Time Series Analysis on Differing Climate Regions

By

Jacob Insley, B.S Mathematics

A thesis submitted to the Graduate Committee of

Ramapo College of New Jersey in partial fulfillment

of the requirements for the degree of

Master of Science in Data Science

May 2024

Committee Members:

Debbie Yuster, Advisor

Stefan Becker, Reader

Matthew Jobrack, Reader

COPYRIGHT

Table of Contents

List of Tables

List of Figures

Abstract

Droughts, hurricanes, tornadoes, and other climate disasters wreak havoc in all corners of the world. Constantly, scientists and mathematicians are working on ways to predict such events and learn more about them. Unfortunately, the weather remains incredibly difficult to predict. If we can learn more about how data science and time series methods work on a variety of climate regions, we can understand how to put them to better use.

Two locations with very different seasonal patterns were looked at: Bergen County, NJ and Napa County, CA. Droughts were classified in both regions and different drought indices were compared in their ability to identify droughts. Time series techniques were used to predict the amount of precipitation in each location. A least squares regression model with a seasonal component and a SARIMA model were created to predict precipitation. We were able to discover some of the strengths and weaknesses of these tools when used on data from different climate regions.

The Standardized Precipitation Index was able identify short term drought well but failed to identify droughts during Napa County summers. Palmer Drought Severity Index identified droughts all year long in both locations, but only identified droughts if they occurred over several months. The SARIMA model decisively portrayed the seasonal pattern of Napa's precipitation, but made more accurate predictions for the more stable climate of Bergen County.

Scientists can use this data to better equip these tools to handle situations which they do not excel at. We can use what we learned about how drought is measured in both locations to find ways to improve the indices we measure with.

Introduction

Droughts are worldwide phenomena that have greatly impacted the livelihoods of billions of people all over the world. Drought can be characterized as a decrease to an area's water availability due to rainfall deficiency [11]. They occur unexpectedly and their effects are felt in different ways depending on climate and socioeconomic factors of the location where the drought occurs. People have dedicated lifetimes to creating and utilizing tools that will allow us to better predict when droughts will occur and come up with ways to handle them. Modeling and machine learning tools are being tested on time series data to better understand how droughts and our climate work such as the Seasonal Autoregressive Integrated Moving Average model [9]. However, a drought often occurs unexpectedly and leaves local populations unprepared to handle its impact.

The goal of this project is to explore the ways in which time series analysis and various prediction methods can be employed to the study of two different climate locations with different seasonal precipitation patterns. Bergen County in New Jersey and Napa County in California were chosen as the locations of study for this project. They are both in the USA, but they have very different climate profiles. We want to see how important the weather pattern is to predict and analyze the precipitation of any given location. To do this, how different drought indices measure drought will be examined in both environments. Time series analysis techniques will be utilized to discover how Napa and Bergen's weather differ. Finally, different prediction models will be created to learn how well they work on areas with different seasonal precipitation patterns. By the end of this project, we should have a better understanding of the importance of climate seasonality when it comes to droughts and precipitation.

Background

This project focuses on precipitation and drought measurements in Bergen County, New Jersey and Napa County, California. Bergen County is the home of Ramapo College. It was chosen since it receives consistent precipitation throughout each year. California has been in the news in recent times for its dry conditions and continuous drought. Napa County is the wine capital of the country and produces an enormous amount of wine that is drunk all over the world. They are intensely vulnerable to the dire effects of droughts. Napa was chosen because it's precipitation fluctuates from receiving little to no precipitation during its summers and receiving heavier amounts of precipitation in its winter.

Extreme Precipitation

Droughts and floods represent opposite sides of extreme precipitation related events. However, both are responsible for billions of dollars in damage and thousands of deaths worldwide [10]. In 2011, droughts in the United States and Mexico caused an estimated $8 billion in damages [11]. In that same year, 35 million people in China were affected by droughts [11].

From 1970 to 2019, floods have caused an estimated 58,700 deaths and $115 billion in damages worldwide [10]. In the USA from 2010 to 2022, the National Weather Service has estimated there to have been 1352 deaths due to flood related events [10]. Several methods have been developed to help mitigate the effects of these events. One such method is in early warning systems, many of which are being developed using artificial intelligence. They use AI on real time geographical data and image validation to report flooding and potential flooding as soon as

risks occur [10]. This has proven to be an effective way to stay just ahead of floods and similar climate events as they happen, but they do not do so well at enabling long-term preparation.

Drought Indices

There are several drought indices being used to measure droughts across the world[1]. Each one tends to measure drought in a different way from the others. Some only include precipitation measurements in their assessment of drought while some incorporate temperature or evapotranspiration [3]. However, two commonly used drought indices are the Standardized Precipitation Index, known as SPI and the Palmer Drought Severity Index, known as PDSI [1].

SPI was developed in 1993 and is obtained by comparing the given precipitation from a specific location to the historical precipitation of that region over a specific time scale [1]. SPI only needs a historic precipitation data set to be calculated. The data set is fitted to a gamma probability density function. The gamma function used is $g(x) = \frac{1}{\beta^{\alpha}\gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}$ where $x > 0$ is the precipitation amount, $\alpha > 0$ is a shape parameter, $\beta > 0$ is a scale parameter, and $\gamma(\alpha) = \int_0^{\infty} y^{\alpha-1}\, e^{-y} dy$ is the gamma function.The parameters, α and β, are solved for using the following maximum likelihood solutions: $\hat{\alpha} = \frac{1}{4A}(1 + \sqrt{1 + \frac{4A}{3}})$ and $\hat{\beta} = \frac{x}{\alpha}$ where $A = lnx - \frac{\sum \ln(x)}{n}$ and n is the number of precipitation observations [5]. The function is used to find the cumulative probability for a precipitation event occurring in a given time scale. Finally, the probability distribution is transformed into a normal distribution with a mean of 0 and standard deviation of 1 [1]. The greater the value of SPI is above 0, the wetter the conditions of the chosen location were for that month and for values below 0, smaller values define dryer conditions. This index was created to better compare drought measurements across regions with

varying climates. A drought is identified if SPI has a value of -1 or below in the region you are measuring [1].

PDSI was developed in 1965 and uses precipitation, soil moisture, temperature, and evapotranspiration to create a water balance model that calculates how dry or wet a particular region is [1]. PDSI also makes use of the condition of the soil moisture from previous months to help in its calculation and does not rely solely on historical data of specific climate variables like SPI and other drought indices [2]. Therefore, it has been effective at categorizing long term droughts. PDSI runs on a scale from -10 to 10, representing dry and wet conditions respectively.

