Time Series Analysis of 50 Years of Precipitation Data in the United States

Special Topic in GIS

Matthew Manni

Semester 2, 2012

Table of Contents

List of Figures

List of Tables

Abstract

In this paper, a time series analysis of 50 years of precipitation data over selected areas in the United States was conducted in order to identify weather patterns, anomalies or cycles in those environments and to model and forecast the data. To find the patterns and anomalies in precipitation from 1960 to 2009, each time series was periodically decomposed, separating the data into trend, seasonal and random components. Autoregressive integrated moving average (ARIMA) models were used to fit functions to the data and forecast values 3 years into the future. A Fourier transform was then used to find cyclical patterns in the data that originally could not be seen with seasonality. Consistent weather patterns were identified in Flint, Michigan and Butte, Montana, both of which have the lowest deviations and the highest and lowest monthly mean precipitation, respectively. Woodward, Oklahoma and Orange County, California have much larger deviations and ranges in precipitation. A 3-year forecast predicted high precipitation in Woodward and low precipitation in Flint and Orange County. Significant cyclical patterns were identified in Woodward and Orange County. Woodward has a 3-year pattern due to occasional severe storms in that climate, while Orange County has a 5-year pattern due to the El Niño-Southern Oscillation (ENSO) cycle. This time series analysis helped identify several precipitation patterns and cycles in the selected locations, as well as an accurate forecast of the data.

1. Introduction

Precipitation plays a vital role in the climate system, affecting the ecosystem of an area and its surrounding environment. This is a significant topic when discussing different ecosystems and their climate variations. Monitoring the changes in seasonal or annual precipitation trends is essential for understanding the recent effects of climate change. Globally, the overall climate cycles have been changing due to an enhanced greenhouse effect from anthropogenic causes such as the use of fossil fuels and increased carbon dioxide emissions (Chen, 2002). This has led to notable variations in several climate-related variables, such as temperature, cloud cover and precipitation. In order to better understand these changes in the climate, one must analyse the data of these variables over time and look for structure, trends, anomalies or cycles in the data.

A time series analysis of climate data, such as precipitation, is often used to help in a climate diagnostic study. Understanding the distribution of precipitation is crucial for modelling and predicting the weather and climate changes that are occurring globally (Huffman, 1997). Certain weather patterns such as the El Niño-Southern Oscillation (ENSO) have been changing recently and can be better understood with precipitation time series analyses (Chen et al., 2002). ENSO and other climate patterns need to be studied in order to continue to analyse climate change.

In this study, precipitation data throughout the continental U. S. was used for a time series analysis to identify any patterns, anomalies or cycles in the data, as well as modelling and forecasting the data. All of these methods could be further used for a cumulative regional climate study in different areas of the United States. Here, four areas of different environments were selected in order to determine any changes dependent of the ecosystem. The selected points are from various regions of the continental U. S. (Table 1, Figure 1).

Table 1: Description of Selected Points

	Point 1	**Point 2**	**Point 3**	**Point 4**
Location Name	Butte, Montana	Flint, Michigan	Woodward, Oklahoma	Orange County, California
Latitude	45.683	43.175	36.175	33.675
Longitude	-112.449	-83.467	-98.967	-117.467
Climate Type	Subarctic	Humid continental	Semi-arid	Mediterranean

Point 1 (Butte, Montana) is located in the North-Northwest region of the U. S. with a subarctic climate that has low precipitation due to the absence of moderating effects of an ocean.

Point 2 (Flint, Michigan) is located in the North-Northeast region of the U. S. with a humid climate characterised by large seasonal temperature differences and well distributed, year-round precipitation (Kottek et al., 2006).

Point 3 (Woodward, Oklahoma) is located in the Southern region of the U. S. with a semi-arid climate that has relatively moderate precipitation. However, it is located between climate zones, causing extreme variations in weather patterns within the state (Kottek et al., 2006).

Point 4 (Orange County, California) is located in the Southwest region of the U. S. and has a Mediterranean climate, similar to a subtropical climate, characterised by warm, dry summers and cool, wet winters (Kottek et al., 2006).

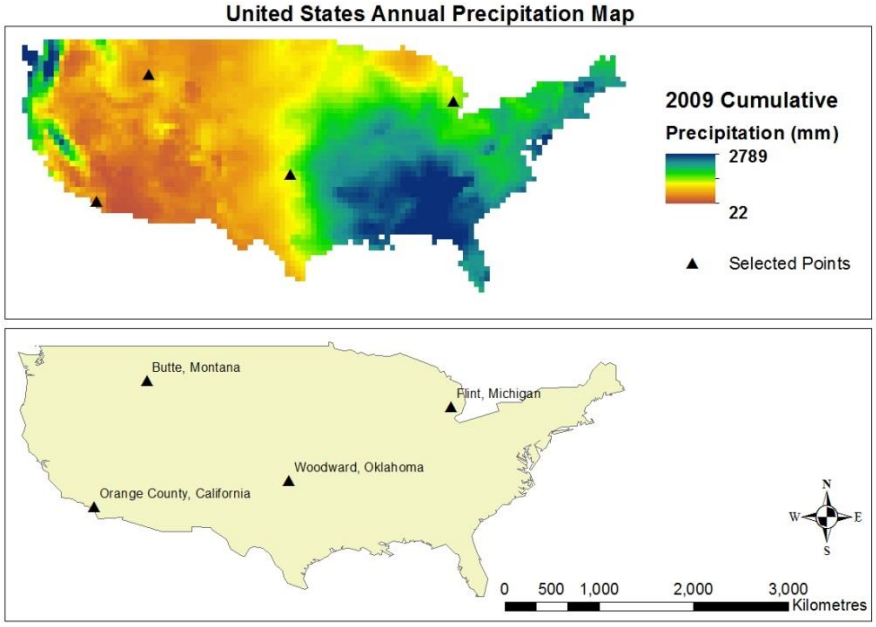

Figure 1: Location Map of Selected Points

The data used in this study includes the monthly and annual total precipitation (in millimetres) of the continental U. S. from 1960 to 2009. In some cases, subsets of data have been used, from 2000 to 2009, for analytical purposes. The CRU, Climatic Research Unit, at the University of East Anglia has provided the CRU TS 3.1 data set free of charge. The data sets originate from observations taken from globally distributed meteorological stations that were integrated onto a high resolution grid (0.5° latitude by 0.5° longitude) covering the Earth's surface (Mitchell, 2002). The time series data sets are monthly variations in climate data which include the following variables: cloud cover, diurnal temperature range, frost day frequency, precipitation, daily mean temperature, monthly average daily maximum temperature, vapour pressure and wet day frequency (New, 2000).

The data provided by the CRU is the first time such a global climate data set has been freely available over such a long period of time (1901-2009). This allows for a comparison of variations in climate with variations in other anomalies at a global scale (Mitchell, 2005). However, for analytical purposes, in this study the data have been clipped to continental U. S. for a 50-year period between 1960 and 2009.

2. Background Theory

The time series analysis was conducted using the R Project computer program for statistical computing and ArcMap for extracting the data provided by the CRU (Shaddick, 2004). The methods in this study are based on those used in time series analyses of monthly precipitation data, which produce information on the structure and the temporal and spatial variability of the data.

2.1 Components of a Time Series:

A time series is defined as a sequence of measurements of at least one variable (precipitation) over time (Ruhf, 2003). They are commonly used in applications such as natural resource studies, agricultural studies, climate modelling, and economic or financial studies. There are certain characteristics that make the time series analysis unique. These include such aspects as a trend over a time period, cycles of various periods in the data, and correlation between observation periods (Rossiter, 2012). A time series can be broken down into four different components that each describes an important characteristic of the data.

Trend – A long term movement in a time series without seasonal and irregular effects. This is the direction (or tendency) and rate of change over time in a time series analysis.

Cyclical – Describes any regular fluctuations (non-seasonal component which varies in a recognisable cycle)

Seasonal – The component of variation in a time series which is dependent on the time of year. It describes any regular fluctuations with a period of less than one year.

> This is useful for *a)* comparing the seasonal effects within the years, from year to year; *b)* removing seasonal effects so that the time series is easier to cope with; and *c)* adjusting a series for seasonal effects using various models (Rossiter, 2012).

Irregular/random – The remaining component when the other components of the series (trend, seasonal, cyclical) have been estimated and removed. This is also known as the residual or remainder component which results from short term fluctuations in a series which are non-systematic and unpredictable.

Another component of time series analysis is the structure, or pattern, of the data, which can be described by both the trend and seasonal variation (Ihaka, 2005). This is usually an obvious feature seen on a time series plot that displays the overall shape of the data.

All of the components in a time series analysis can be described individually by decomposing the time series and separating the trend, seasonal and irregular components (Figure 2).

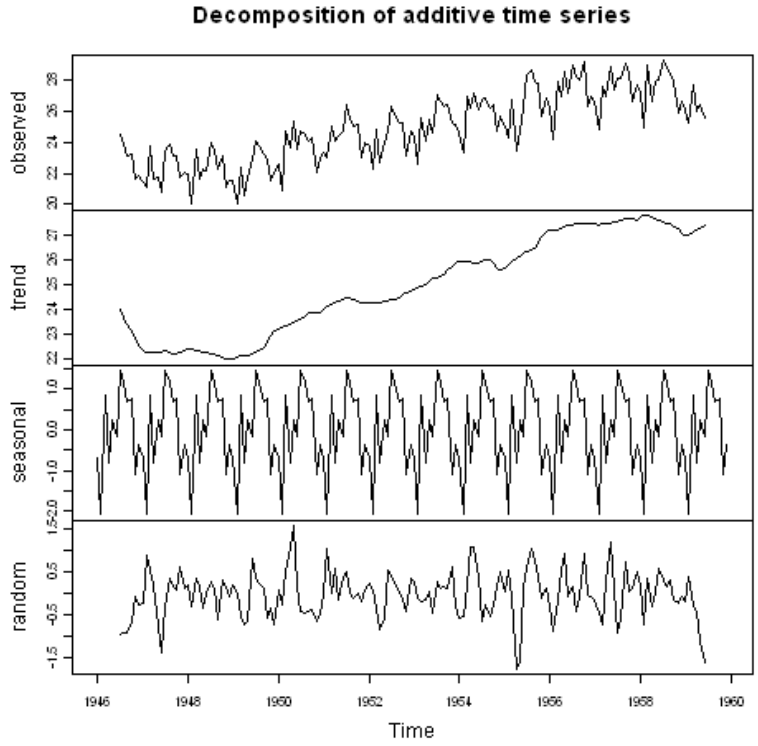

Figure 2: Example of Time Series Decomposition [*Source: Coghlan (2011)*]

2.2 Identifying Patterns

The two main objectives of a time series analysis are to identify the nature of the phenomenon and forecast possible future values (Ihaka, 2005). Once a pattern is identified, the data can be interpreted and possibly extrapolated in order to predict future events (forecasting). Most time series data contain a systematic pattern and some random noise, or error, which makes the pattern difficult to recognize. One analysis technique uses filtering, or smoothing, to remove the random noise and make the pattern easier to identify (Chatfield, 1996).

