016	40.471	-74.436	44995
12. New Milford	01/01/1919	03/12/2016	40.961	-74.015	33877
13. Newark	05/01/1935	11/30/2016	40.682	-74.169	29777
14. Pemberton	08/01/1929	03/09/2016	39.916	-74.578	27061
15. Plainfield	01/01/1893	11/02/2016	40.603	-74.402	42516
16. Somerville	01/01/1893	04/13/2012	40.623	-74.669	44850
17. Sussex	01/01/1893	03/09/2016	41.325	-74.644	42948
18. Toms River	01/01/1893	09/16/2014	39.950	-74.216	43136
19. Woodstown	10/01/1946	03/12/2016	39.546	-75.164	22870
				Total	722774

TABLE 2.1: New Jersey Site Overview

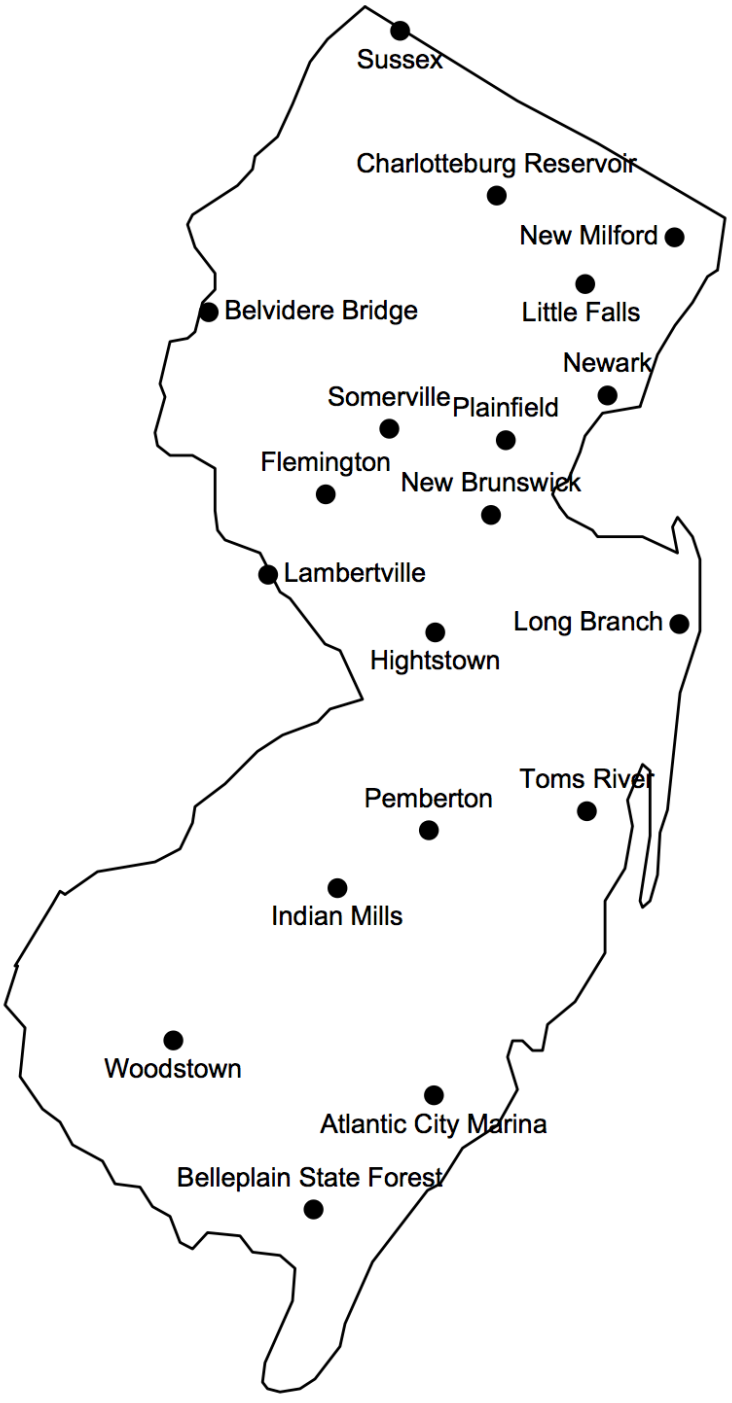


FIGURE 2.1: Graphical Map of New Jersey highlighting locations of 19 selected station sites plotted using corresponding latitude and longitude coordinates.

The requested data was delivered in comma separated values (CSV) files containing ".csv" extensions optimal for use in Microsoft Excel spreadsheets. Each site was delivered in a separate file with record information for all listed years with each containing a set of uniform variables. Table 2.2 lists the variables names with their descriptions and corresponding unit of measure for each variable as provided in the GHCN-Daily documentation file with the data. Selected variables were stripped from the original data files and used to construct a masterfile used for analysis in this project. The masterfile contains a complete compilation of all record observation for the 19 selected sites with a total of 722774 rows or observations.

Variables	**Description and Unit of Measure**
Station Name	Station name (usually city/airport name)
Elevation	Elevation above mean sea level (thousands of meters)
Latitude	Decimated degrees with N hemisphere values > 0, S hemisphere < 0
Longitude	Decimated degrees with E hemisphere values > 0, W hemisphere < 0
Record Date	Date of record
PRCP	Precipitation (tenths of mm, inches)
SNOW	Snowfall (mm, inches)
SNWD	Snow depth (mm, inches)
MDPR	Multiday precipitation total (tenths of mm)
MDSF	Multiday snowfall total
DAPR	Number of days included in MDPR
DASF	Number of days included in MDSF
TMIN	Minimum temperature (Celsius)
TMAX	Maximum temperature (Celsius)
TOBS	Temperature at the time of observation (Celsius)
TAVG	Average of hourly temperature (Celsius)

TABLE 2.2: Summary of Variables in NOAA datset

Chapter 3

Methods

3.1 Data Cleaning

Clean data is a necessary prerequisite for running statistical analyses on datasets of interest (Rahm and Do, 2000). Data quality problems often exist in data collections due to outliers, unusual values, missing information, and other forms of invalid data (Rahm and Do, 2000). Most of these errors are usually due to human error, such as incorrect logging of data or inaccurate measurement recordings (Van-den Broeck et al., 2005). Invalidity of a single measurement and data point might be acceptable however multiple or repeated errors may cause serious issues over time and need to be sorted out (Van-den Broeck et al., 2005). Such problems present in a data set from a single source are usually intensified many times when multiple data files from several sources need to be integrated (Rahm and Do, 2000). Each source usually contains data in a format that serves the specific needs of that particular database and often times, there are several formats accepted within a single database. The compilation of these multiple sources usually results in a large amount of variation and heterogeneity within

the dataset, and need to be corrected before performing any analysis (Rahm and Do, 2000). This increases the need and importance for data cleaning significantly. Data cleaning deals with detecting and correcting these errors and inconsistencies to improve the quality of the data and to minimize their impact on study results (Rahm and Do, 2000; Van-den Broeck et al., 2005).

In general, the process of data cleaning deals with errors once they have occurred (Van-den Broeck et al., 2005) and is usually carried out in a series of steps. In this project, data cleaning was an integral part of the project before running analysis on the rainfall data. The data was first subjected to an initial manual inspection of the data to identify errors and irregularities that might be present within the dataset. The following questions served as a guideline for the check:

1. Is there missing data?
2. Are the precipitation measurements in a single uniform unit of measurement?
3. Do the precipitation measurements fall within an acceptable range of values?
4. Are there any unusual values or potential outliers?

The detected errors were either corrected or removed entirely from the data set before running analysis and constructing time series graphs.

3.1.1 Missing Values

During the initial check for missing values, there were two forms of missing values identified within the data. The first involved missing precipitation measurements for existing record

dates. There was a total of 567 missing rainfall observations which may have been a result of human error, lost precipitation records, incorrect transfer of data, or other various reasons. For this study, these missing values were eliminated since they did not account for a significant proportion (less than 0.1%) of observations when compared to the total number of observations. The second form of missing values were blocks of missing record dates along with their corresponding precipitation measurements. Blocks of sequential years along with precipitation records were missing entirely from several sites. The sites with these missing blocks of years were easily identifiable through the time series graphs constructed for each site. Larger blocks of missing values were easily seen in the time series graphs from the missing points and gaps in the plot.

3.1.2 Precipitation Records

The integrated data set contained precipitation record observations in different units of measure. Several sites had precipitation observations recorded in inches while others had record measurements in millimeters. For homogeneity in the data set, the precipitation records were all converted to a single unit of measure. The observations recorded in millimeters were converted into inches. The rainfall records were then checked to make sure they were within an acceptable range of values. Any abnormal or unexpected values outside the range such as recorded measurements below zero which may have been logged incorrectly were eliminated from the data. The remaining measurements were retained for analysis.

For examination of rainfall behavior, only days with precipitation were extracted from the data while the days with no observed precipitation were omitted. By coding days with precipitation as "1" and the days with no observed rainfall as "0", the desired observations

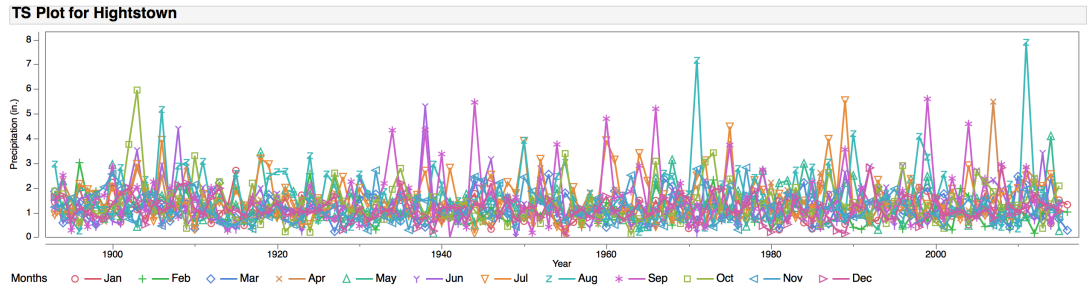
or "successes" were isolated. Table 3.1 shows the total number of observations (N) for each category. As can be seen from Table 3.1, 68.21% of the original data was eliminated to capture the days with precipitation.

k	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	492645	68.21	492645	68.21
1	229563	31.79	722207	100.00

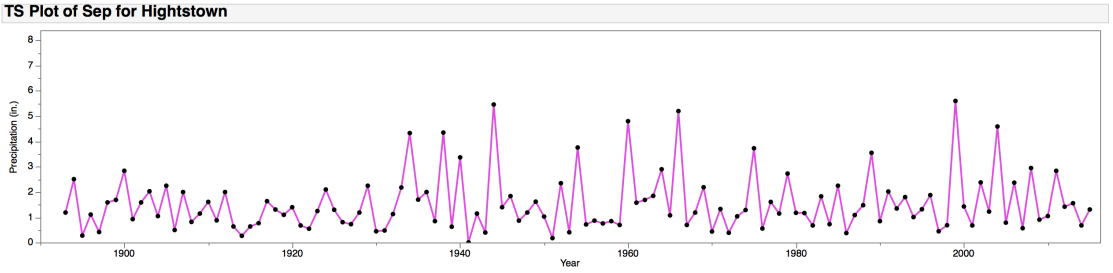
TABLE 3.1: Percentage of precipitation days in NJ: 0=none, 1=precipitation

3.2 Time Series

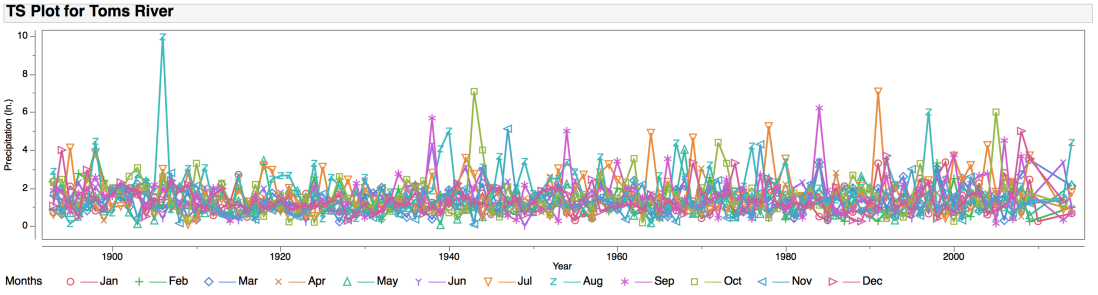
A time series is a sequence of observations taken over a sequential period of time such as daily precipitation measurements (Box et al., 2008). A unique characteristic of a time series is that ordering of the sequence of observations matters since adjacent observations are typically dependent. Unlike regression data, changing the time index would make the analysis of the time series data meaningless. A time series is plotted with time along the horizontal axis and the dependent variable along the vertical axis. Time series (TS) plots can reveal patterns such as trends, unusual observations, potential outliers, period or cycles, or a combination of patterns (Montgomery et al., 2015). Displayed in Figure 3.1 are time series graphs generated in JMP (SAS Institute Inc., 2016) for two of the 19 sites with time in years on the horizontal axis and observed maximum monthly precipitation measurements on the vertical axis with the months overlaid on a single plot. Each month has a different marker with a line connecting through the points. Illustrated in Figure 3.1(a) is the time series plot for Hightstown for all months of the year and a single month of September in Figure 3.1(b).



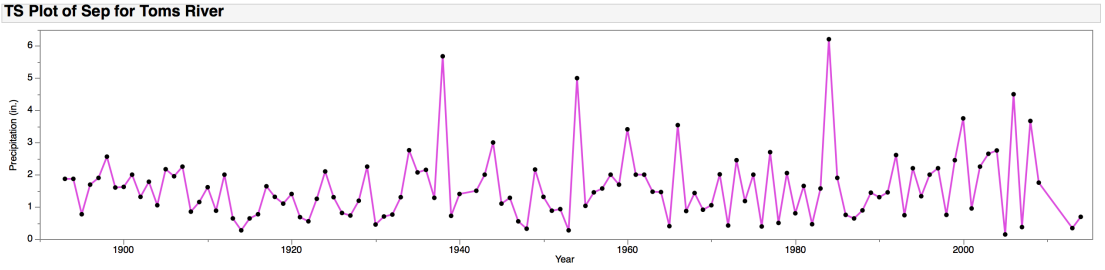
(a)



(b)



(c)



(d)

FIGURE 3.1: Time series plots of monthly maximum precipitation for the sites Hightstown and Toms River in NJ.

The plot shows several high peaks in some years with no noticeable pattern or consistent trend. This behavior was seen in several of the sites through the time series graphs. Displayed in Figure 3.1(c) is the time series plot for Toms River for all months. There is a clear increase of consecutive high peaks after the year 1940 with the exception of a singular high peak around the year 1905 during the month of August. The month of September is again isolated in Figure 3.1(d) where this behavior is emphasized.

3.3 Rain Intensity

Intensity was studied in this project by grouping the monthly maximum precipitation for each month into eight separate categories with each corresponding to a certain sequential range of rainfall amounts. The monthly maximum amounts were categorized rather than the monthly averages since intense rain is more likely to occur during days with the highest amount of rainfall or the monthly maximum. The range for each categorical group is listed in Table 3.2 along with the total number of observations (N) for each group for all selected sites in the state of New Jersey.

Group	Rain intensity	N
1	Max prcp 0.00 - 0.49 in.	1861
2	Max prcp 0.50 - 0.74 in.	3108
3	Max prcp 0.75 - 0.99 in.	4078
4	Max prcp 1.00 - 1.24 in.	3960
5	Max prcp 1.25 - 1.49 in.	3147
6	Max prcp 1.50 - 1.99 in.	3976
7	Max prcp 2.00 - 2.99 in.	2684
8	Max prcp > 3.00 in.	1025
	Total	23840

TABLE 3.2: Summary of Rain Intensity Groups.

To examine the distribution of the groups, a level plot was constructed displaying the information with time in years as the dependent variable and time in months as the independent variable. A level plot aids in comparing the intensities between the years for a particular site. Displayed in Figure 3.2 is the level plot for the sites of Hightstown and Toms River with the time in months on the horizontal axis and the time in years on the vertical axis, with a different panel for each site. Each cell or block corresponds to a certain month of the time series and is color coded based on the group category. As seen in Figure 3.2, there are more frequent blocks of darker shades for months 7 and 8 for both sites. This signifies that intense rain occurs usually during the months of July and August.

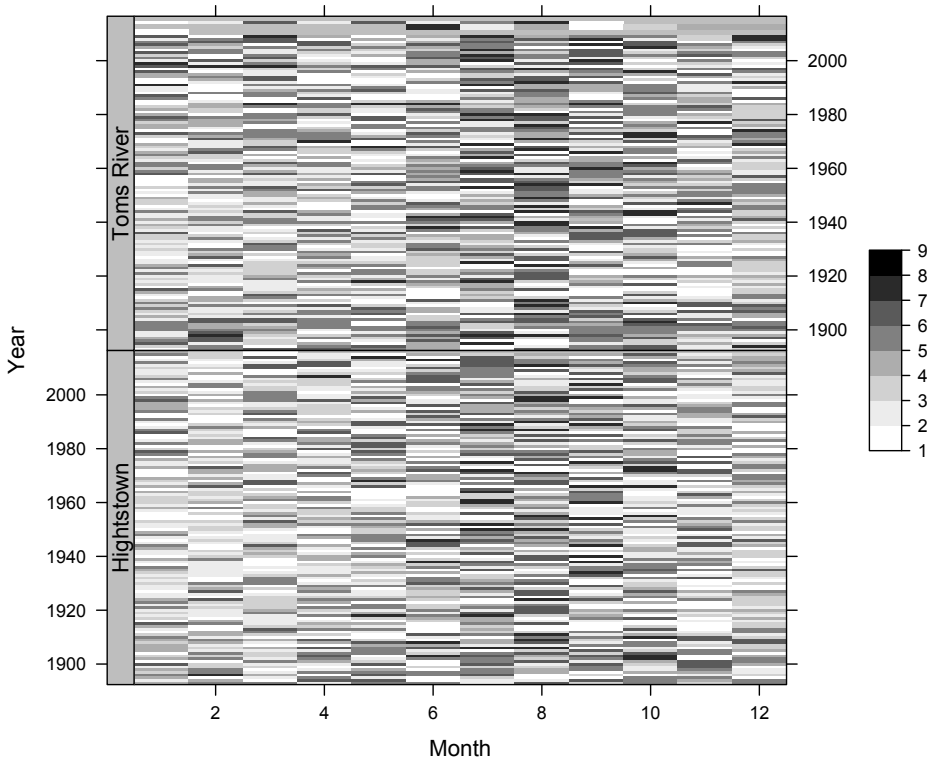


FIGURE 3.2: Level Plot comparing intensities between Hightstown and Toms River.