Methodology

The precipitation data for this thesis project was collected from the PRISM climate group [12]. PRISM stands for Parameter-elevation Regressions on Independent Slopes Model. PRISM records localized climate data from a collection of nearby climate recording station locations and weighs them to come up with accurate precipitation measurements for the area you are focusing on [3]. The data comes from a gridded data set. We collected monthly precipitation data for Napa County, California and Bergen County, New Jersey from January 1895 to December 2023. The data for each location is measured at a single grid cell within each county at 4km resolution. The Napa County data was measured at latitude of 38.5064 and longitude of -122.3305. The Bergen County data was measured at latitude of 40.9617 and longitude of -74.0782 [12]. Each data point is the total amount of monthly precipitation that was measured in the selected location. All the data was compiled into Python and explored using Python and various python packages.

The SPI values for this project were obtained by using the National Drought Mitigation Center's SPI Generator [4]. The generator uses the gamma distribution mentioned above to convert monthly precipitation data into SPI values. We used the PRISM precipitation data for this calculation, and we output monthly SPI values as measures of wetness of the region in order to determine if either location was in a drought. The PDSI values were taken from the National Oceanic and Atmospheric Administration's National Climate Data Center [13]. The dataset consists of monthly PDSI values from January 1895 to December 2023. The values are representative of United States climate divisions [13]. A division's measurement is gained by weighing various station data within that division to come up with a monthly average

measurement. The 1st climate divisions of California and New Jersey contain the two counties used in this project and are where the PDSI measurements come from [13].

Exploratory Data Analysis

We look at precipitation data for Napa County, California and Bergen County, New Jersey to see how the two locations may differ.

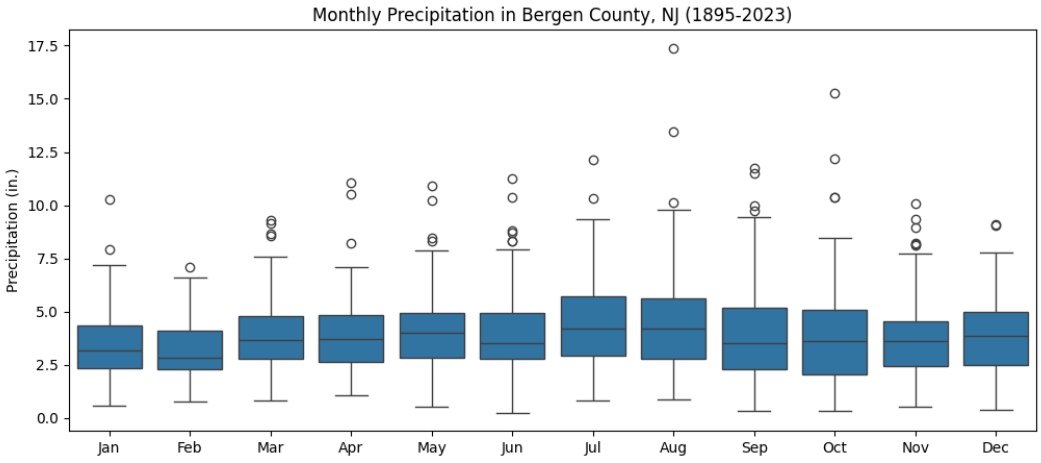

Figure 1: Boxplot of Monthly Precipitation Totals in Bergen County, NJ. Bergen County has a mostly even distribution of precipitation throughout the months of the year. There is no immediately visible seasonal trend that can be noticed. There is a peak in the data with precipitation in July and August.

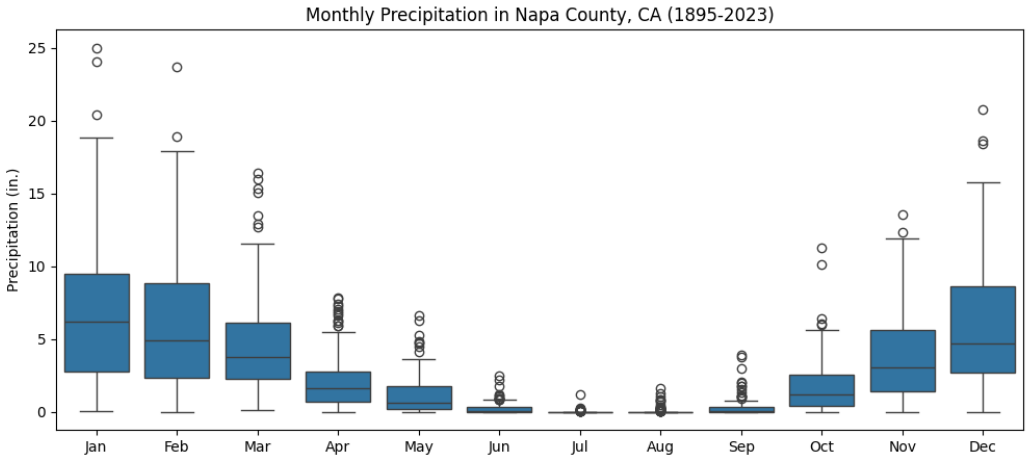

Figure 2: Boxplot of Monthly Precipitation Totals in Napa County, CA. Napa County has a significant seasonal trend to its precipitation data. The months of November through March get significantly more precipitation than the other months. June through September, specifically, get little to no precipitation over the entire 128-year dataset.

California sees significant changes in its levels of precipitation from the winter to summer. If we look at July specifically, Napa County receives an average of 0.02 inches of precipitation during July with a maximum of 1.19 inches of precipitation in July 1974. Also, both July and August have a median precipitation of 0.00 and June and September have a median precipitation of 0.05. Those numbers indicate that precipitation should not be expected during those summer months.

When looking at the data for Bergen County, every month averages between 3 and 5 inches of precipitation. There is much less of a need to prepare for large changes in precipitation throughout the year in Bergen County. However, there can still be unexpected changes in the weather. For example, the most amount of precipitation was recorded in August 2011 with 17.4 inches. This coincided with the arrival of Hurricane Irene, which showed how volatile and

unpredictable the weather can be. In figures 3 and 4, we can see how precipitation has changed from 1895 to 2023 in both locations.