The first step in identifying a pattern is to analyse the trend components in the time series data. If the trend is inconsistent and contains error then the data needs to be smoothed. Smoothing is used to reduce irregularities in data, providing a clearer view of the true underlying behaviour of series. This involves a form of local averaging of data in which the irregular components of each observation cancels each other out (Chatfield, 1996). It can remove seasonality and random fluctuations, thus making long term fluctuations in a time series stand out more clearly. The type of smoothing depends on the type seasonality in the series.

The most common smoothing technique for time series analysis is the moving average smoothing, which replaces each component of the series by the average of *n* surrounding components (Zucchini, 2011). This type of smoothing is adjusted to allow for seasonal or cyclical components of a time series, either of which can be eliminated from the variation by taking a moving average. After the data is smoothed, it is possible to fit a function to the time series data by using a linear, logarithmic, exponential, or polynomial function (Chatfield, 1996).

The second step in identifying a pattern is to analyse the seasonal component of the time series data. The seasonality can be identified as a pattern in the data that repeats at every lag, a time period between two measurements, if the error is small enough (McLeod, 2011). It is measured by the autocorrelation, or the similarity of a variable with itself. Autocorrelation is best described as the correlation or similarity between observations as a function of the time difference between them (Chatfield, 1996).

The seasonal patterns can be observed via correlograms, or the autocorrelation plot, which plots the autocorrelations versus time lags. By examining the strong autocorrelations, one can determine when the significant seasonal pattern occurs (McLeod, 2011). Another similar method is to examine the partial autocorrelation. The difference with partial autocorrelation is that the correlations with all the components within the lag are partialled out, giving a unique contribution of each time lag (Chatfield, 1996). Cowpertwait (2006) explains this concept best by stating that it yields the correlation after removing the effect of any correlations due to previous terms.

Differencing a series can also be used because it removes the trend from a time series. This can help with identifying some other seasonality more easily and it can make the series stationary, which will be necessary for the following methods. In order for the data to be stationary, the time series should have a constant mean, variance and autocorrelation through time (Soltani et al., 2007).

2.3 Modelling

Modelling and forecasting are commonly used for time series analyses to identify a function that can be used to predict future values in the time series. They are major tools used in the decision making of water resources and studies on climate change. Stochastic models can be developed to establish linear prediction formulas that fit a function to the data (Soltani et al., 2007). The three types of linearly dependent models are autoregressive (AR) models, integrated (I) models and moving average models (MA). The models can also be combined to produce either autoregressive moving average (ARMA) models or autoregressive integrated moving average (ARIMA) models, the latter of which is the most commonly used time series model for climate data sets (Ruhf, 2003).

This study uses the ARIMA model for the precipitation time series analysis. The model was first introduced by Box and Jenkins (1976) and includes several parameters that need to be explained before understanding the process (Zucchini, 2011). The three types of parameters in the ARIMA model include: the autoregressive parameter (p), the number of differencing passes (d) and the moving average parameter (q) (Zucchini, 2011). The notation of the model is summarised as ARIMA (p, d, q) in which all parameters are non-seasonal. Another form of the ARIMA model is known as the multiplicative ARIMA: (p, d, q) x (P, D, Q), where P is the seasonal autoregressive parameter, D is the number of seasonal differencing passes, and Q is the seasonal moving average parameter (Soltani et al., 2003). This model is often used for patterns that display seasonal repetition over time.

The time series modelling process involves three steps with ARIMA: parameter identification, estimation and evaluation (Soltani et al., 2003). As mentioned before, it is essential that the series is stationary before being input into ARIMA. This can be accomplished by differencing the series, and sometimes log transforming, a certain number of times, d, until it is stationary. Examining the autocorrelation (ACF) and partial autocorrelation functions (PACF) allows one to determine if the series is stationary (Zucchini, 2011). An example of both ACF and PACF plots is seen in Figure 3.

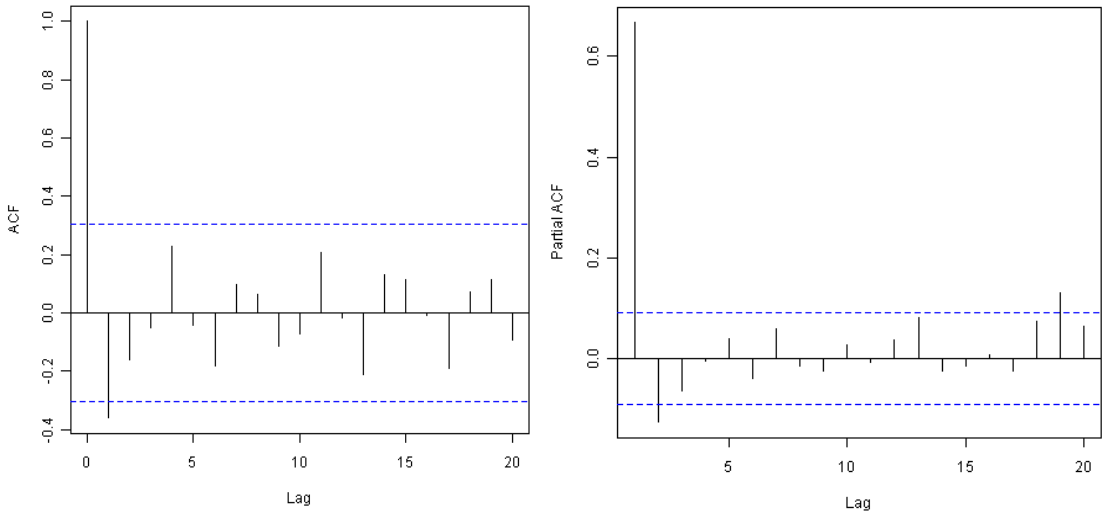

Figure 3: Example of Autocorrelation and Partial Autocorrelation Plots [*Source: Coghlan (2011)*]

The autoregressive (*p*) and moving average (*q*) parameters need to be estimated at the next stage of modelling. These parameters should be decided based on the shape of both the autocorrelation and partial autocorrelation plots, where the sum of squared residuals is minimised for the time series (Soltani, 2007). The sum of squares of residuals can be computed by using a form of the maximum likelihood method. The parameter estimation process is usually performed on differenced data and is used for the next steps for forecasting and validation through goodness of fit tests (McLeod, 2011).

Significance tests are performed in order to test the quality of the model parameters used in the previous steps. The standard error of estimation is used to test the statistical significance of all of the parameter values.

2.4 Exponential Smoothing (Forecasting)

Another way to predict or forecast values for the time series is to use exponential smoothing. In this method, the future values are predicted based on an average of the previous observations, which are given weights with more recent values receiving the highest weights. This type of moving average method differs from the ARIMA process such that the preceding observations are given weights such that the youngest observation is assigned the greatest weight and older observations are assigned exponentially smaller weights (Zucchini, 2011). In R, the forecast procedure is a tool used for applying an exponential smoothing technique to the trend term and the seasonal term.

2.5 Spectral Analysis

Besides the aforementioned trend and seasonal analysis and their forecasting, another analytical tool for time series data is the spectrum analysis. Spectral analysis deals with the examination of cyclical patterns of data that may not be related to seasonality (Galford et al., 2008). This is commonly used for studying the weather patterns in certain areas by decomposing the time series with cyclical components. It is also known as a Fourier analysis and is a mathematical technique used to decompose the cyclical component, or any time series signal, into several sine and cosine functions, each characterised by specific amplitudes and phase angles (Quiroz, 2011). The Fourier analysis can help identify underlying recurring cycles of different periods within the time series that could not be seen beforehand. As opposed to ARIMA modelling or exponential smoothing, where the seasonal component is usually known and constant, the spectrum analysis is used to identify the seasonal fluctuations of different periods (Martinez, 2009).

The Fourier transform, the decomposition process of the Fourier analysis, basically yields a curve-fitting algorithm used to convert a function from the time spectrum into the frequency spectrum (Lau,

1995). As mentioned before, the function is decomposed into sinusoids of different frequencies, from the time domain, and thus cannot provide information about when frequency events occur. The transform can be used to analyse the frequency content of the time series as a superposition of sine and cosine basis functions (Torrence, 1998).

3. Methods

In this study, the precipitation data from all four points were analysed using common time series analysis methods to identify patterns and cycles in the data and to model and forecast the data based on the structure of the times series. The general methods used for this analysis include decomposing the time series, modelling and forecasting the data, and a spectral analysis.

3.1 Obtaining the Data and Converting it into a Time Series

The original global data sets were first downloaded from the CRU website in ASCII format. A script was written in order to extract the data and convert them into ESRI ASCII grid format, which then needed an AML to convert them into ArcGIS grid format. The data was then brought into ArcMap in raster format where the global precipitation data was clipped for the United States only. From here, four points were selected and sampled to obtain the numerical values of the monthly precipitation data from 1960 to 2009. The data was then copied into Microsoft Excel before being converted into DAT files which were used within R.

Once loaded into R, the data was converted into a time series function using the **ts()** tool in R. Thereafter, the time-series-formatted data was used for the analysis using other tools in R. Each point was plotted to show the variation of monthly precipitation data from 1960 to 2009. The mean monthly precipitation of each point was also calculated to reduce volume. The descriptive statistics were taken in order to obtain the exact values of such statistics as mean, median and standard deviation.

3.2 Decomposing the Time Series

The best method to visualise all of the components in a time series data set is to periodically decompose the data using the **stl()** function in R. This function was used to break down the time series into seasonal, trend and irregular (random) components, each of which were plotted separately onto a single aggregate plot along with the original data. This allows for an initial analysis of the data in order to find patterns in each component from the original data. A plot of the per-month variability was also created to analyse the deviation of monthly precipitation from the annual mean at each individual month.

3.3 Modelling and Forecasting

In order to identify a pattern in the data and model the time series data to a fitting function, the data first needs to become stationary. All points for the time series data set were differenced for the best results of stationarity. Plotting the autocorrelation function plot of the differenced time series showed whether the autocorrelation is constant through time. The **acf()** and **pacf()** functions were used to produce the autocorrelation function and partial autocorrelation function plots, respectively. These were plotted for both the original data, used for the seasonal component, and the differenced data, used for the trend component, for each point. The plots show the autocorrelations at each month, with each lag representing a year.

An ARIMA model was used to model and then forecast the data in this study. The parameters used in the models were estimated from the analysis of the differenced ACF and PACF plots for each individual point. For this seasonal model, both non-seasonal (p, d, q) and seasonal (P, D, Q) parameters were selected for the ARIMA models with a seasonal lag of 12. The differencing components, d and D, were easily

determined from the number of times the data was differenced, either once or twice. The non-seasonal moving average parameters, *q*, were determined from the significant spikes in the ACF plots and the non-seasonal autoregressive parameters, *p*, were determined from the shape pattern in the PACF plots (Hyndman, 2012). The seasonal parameters for both the moving average (*Q*) and the autoregressive (*P*) components was determined by examining the significant spikes at seasonal lags, but were more difficult to determine. Tests were then used to evaluate, or validate, the accuracy of the models based on the estimated parameters. The models with the lowest AICc value, or the corrected Akaike information criterion – a measure of statistical goodness of fit, were selected, with almost all of the spikes in the residual plots fitting within the 5% significance limits (Hyndman, 2012).