3.4 Spectral Envelope

Categorical processes, in which harmonic analysis is of interest, occur in medical, behavioral and genetic sciences. Since a categorical process is qualitative, Stoffer et al. (1993) introduce an approach for the spectral analysis and scaling of categorical time series which they refer to as the spectral envelope. The concept of the spectral envelope is generalized in McDougall et al. (1997) to include real-valued time series and Stoffer and Tyler (1998) present an extension of this methodology for the problem of matching categorical sequences. For full details, the reader is referred to the above papers.

3.4.1 Categorical Time Series

Let X_t be a categorical time series with finite state space $\mathscr{C} = \{c_1, c_2, \ldots, c_k\}$. Denote by $X_t(\beta)$ the real-valued time series corresponding to the scaling that assigns state c_j the value β_j ($j = 1, 2, \ldots, k$) where $\beta = (\beta_1, \ldots, \beta_k)' \in \mathscr{R}^k$. Assuming $X_t(\beta)$ has a continuous spectral density $f(\omega;\beta)$, $-\pi < \omega \leq \pi$, the spectral envelope provides a criterion for obtaining scalings that make the spectrum "interesting" in some sense. Specifically, β is chosen to maximize the relative power (variance) across frequencies $\omega \in (-\pi, \pi]$. Rather than work with $X_t(\beta)$ directly, we may use the k-dimensional time series Y_t defined by $Y_t = e_j$ when $X_t = c_j$, where e_j is the $k \times 1$ unit vector having a one in the jth position. Assuming the *basis* vector series Y_t has continuous spectral density $f(\omega)$ it follows that $f(\omega;\beta) = \beta' f(\omega)\beta$ and $\mathrm{Var}[X_t(\beta)] = \beta' V \beta$ where $V = \mathrm{Var}[Y_t]$. Thus the spectral envelope criterion may be expressed as

$$\lambda(\omega) = \sup_{\beta}\{\beta' f(\omega)\beta \,/\, \beta' V \beta\} \quad , \quad -\pi < \omega \leq \pi \tag{3.1}$$

This corresponds to the problem of finding the largest eigenvalue $\lambda(\omega)$ associated with the eigensystem

$$f(\omega)\beta(\omega) = \lambda(\omega)V\beta(\omega) \tag{3.2}$$

over $\beta(\omega)$ not proportional to 1_k, the $k \times 1$ vector of ones. As only real-valued scalings are to be considered, $f(\omega)$ can be replaced by $f^{\text{re}}(\omega)$ in (3.1,3.2) since the imaginary part of a Hermitian matrix is skew symmetric. In addition, $f^{\text{re}}(-\omega) = f^{\text{re}}(\omega)$ implies $\lambda(\omega)$ need only be considered over the positive frequencies.

A graph of the spectral envelope $\lambda(\omega)$ over $0 \leq \omega \leq \pi$, can be readily interpreted as the largest proportion of total power that can be attributed to the frequencies $\omega d\omega$ for any scaled process $X_t(\beta)$ — the maximum being achieved by the scaling $\beta(\omega)$. Scalings for which $\lambda(\omega)$ is relatively large warrant attention and this can be objectively assessed by noting that $\lambda(\omega) = \sigma^2 / 2\pi$ when X_t is a white noise process. Thus, $\lambda(\omega)$ is consistent with the usual notion of a white noise spectrum and implies that non-existent periodicities will not be artificially introduced by the spectral envelope procedure.

3.4.2 Constrained Optimization

In the categorical case, $\text{rank}(V) = k - 1$ so $\lambda(\omega)$ cannot be viewed as the largest root of the determinantal equation $|f^{\text{re}}(\omega) - \lambda V| = 0$ since this is zero for any λ. For computational purposes it is convenient to introduce a constraint matrix A whose columns are linearly independent of 1_k and such that $A'VA$ is positive definite. Then $\lambda(\omega)$ can be defined as the largest eigenvalue of the determinantal equation

$$|A' f^{\text{re}}(\omega) A - \lambda A'VA| = 0$$

where $\lambda(\omega)$ does not depend on the choice of A. Although the corresponding eigenvector $b(\omega)$ will depend on A, the equivalence class of scalings associated with $\beta(\omega) = Ab(\omega)$ does not. Given the multinomial form of Y_t, a simple choice is $A' = [I_{k-1} : 0]$ which corresponds to setting the last component of $b(\omega)$ to zero.

3.4.3 Estimating the Spectral Envelope

The periodogram of the observed vector time series $Y_t,\ t = 1,\ldots,T$ is given by

$$I_T(\omega) = (2\pi T)^{-1} d_T(\omega) d_T^*(\omega) \quad , \quad -\pi < \omega \leq \pi \tag{3.3}$$

where $d_T(\omega) = \sum_{t=1}^{T} Y_t \exp\{-i\,t\omega\}$ is the finite Fourier transform of Y_t and $d_T^*(\omega)$ denotes the complex conjugate transpose. The asymptotic properties of $I_T(\omega)$ as an estimator of the spectral density $f(\omega)$ are well-established and we refer the reader to Brillinger (1981) for details. The key point is that large values of the sample spectral envelope $\lambda_T(\omega)$ based on the periodogram, warrant attention against the assumption that the categorical process X_t is white noise. In practice, $I_T(\omega_j)$ is computed at the discrete frequencies $\omega_j = 2\pi j/T$ for $j = 1,\ldots,[T/2]$ where $[T/2]$ denotes the integer part of $T/2$.

3.5 Functional Data

Functional data analysis provides a unique approach to study complex data in statistics. In a functional data set, the observations are functions in which the basis unit of information is the entire observed function rather than a string of numbers (Sun and Genton, 2011). Each observation is a real function $Y_i(t)$ where $i = 1,\ldots,n$ is an index of the i-th curve from

a total of n curves and $t \in \tau$ where τ is an interval in $\mathscr{R}$. The data values are denoted as $Y_1(t), Y_2(t),...,Y_n(t)$ where $Y_1(t)$ is the first curve in the data set and $Y_n(t)$ is the last curve. For this project, the rainfall data was transformed into functional data set by creating an $n \times p$ matrix for each site with n being the number of curves or years and p being the number of points for each curve for the maximum monthly precipitation record observation for each month of the year. Depicted in Figure 3.3 is a functional data (FD) plot constructed for the site of Hightstown by plotting the recorded maximum monthly precipitation observation for each month for all observed years on a single graph. For this site, $i = 1,\ldots,124$ and $t = 1,\ldots,12$ with $Y_i(t)$ representing a function for the maximum precipitation measurement for month t in year i. There are a total of 124 years or curves matching the total number of available years for the site.

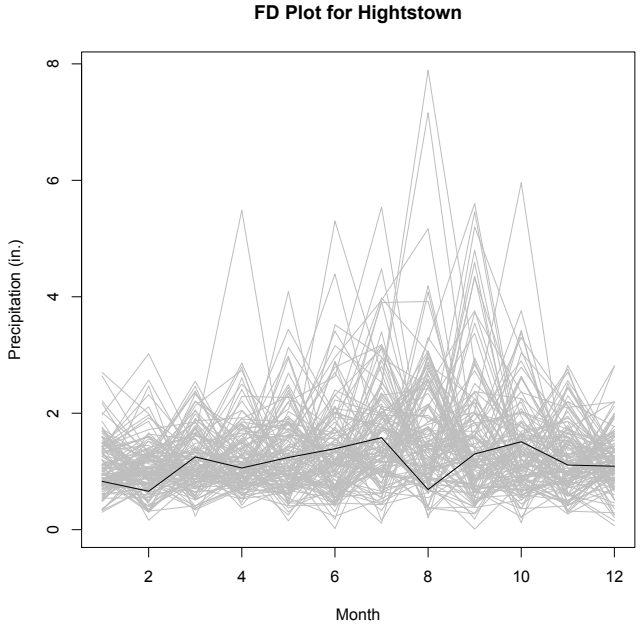


FIGURE 3.3: Functional Data plot for Hightstown with the functional observation for year 1925 highlighted in black.

3.5.1 Outliers

Outliers are observations that deviate significantly from the majority of observations and are often present in data. They can be generated from instrument or human related error and can lead to an incorrect or skewed analysis of the data (Liu et al., 2004). To detect outliers present in the functional datasets, the functional boxplot (FB plot) method was used. In exploratory data analysis, boxplots provide a straightforward but informative method for data visualization (Sun and Genton, 2011). A boxplot is a graphical method to display an overall summary of the data such as the minimum, median, maximum, and the first and third quartile and help highlight any outliers present within the data. Similarly, a functional boxplot is an extension of classical boxplots (Sun and Genton, 2011). They help display summary statistics for functional data and detect potential outlier curves and therefore can be used as an outlier identification method.

López-Pintado and Romo (2009) introduced the notion of band depth (BD) or modified band depth (MBD) for functional data which orders the curves and defines a measure to detect functional quantiles and present outliers (Sun and Genton, 2011). The sample curves are ordered from the center outward: $Y_{[1]}(t), \ldots, Y_{[n]}(t)$ with $Y_{[1]}(t)$ being the most central curve or median curve with the greatest band depth and $Y_{[n]}(t)$ as the most outlying curve and consequently with the lowest band depth. Potential outlier curves which are determined by the $1.5 \times IQR$ (interquartile range) rule of common boxplots applied to functional boxplots that have low band depths (Sun and Genton, 2011). Functional boxplots were constructed for each site using the `fbplot` function from the R package `fda` (Ramsay et al., 2017).

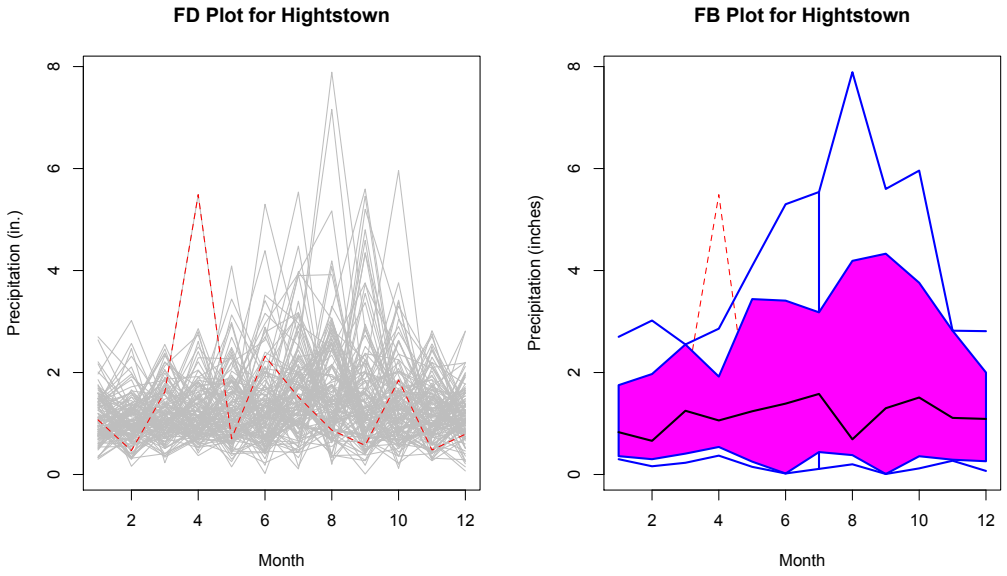


FIGURE 3.4: Functional Boxplot for Hightstown with outlier curve (2007) highlighted in red.

Illustrated in Figure 3.4 is the functional boxplot for the functional data plot for Hightstown. The central 50% region of the data despicted by the shaded region is defined as the envelope (Sun and Genton, 2011) and represents the box of the common boxplot. The envelope is analogous to the IQR and can be used for a general view of the spread of the central curves since the IQR is resistant to the effect of outliers or extreme values. The median curve, $Y_{[1]}(t)$, is highlighted in black inside the envelope. The median curve can be used as a measure of centrality of the central curves. The whiskers are the vertical lines and the attached curves are the most maximum and minimum curves. The identified outlier curve is outlined by the dashed line. For this site, the functional boxplot method detected 1925 as the median year and 2007 as the functional outlier year.

Site	Outlier Year	Median Year	Total Curves
Atlantic City		1994	69
Belleplain State Forest		1982	85
Belvidere Bridge		1914	124
Charlotteburg Reservoir		1959	124
Flemington		1991	115
Hightstown	2007	1925	124
Indian Mills		1974	116
Lambertville	1894	1956	120
Little Falls	1970, 2003	1955	109
Long Branch		1933	102
New Brunswick	2007	1925	124
New Milford	2007	1969	94
Newark	1976	1944	82
Pemberton		1936	80
Plainfield		1906	119
Somerville		1945	124
Sussex		1939	120
Toms River		1933	120
Woodstown		1989	65

TABLE 3.3: Summary Statistics for the functional boxplots of the functional data sets.

Functional boxplots were constructed for each site to detect functional outliers. A summary of the results is presented in Table 3.3 with the site, the outlier year if present, the median year, and the total number of the curves (N) for each site. A few sites have a single outlying curve such as Lambertville and New Brunswick while Little Falls has two outlying curves.

Chapter 4

Results for New Jersey

4.1 Analysis of Regional Rainfall Trends

The plot of New Jersey was split into four regions for this project by grouping nearby sites together. The first region includes the sites in the northern part of New Jersey, the second consists of sites in the central Jersey, the third is composed of the sites in South Jersey, and the fourth are the sites on or near the Jersey coast. A map of this splitting can be seen in Figure 4.1. The sites in each of the regions are listed in Table 4.1. To obtain a single data set for each region, the maximum precipitation record was taken from the monthly maximum records of the sites within each region for each month for all recorded years. This yielded four complete rainfall data sets with no missing years or missing records. A summary of the four data sets is provided in Table 4.2 with the start year, end year, and the total number of observations (N) for each region.

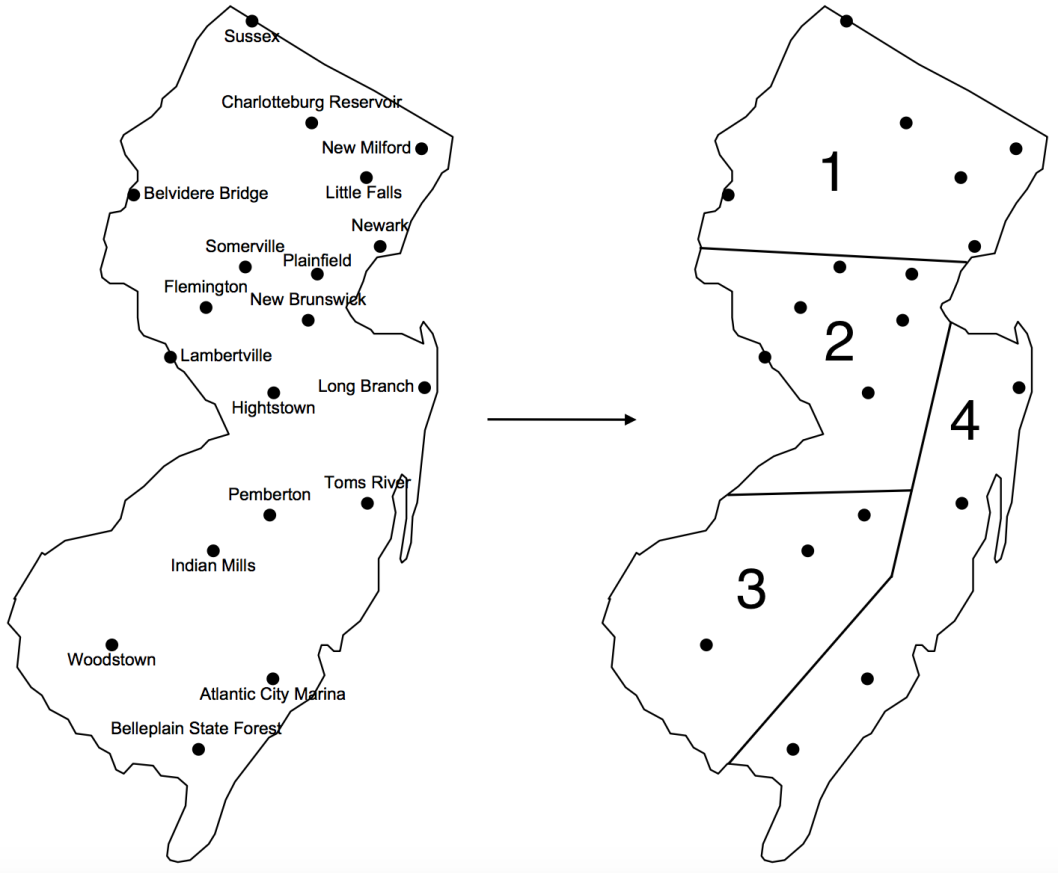


FIGURE 4.1: Map of NJ split into four regions.