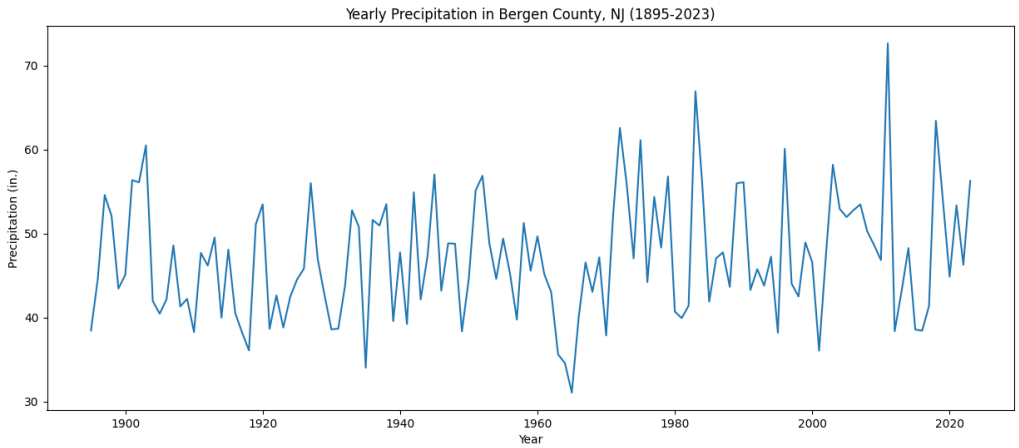

Figure 3: Yearly Line Plot of Bergen County, NJ Precipitation: This plot shows the total yearly precipitation in Bergen County from 1895 to 2023. The total yearly precipitation fluctuates within the range of 31.05 and 72.63 inches of precipitation throughout the dataset.

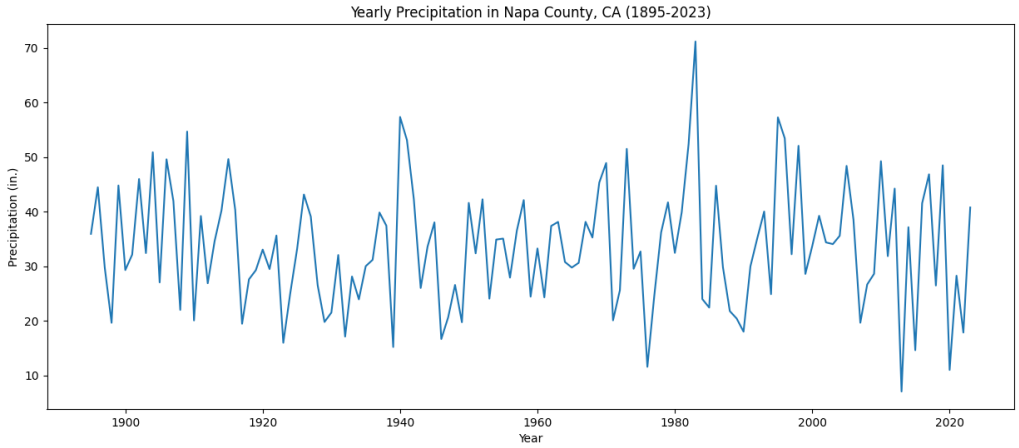

Figure 4: Yearly Line Plot of Napa County, CA Precipitation: This visualization shows the total yearly precipitation in Napa County from 1895 to 2023. The image shows just how much

precipitation can change year to year. Historical precipitation data does not generate a smooth curve.

Trend analyses were conducted in both locations to determine if either data set exhibits any kind of trend. The Mann-Kendall test was used to determine if a trend exists in Bergen County or Napa County's precipitation. A hypothesis test is made where the null hypothesis indicates that there is no trend and a significance value of 0.05 was used. The test produced a p-value of 0.0715 for Bergen County, NJ and a p-value of 0.9464 for Napa County, CA. In both cases, we fail to reject the null hypothesis indicating neither data set to have a significant trend. Figures 3 and 4 show changes from years with incredibly high amounts of precipitation to years with extremely low precipitation only a few years apart. Interestingly, this jagged image of yearly precipitation appears in both Bergen County and Napa. Despite Bergen's months having a range of 3.16 to 4.53 inches of average precipitation, the total amount from year to year has a range of 41.58 inches. California's yearly precipitation ranges from 7.06 to 71.18 inches of precipitation. Both locations have years that are greater than 40 inches of precipitation apart. Even though California and New Jersey have different patterns of precipitation, they both have no long-term trend in their data with erratic changes year to year. It is understandable why predicting precipitation is a difficult task.

Comparing Drought Indices

This section will look at how two different drought indices identify drought in Bergen County and Napa County. We use a value of -2 or below to represent drought from PDSI. On Palmer's drought scale, -2 and below represents at least a moderate drought which is significant enough to cause concern [1]. For SPI, we identify drought when SPI is at or below -1 as this is

the commonly used drought point for this index [1]. We start by seeing how often SPI and PDSI have record droughts in each region. We can look to see if there exists a seasonal trend for drought identification.