After the ARIMA models were determined for the time series data, predictions were made based on each model. The predicted values for the future derived from the models are based on the recent trends in the data. Each point was forecasted three years into the future using a forecasting function in R. A separate forecasting method based on exponential smoothing was also used to predict future values and to compare with the ARIMA models. A subset of the data, 2000 to 2012, was used to better view the results of the forecasting.

3.4 Spectral Analysis

A spectrum analysis was performed on the time series data in order to determine any cyclical patterns in the precipitation data set. These cyclical patterns were found at longer lags, or frequencies, than the seasonal component of the data. The Fourier analysis produced a sinusoidal function that was used to describe long-term cycles in the data that could not be recognised beforehand (Galford, 2008). The fast Fourier transform (FFT) algorithm was used in this study.

The spectrum tool in R was used to plot the spectral density of the time series by removing a fitted linear trend before calculating the spectrum (Cowpertwait, 2006). This function produced a smoothed periodogram using a FFT, which reduced the effect of measurement noise. The periodogram distributes the variance of the data over the frequency and is obtained by summing the squared coefficients for each frequency and plotting against the frequency. The periodogram was analysed by identifying peaks at certain frequencies, with each peak representing a significant pattern in precipitation (Cowpertwait, 2006).

Examining the first peak in the periodogram helped identify the frequency at which the long-term cycles occur. The cycle length of less frequent events was calculated from the frequency of the first peak by taking the inverse of that frequency value and multiplying by 12. This cycle length is calculated in months and determines the cyclic pattern at each location. Peaks that occur at frequencies of one and greater are the seasonal cycles seen throughout the year.

4. Results and Discussion

4.1 Periodic Decomposition of the Time Series

The monthly precipitation data of four selected points in the continental U. S. from 1960 to 2009 all have quite a wide variety of precipitation levels, as seen in Figure 4. Examining the raw data, Flint, Michigan seems to have the most consistent weather patterns, with moderate rainfall throughout the year over the entire time period. Orange County, California on the other hand has the most variation in precipitation levels, yet seems to have a large amount of low rainfall months. This point's local environment, located near the coast in Southern California, is obviously the cause of the extreme values seen here. The occasional significant increase in precipitation is most likely an effect of El Niño, or the El Niño-Southern Oscillation (ENSO), a climatic pattern that causes wetter winters to occur in that region roughly every five years (Chen et al., 2002). Another note on the raw data is the low range of precipitation levels for Butte, Montana, located in a drier area of the United States.

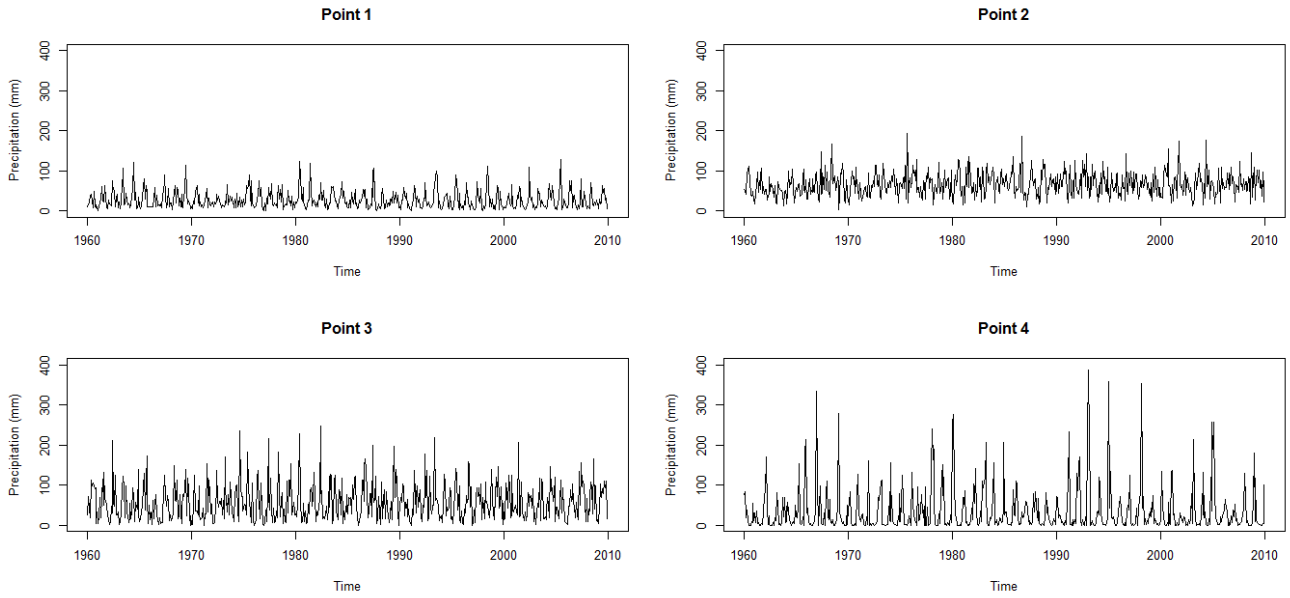

Figure 4: Time Series Plots of Raw Data

The raw time series data can also be represented by plotting the mean monthly precipitation of all the points. These plots (Figure 5) remove the high volume of data from the previous plots, making it much easier to visualise. The overall trend of the time series data is comprehendible from the mean monthly plots, as well as some seasonality within all four points. Butte and Flint both seem to show consistent low and high mean precipitation, respectively, in this 50-year span. This could indicate that there were no anomalies or extreme weather events during that time. In Woodward, there seems to be high precipitation during the mid-1970s, mid-1980s, and in 2008. These are most likely years that had more storm occurrences than usual. In Orange County, there are several years that have high precipitation, including mid-1960s, 1970, late-1970s, mid-1980s, most of 1990s, and mid-2000s. These are all caused by El Niño events, where precipitation is much higher than usual, most notably around 1980 and throughout the 1990s.

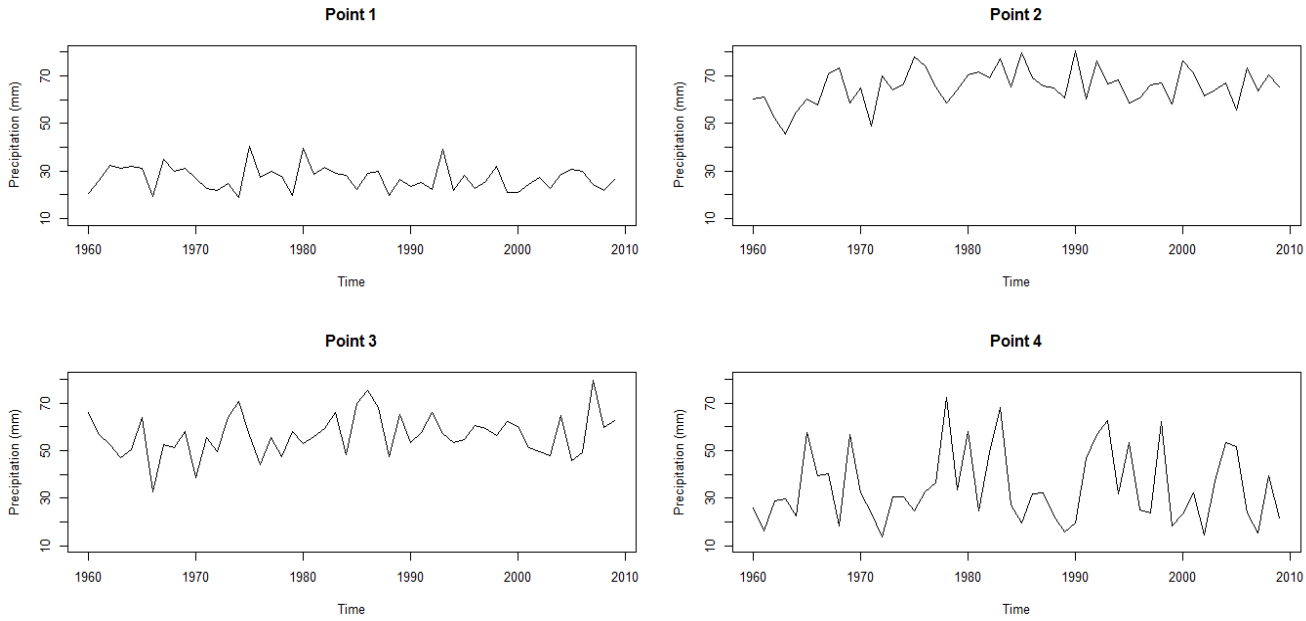

Figure 5: Mean Monthly Time Series Plots of Raw Data

Numerical values of the basic descriptive statistics of the raw data give a different view on the time series data. Table 2 shows the values of certain statistics of the precipitation data, including mean and

standard deviation. One interesting statistic from the data is from Flint, which contains a relatively low range of precipitation yet has the highest mean values. However, Orange County has the highest range of precipitation values, but a relatively low mean (and the lowest median). Again, Butte is represented by values that one would expect for that environment. Woodward, Oklahoma however, has high values in all categories (range, standard deviation, mean and median), which would not always be expected in semi-arid regions.

The descriptive statistics for both Butte and Orange County also show that there are some extreme precipitation events that skew the mean to a higher value. The monthly median is lower than the monthly mean for both areas, especially for Orange County. The large difference between the monthly mean and median, as well as the high standard deviation, further emphasizes the effects that ENSO has on the precipitation in Southern California.

Table 2: Descriptive Statistics of Raw Data

All values are in mm	**Annual Mean**	**Monthly Mean**	**Monthly Median**	**Standard Deviation**	**Minimum Value**	**Maximum Value**
Butte	325.2	27.1	19.9	22.63	0	128.6
Flint	779.2	65.54	61.85	29.33	3.1	194
Woodward	676.2	56.59	45.35	44.95	0	248.8
Orange County	388.5	34.53	10.05	56.8	0	389

Figure 6 through Figure 9 represent the plots of the periodic decomposition of time series of Point 1 through Point 4, respectively. The distinct sinusoidal shape of the seasonal component is caused by annual seasonal changes of precipitation throughout the year. Precipitation is high during the spring months and much lower in the summer time in Orange County; however this can change depending on location. This data is known as a periodic series and was decomposed as such because the precipitation varies across the seasons in a cyclic manner. The trend and irregular components slightly follow the pattern seen in the original data, but do not represent much more than that at this point.