Region1	Region 2	Region 3	Region 4
Belvidere Bridge	Flemington	Indian Mills	Atlantic City
Charlotteburg	Hightstown	Pemberton	Belleplain
Newark	Lambertville	Woodstown	Long Branch
New Milford	New Brunswick		Toms River
Little Falls	Plainfield		
Sussex	Somerville		

TABLE 4.1: NJ Sites split into four regions

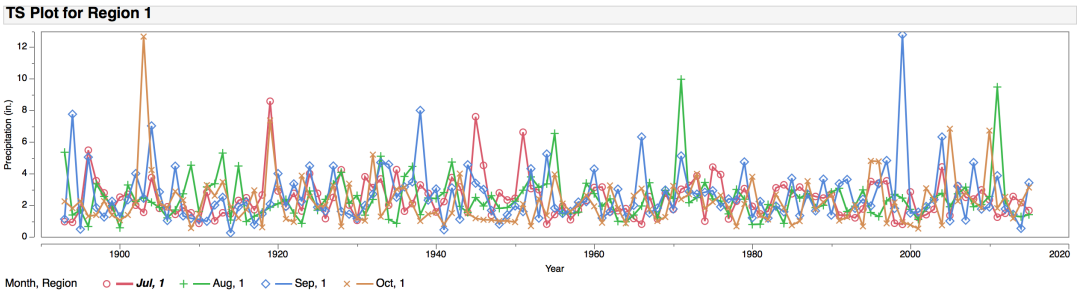
Region	Start Year	End Year	N
Region 1	1893	2016	1479
Region 2	1893	2016	1479
Region 3	1901	2016	1379
Region 4	1893	2016	1479

TABLE 4.2: New Jersey Regions Overview

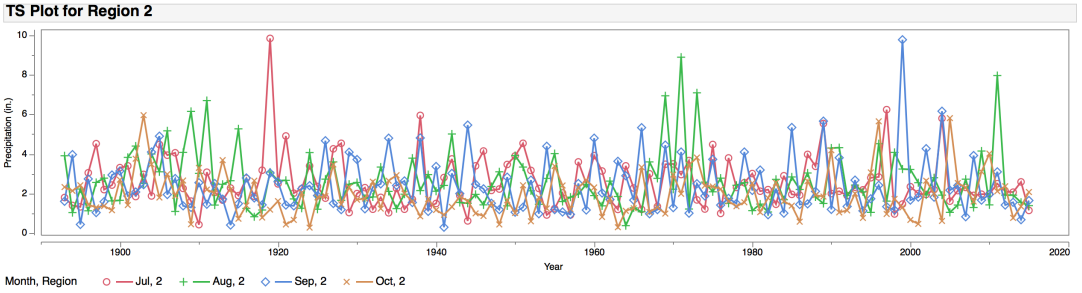
Time series graphs were constructed for each region using the regional monthly maximum data sets to present the data and to examine the rainfall patterns. A logarithmic transformation was applied to the data set to transform the data for possible patterns or underlying interesting behavior. Displayed in Figure 4.2 are the time series plots for months July to October for each region. The time series graphs for Regions 3 and 4 show a significant increase of high peaks after the year 1940. The transformation did not show a clear pattern or trend for Regions 1 and 2 but did show a slight increase in precipitation for Regions 3 and 4.

Individual time series for months July to October are presented in Figures 4.3-4.6. The plots include a smoothed spline outlined in black to highlight the trend of the data. Smooth splines aim to provide a means to smooth a noisy function such as the precipitation data time series. Time series graphs for Regions 3 shows a slight increasing trend for September and October.

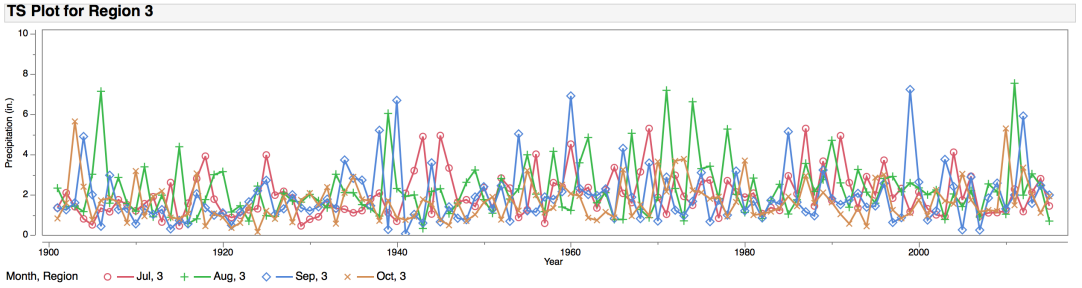
Functional data sets were created for each of the 4 regional monthly maximum data sets and functional boxplots were produced for a summary of the data. Displayed in Figure 4.7 are the functional boxplots for each region. There were no functional outliers detected from the functional boxplot method evident by the absence of dashed lines. The functional regional data appears to be highly variable especially for the later months of the year from July to December. A summary of the results obtained from the functional boxplots is presented in Table 4.3. Each region has a total of 124 curves or years with the exception or Region 3 with a total of 115 curves. Region 3 has a median of the oldest year and Region 1 and 4 with the more recent year.



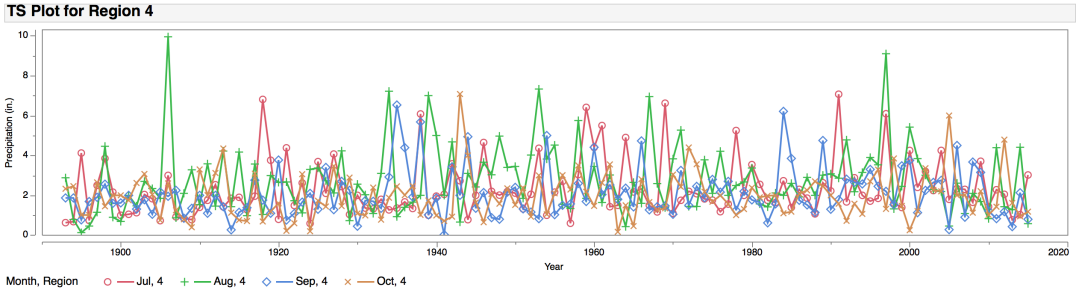
(a)



(b)



(c)



(d)

FIGURE 4.2: Time series plots of regional monthly maximum precipitation for each Region.

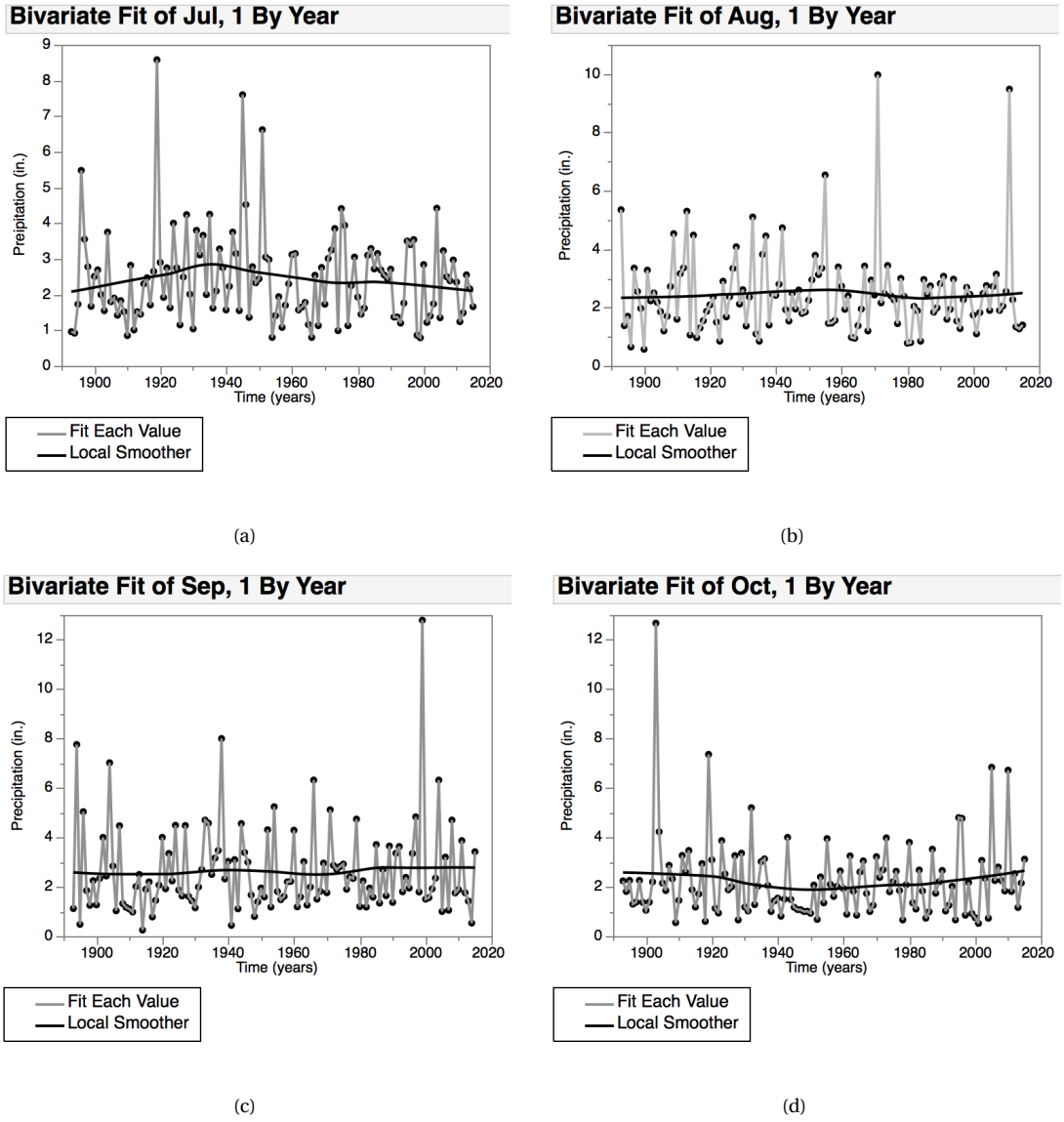


FIGURE 4.3: Individual time series plots for months July-October for Region 1.

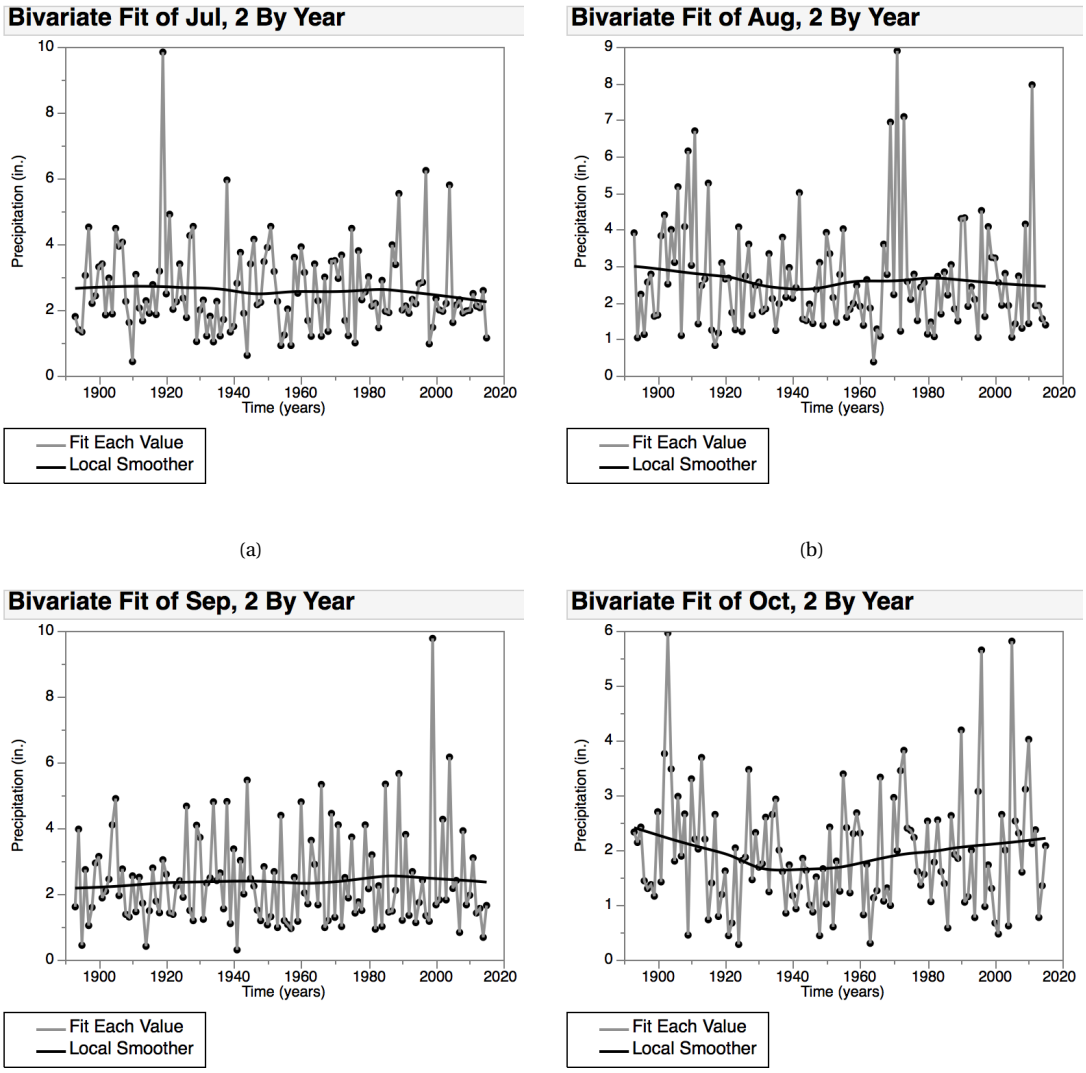


FIGURE 4.4: Individual time series plots for months July-October for Region 2.

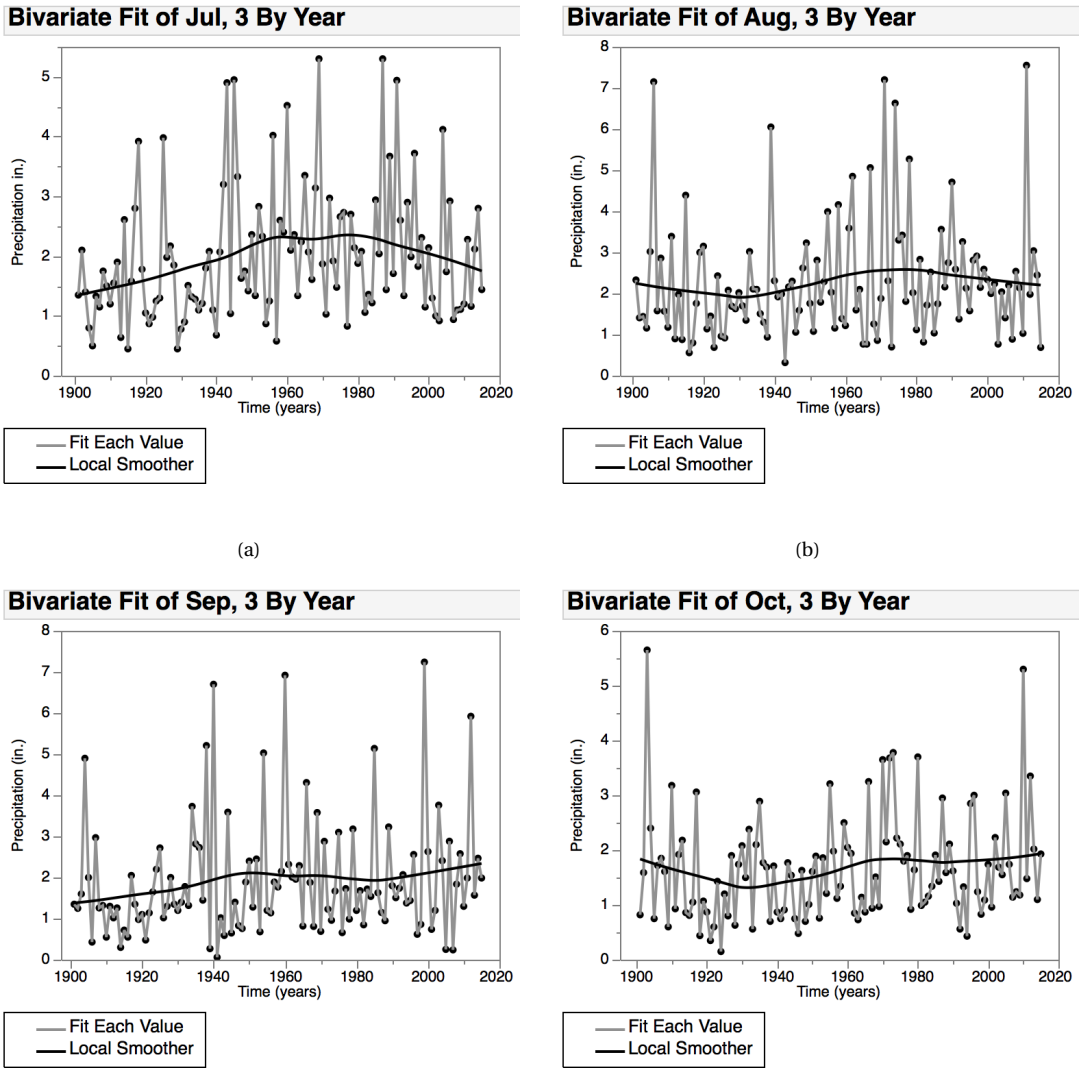


FIGURE 4.5: Individual time series plots for months July-October for Region 3.