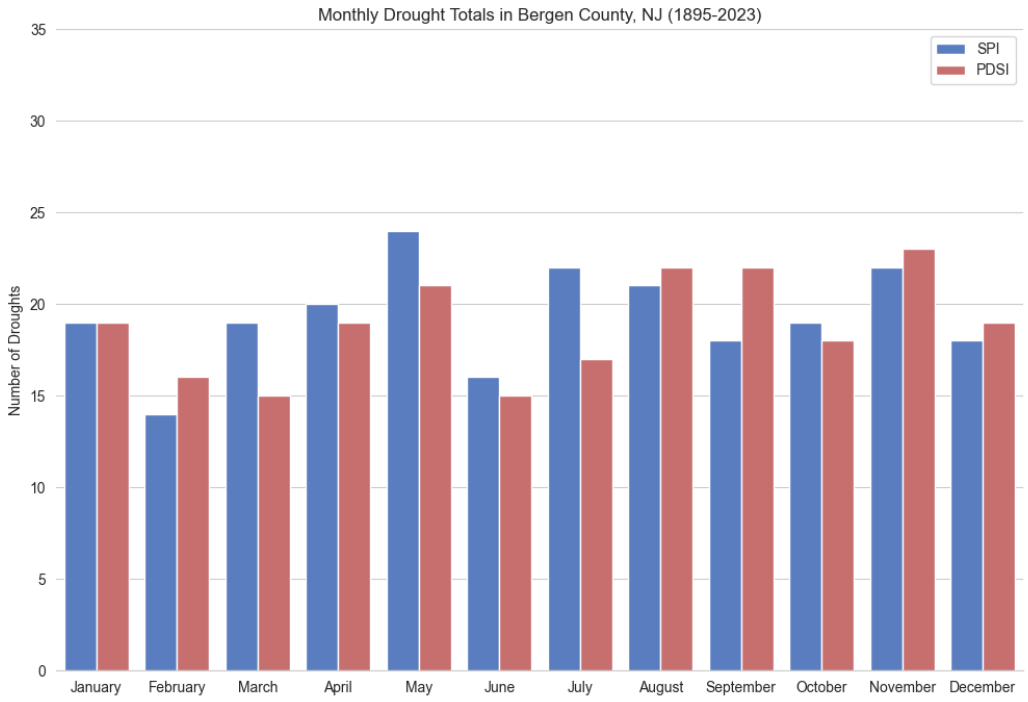

Figure 5: Number of Droughts Recorded Each Month in Bergen County, NJ. This plot shows the number of times SPI and PDSI have recognized a drought during each month throughout the dataset. The largest difference in total drought identifications between each index is 4 total droughts during the months of March and September.

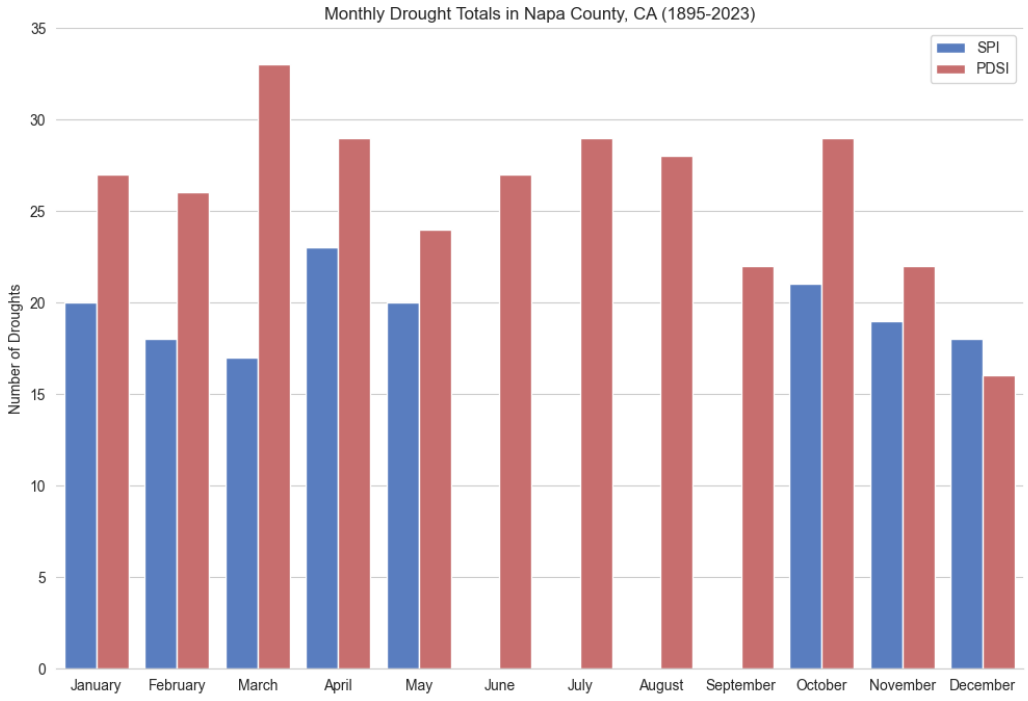

Figure 6: Number of Droughts Recorded Each Month in Napa County, CA. This plot shows the monthly drought totals for each index in Napa County. Similarly to the Bergen County data, PDSI records at least 16 droughts in each month. However, SPI did not identify any droughts that occurred from June through September months.

Since those summer months in Napa County hardly gain precipitation, SPI fails to detect any significant departure from the months' long term average precipitation. Therefore, SPI does not show any droughts to have taken place from June to September. PDSI identified 312 droughts from 1895 to 2023 in Napa whereas SPI identified 156 droughts. SPI's inability to identify droughts during the summer likely accounts for the discrepancy between the two. Here we can see how PDSI may be a stronger drought index to detect droughts in areas which have a season that gets little to no precipitation. Since PDSI makes use of other factors besides

precipitation, it can identify drought as occurring in months that typically do not get much rain or snow.

When looking at the months in Bergen County, SPI and PDSI recorded a number of droughts no greater than 4 droughts apart from each other. Both indices seem to have no problem dealing with the even distribution of precipitation that occurs throughout the year in New Jersey. PDSI identified 226 droughts in Bergen and SPI identified 232 droughts. Another way to see how SPI and PDSI compare to each other is by looking at their yearly distribution.