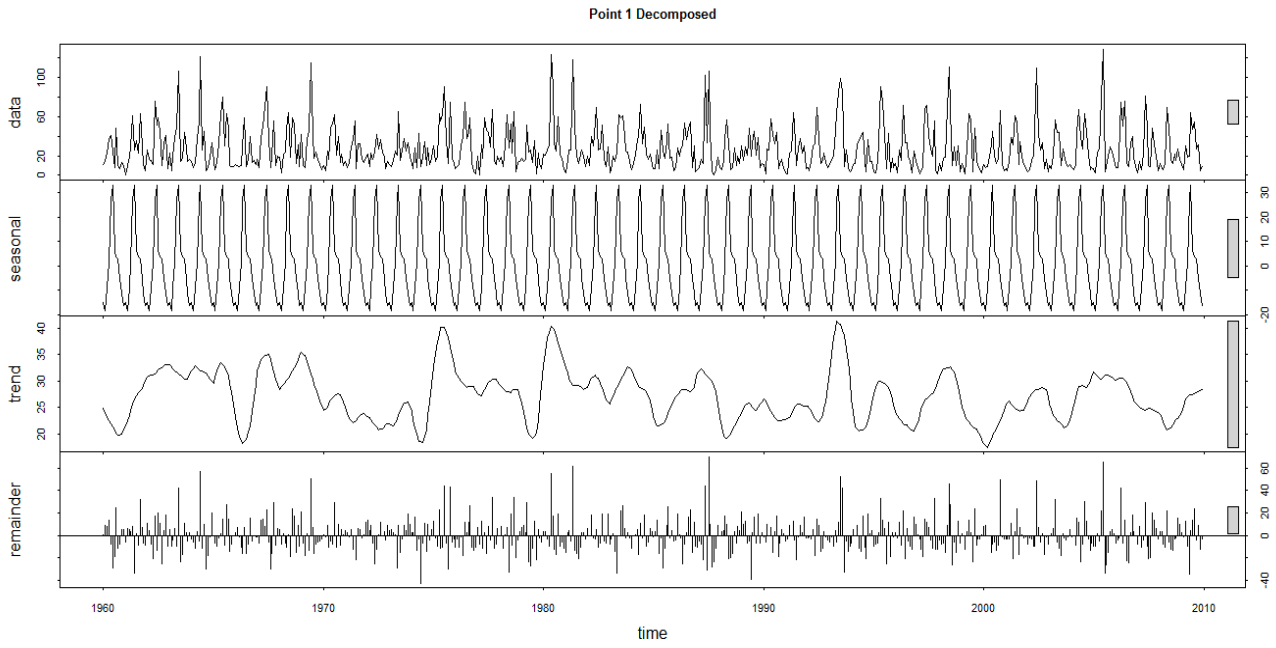

Figure 6: Periodic Decomposition of Butte, Montana

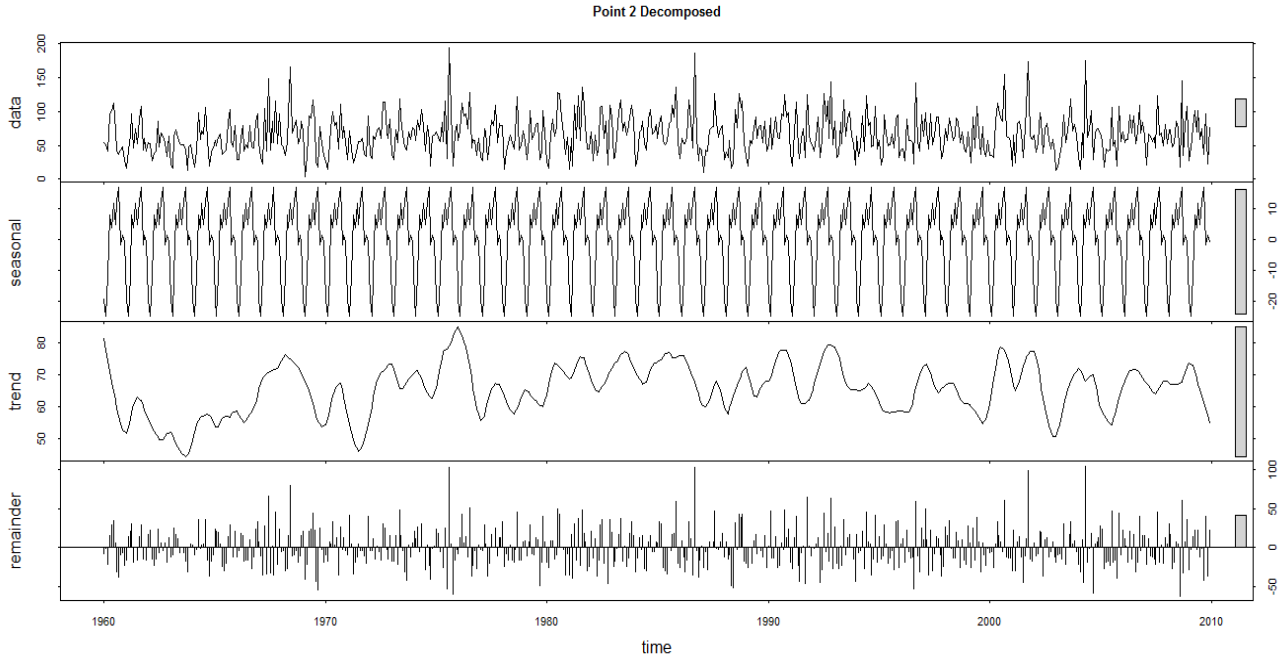

Figure 7: Periodic Decomposition of Flint, Michigan

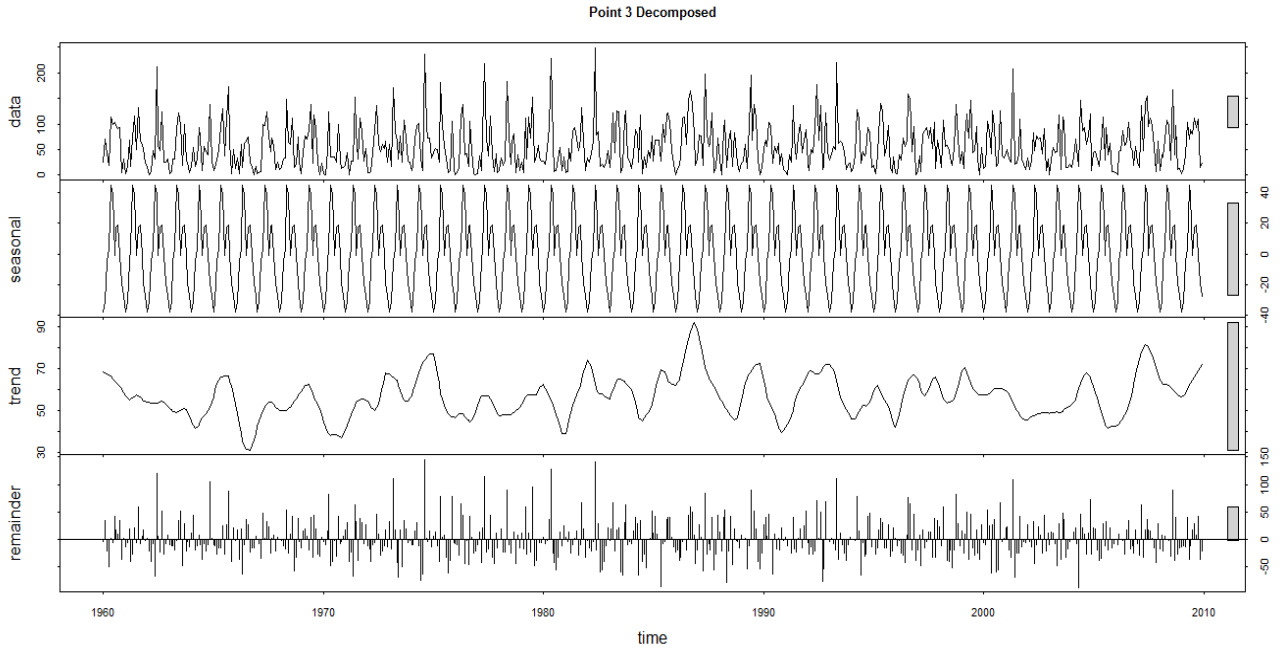

Figure 8: Periodic Decomposition of Woodward, Oklahoma

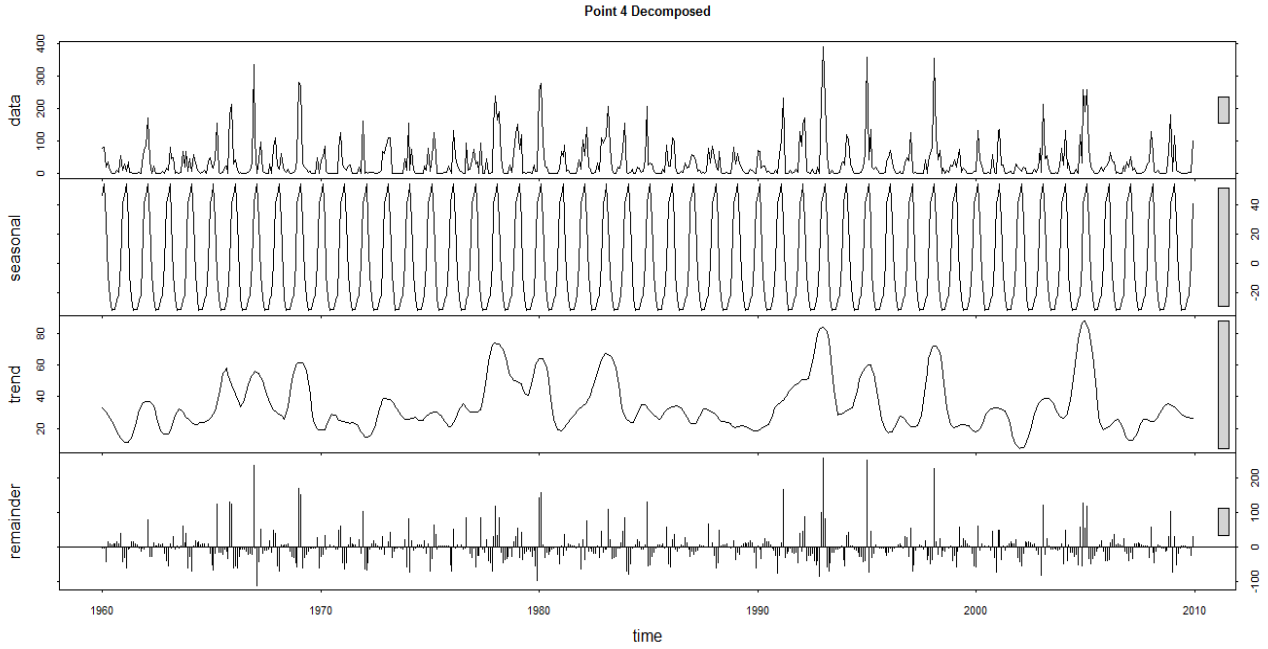

Figure 9: Periodic Decomposition of Orange County, California

The deviations of monthly precipitation from the mean annual precipitation on a month-to-month basis are shown in Figure 10. These plots show the per-month variability of rainfall with respect to the mean values of annual precipitation by month. Orange County has very little range and small deviations from the annual mean in the summer months, meaning that the summers are consistently dry in that area. Butte and Woodward have lower range and deviations in the winter months instead, as would be expected for semi-arid regions. However, Flint has a much larger range, yet still maintains little deviation from the annual mean all year round.