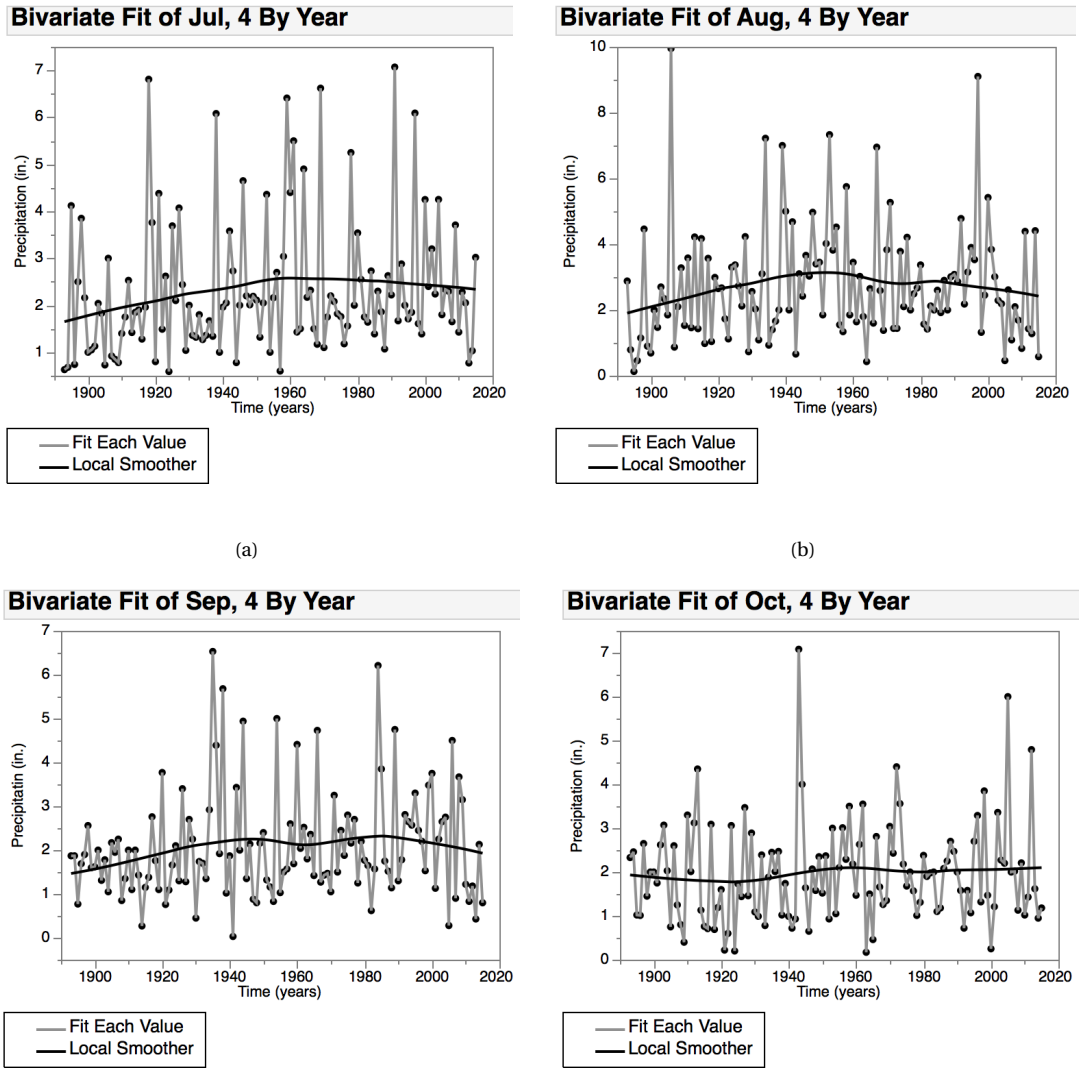


FIGURE 4.6: Individual time series plots for months July-October for Region 4.

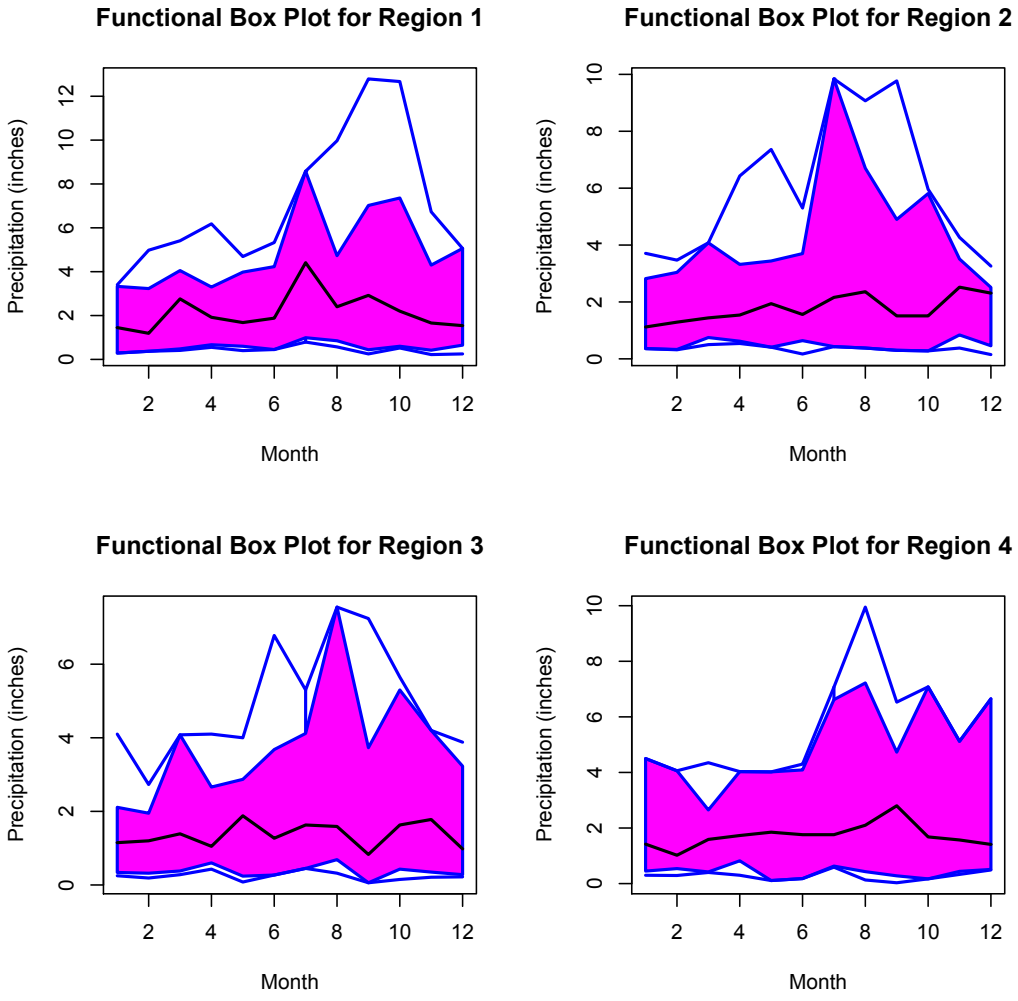


FIGURE 4.7: Functional boxplot for each Region.

Region	Median Curve	Total Curves
Region 1	1975	124
Region 2	1947	124
Region 3	1938	115
Region 4	1975	124

TABLE 4.3: Summary of functional boxplots for each Region.

4.2 Time Series Model

Time series graphs were constructed for each region by plotting regional monthly maximum observations for each month for all years on the vertical axis and time in months for all years on the horizontal axis. Displayed in Figure 4.8 is the time series for Region 1. The time series plot shows no clear upward trend, however there is a large number of high peaks in the later years after 1970. A seasonal ARIMA (AutoRegression Integrated Moving Average) model was considered to describe the patterns of the time series. Seasonality in a times series is a change that repeats at S time periods where S is the number of time periods before the pattern repeats again. A seasonal ARIMA model includes both a non-seasonal and seasonal factor in a single model. The model form is given by: $ARIMA(p,d,q) \times (P,D,Q)_S$ where p is the non-seasonal AR order, d is the non-seasonal differencing, q is the non-seasonal MA order, P is the seasonal AR order, D is the seasonal differencing, Q is the seasonal MA order, and S is the repeating seasonal period (number of observations per season).

By examining the autocorrelation function (ACF) which gives the correlations between the series at x_t and the x_{t-k} for $k = 1,2,...$(lagged values). The data for regions appeared to be non-stationary as many lags were significant. This indicated that a difference might be necessary to create a stationary series. The difference between two time points, $x_t - x_{t-1}$, can be expressed as $(1-B)x_{t-1}$. For a seasonal difference, the difference is between the value and a value with a lag that is a multiple of S. For a seasonality with $S = 12$, the difference is $x_t - x_{t-12}$ and can be expressed as $(1-B^{12})x_t$. The ACF of the seasonally differenced data displayed a significant spike at the 12^{th} lag before cutting off while the partial autocorrelation function

(PACF) plot displayed a tapering pattern at multiples of twelve. However, the seasonal difference resulted in unstable ARIMA parameters. For optimal models, the `auto.arima` function from the `forecast` (Hyndman and Khandakar, 2008) library in R was used to generate a model that would fit the time series. A summary of the optimal ARIMA models estimated from R along with the associated Akaike's Information Criterion (AIC) is provided in Table 4.4.

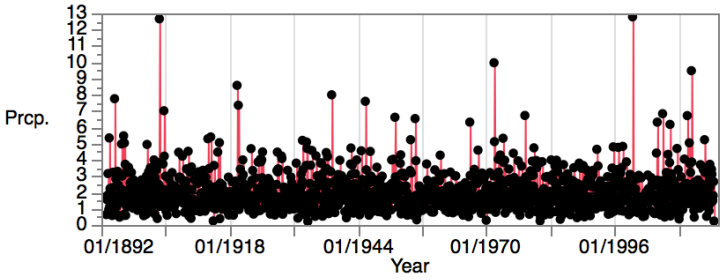
Region	TS Model	AIC
Region 1	$ARIMA(1,0,0)(2,0,0)_{12}$	4634.76
Region 2	$ARIMA(2,0,3)(0,0,2)_{12}$	4308.68
Region 3	$ARIMA(3,1,0)(0,0,2)_{12}$	4047.87
Region 4	$ARIMA(1,1,0)(1,0,0)_{12}$	4917.58

TABLE 4.4: Summary of Time Series ARIMA models for NJ Regions.

For region 1, the optimal estimated model is $ARIMA(1,0,0)(2,0,0)_{12}$ which has a periodic seasonality of 12, a non-seasonal AR order of 1, and a seasonal order of 2. Figure 4.8 shows the parameter estimates for the estimated model and the model summary. Presented in Figure 4.9 is the white noise test for the residuals for Region 1 where the null hypothesis tests whether the series is white noise process. A white noise process is a random process where the observations are uncorrelated to one another. It is expected that the residuals should be uncorrelated and the series should be a white noise process. However based on the small p -value, the null hypothesis is rejected indicating that the series is not a white noise process. Although the residuals show significance, the models provide a means to examine the overall rainfall trend and a rough estimated forecast for the future trend. The predicted time series plot for Region 1 is displayed in 4.10. The plot displayed the predicted values on the vertical axis along with a forecast for the next 25 years plotted past the vertical line. The models for Regions 2, 3, and 4 are provided in Figures 4.11, 4.14, and 4.17 with the predicted time series in Figures 4.13, 4.16, and 4.19. The white noise tests for Regions 2, 3, and 4 are

displayed in Figures 4.12, 4.15, and 4.18. Region 3 (Southern Jersey) shows increased rainfall events after 1940. No clear pattern was seen for the other regions.

Region 1



Mean	2.0056891
Std	1.1664431
N	1479
Zero Mean ADF	-13.8405
Single Mean ADF	-35.07145
Trend ADF	-35.0597

Model: Seasonal ARIMA(1, 0, 0)(2, 0, 0)12

Model Summary

DF	1475	Stable	Yes
Sum of Squared Errors	1977.13682	Invertible	Yes
Variance Estimate	1.34043174		
Standard Deviation	1.15777016		
Akaike's 'A' Information Criterion	4634.76405		
Schwarz's Bayesian Criterion	4655.96053		
RSquare	0.01722774		
RSquare Adj	0.01522888		
MAPE	53.469272		
MAE	0.79935592		
-2LogLikelihood	4626.76405		

Parameter Estimates

Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob>|t|	Constant Estimate	Mu
AR1,1	1	1	0.0816377	0.0262306	3.11	0.0019*	2.00580988	
AR2,12	2	12	0.0387511	0.0262228	1.48	0.1397	1.60878025	
AR2,24	2	24	0.0878897	0.0261744	3.36	0.0008*		
Intercept	1	0	2.0058099	0.0374068	53.62	<.0001*		

FIGURE 4.8: Optimal ARIMA model summary and parameter estimates for Region 1.

White Noise Test: Region 1

Fisher's Kappa	28.17922
Prob > Kappa	2.559e-10
Bartlett's Kolmogorov-Smirnov	0.0388451

FIGURE 4.9: White noise test for Region 1.

Forecast

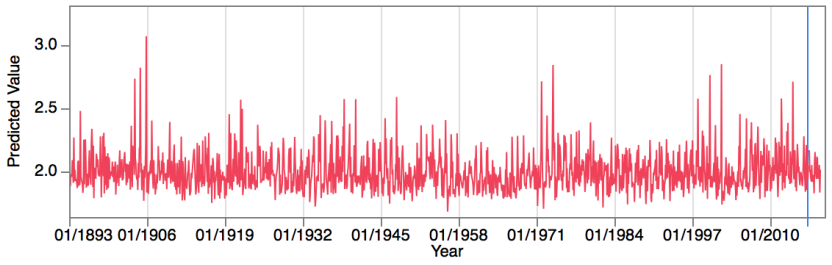
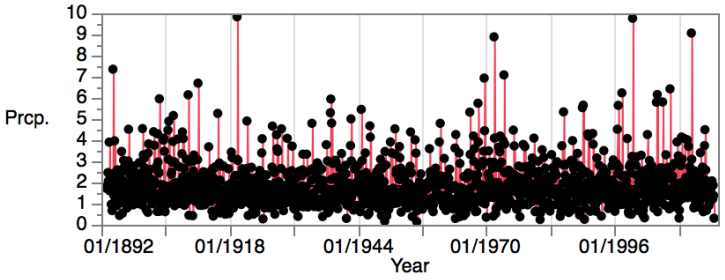


FIGURE 4.10: Time series of predicted observations from estimated ARIMA model for Region 1.

Region 2



Mean	1.876775
Std	1.0656673
N	1479
Zero Mean ADF	-13.17641
Single Mean ADF	-33.39505
Trend ADF	-33.38444

Model: Seasonal ARIMA(2, 0, 3)(0, 0, 2)12

Model Summary

DF	1471	Stable	Yes
Sum of Squared Errors	1576.9808	Invertible	Yes
Variance Estimate	1.07204677		
Standard Deviation	1.03539691		
Akaike's 'A' Information Criterion	4308.68448		
Schwarz's Bayesian Criterion	4351.07745		
RSquare	0.06058094		
RSquare Adj	0.05611056		
MAPE	52.4689608		
MAE	0.72987965		
-2LogLikelihood	4292.68448		

Parameter Estimates

Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob>|t|	Constant Estimate	Mu
AR1,1	1	1	0.144823	0.2295957	0.63	0.5283	0.72397377	1.87714905
AR1,2	1	2	0.469499	0.1601014	2.93	0.0034*		
MA1,1	1	1	0.040453	0.2304675	0.18	0.8607		
MA1,2	1	2	0.452259	0.1514693	2.99	0.0029*		
MA1,3	1	3	0.146836	0.0266539	5.51	<.0001*		
MA2,12	2	12	-0.057302	0.0263004	-2.18	0.0295*		
MA2,24	2	24	-0.145711	0.0241561	-6.03	<.0001*		
Intercept	1	0	1.877149	0.0301388	62.28	<.0001*		

FIGURE 4.11: Optimal ARIMA model summary and parameter estimates for Region 2.

White Noise Test: Region 2

Fisher's Kappa	43.796382
Prob > Kappa	1.94e-17
Bartlett's Kolmogorov-Smirnov	0.0415409

FIGURE 4.12: White noise test for Region 2.

Forecast

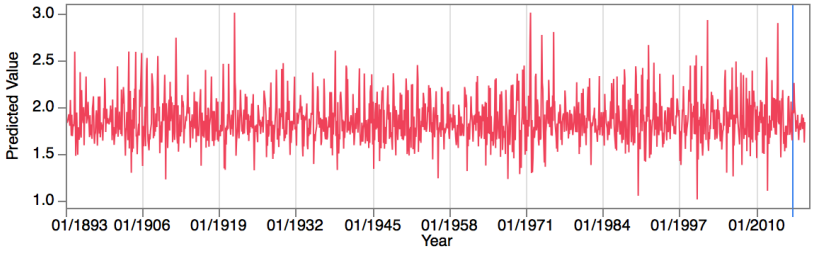
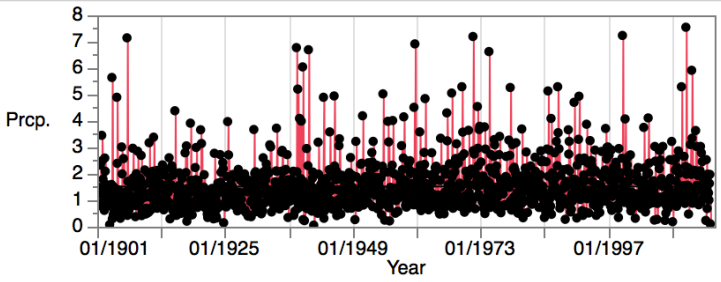


FIGURE 4.13: Time series of predicted observations from estimated ARIMA model for Region 2.

Region 3



Mean	1.5550274
Std	0.9542669
N	1379
Zero Mean ADF	-13.95868
Single Mean ADF	-33.70036
Trend ADF	-34.20767

Model: Seasonal ARIMA(3, 1, 0)(0, 0, 2)12

Model Summary

DF	1372	Stable	Yes
Sum of Squared Errors	1507.90997	Invertible	Yes
Variance Estimate	1.09905974		
Standard Deviation	1.0483605		
Akaike's 'A' Information Criterion	4047.50262		
Schwarz's Bayesian Criterion	4078.87295		
RSquare	-0.2010526		
RSquare Adj	-0.2054296		
MAPE	66.7570517		
MAE	0.7306727		
-2LogLikelihood	4035.50262		

Parameter Estimates

Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob>|t|	Constant Estimate	Mu
AR1,1	1	1	-0.7031792	0.0267168	-26.32	<.0001*	-0.0006554	-0.0002715
AR1,2	1	2	-0.4758147	0.0304195	-15.64	<.0001*		
AR1,3	1	3	-0.2346569	0.0263389	-8.91	<.0001*		
MA2,12	2	12	-0.0575086	0.0278162	-2.07	0.0389*		
MA2,24	2	24	-0.0562900	0.0244431	-2.30	0.0214*		
Intercept	1	0	-0.0002715	0.0008624	-0.31	0.7529		

FIGURE 4.14: Optimal ARIMA model summary and parameter estimates for Region 3.

White Noise Test: Region 3

Fisher's Kappa	48.052919
Prob > Kappa	1.719e-19
Bartlett's Kolmogorov-Smirnov	0.121468

FIGURE 4.15: White noise test for Region 3.