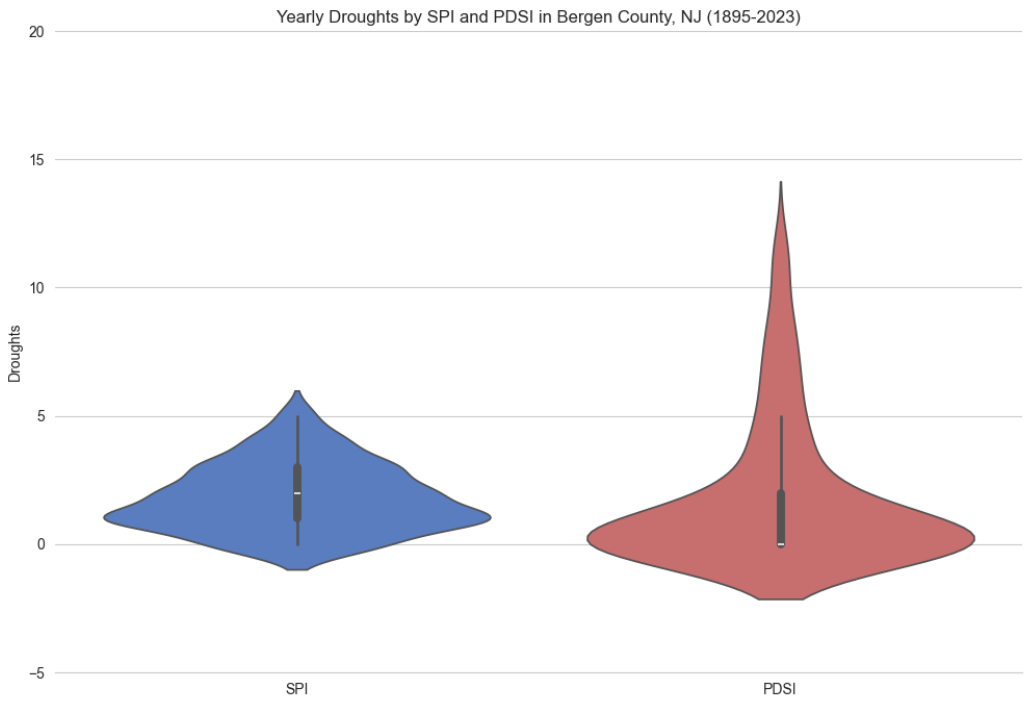

Figure 7: Violin Plot of Yearly Droughts in Bergen County, NJ. This violin plot shows the distribution of droughts identified in Bergen County each year by SPI and PDSI. PDSI has

multiple years where 10 or more droughts were identified whereas SPI never identified more than 6 droughts in a single year.

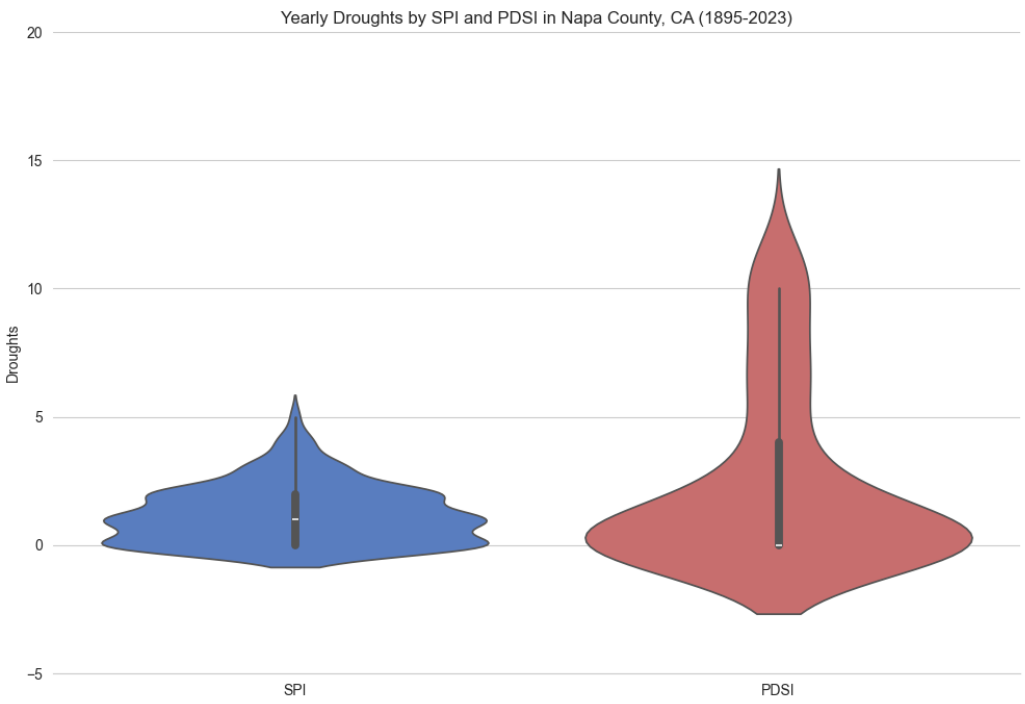

Figure 8: Violin Plot of Yearly Droughts in Napa County, CA: This plot shows the distribution of droughts that each index classified in Napa in each year of the dataset. SPI has a balanced distribution of drought identifications versus PDSI which has a skewed distribution.

In both locations, PDSI had a median of 0 yearly droughts. In Bergen County, SPI had a median of 2 and in Napa County, SPI had a median of 1. PDSI had 11 years in Napa County and 5 years in Bergen County where it identified 10 or more droughts in a year. SPI did not identify more than 5 droughts in any year in either location. PDSI identified 0 droughts in 70 years of the Bergen dataset and 66 years of the Napa dataset. SPI identified 0 droughts in 19 years of the

Bergen dataset and 42 years of the Napa dataset. As shown, both indices have their positives and negatives and you have to be considerate about when is the best time to use each.

While looking at drought in both locations, it is important to see if there is a trend to the number of droughts occurring each year. Time series analysis of drought data will show how drought occurrences may be changing over the course of the dataset.

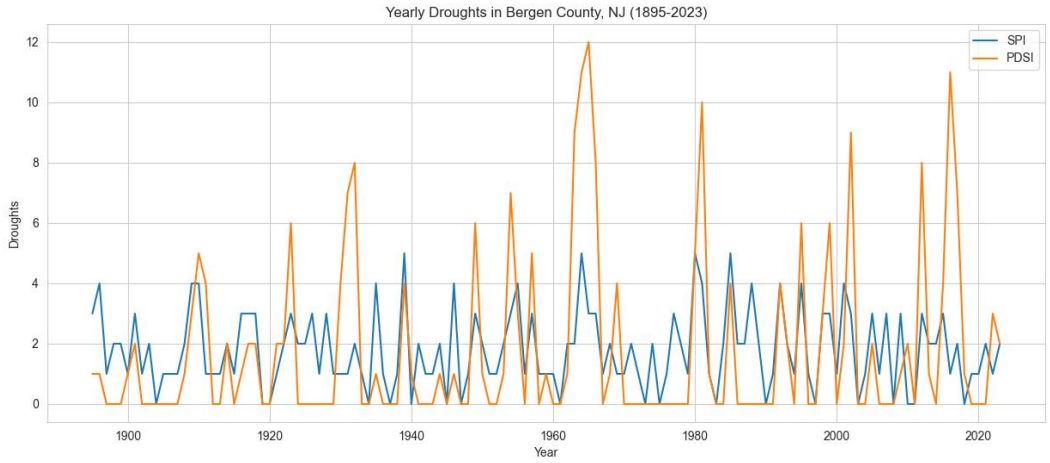

Figure 9: Line Plot of Droughts Identified by SPI and PDSI in Bergen County, NJ: This plot shows how many droughts each index identified throughout the years in Bergen County.