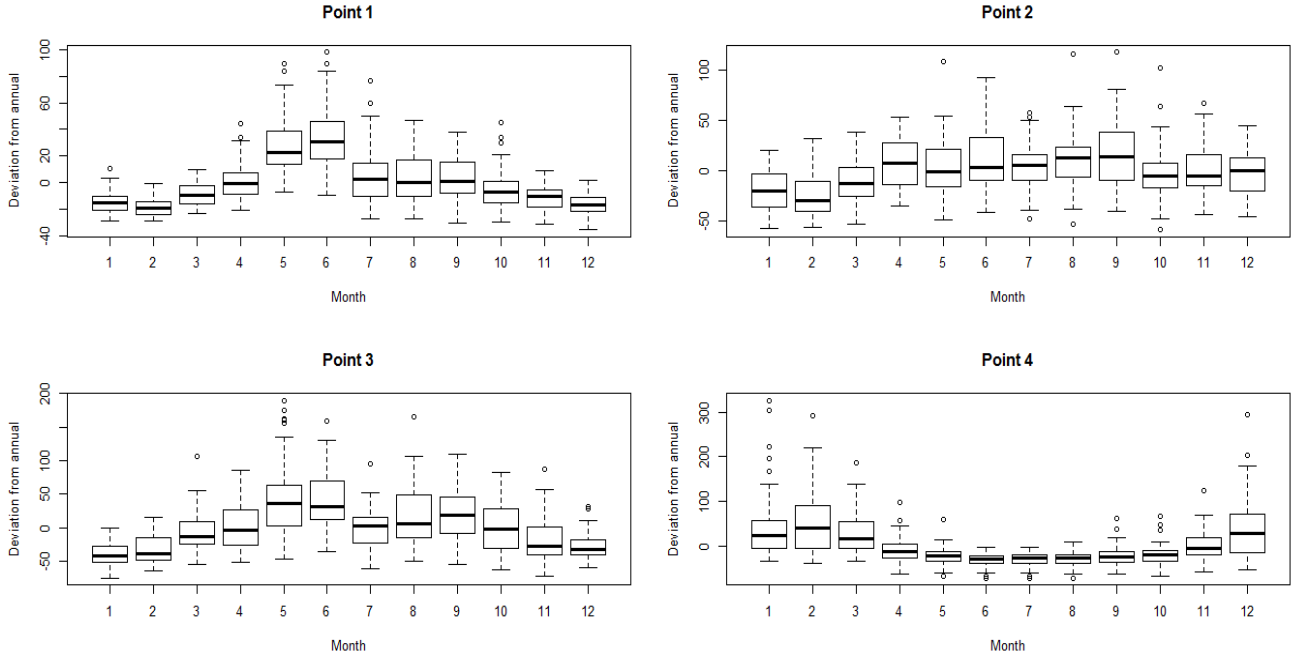

Figure 10: Per-Month Variability (monthly deviation from annual mean)

4.2 Autocorrelation and ARIMA Modelling

Before the time series has been transformed, the trend and seasonal components are examined to find any patterns. The trend component needs to be smoothed with a moving average smoothing, which

will occur in the ARIMA modelling process. For the seasonal component, the ACF and PACF plots of the original data (Figure 11 & Figure 12, respectively) clearly show the seasonal patterns of each point. The autocorrelation function reverses every 6 months due to the effects of seasonal changes throughout the year; however these seasonal trends were expected for precipitation time series data. The partial autocorrelation function shows very similar results. The ACF plots of Butte and Orange County have more highly auto-correlated seasonal data, which in turn means they will have higher predictability after modelling (NIST, 2003). All ACF and PACF plots clearly indicate the seasonality of rainfall, representing the fact that the same months are correlated with each other during consecutive years (Soltani et al., 2007).

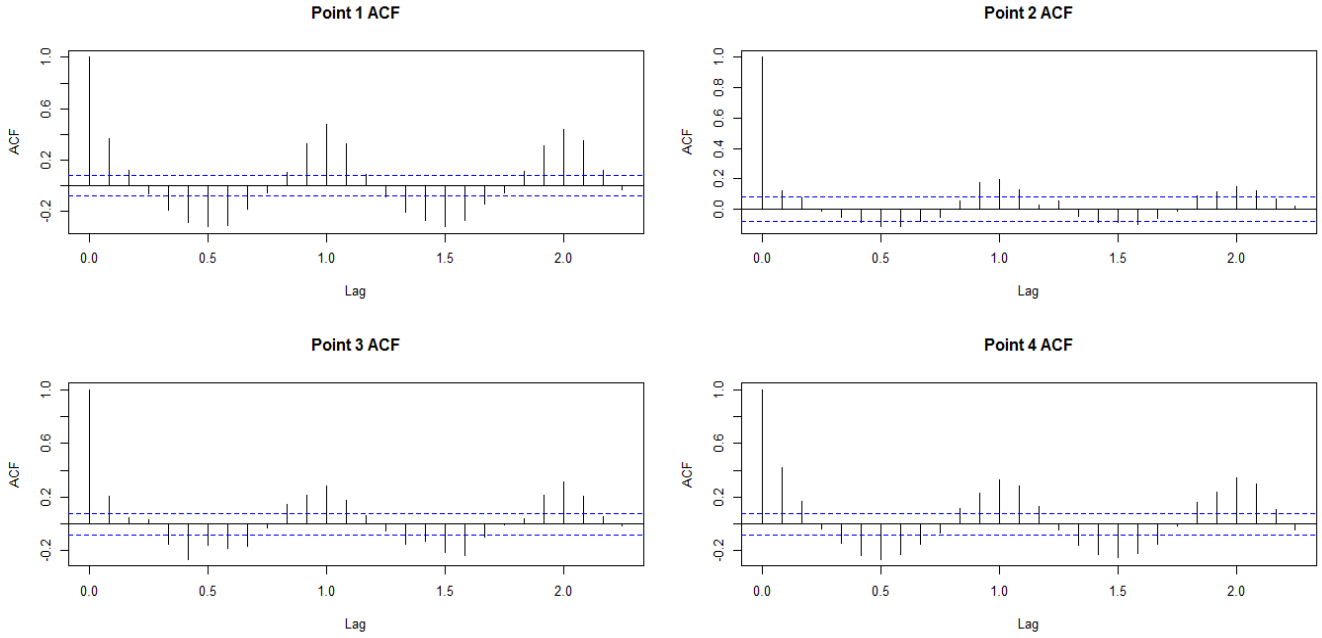

Figure 11: ACF Plots of Original Data

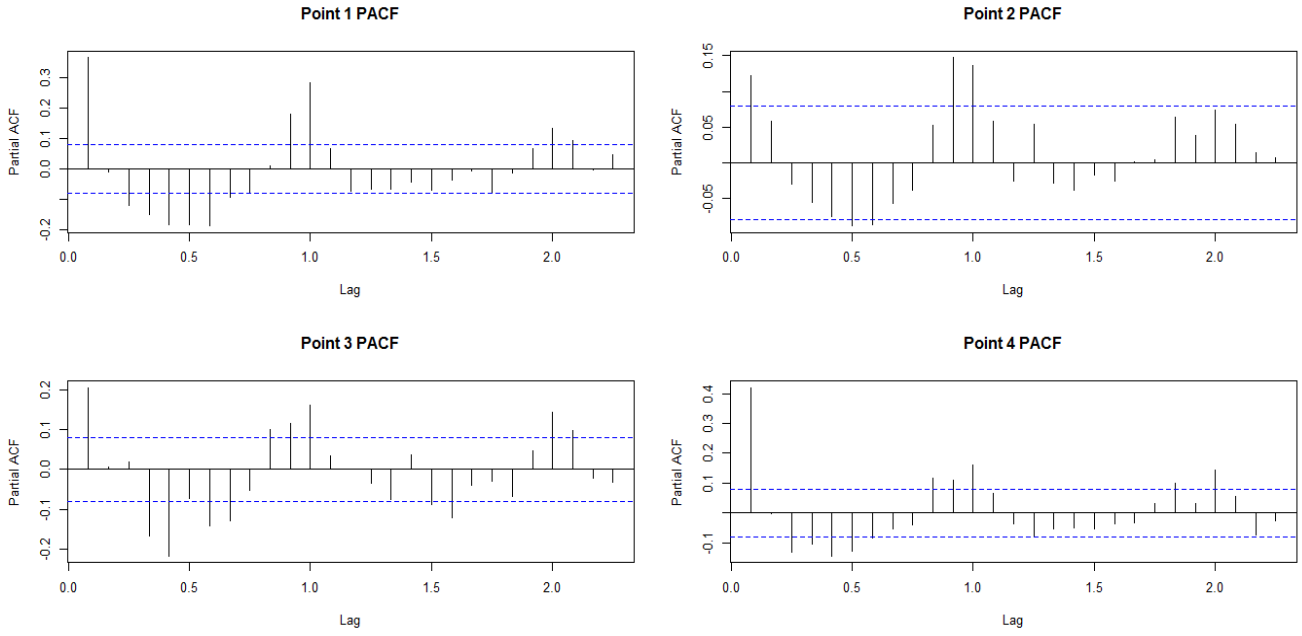

Figure 12: PACF Plots of Original Data

The data needed to be stationary before any modelling or further analysis could take place. To become stationary, the data needs to be differenced and the autocorrelations need to be examined for constant autocorrelation through time. The differenced ACF plots for all points (Figure 13) represent the stationary

time series which can further be modelled and forecasted. The differenced PACF plots (Figure 14) were also produced at this point to be used later in parameter selection for the ARIMA models.