Forecast

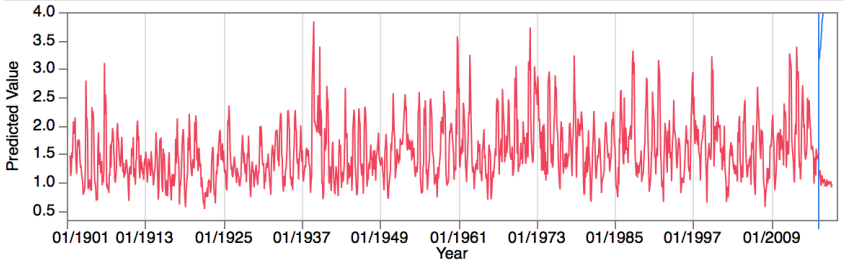
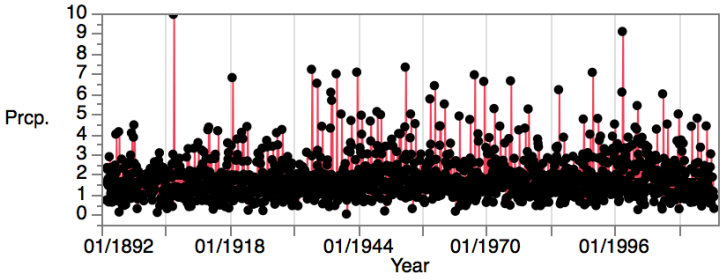


FIGURE 4.16: Time series of predicted observations from estimated ARIMA model for Region 3.

Region 4



Mean	1.8208365
Std	1.0682745
N	1479
Zero Mean ADF	-13.74012
Single Mean ADF	-34.24453
Trend ADF	-34.42284

Model: Seasonal ARIMA(1, 1, 0)(1, 0, 0)12

Model Summary

DF	1475	Stable	Yes
Sum of Squared Errors	2375.0276	Invertible	Yes
Variance Estimate	1.6101882		
Standard Deviation	1.26893191		
Akaike's 'A' Information Criterion	4901.6882		
Schwarz's Bayesian Criterion	4917.58354		
RSquare	-0.4082497		
RSquare Adj	-0.4101592		
MAPE	70.7190884		
MAE	0.89057035		
-2LogLikelihood	4895.6882		

Failed: Cannot Decrease Objective Function Hessian is not positive definite.

Parameter Estimates

Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob>|t|	Constant Estimate	Mu
AR1,1	1	1	-0.4534524	0.0232323	-19.52	<.0001*		
AR2,12	2	12	0.0514985	0.0258478	1.99	0.0465*	-0.0004755	-0.0003449
Intercept	1	0	-0.0003449	.	.	.		

FIGURE 4.17: Optimal ARIMA model summary and parameter estimates for Region 4.

White Noise Test: Region 4

Fisher's Kappa	18.602338
Prob > Kappa	4.9738e-6
Bartlett's Kolmogorov-Smirnov	0.1682165

FIGURE 4.18: White noise test for Region 4.

Forecast

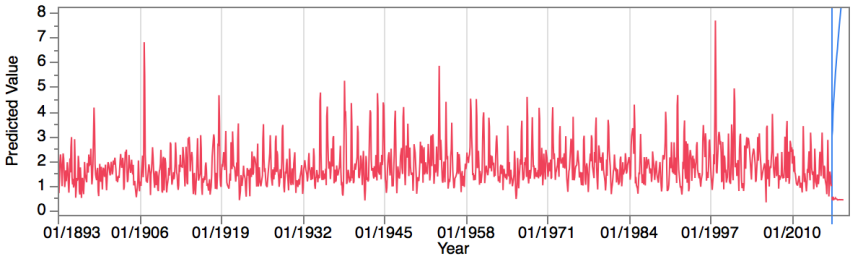


FIGURE 4.19: Time seire of predicted observations from estimated ARIMA model for Region 4.

4.3 Regional Rain Intensity

The regional monthly maximum data sets were grouped to explore intensity. Depicted in Figure 4.20 is the level plot comparing Regions 1 and 2. There are a large number of dark shaded blocks for most months with the highest amount for the months of July and August signifying intense precipitation occurred most months of the year. Similarly, displayed in Figure 4.21 is the level plot comparing Regions 3 and 4.

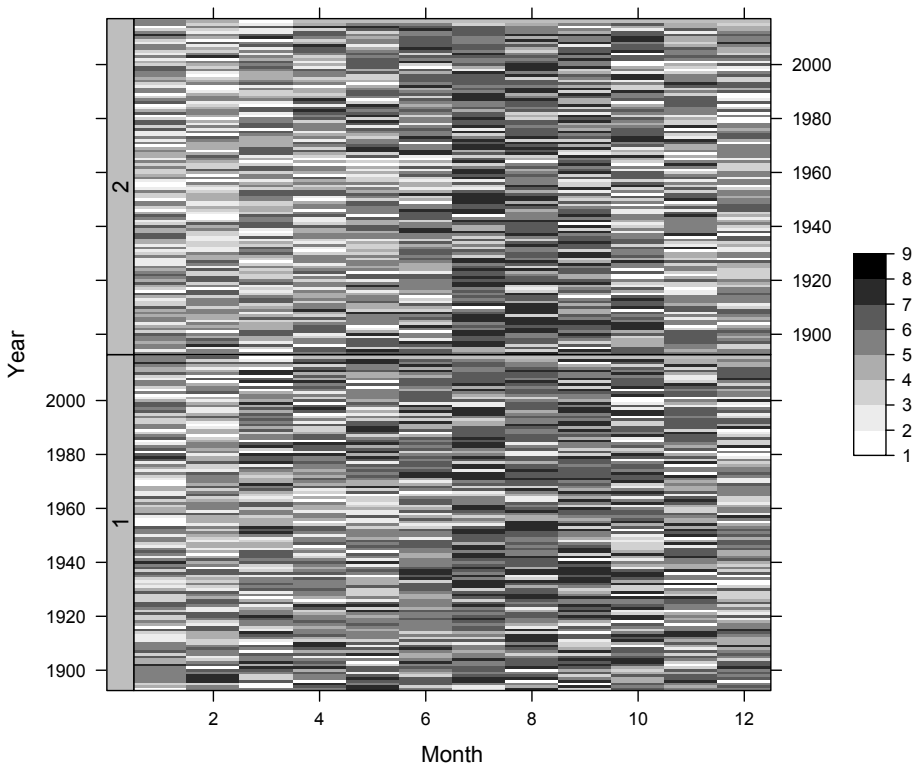


FIGURE 4.20: Level plot comparing intensities for Regions 1 and 2.

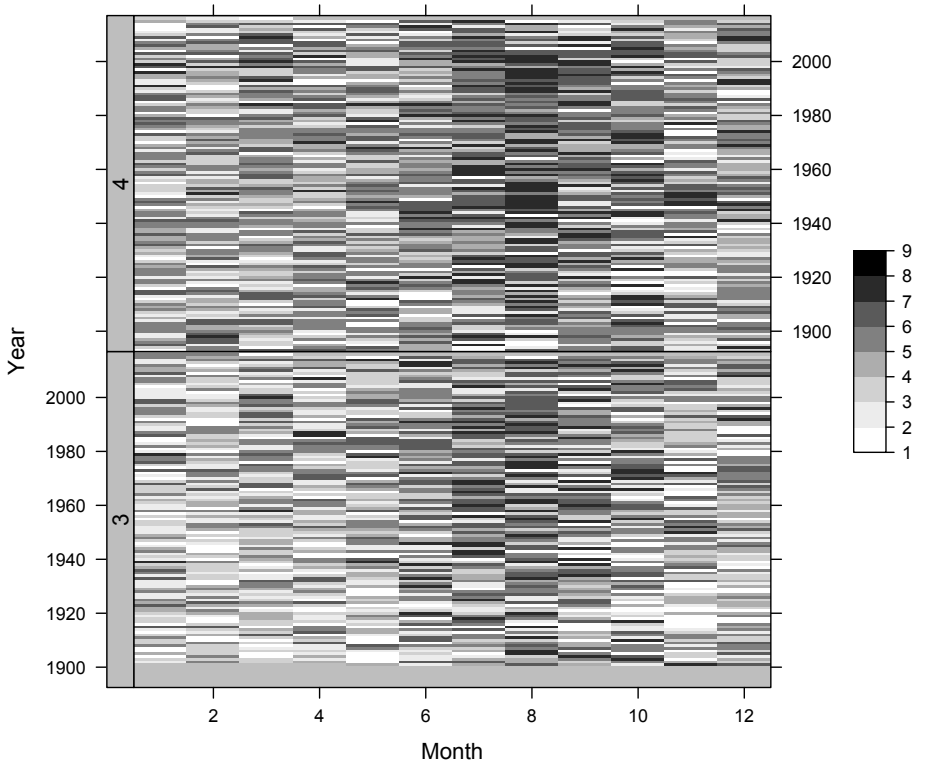


FIGURE 4.21: Level plot comparing intensities for Regions 3 and 4.

There appears to be more blocks of lighter shades for Region 3 compared to the other Regions especially for the earlier months of the year, however there are darker blocks in the later years. In Region 4, there are darker shades showing more intense rainfall than Region 3 while notably less than Region 1 and 2.

4.3.1 Categorical Time Series

Displayed in Figure 4.22 is the categorical time series for Region 3. The intensities are plotted on the vertical axis and the time in years on the horizontal axis. The spectral envelope plot for Region 3 is shown in Figure 4.23. The periodogram shows a significant peak at approximately .083, the period for this value is $1/0.083 \approx 12$. This shows that there is a complete cycle

or repeating pattern every 12 months as is expected. The spectral envelopes for the remaining three regions also had similar spectral envelopes with a dominant peak approximately around 0.083. The approximate 0.00001 null significance threshold for white noise is shown by the dashed line. The optimal scaling $\beta(\omega)'Y_t$ for this peak is

$$\beta = (0.29, 0.32, 0.32, 0.33, 0.27, 0.24, 0.13)$$

which is essentially a weighted average of the low to moderate rainfall categories with less weight assigned to high rainfall events (recall that group 8 was zeroed).

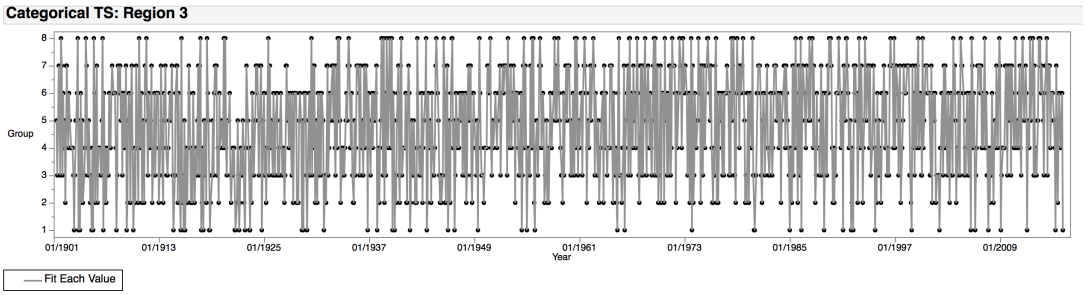


FIGURE 4.22: Categorical Time Series for Region 3 with the group on the vertical axis and time on the horizontal axis.

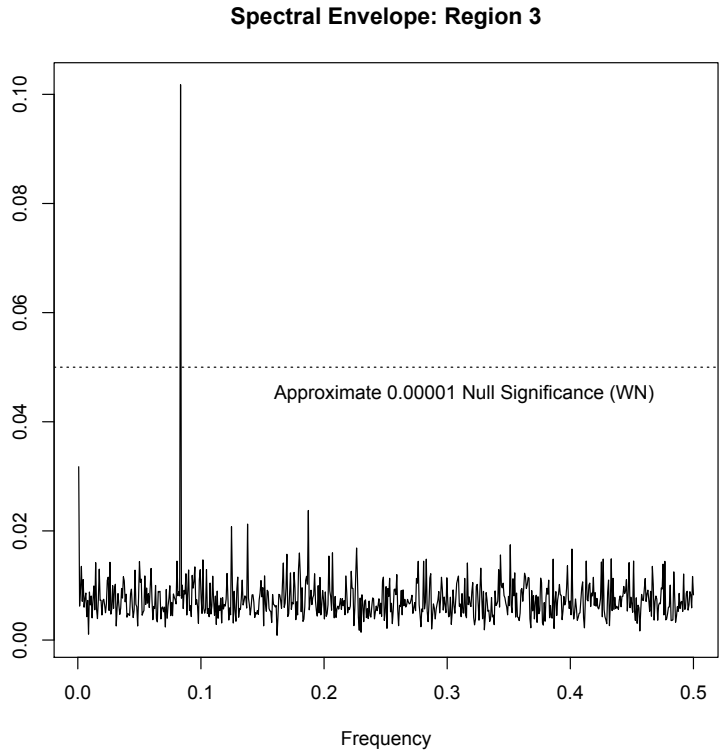


FIGURE 4.23: Spectral Envelope periodogram for Region 3 shows a dominant peak at approximately 0.083.

Chapter 5

Extensions

5.1 Introduction

Examination of rainfall patterns was extended to selected states across the United States (U.S.). A complete list of the states is provided in Table 5.1. The combined data file for the selected sates contained a total of 1607668 observations. Plots of the states with highlighted sites are illustrated in Figures 5.1-5.6, along with the site overview in Tables 5.2-5.7.

State	Total Sites	N
Alabama	9	285094
California	11	358308
Florida	8	230448
Illinois	5	106650
Louisiana	8	198372
Missouri	8	204206
Texas	8	224590
		1607668

TABLE 5.1: Selected states overview.

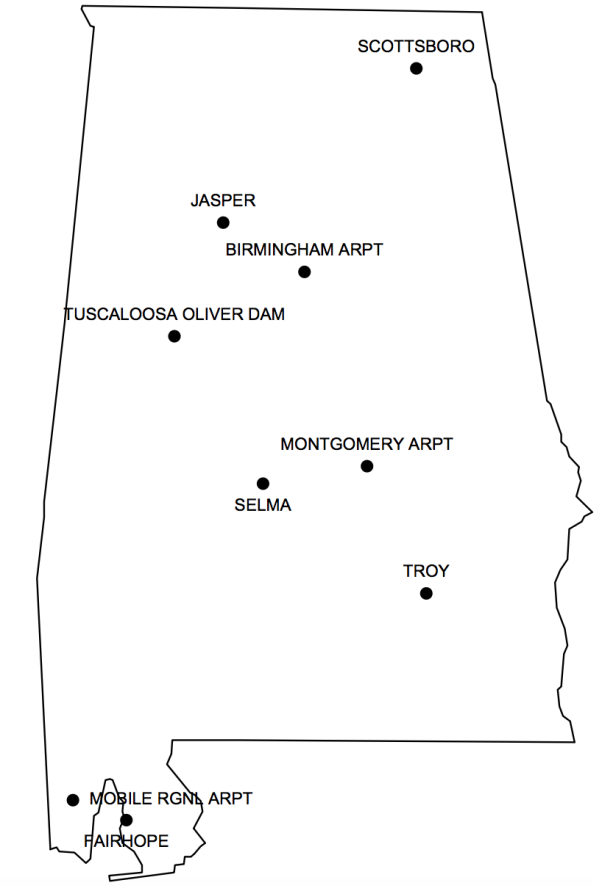


FIGURE 5.1: Graphical Map of Alabama highlighting locations of selected station sites.

Site	Start Year	End Year	Latitude	Longitude	N
1. Birmingham Arpt	01/01/1930	08/06/2017	33.562	-86.754	13019
2. Fairhope	08/01/1917	08/04/2017	30.572	-87.900	36184
3. Jasper	10/01/1891	08/06/2017	33.831	-87.277	21004
4. Mobile Rgnl Arpt	01/01/1948	08/06/2017	30.681	-88.244	25412
5. Montgomery Arpt	01/01/1948	08/06/2017	32.502	-86.351	25412
6. Scottsboro	10/08/1891	08/05/2017	34.672	-86.034	40868
7. Selma	01/01/1895	08/06/2017	32.407	-87.021	42600
8. Troy	06/01/1908	07/23/2017	31.808	-85.969	38917
9. Tuscaloosa Dam	0/01/1900	08/06/2017	33.211	-87.591	41678
				Total	285094

TABLE 5.2: Alabama Site Overview

FIGURE 5.2: Graphical Map of California highlighting locations of selected sites.