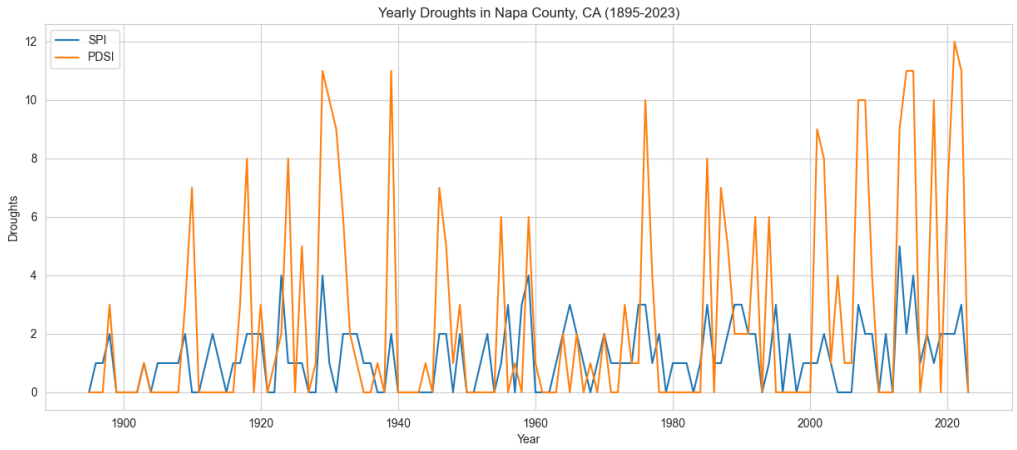

Figure 10: Line Plot of Droughts Identified by SPI and PDSI in Napa County, CA: This plot shows the yearly drought totals in Napa County. PDSI has increased in the number of droughts identified starting in 2000.

Mann-Kendall tests were run on the SPI and PDSI data for both Bergen and Napa County to identify if there are trends in these datasets with a significance value of 0.05. For Bergen County, SPI had a p-value of 0.906 and PDSI had a p-value of 0.343. Therefore, the null hypothesis was not rejected in both cases. There is no significant trend for SPI and PDSI in Bergen County. For Napa County, SPI had a p-value of 0.006 and PDSI had a p-value of 0.005. The null hypothesis was rejected in both cases and there are trends in Napa's drought datasets. For Napa, SPI had a z-score of 2.74 and PDSI had a z-score of 2.77. This indicates that both drought indices agree that there is an increasing trend of droughts in Napa County.

Time Series Analysis

There was no significant trend for precipitation either Bergen County or Napa County. There was a clear seasonality to the Napa precipitation data and a possible seasonality for Bergen County. The first thing we need to do is isolate and analyze the seasonality of the data. In order to do that, we will first look at autocorrelation plots of the monthly precipitation data to see if there are clear signs of seasonality. Autocorrelation functions can tell you if certain values in a data set influence previous or nearby values. They create autocorrelation coefficients, r_l where l is the lag point that the coefficient represents. The value of a coefficient at lag, l, represent the strength of the relationship between a point in the dataset and another point l spaces away in the dataset. Now, we can look at an autocorrelation plot for Bergen County.

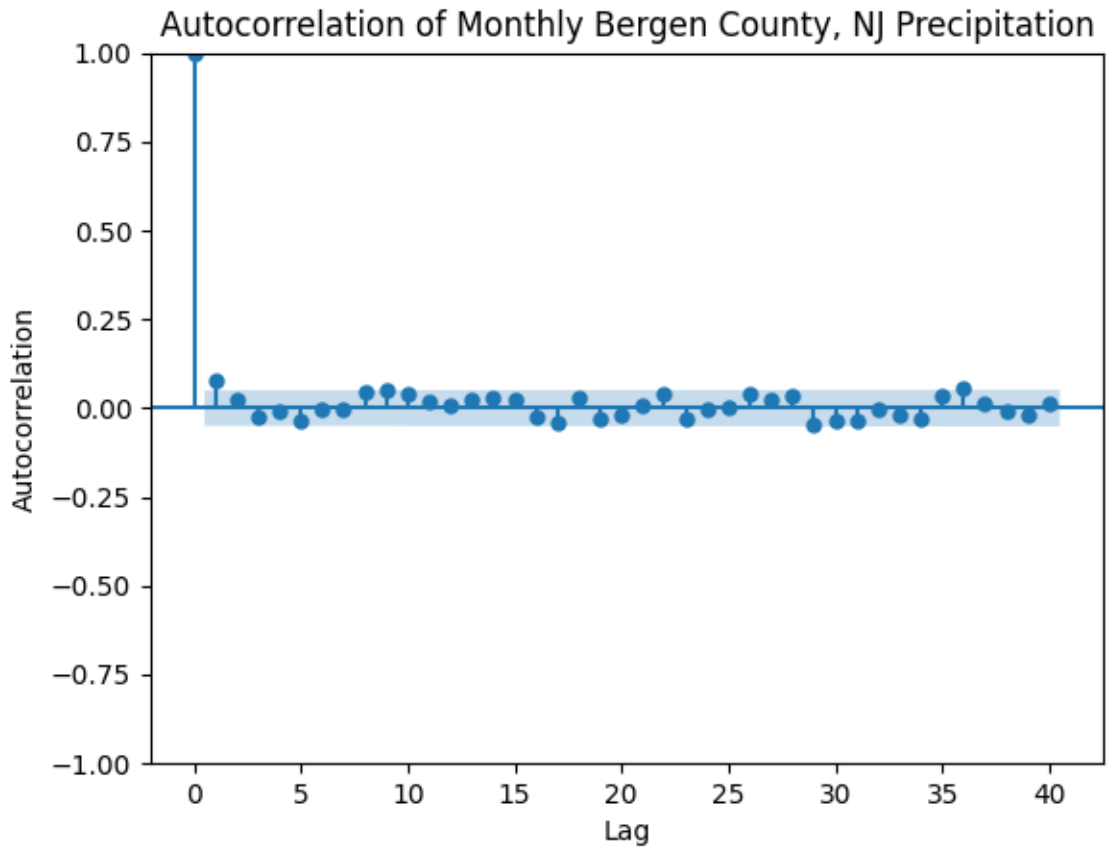

Figure 11: Autocorrelation plot of Bergen County, NJ Precipitation. The plot does not show any significant lag points. All the bars after lag 2 are close to 0 and within the 95% confidence interval represented by the shaded region.