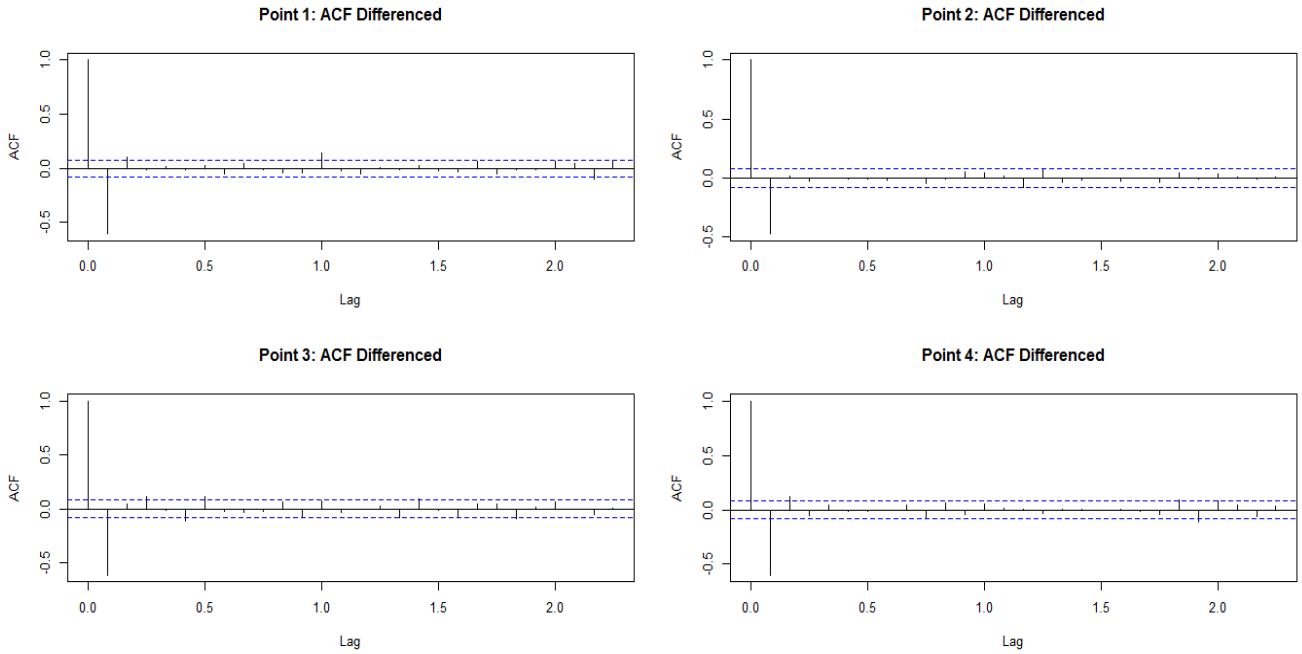

Figure 13: ACF Plots of Differenced Data

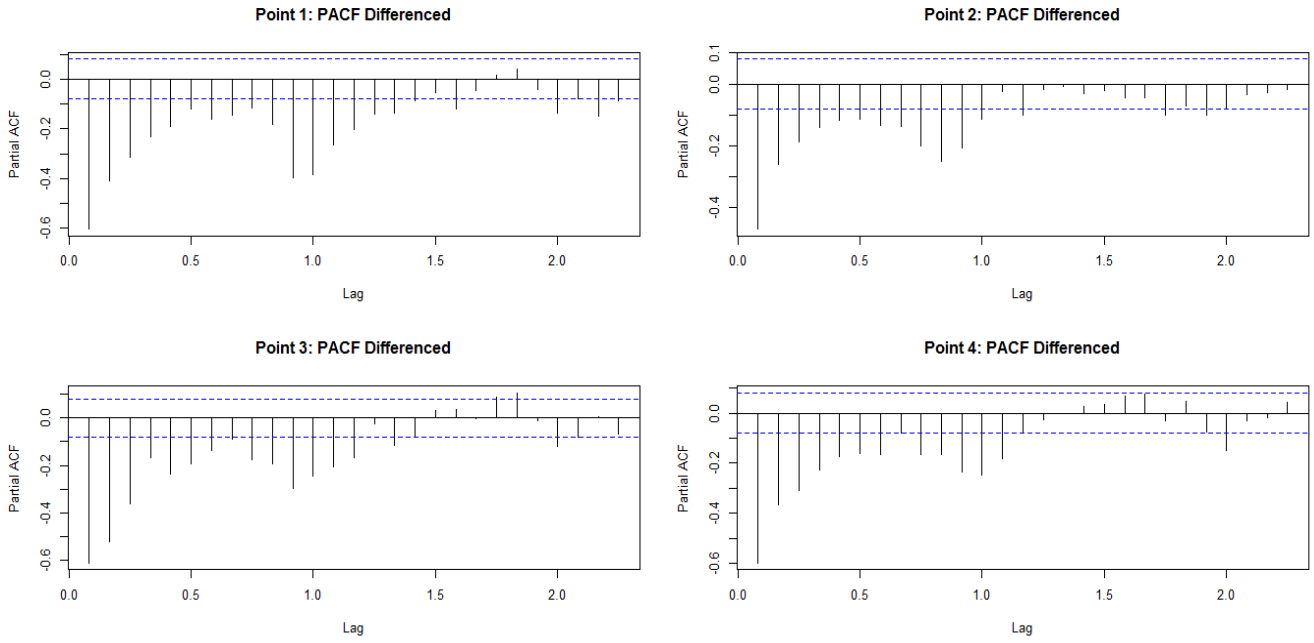

Figure 14: PACF Plots of Differenced Data

The parameters to be estimated in the ARIMA models were chosen based on the analysis of both the differenced ACF and PACF plots. The following parameters were selected for each point:

Butte:	ARIMA $(1, 1, 1)\ (0, 1, 1)_{12}$
Flint:	ARIMA $(2, 1, 1)\ (0, 1, 1)_{12}$
Woodward:	ARIMA $(2, 1, 1)\ (1, 2, 2)_{12}$
Orange County:	ARIMA $(2, 1, 1)\ (1, 2, 2)_{12}$

These parameters were chosen from the shape pattern of the plots and the location of significant spikes at certain lags on both the ACF and PACF plots. Several ARIMA models and their residuals were examined before selecting the values above.

The tests used to evaluate the accuracy of the models, based on the estimated parameters listed above, resulted in almost all of the spikes in the residual plots fitting within the 5% significance limits. Figure 15 through Figure 18 show an in-depth breakdown of the ACF and PCF plots for the differenced data of Point 1 through Point 4, respectively. These plots were used to determine the parameters.