Site	Start Year	End Year	Latitude	Longitude	N
1. Bakersfield Arpt	10/01/1937	08/06/2017	35.423	-119.039	29156
2. Fresno Intl Arpt	12/04/1941	08/06/2017	36.775	-119.718	26887
3. Los Banos	03/01/1906	08/06/2017	37.058	-120.849	35873
4. Palm Springs	03/01/1906	08/03/2017	33.830	-116.545	37294
5. Redding Arpt	09/01/1986	08/06/2017	40.509	-122.293	11289
6. Sacramento	07/11/1877	08/04/2017	38.574	-121.550	42865
7. San Diego Intl Arpt	07/01/1939	08/06/2017	32.732	-117.197	28490
8. San Francisco Dwtn	01/01/1921	08/04/2017	37.794	-122.399	35270
9. San Francisco Intl Arpt	07/01/1945	08/06/2017	37.615	-122.392	26287
10. Santa Barbara	01/01/1893	08/03/2017	34.434	-119.719	42181
11. Santa Cruz	01/01/1893	08/05/2017	36.974	-122.030	42716
				Total	358308

TABLE 5.3: California Site Overview

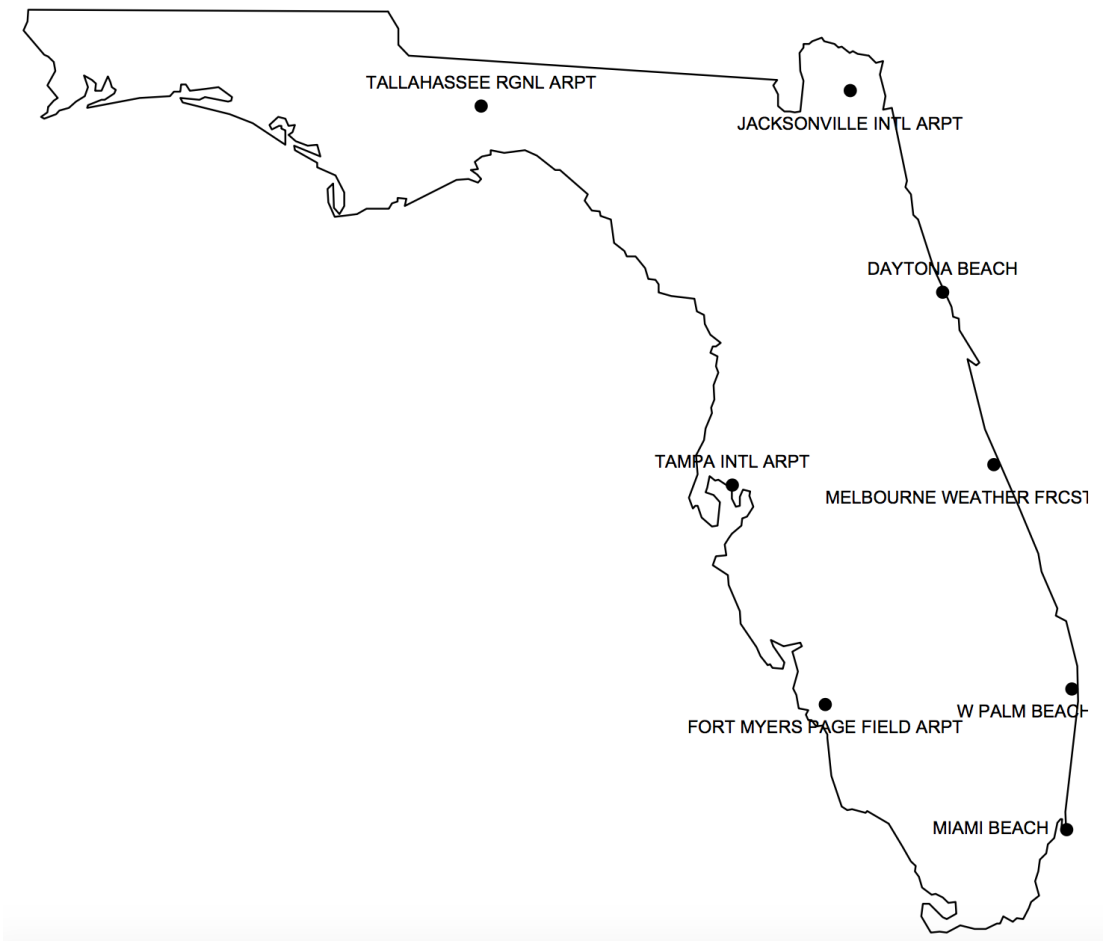


FIGURE 5.3: Graphical Map of Florida highlighting locations of selected station sites.

Site	Start Year	End Year	Latitude	Longitude	N
1. Daytona Beach	02/01/1923	08/08/2017	29.210	-81.022	20176
2. Fort Meyers Arpt	01/01/1892	08/03/2017	26.586	-81.867	40913
3. Jacksonville Intl Arpt	04/01/1938	08/05/2017	30.494	-81.687	28913
4. Melbourne	07/21/1937	08/05/2017	28.113	-80.654	28350
5. Miami Beach	01/05/1927	08/04/2017	25.790	-80.130	28028
6. Tallahassee Rgnl Arpt	05/01/1942	08/05/2017	30.395	-84.345	27066
7. Tampa Intl Arpt	02/01/1939	08/05/2017	27.983	-82.537	28668
8. West Palm Beach	07/01/1938	08/05/2017	26.685	-80.092	28334
				Total	230448

TABLE 5.4: Florida Site Overview

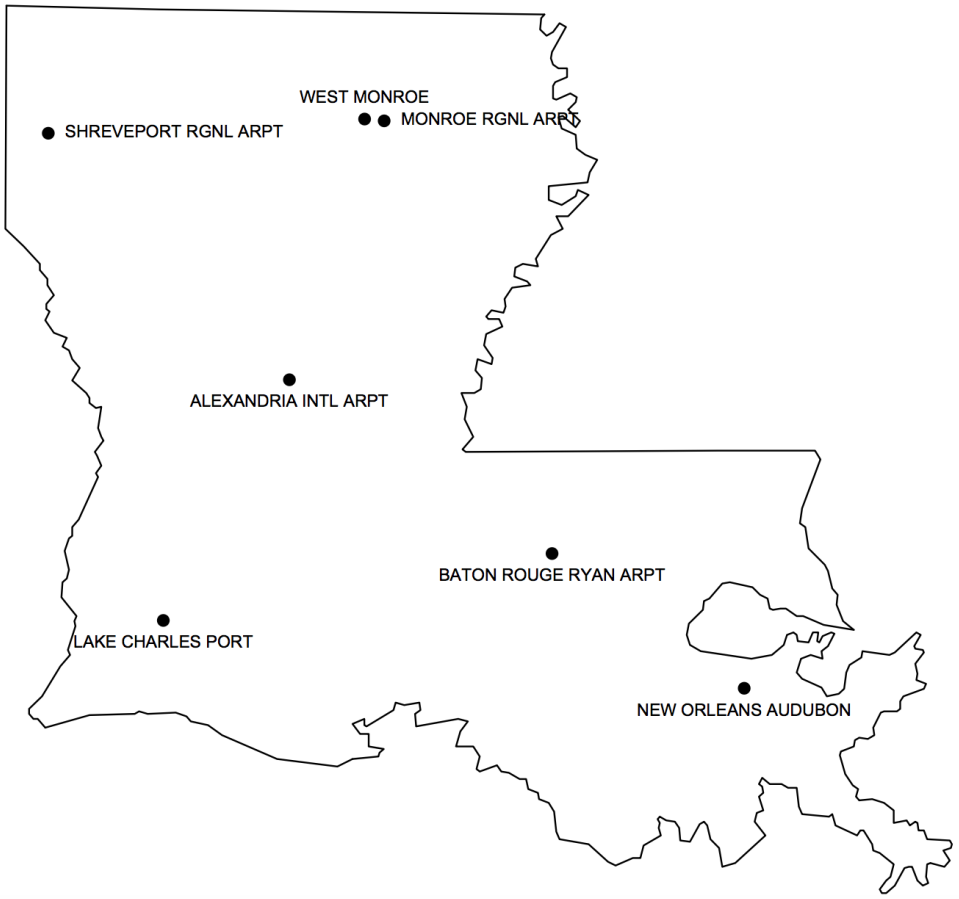


FIGURE 5.4: Graphical Map of Louisiana highlighting locations of selected station sites.

Site	Start Year	End Year	Latitude	Longitude	N
1. Alexandria Intl Arpt	01/01/1948	08/03/2017	31.326	-92.547	8162
2. Baton Rouge Arpt	01/01/1930	08/03/2017	30.532	-91.151	31984
3. Lake Charles Port	02/02/1955	08/05/2005	30.226	-93.217	9109
4. Lake Providence	01/01/1893	08/05/2017	32.804	91.170	33285
5. Monroe Rgnl Arpt	01/01/1930	08/03/2017	32.510	-92.043	28666
6. New Orleans	01/02/1893	08/03/2017	29.916	-90.130	44257
7. Shreveport Rgnl Arpt	07/01/1939	08/05/2017	32.453	-93.828	28518
8. West Monroe	01/01/1938	08/03/2017	32.518	-92.147	14391
				Total	198372

TABLE 5.5: Louisiana Site Overview

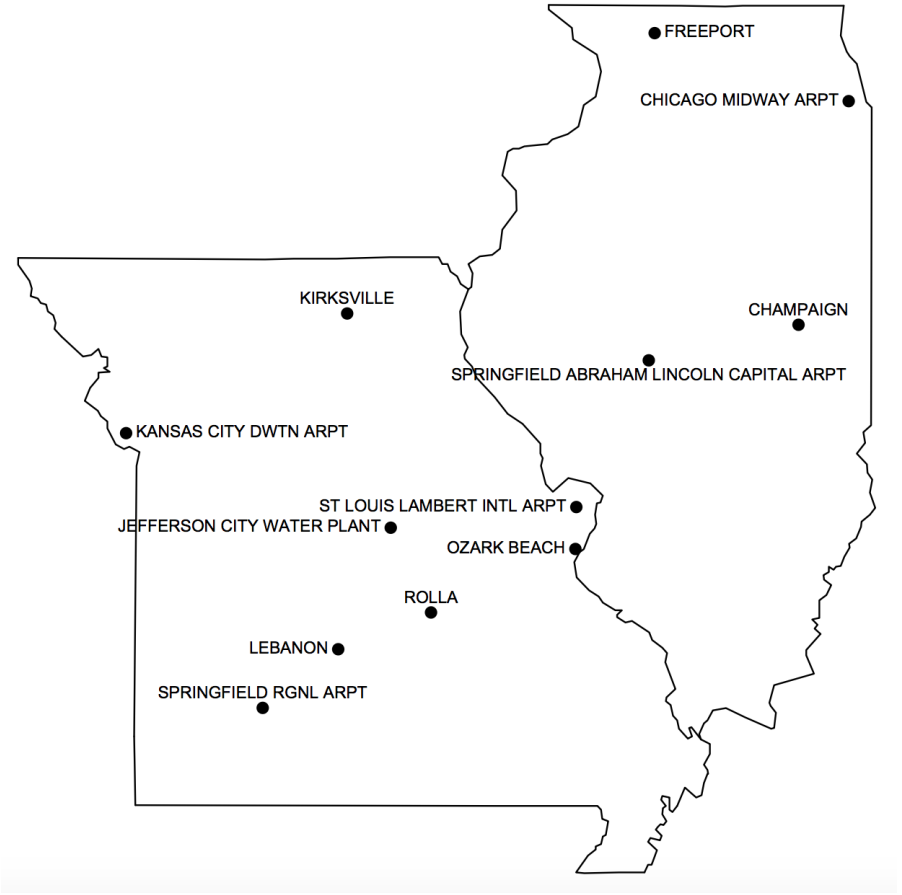


FIGURE 5.5: Graphical Joint Map of Missouri and Illinois highlighting locations of selected station.

Site	Start Year	End Year	Latitude	Longitude	N
1. Champaign	08/01/1902	08/06/2017	40.11025	-88.235385	32658
2. Chicago Midway Arpt	02/29/1928	08/05 2017	41.78678	-87.752188	31266
3. Freeport	06/01/1948	08/06/2017	42.29669	-89.621227	10068
4. Jefferson City Plant	01/01/1893	08/07/2017	38.5888	-92.163541	10060
5. Kansas City Dwtn Arpt	01/01/1934	08/05/2017	39.2977	-94.712949	9643
6. Kirksville	02/01/1893	08/07/2017	40.19475	-92.58325	10019
7. Lebanon	01/01/1893	06/03/2017	37.67775	-92.669536	42964
8. Ozark Beach	07/01/1924	08/07/2017	38.42932	-90.384741	31992
9. Rolla	01/01/1893	08/07/2017	37.95345	-91.775573	42429
10. Springfield A.L. Arpt	01/01/1901	08/06/2017	39.84351	-89.678096	32658
11. Springfield Rgnl Arpt	08/01/1940	08/07/2017	37.23808	-93.397096	28129
12. St. Louis Intl Arpt	04/01/1938	08/06/2017	38.74373	-90.375881	28970
				Total	310856

TABLE 5.6: Missouri and Illinois Site Overview

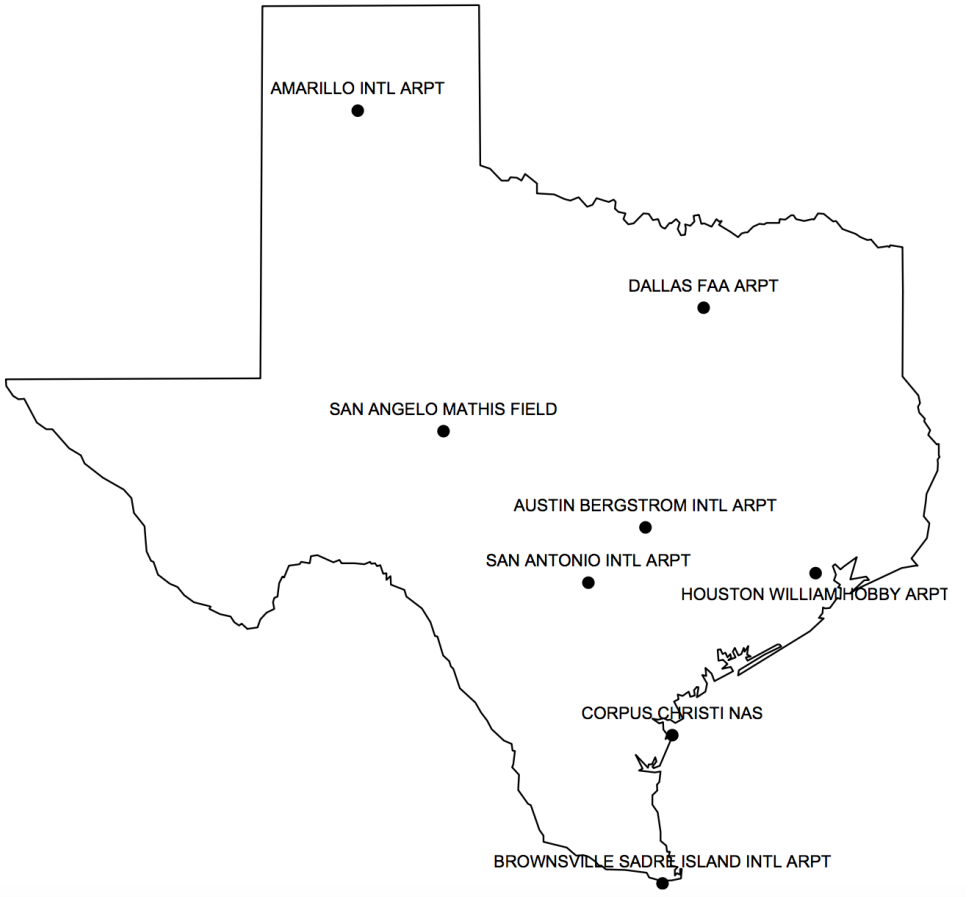


FIGURE 5.6: Graphical Map of Texas highlighting locations of selected station sites.

Site	Start Year	End Year	Latitude	Longitude	N
1. Amarillo Intl Arpt	03/01/1943	08/05/2017	35.220	-101.707	26359
2. Austin Intl Arpt	03/15/1948	08/05/2017	30.197	-97.666	15248
3. Brownsville Intl Arpt	12/01/1898	08/05/2017	25.906	-97.426	34388
4. Corpus Christi	03/01/1945	08/03/2017	41.786	-87.752	24941
5. Dallas FAA Arpt	08/01/1939	08/05/2017	32.845	-96.849	28475
6. Houston Arpt	05/01/1930	08/05/2017	29.646	-95.276	30339
7. San Angelo	08/01/1907	08/05/2017	31.357	-100.502	38923
8. San Antonio Intl Arpt	08/14/1946	08/05/2017	29.531	-98.468	25917
				Total	224590

TABLE 5.7: Texas Site Overview

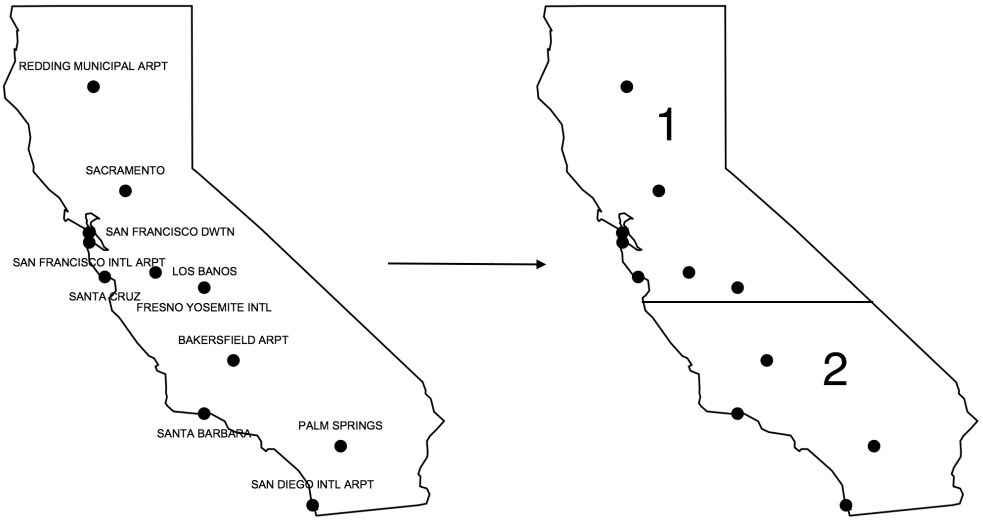


FIGURE 5.7: Graphical Map of California highlighting locations of selected sites.

Region 1	Region 2
Fresno Intl Arpt	Bakersfield Arpt
Los Banos	San Diego Intl Aprt
Redding Arpt	Santa Barbara
Sacramento	Palm Springs
San Francisco Dwtn	
San Francisco Intl Arpt	
Santa Cruz	

TABLE 5.8: Sites in each region for the state of California.

The data for the states was provided from NOAA as was for New Jersey. The same data cleaning methods from the New Jersey data set were employed. One notable difference from the New Jersey data set was that days with precipitation measurements equal to zero were retained for the states data sets. Each state was treated as a single region with the exception of California which was split into a northern region numbered as 1 and a southern region numbered as 2. An illustration of this splitting is provided in Figure 5.7 with the sites in each region in Table 5.8.

Optimal ARIMA models were estimated for each state using `auto.arima` function in R.

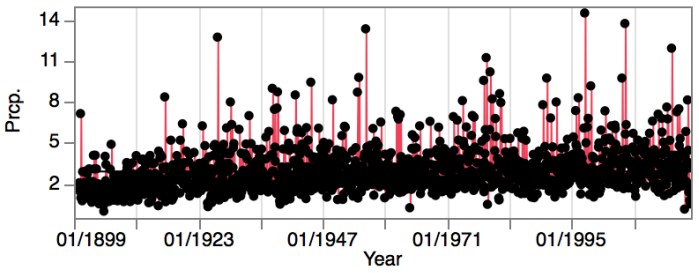
The state time series plots, estimated model summaries, and parameter estimates are provided in Figures 5.8 through 5.22. A summary of the estimated models and the AIC scored are provided in Table 5.9. Illinois has the smallest AIC while Alabama has the highest AIC although the residuals for the Alabama model show no significance indicating that the model is a good fit. Time series graphs for Louisiana and Missouri both contain a gap in the plot due to missing data.