The autocorrelation plot for Bergen County appears to represent a lack of significant seasonality in the data set. We saw this when looking at the box plot of Bergen's precipitation by month. Because precipitation is spread mostly evenly throughout the year, knowing how much precipitation there was in an unknown month would not help you to guess what month's data you are looking at. The shaded blue region in Figure 9 represents the confidence interval for the plot. Since the majority of the points fall within that confidence interval, we can be sure that

seasonality is not a significant factor in the Bergen County dataset. However, it is likely that will not be the case for Napa County.

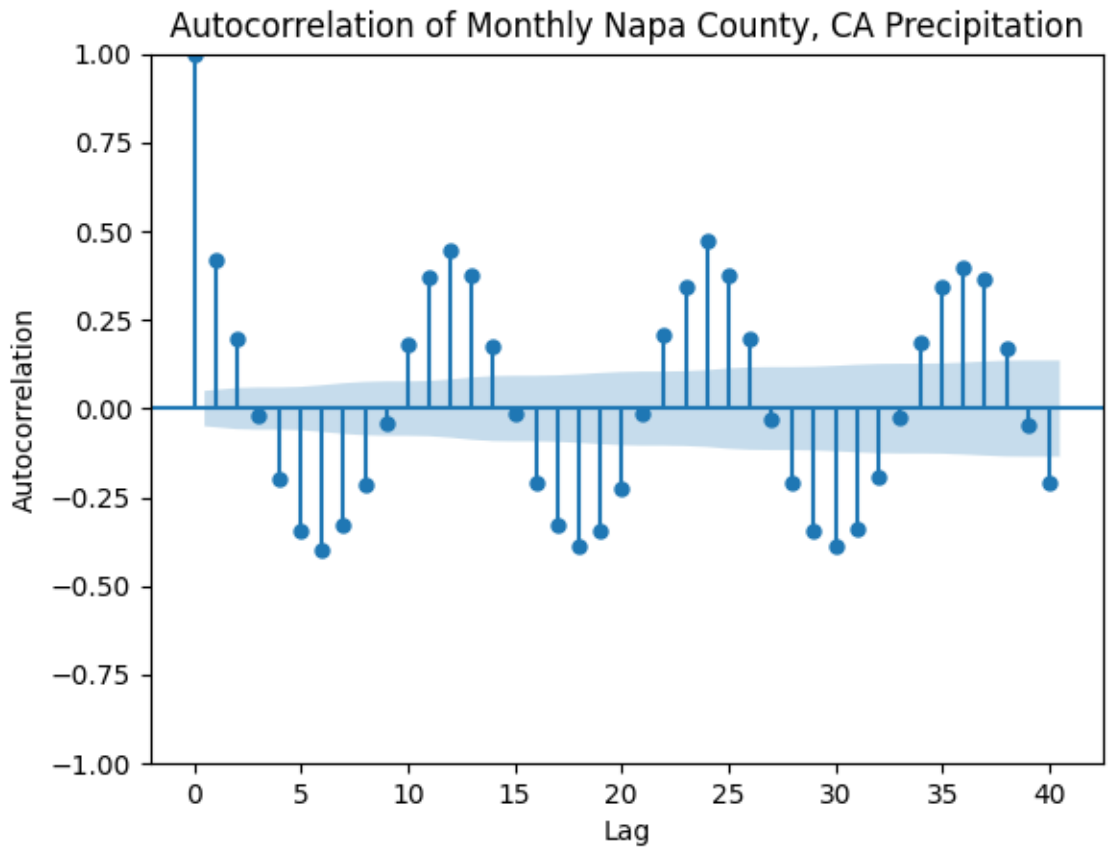

Figure 12: Autocorrelation plot for Napa County, CA Precipitation. The plot shows bars reaching 0.5 at lags 12, 24, 36, etc. and -0.5 at 6, 18, 30, etc. This suggests that monthly precipitation values are positively correlated with the same month's values in the previous year. The negative values suggest that each month's values are negatively related to values from 6 months ago.

This plot presents a very strong case for Napa's precipitation to have seasonality. The coefficients peak at every 12 lags. This is saying that precipitation in January is very likely to be

related to the precipitation in the previous January. Now we know that the Napa data set follows a typical 12-month seasonal pattern. Monthly precipitation will be similar to the values of the same month in previous years. The coefficients also dip at a lag of 6 months and then every 12 months after. This makes perfect sense when you consider the drastic difference in precipitation between summer months and winter months.

Now we know that Napa County has a strong seasonal component to its precipitation. The next thing we will do is visualize the seasonality aspect of it so we can analyze it. Even though Bergen County's seasonality was not apparent, it is still worth visualizing as well. To do that, we have to find the trend of the dataset. For our data set, we will take a moving average in order to gauge the trend. We know the seasonal frequency of our data is 12 months so that will be the length of our moving average. We want to work on integer times, so we will use a centered moving average to accomplish this. With the following equation,

$$\widehat{T_t} = \frac{(Y_{t-6}+\ldots+Y_{t+5})/12 + (Y_{t-5}+\ldots+Y_{t+6})/12}{2},$$

where $t = 7, \ldots, T-6$ will give us the trend for our dataset [6]. Next, we subtract the trend from the actual precipitation data to receive the detrended data. Lastly, to abstract the seasonality aspect of the data set we will group the detrended data by the month and take the average for each month's values. After that we end up with the following plots.