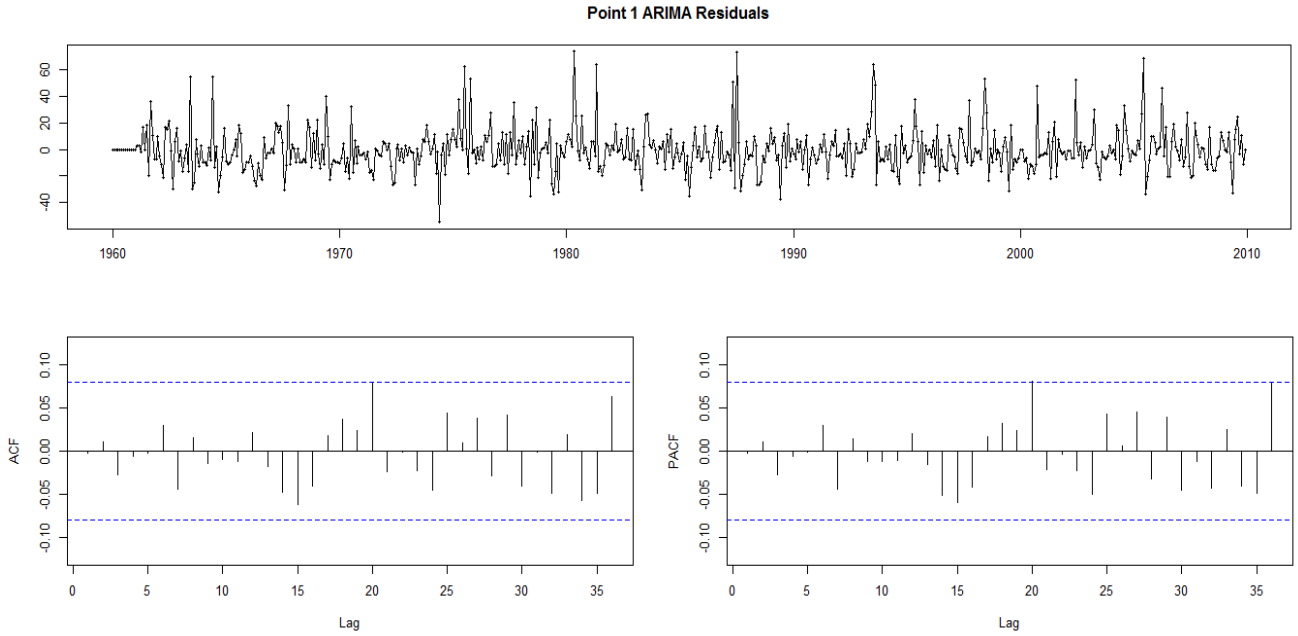

Figure 15: Butte, Montana ARIMA Parameter Analysis

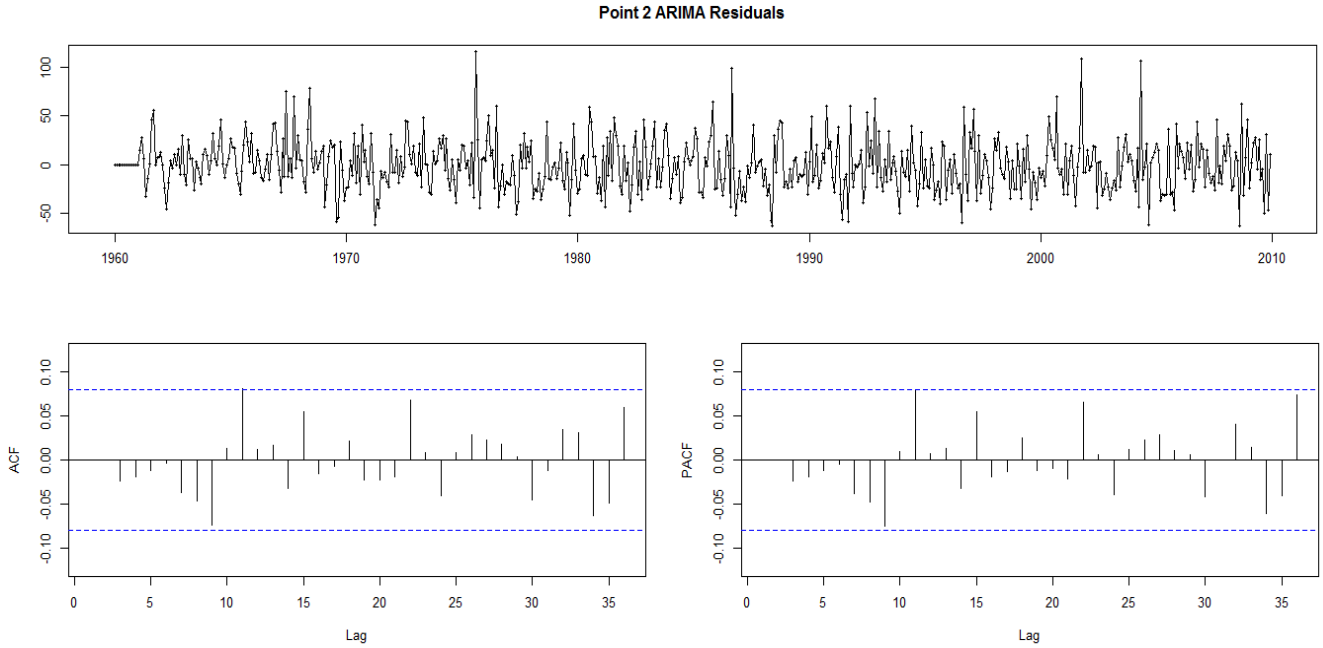

Figure 16: Flint, Michigan ARIMA Parameter Analysis

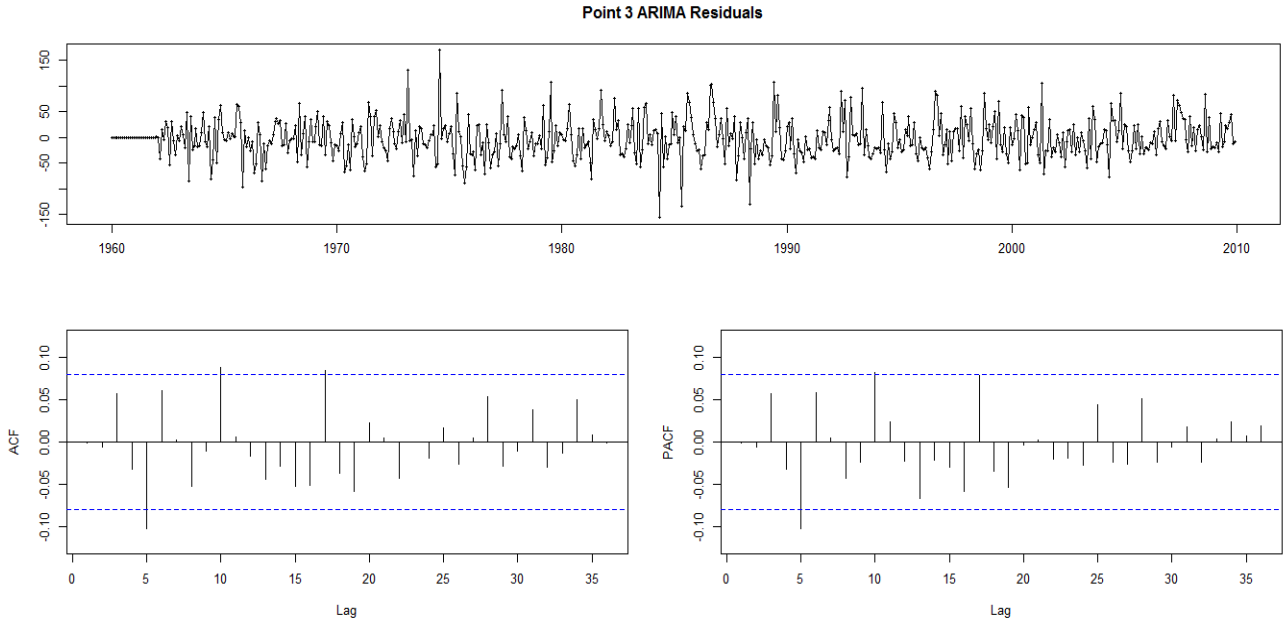

Figure 17: Woodward, Oklahoma ARIMA Parameter Analysis

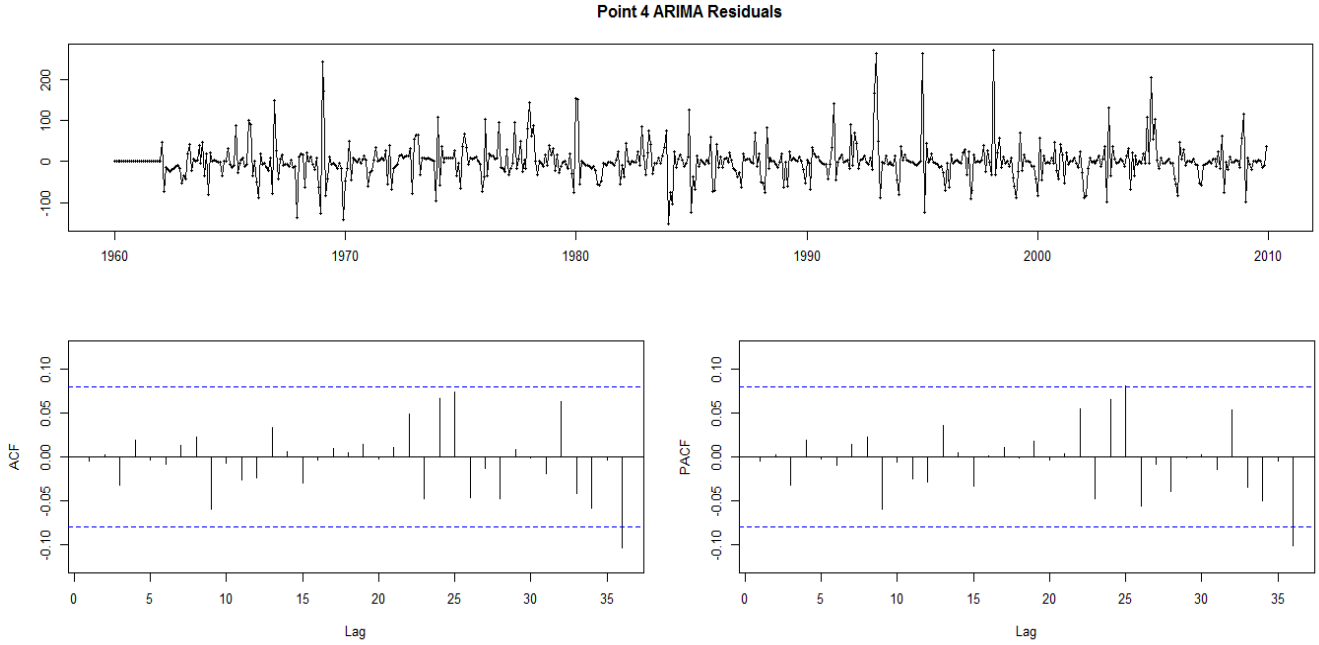

Figure 18: Orange County, California ARIMA Parameter Analysis

4.3 Forecasting

Now that there are seasonal ARIMA models for the time series, the data can be used for forecasting. In Figure 19 each point is forecasted three years into the future with the forecast shown as a blue line, the 80% prediction intervals as an orange shaded area, and the 95% prediction intervals as yellow shaded area (Coghlan, 2011). The values predicted in the plots below do not account for any possibility of extreme weather events, but instead seem to forecast a safe window of values based on recent precipitation values. The forecast predicts high precipitation, relative to the mean, in Woodward and low precipitation in Flint and Orange County.

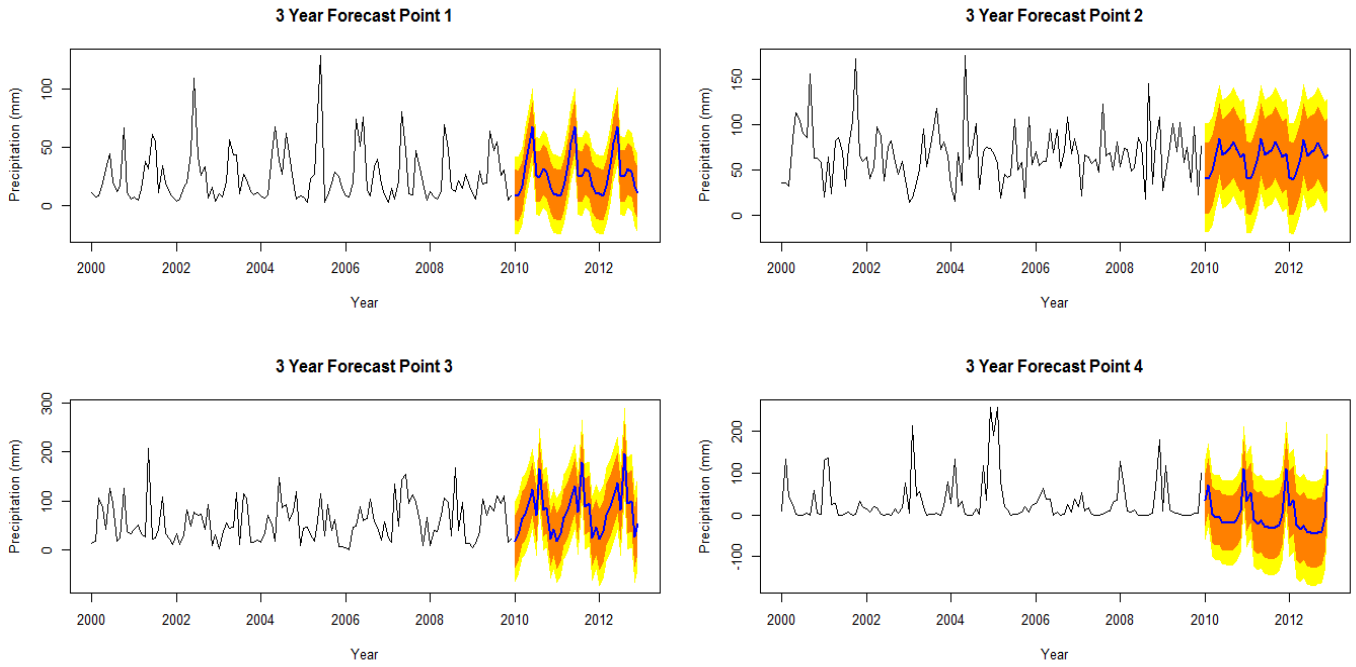

Figure 19: 3 Year Forecast using ARIMA Models

Another method to forecast values for this time series is exponential smoothing. The forecast tool in R was used to predict the same time period for an exponential smoothing models (Figure 20).