State	TS Model	AIC
Alabama	$ARIMA(1,1,1)(1,0,0)_{12}$	5298.29
California: Region 1	$ARIMA(0,1,1)(0,0,1)_{12}$	3900.02
California: Region 2	$ARIMA(2,0,0)(2,0,0)_{12}$	4541.50
Florida	$ARIMA(1,1,3)(0,0,2)_{12}$	5224.25
Illinois	$ARIMA(1,1,0)(0,0,2)_{12}$	3204.97
Louisiana	$ARIMA(0,1,0)(2,0,1)_{12}$	5678.51
Missouri	$ARIMA(2,1,2)(0,0,2)_{12}$	3569.97
Texas	$ARIMA(0,0,0)(2,0,0)_{12}$	5285.28

TABLE 5.9: Summary of Time Series ARIMA models for States.

Time series plots of predicted values for corresponding fitted ARIMA models for the states are also displayed below. The time series also include a rough estimated forecast for the next 25 years based on the fitted models. There is a clear increasing trend in precipitation amounts seen from the predicted time series models for the states of Alabama, Florida, Missouri, and Texas. It is worth noting that the increases in precipitation were primarily observed across the southern states (Missouri being the exception).

Alabama



Mean	2.9695548
Std	1.6305237
N	1415
Zero Mean ADF	-12.84863
Single Mean ADF	-34.0635
Trend ADF	-36.29586

Model: Seasonal ARIMA(1, 1, 1)(1, 0, 0)12

Model Summary

DF	1410	Stable	Yes
Sum of Squared Errors	3480.50293	Invertible	Yes
Variance Estimate	2.4684418		
Standard Deviation	1.57112756		
Akaike's 'A' Information Criterion	5298.29151		
Schwarz's Bayesian Criterion	5319.30822		
RSquare	0.07191844		
RSquare Adj	0.0699438		
MAPE	.		
MAE	1.10415329		
-2LogLikelihood	5290.29151		

Parameter Estimates

Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob>|t|	Constant Estimate	Mu
AR1,1	1	1	0.02417728	0.0312707	0.77	0.4396		0.00114359
AR2,12	2	12	0.06781799	0.0270086	2.51	0.0122*	0.00104026	
MA1,1	1	1	0.99065994	0.0040480	244.73	<.0001*		
Intercept	1	0	0.00114359	0.0004640	2.46	0.0138*		

FIGURE 5.8: Optimal ARIMA model summary and parameter estimates for Alabama.

Forecast

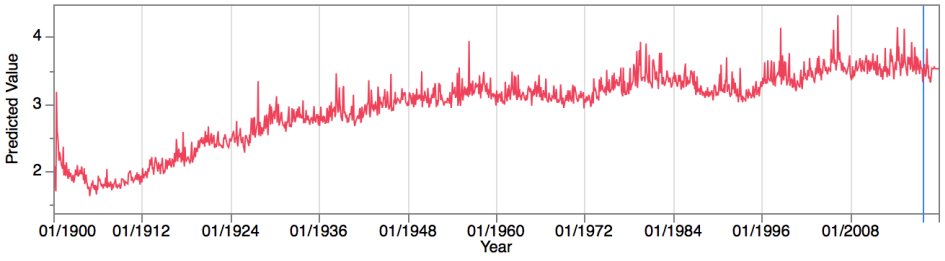
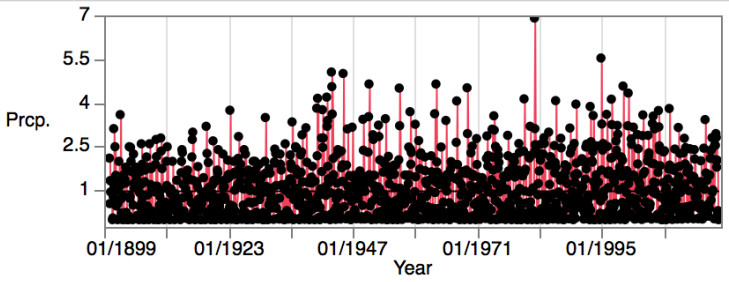


FIGURE 5.9: Time series of predicted observations from estimated ARIMA model for Alabama.

California: Region 1



Mean	1.0026577
Std	1.0056283
N	1411
Zero Mean ADF	-15.14287
Single Mean ADF	-23.2732
Trend ADF	-23.49218

Model: Seasonal ARIMA(0, 1, 1)(0, 0, 1)12

Model Summary

DF	1407	Stable	Yes
Sum of Squared Errors	1305.43489	Invertible	Yes
Variance Estimate	0.92781442		
Standard Deviation	0.96323124		
Akaike's 'A' Information Criterion	3900.01804		
Schwarz's Bayesian Criterion	3915.77208		
RSquare	0.08332422		
RSquare Adj	0.0820212		
MAPE	.		
MAE	0.72213878		
-2LogLikelihood	3894.01804		

Parameter Estimates

Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob>|t|	Constant Estimate	Mu
MA1,1	1	1	0.5495682	0.0492019	11.17	<.0001*		-0.0007655
MA2,12	2	12	-0.2684663	0.0258486	-10.39	<.0001*	-0.0007655	
Intercept	1	0	-0.0007655	0.0032405	-0.24	0.8133		

FIGURE 5.10: Optimal ARIMA model summary and parameter estimates for California Region 1.

Forecast

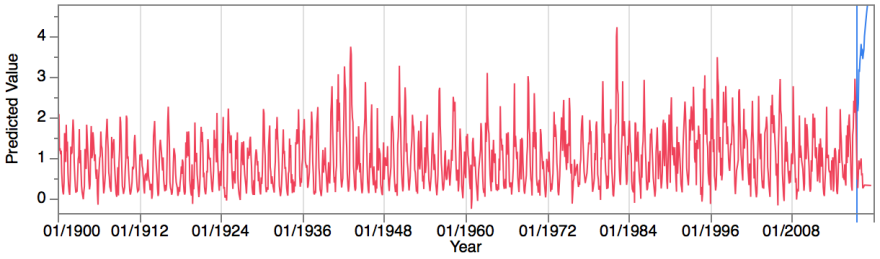
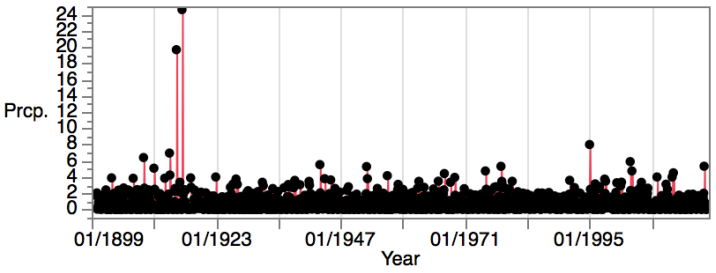


FIGURE 5.11: Time series of predicted observations from estimated ARIMA model for California Region 1.

California: Region 2



Mean	0.8167744
Std	1.2723222
N	1392
Zero Mean ADF	-24.55123
Single Mean ADF	-32.27918
Trend ADF	-32.30847

Model: Seasonal ARIMA(2, 0, 0)(2, 0, 0)12

Model Summary

DF	1387	Stable	Yes
Sum of Squared Errors	2111.7792	Invertible	Yes
Variance Estimate	1.52255169		
Standard Deviation	1.23391721		
Akaike's 'A' Information Criterion	4541.50091		
Schwarz's Bayesian Criterion	4567.69339		
RSquare	0.06257745		
RSquare Adj	0.059874		
MAPE	.		
MAE	0.69153874		
-2LogLikelihood	4531.50091		

Parameter Estimates

Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob>|t|	Constant Estimate	Mu
AR1,1	1	1	0.03263592	0.0295862	1.10	0.2702		
AR1,2	1	2	0.05770234	0.0294481	1.96	0.0503	0.50412678	0.81832711
AR2,12	2	12	0.13671691	0.0336951	4.06	<.0001*		
AR2,24	2	24	0.18605821	0.0297321	6.26	<.0001*		
Intercept	1	0	0.81832711	0.0530094	15.44	<.0001*		

FIGURE 5.12: Optimal ARIMA model summary and parameter estimates for California Region 2.

Forecast

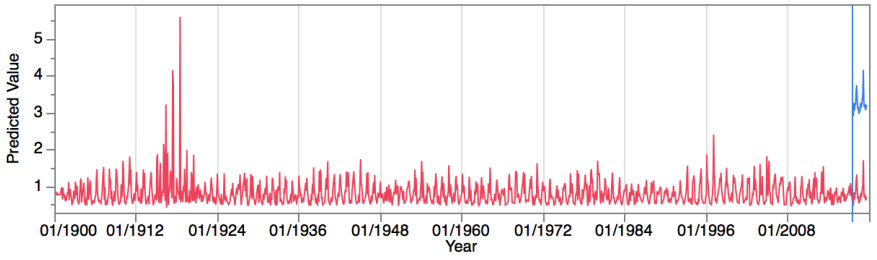
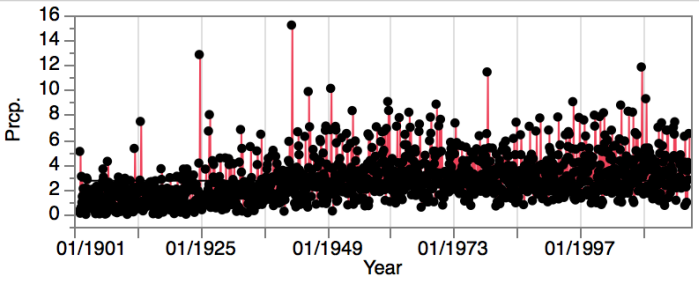


FIGURE 5.13: Time series of predicted observations from estimated ARIMA model for California Region 2.

Florida



Mean	2.7731556
Std	1.7976144
N	1388
Zero Mean ADF	-12.71497
Single Mean ADF	-27.73116
Trend ADF	-31.88837

Model: Seasonal ARIMA(1, 1, 3)(0, 0, 2)12

Model Summary

DF	1380	Stable	Yes
Sum of Squared Errors	3466.15859	Invertible	Yes
Variance Estimate	2.51170912		
Standard Deviation	1.58483725		
Akaike's 'A' Information Criterion	5224.25037		
Schwarz's Bayesian Criterion	5260.89466		
RSquare	0.22273692		
RSquare Adj	0.21935751		
MAPE	.		
MAE	1.14631607		
-2LogLikelihood	5210.25037		

Parameter Estimates

Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob>|t|	Constant Estimate	Mu
AR1,1	1	1	-0.9331580	0.1089602	-8.56	<.0001*		0.00169628
MA1,1	1	1	-0.0382750	0.1106275	-0.35	0.7294	0.00327918	
MA1,2	1	2	0.9065353	0.1134429	7.99	<.0001*		
MA1,3	1	3	0.1080604	0.0270588	3.99	<.0001*		
MA2,12	2	12	-0.1251877	0.0280421	-4.46	<.0001*		
MA2,24	2	24	-0.0994054	0.0249807	-3.98	<.0001*		
Intercept	1	0	0.0016963	0.0006821	2.49	0.0130*		

FIGURE 5.14: Optimal ARIMA model summary and parameter estimates for Florida.

Forecast

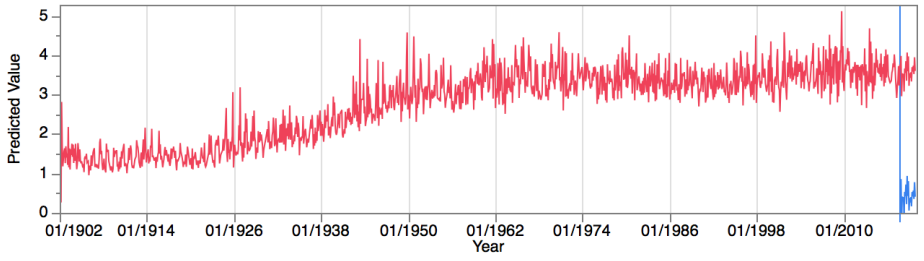


FIGURE 5.15: Time series of predicted observations from estimated ARIMA model for Florida.

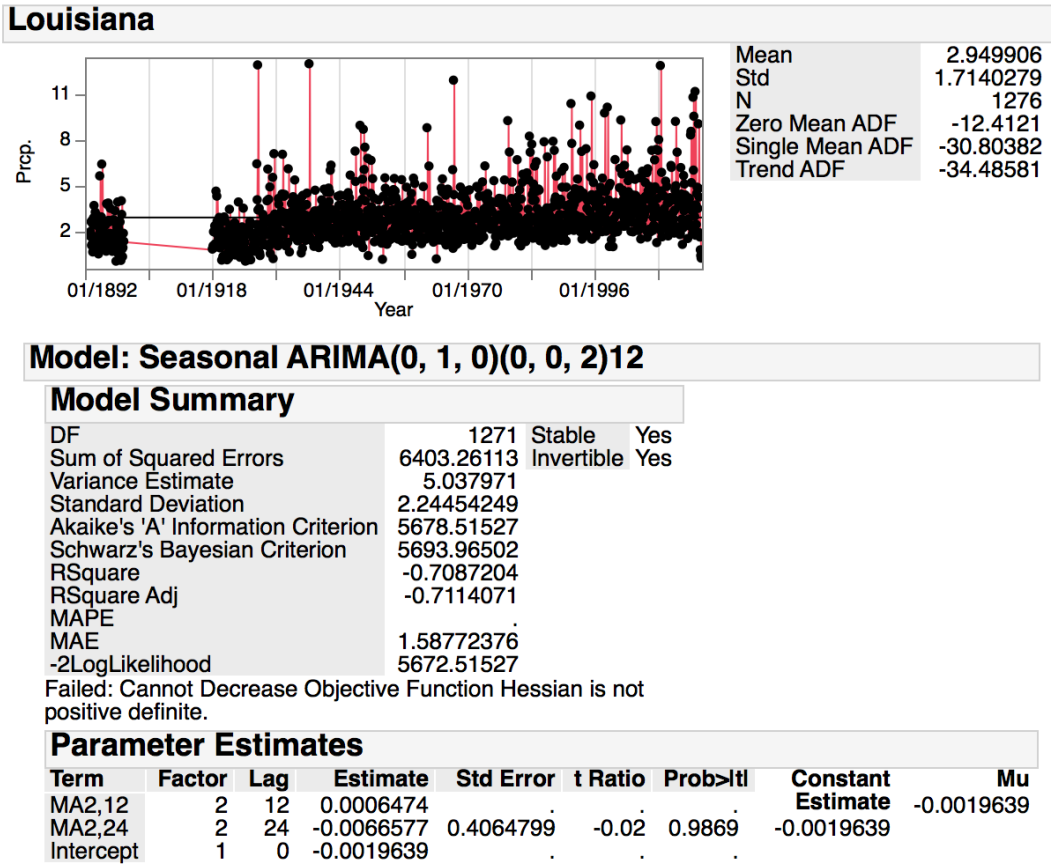


FIGURE 5.16: Optimal ARIMA model summary and parameter estimates for Louisiana.

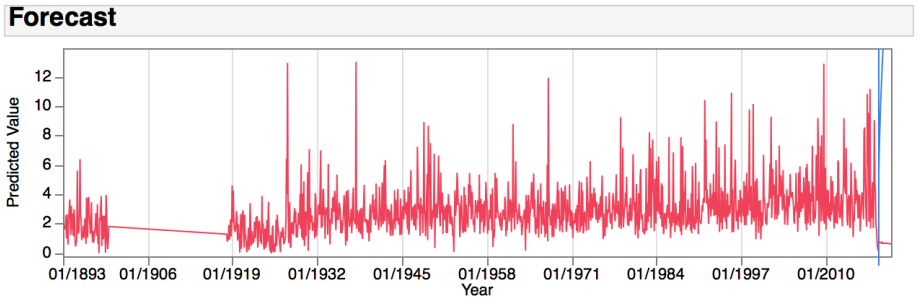
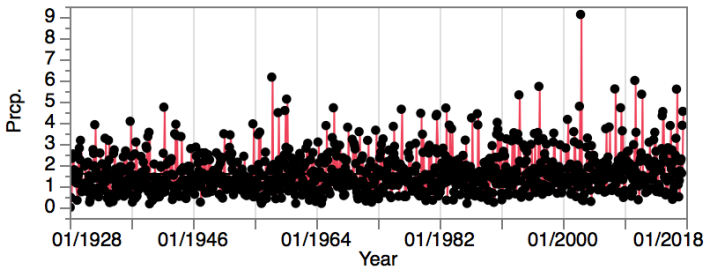


FIGURE 5.17: Time series of predicted observations from estimated ARIMA model for Louisiana.