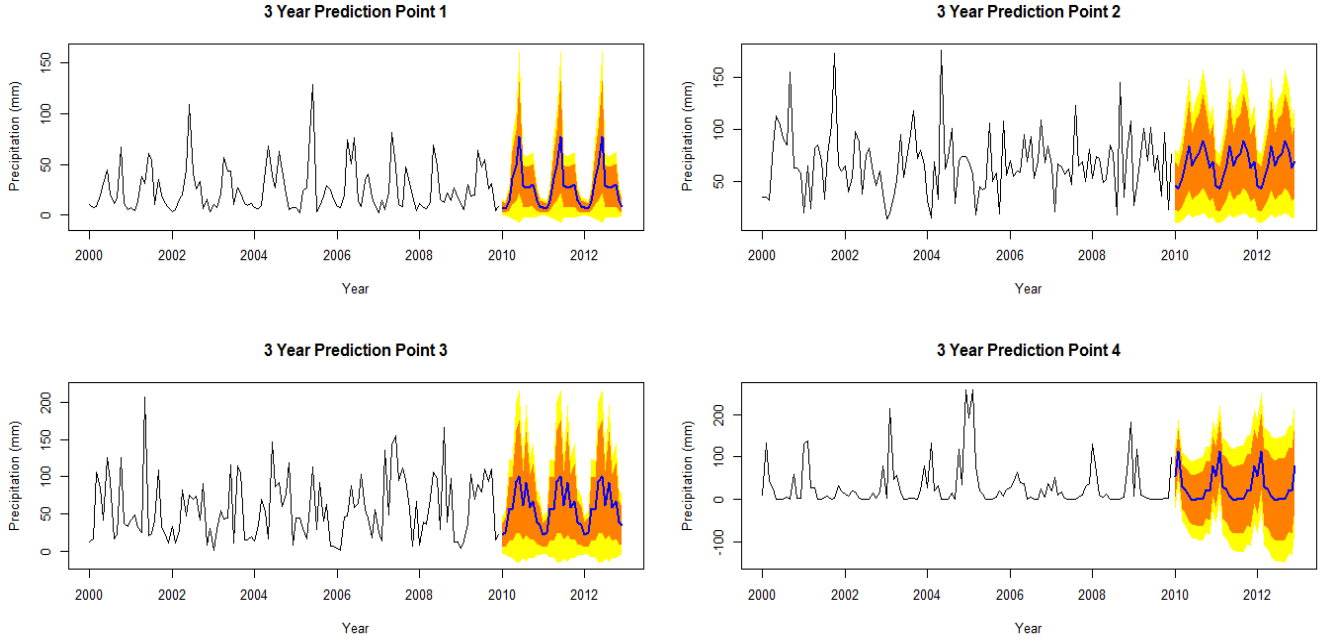

Figure 20: 3 Year Forecast using Exponential Smoothing

The two models used for forecasting look similar at first glance, but the accuracy of the forecasting models will determine which method is better. Table 3 relates the root mean square errors (RMSE) of all points for both methods. This is also known as the estimated white noise standard deviation in ARIMA analysis and it is the statistic whose value is minimised during the parameter estimation process (Hyndman, 2012). Overall, the ARIMA models are just slightly better than the exponential smoothing models in terms of forecasting over a 3 year period.

Table 3: Accuracy of Forecasting Methods

	ARIMA RMSE	Exponential Smoothing RMSE
Butte, Montana	15.108	15.614
Flint, Michigan	27.965	27.483
Woodward, Oklahoma	32.369	33.077
Orange County, California	36.337	37.818

4.4 Fourier Analysis

The spectrum tool plotted smoothed periodograms which are used to identify the dominant frequencies of the time series (Figure 21). This lead to identifying the dominant cyclical behaviour in the series at certain frequencies.

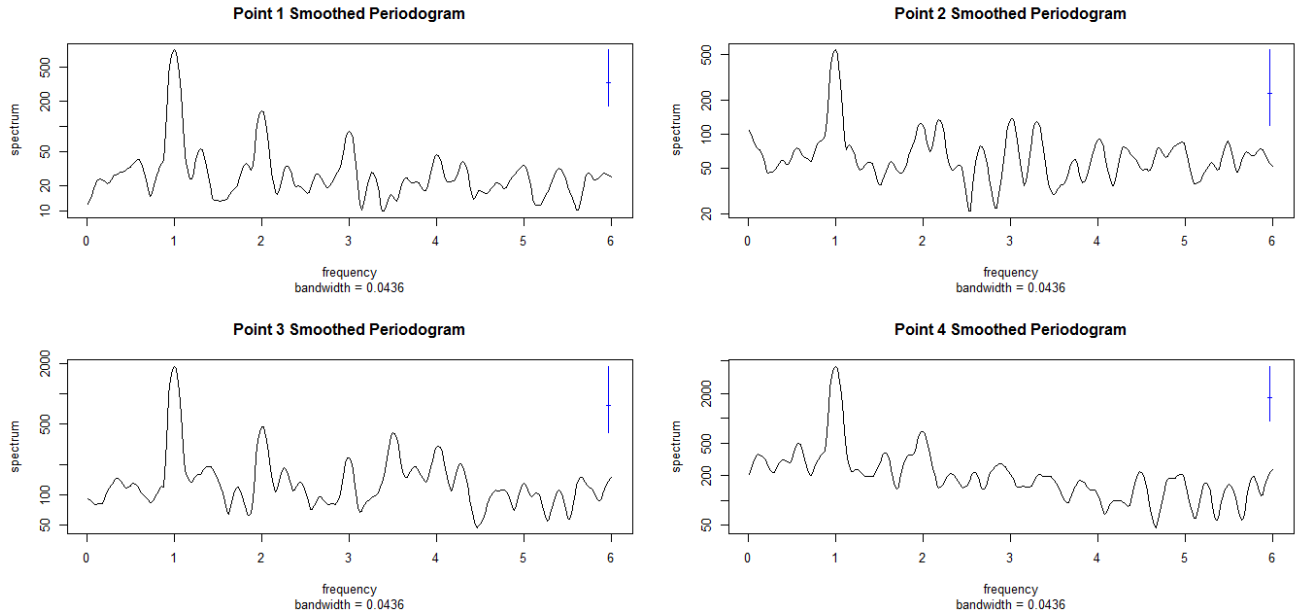

Figure 21: Smoothed Periodograms

The periodograms show that there is an obvious cyclical pattern every year, as well as other patterns due to seasonality. The first peak in the periodogram will provide information on the long-term cyclical pattern for each area of interest. Table 4 identifies the length of the cyclic pattern for all points.

Table 4: Cyclic Pattern Length

	Years
Butte, Montana	1.636
Flint, Michigan	1.714
Woodward, Oklahoma	3.000
Orange County, California	5.000

The cyclic patterns for both Butte and Flint are not significant enough to examine further. Woodward and most notably Orange County appear to have stronger cyclic patterns. The pattern for Woodward represents a consistent precipitation pattern approximately every 36 months. As mentioned before, this area is located between climate zones, so the cyclic pattern could be caused by extreme weather events, such as storms or monsoons, which occur at that cycle length. Severe storms, like thunderstorms, are not uncommon in this location, but seem to occur at a 3-year cycle. The pattern for Orange County is easily identified as the ENSO cycle, which occurs approximately every 5 years. This is exactly what one would expect to see when identifying cyclic patterns in Southern California. However, the spectral analysis only

justifies the separation time between El Niño events in this area, yet cannot determine exactly when the event will occur in a certain year.

5. Conclusion

This 50-year study of precipitation data identified several patterns and anomalies in precipitation events of four selected areas in the continental United States. By decomposing the time series data, trends and seasonal cycles were detected for different environments. The range of precipitation levels varied from point to point, with Orange County, California having the most variation and Flint, Michigan having the least. The periodic decomposition also revealed some extreme weather patterns found in Woodward, Oklahoma, due to severe storms, and Orange County, due to El Niño events. The autocorrelation and ARIMA modelling processes of the analysis helped determine a method of fitting a model to each time series for further forecasting. This forecasting method and an exponential smoothing technique allowed for 3 years of prediction for precipitation levels. A spectral analysis of this data was used to find long-term cyclical patterns at each location. Butte, Montana and Flint did not have any significant patterns; however in Woodward and Orange County, 3-year and 5-year cycles, respectively, could be identified with the analysis of periodograms. The effect of El Niño events in Southern California is clear in both the periodic decomposition and the spectral analysis. These patterns in both locations could slightly be recognised in the decomposition analysis, but the length of the cycle was calculated with the Fourier analysis.

This study can possibly be extrapolated over similar environments for determining the precipitation statistics in those areas as well. Still, more information on that environment would be needed for an accurate analysis. Some studies suggest that these climatic patterns can be identified on a global scale with the right amount of data and data accuracy in the future (Chen et al., 2002).

6. References

Chatfield, C. (1996). The Analysis of Time Series. An Introduction. 5th edition, New York, Chapman & Hall.

Chen, M., Xie, P., et al. (2002). “Global land precipitation: A 50-yr monthly analysis based on gauge observations.” Journal of Hydrometeorology **3**(10): 249-266.

Coghlan, A. (2011). A Little Book of R for Time Series. Release 0.1, Ireland, Creative Commons Attribution.

Cowpertwait, P. S. P. (2006). Introductory Time Series with R. New York, Springer Science & Business Media. LLC.

Galford, G. L., J. F. Mustard, et al. (2008). "Wavelet analysis of MODIS time series to detect expansion and intensification of row-crop agriculture in Brazil." Remote Sensing of Environment **112**(2): 576-587.

Huffman, G. J., R. F. Adler, et al. (1997). “The global precipitation climatology project (GPCP) combined precipitation dataset.” Bulletin of American Meteorological Society **78**(1): 5-20.

Hyndman, R .J. and G. Athanasopoulos (2012). Forecasting: Principles and Practice. Available at: http://otexts.com/fpp/, Accessed on: 27/9/2012.

Ihaka, R. (2005). Time Series Analysis. University of Auckland, New Zealand. Available at: www.stat.auckland.ac.nz/~**ihaka**/726/notes.pdf

Kottek, M., J. Grieser, et al. (2006). “World map of the Koppen-Geiger climate classification updated.” Meteorologische Zeitschrift **15**(3): 259-263.

Lau, K. and H. Wang (1995). "Climate signal detection using wavelet transform: How to make a time series sing." Bulletin of the American Meteorlogical Society **76**(12): 2391-2402.

Martinez, B. and M. A. Gilabert (2009). "Vegetation dynamics from NDVI time series analysis using the wavelet transform." Remote Sensing of Environment **113**(9): 1823-1842.

McLeod, A. I., H. Yu, et al. (2011). "Time sereis analysis with R." Handbook of Statistics **30**.

Mitchell, T. D., M. Hulme, et al. (2002). "Climate data for political areas." Area **34**(1): 109-112.

Mitchell, T. D. and P. D. Jones (2005). "An improved method of constructing a database of monthly climate observations and associated high-resolution grids." International Journal of Climatology **25**(6): 693-712.

New, M., M. Hulme, et al. (2000). "Representing twentieth-century space-time climate variability. Part II: Development of 1901-96 monthly grids of terrestrial surface climate." Journal of Climate **13**(13): 2217-2238.

NIST/SEMATECH (2003) Engineering Statistics - Handbook of Statistical Methods. Available at: http://www.itl.nist.gov/div898/handbook/, Accessed on 7/9/2012.

Rossiter, D. G. (2012). "Time series analysis in R." Technical Report ITC, Enschede, NL. Version 1.0.

Ruhf, R. J. and E. M. C. Cutrim (2003). "Time series analysis of 20 years of hourly precipitation in southwest Michigan." Journal of Great Lakes Research **29**(2): 256-267.

Shaddick, G. (2004). Using R (with applications in Time Series Analysis). University of Bath, England.

Soltani, S., R. Modarres, et al. (2007). "The use of time series modeling for the determination of rainfall climates of Iran." International Journal of Climatology **27**(6): 819-829.

Torrence, C. and G. P. Compo (1998). "A pracitcal guide to wavelet analysis." Bulletin of the American Meterologial Society **79**(1): 61-78.

Quiroz, R., C. Yarleque, et al. (2011). "Improving daily rainfall estimation from NDVI using a wavelet transform." Environmental Modelling & Software **26**(2): 201-209.

Zucchini, W. and O. Nenadic (2011). Time Series Analysis with R - Part I. University of Goettingen, Germany. Available at: http://bit.ly/HsiVH