Illinois



Mean	1.6272998
Std	0.9726991
N	1074
Zero Mean ADF	-10.90891
Single Mean ADF	-26.05735
Trend ADF	-26.59764

Model: Seasonal ARIMA(1, 1, 0)(0, 0, 2)12

Model Summary

DF	1069	Stable	Yes
Sum of Squared Errors	1235.79043	Invertible	Yes
Variance Estimate	1.15602472		
Standard Deviation	1.0751859		
Akaike's 'A' Information Criterion	3204.96996		
Schwarz's Bayesian Criterion	3224.88282		
RSquare	-0.2194348		
RSquare Adj	-0.222857		
MAPE	68.0684864		
MAE	0.80422397		
-2LogLikelihood	3196.96996		

Parameter Estimates

Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob>|t|	Constant Estimate	Mu
AR1,1	1	1	-0.4646189	0.0271702	-17.10	<.0001*		0.00414462
MA2,12	2	12	-0.0980382	0.0309842	-3.16	0.0016*	0.00607029	
MA2,24	2	24	-0.0200616	0.0301897	-0.66	0.5065		
Intercept	1	0	0.0041446	0.0186129	0.22	0.8238		

FIGURE 5.18: Optimal ARIMA model summary and parameter estimates for Illinois

Forecast

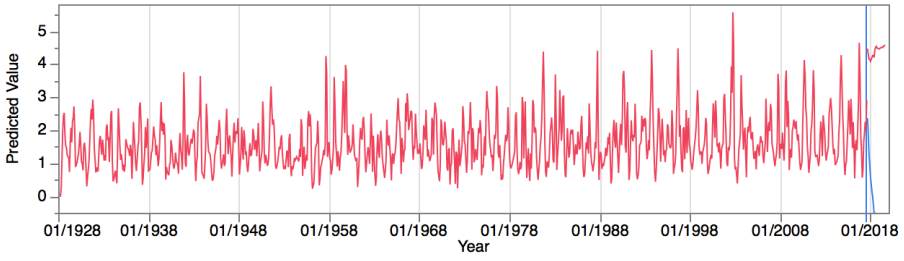
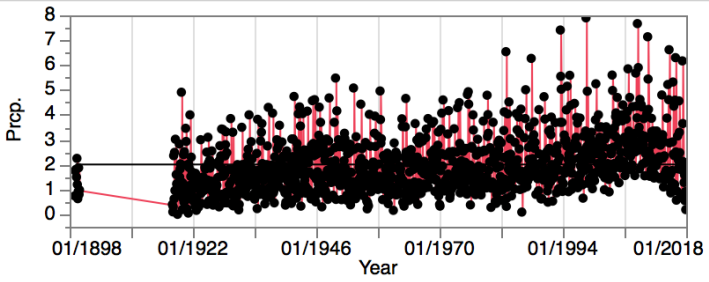


FIGURE 5.19: Time series of predicted observations from estimated ARIMA model for Illinois.

Missouri



Mean	2.0391667
Std	1.1555044
N	1212
Zero Mean ADF	-10.59268
Single Mean ADF	-25.4694
Trend ADF	-28.35996

Model: Seasonal ARIMA(2, 1, 2)(0, 0, 2)12

Model Summary

DF	1204	Stable	Yes
Sum of Squared Errors	1332.14283	Invertible	Yes
Variance Estimate	1.10643092		
Standard Deviation	1.0518702		
Akaike's 'A' Information Criterion	3569.97301		
Schwarz's Bayesian Criterion	3605.66743		
RSquare	0.17441502		
RSquare Adj	0.17030081		
MAPE	.		
MAE	0.79758345		
-2LogLikelihood	3555.97301		

Parameter Estimates

Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob>|t|	Constant Estimate	Mu
AR1,1	1	1	-0.7688227	0.0773918	-9.93	<.0001*		0.00118095
AR1,2	1	2	0.1662219	0.0306597	5.42	<.0001*	0.00189259	
MA1,1	1	1	0.0590598	0.0731225	0.81	0.4194		
MA1,2	1	2	0.9264640	0.0722949	12.82	<.0001*		
MA2,12	2	12	-0.0865561	0.0303143	-2.86	0.0044*		
MA2,24	2	24	-0.0778132	0.0280609	-2.77	0.0056*		
Intercept	1	0	0.0011810	0.0003577	3.30	0.0010*		

FIGURE 5.20: Optimal ARIMA model summary and parameter estimates for Missouri.

Forecast

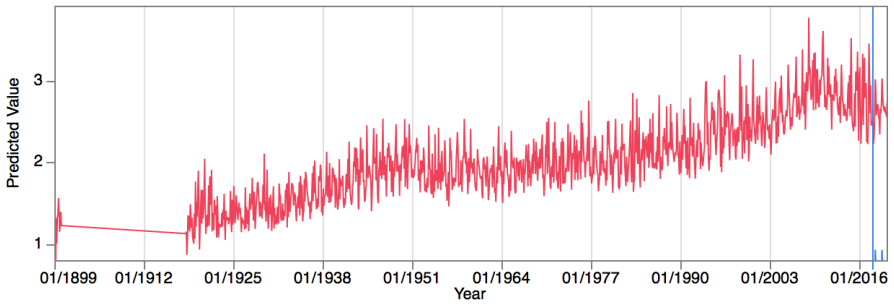


FIGURE 5.21: Time series of predicted observations from estimated ARIMA model for Missouri.

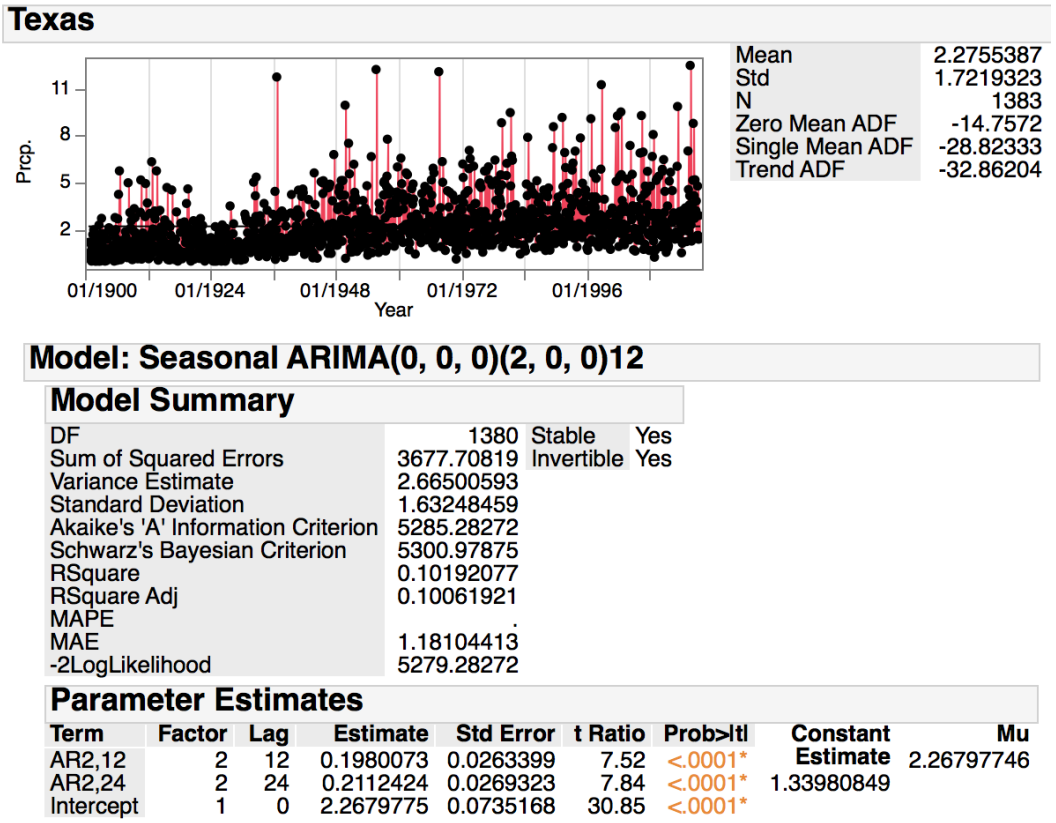


FIGURE 5.22: Optimal ARIMA model summary and parameter estimates for Texas.

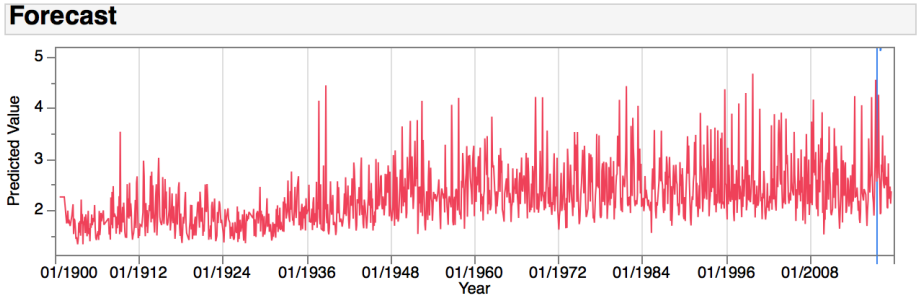


FIGURE 5.23: Time series of predicted observations from estimated ARIMA model for Texas.

Chapter 6

Conclusions

The analysis in this project was performed on the data set provided by the NOAA with no alterations made to the record observations other than unit conversion and the results obtained are representative of them. Due to the large amount of available data, and the highly variable properties of precipitation, the time series graphs and functional data plots were extremely noisy. The behavior of rainfall in the graphs was unpredictable and overwhelmingly erratic. In summary, the time series plots for the regions displayed several high peaks approximately after the year 1940 for Region 3 and 4 although the forecast models did not show an obvious increase. Similarly several states, Alabama, Florida, Missouri, and Texas, also exhibited increased rainfall amounts clearly seen in the forecast models. These changes were expected from the IPCC model predictions. We expected that California would show decreasing precipitation amounts however, the time series plots both regions of California showed constant precipitation amounts. This indicates that there is some activity or change in precipitation patterns.

6.1 Limitations

If rain gauge measurements are made for extended periods of time, it is highly probable that the conditions at the site where the rain gauge is operated will change over time. A change in environmental factors such as an increase in the growth or number of trees may affect the site. Human interaction, industrialization, and growth of towns and cities may also play a role in affecting the site of the rain gauge. Often times, due to a combination of different factors and reasons, the location of the site may often be moved to a nearby location or be eliminated entirely. Due to this, it is highly uncommon for precipitation recording sites to remain unchanged for long periods of time.

The conditions of the rain gauge tool may also change over time. The gauge may become damaged at a certain point and need to be replaced, or a newer type or model of gauge may replace the older model with newer ones needing to be recalibrated more frequently. Such changes all affect precipitation measurements, sometimes gradually and other times rather suddenly. As mentioned in chapter 2, older precipitation records used older manual mechanical rain gauges which are more prone to error and may not be as reliable as newer models. For precipitation records that are used for detection of trends, this can cause problems and can lead to incorrect analysis. It is therefore extremely rare that a good record is kept for extended periods of time, and this makes detection of changes in precipitation difficult (Strangeways, 2006).

Record years using older models of rain gauges may have higher or lower reported measurements of precipitation than the actual amount and this can lead to incorrect observation and analysis. Due to these reasons, some of the conclusions made in this thesis can be a little

biased. Perhaps some of the results seen that are different from the expected models may not be as accurate.

6.2 Implications

The splitting of the states into regions this project was not based on any models, rather the sites adjacent to each other were clustered into separate regions to examine overall regional effects for the states. For the New Jersey regions, the spectral envelope and functional data plots did not detect natural phenomena and occurrences such as the NAO or the El Nino Southern Oscillations. This is interesting in the sense the data we examined did not show these well-known periodic effects. There was a considerable amount of variation for the second half of the year for the New Jersey regional functional boxplots. This may simply be particular to the New Jersey climate and does provide temporal information in contrast to the ARIMA model analyses. However, taking the maximum for the monthly maximum of the sites within each region, may have dampened the resolution to the extent that interesting phenomena were not able to be detected.

What we were able to show based on the entire historical records available is evidence of increasing trends and volatility in precipitation for southern states over northern states with the exception of Missouri. This is clearly a concern for southern coastal populations, particularly in view of the devastating effects of major Category 4 hurricanes in 2017 (Harvey, Texas ; Irma, Florida ; Maria, Puerto Rico).

6.3 Future Work

The impetus for this work was provided by the previous Montclair State University Director of the Passaic River Institute, Professor Kirk Barrett (private communication) who suggested rainfall events were becoming increasingly more frequent and intense over time. In this Thesis we have provided some statistical evidence of this supposition. Future research would involve GARCH (Generalized AutoRegressive Conditional Heteroskedasticity) time series models (Engle, 1982) to account for the non-stationary residual variance and spatial-temporal analyses. This could be combined with “long memory" models where Y_t is stationary, but the autocorrelations remain significant for large lags suggesting a seasonal and/or simple difference $\nabla^d Y_t$ is needed. Evidence of long memory processes was first documented by Hurst (1951) in hydrology with regard to long term reservoir storage capacities. Hence the connection to our study of precipitation where the ARIMA models do not pass the residual white noise test. Long memory processes can be modeled by using a $\widehat{\text{f}}$ractional difference $d \in (-0.5, 05)$. We refer the reader to Brookwell and Davis (1991) for details and additional references.

A simple approach to increase resolution of precipitation event would potentially to be compute weekly maximums, provided missing observations or blocks do not become problematic. We also believe that increased resolution through spatio-temporal statistical analysis (Cressie and Wikle, 2015) over more sites could provide more insight into precipitation events. Finally, climate “change" involves many factors and we would like incorporate some of these covariates in future research.

Bibliography

Box, G., Jenkins, G., and Reinsel, G. (2008). *Time Series Analysis: Forecast and Control. Wiley.* Hoboken, NJ, 4th edition.

Brillinger, D. (1981). *Time Series: Data Analysis and Theory.* Holdan-Day, San Fransisco.

Brookwell, P. J. and Davis, R. (1991). *Time Series: Theory and Methods.* Springer-Verlag, New York, 2nd edition.

Chang, N.-B., Imen, S., Bai, K., and Yang, Y. J. (2017). The impact of global unknown teleconnection patterns on terrestrial precipitation across north and central america. *Atmospheric Research*, 193:107–124.

Cressie, N. and Wikle, C. K. (2015). *Statistics for Spatio-Temporal Data.* John Wiley & Sons.

Dore, M. H. (2005). Climate change and changes in global precipitation patterns: what do we know? *Environment International*, 31(8):1167–1181.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society*, pages 987–1007.

Huntington, T. G. (2006). Evidence for intensification of the global water cycle: review and synthesis. *Journal of Hydrology*, 319(1):83–95.

Hurrell, J. W. and Deser, C. (2010). North atlantic climate variability: the role of the north atlantic oscillation. *Journal of Marine Systems*, 79(3):231–244.

Hurrell, J. W. et al. (1995). Decadal trends in the north atlantic oscillation: regional temperatures and precipitation. *Science-AAAS-Weekly Paper Edition*, 269(5224):676–678.

Hurst, H. E. (1951). Long term storage capacity of reservoirs. *Trans. Amer. Soc. Civil Engrs.*, 116:778–808.

Hyndman, R. J. and Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, 26(3):1–22.

IPCC: Core Writing Team, Pachauri, R. and Reisinger, A. (2007). Climate change 2007: Synthesis report. contribution of working groups i, ii and iii to the fourth assessment report of the intergovernmental panel on climate change. ipcc, geneva, switzerland, 104 pp.

IPCC: Core Writing Team, Pachauri, R. and Reisinger, A. (2014). Climate change 2014: Synthesis report. contribution of working groups l, ll, and lll, to the fifth assessment report of the intergovernmental panel on climate change [core writing team, r.k. pachauri and l.a. meyer (eds.)]. ipcc, geneva, switzerland, 151 pp.

Jones, P., Jonsson, T., and Wheeler, D. (1997). Extension to the north atlantic oscillation using early instrumental pressure observations from gibraltar and south-west iceland. *International Journal of climatology*, 17(13):1433–1450.

Kidd, C., Becker, A., Huffman, G. J., Muller, C. L., Joe, P., Skofronick-Jackson, G., and Kirschbaum, D. B. (2017). So, how much of the earth's surface is covered by rain gauges? *Bulletin of the American Meteorological Society*, 98(1):69–78.

Liu, H., Shah, S., and Jiang, W. (2004). On-line outlier detection and data cleaning. *Computers & chemical engineering*, 28(9):1635–1647.

López-Pintado, S. and Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486):718–734.

McDougall, A., Stoffer, D., and Tyler, D. (1997). Optimal transformations and the spectral envelope for real-valued time series. *Journal of statistical planning and inference*, 57(2):195–214.

Melillo, J. M., Richmond, T., and Yohe, G. (2014). Climate change impacts in the united states. *The Third National Climate Assessment. U.S. Global Change Research Program*.

Menne, M.J., I. D. R. V. B. G. and Houston, T. (2012). An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology*, 29:897–910.

Michaelides, S., Levizzani, V., Anagnostou, E., Bauer, P., Kasparis, T., and Lane, J. (2009). Precipitation: Measurement, remote sensing, climatology and modeling. *Atmospheric Research*, 94(4):512–533.

Montgomery, D. C., Jennings, C. L., and Kulahci, M. (2015). *Introduction to time series analysis and forecasting*. John Wiley & Sons.

O'Gorman, P. A. (2015). Precipitation extremes under climate change. *Current climate change reports*, 1(2):49–59.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rahm, E. and Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, 23:2000.

Ramsay, J. O., Wickham, H., Graves, S., and Hooker, G. (2017). *fda: Functional Data Analysis*. R package version 2.4.7.

SAS Institute Inc. (2016). *JMP Pro, Version 13.2*. SAS Institute Inc., Cary, NC.

Seinfeld, J. H. and Pandis, S. N. (2016). *Atmospheric chemistry and physics: from air pollution to climate change*. John Wiley & Sons.

Stoffer, D. S. and Tyler, D. E. (1998). Matching sequences: Cross-spectral analysis of categorical time series. *Biometrika*, 85(1):201–213.

Stoffer, D. S., Tyler, D. E., and McDougall, A. J. (1993). Spectral analysis for categorical time series: Scaling and the spectral envelope. *Biometrika*, 80(3):611–622.

Strangeways, I. (2006). *Precipitation: theory, measurement and distribution*. Cambridge University Press.

Sun, Y. and Genton, M. G. (2011). Functional boxplots. *Journal of Computational and Graphical Statistics*, 20(2):316–334.

Tapiador, F., Navarro, A., Levizzani, V., García-Ortega, E., Huffman, G., Kidd, C., Kucera, P., Kummerow, C., Masunaga, H., Petersen, W., et al. (2017). Global precipitation measurements for validating climate models. *Atmospheric Research*.

Trenberth, K. E. (2012). Framing the way to relate climate extremes to climate change. *Climatic change*, 115(2):283–290.

Trenberth, K. E., Dai, A., Rasmussen, R. M., and Parsons, D. B. (2003). The changing character of precipitation. *Bulletin of the American Meteorological Society*, 84(9):1205–1217.

Van-den Broeck, J., Argeseanu Cunningham, S., Eeckels, R., and Herbst, K. (2005). Data cleaning: Detecting, diagnosing, and editing data abnormalities. *PLOS Medicine*, 2(10).