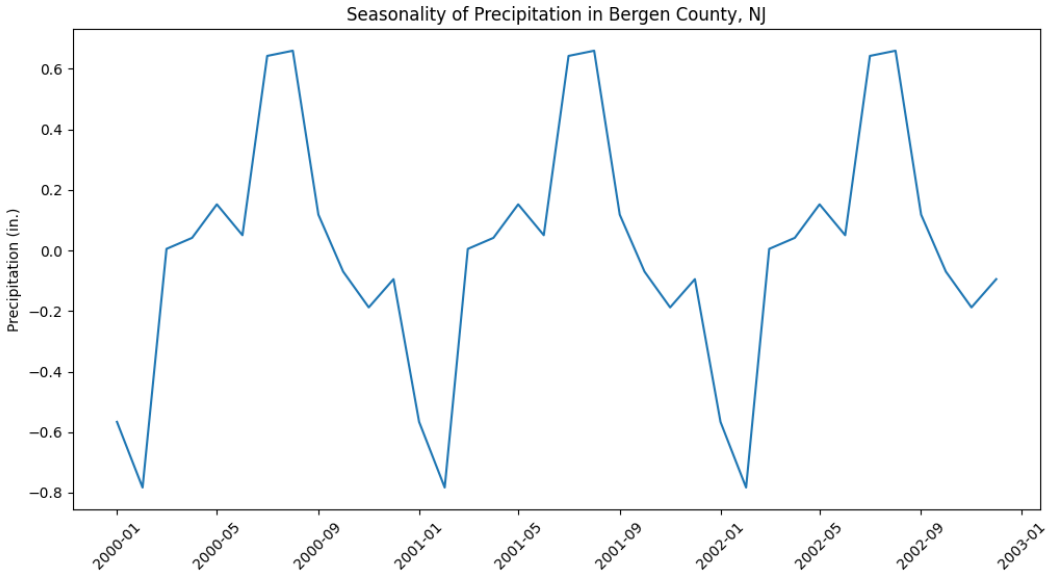

Figure 13: Seasonality of Precipitation in Bergen County, NJ. This plot is the seasonal pattern for Bergen County. The pattern is not smooth and is smaller in scale as it only goes from 0.6 to -0.8. The pattern takes a sharp dip at the beginning of each year and peaks around July.

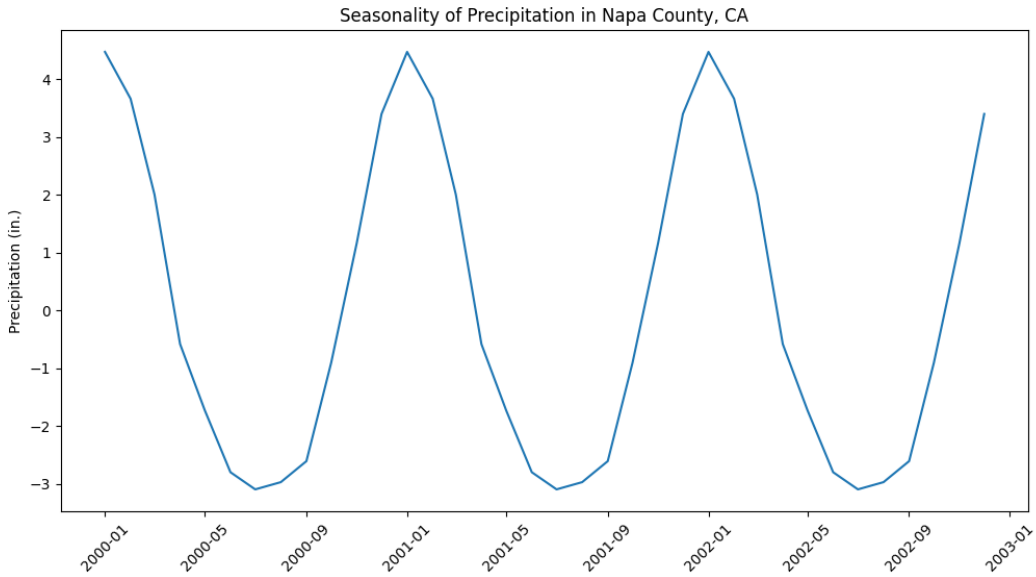

Figure 14: Seasonality of Precipitation in Napa County, CA: This plot shows the seasonal pattern for Napa County. It is a smooth and prominent continuous wave that represents what Napa's seasonal pattern looks like. It peaks during winter months and dips during the summer.

Both Bergen County and Napa County's precipitation data show repeated patterns representing seasonality in both datasets. Figure 10 shows the pattern dip during the summer months and peak during winter, which matches perfectly with the distribution we saw in Figure 2. However, Bergen County's pattern peaks during the summer and has a sharp decline during the beginning of the year. We can also see that the seasonality is smaller in scale for Bergen County than Napa County. This is likely due to Bergen County having an insignificant seasonal component compared to Napa.

In order to see how well each dataset can be used for prediction models, we will use modeling that incorporates the seasonality of the data. One of the ways we can do that is by combining the trend of the data with an inclusion of its seasonal component. We can do this in a least squares regression estimation. In a typical least squares regression, you would fit the data using the sum of a constant and linear component. However, the precipitation data has a seasonal element to it. An ordinary trend line will likely miss the mark on forecasting the data. We can add a seasonal component to the least squares instead to compensate for the periodic elements of our data set.

Our new least squares estimate will look like: $\hat{y}^{d} = \hat{y}^{lin} + \hat{y}^{seas}$ where

$$\hat{y}^{\text{lin}} = \theta_1 \begin{bmatrix} 1 \\ 2 \\ \vdots \\ N \end{bmatrix}, \qquad \hat{y}^{\text{seas}} = \begin{bmatrix} \theta_{2:(P+1)} \\ \theta_{2:(P+1)} \\ \vdots \\ \theta_{2:(P+1)} \end{bmatrix}$$

[7].

The seasonal component has period P = 12 since we know that is the seasonal frequency of our data set. It has a pattern of $\theta_2, \ldots, \theta_{P+1}$ that is repeated $\frac{N}{P}$ times where N is the length of the time series [7]. This repeating seasonal part adds to the linear trend to keep the prediction consistent with the actual seasonal trend of the data set. Therefore, the least squares model should more accurately forecast the data. To compute the least squares fit, we have to minimize $\left||A\theta - y^{d}|\right|^{2}$ where θ is a $(P + 1)$ vector and where A is an N×(P+1) matrix that consists of a first column of 1 through N concatenated with $\frac{N}{P}$ P×P identity matrices [7]. This will ensure that each element of the seasonal pattern (in our case the months of the year) is weighted into the final least squares solution. In other words, the predictions consist of the linear trend adjusted by an offset for each month. The predicted value, $\hat{y}(n)$, in month n $(n = 1, \ldots, N)$ is

$$\hat{y}(n) = \theta_1 n + \theta_{n \bmod 12 + 1}$$

$\theta_1 n$ is the linear trend of the dataset and $\theta_2, \ldots, \theta_{13}$ are the monthly offsets corresponding to January through December.

We ran this least squares estimation for both the Bergen County and Napa County data and obtained the following results.

Parameter	**Description**	**Bergen County** **Least Squares Model**	**Napa County** **Least Squares Model**
$\boldsymbol{\theta_1}$	Linear trend component	0.024	-0.001
$\boldsymbol{\theta_2}$	January offset	1.845	5.059
$\boldsymbol{\theta_3}$	February offset	1.733	7.212
$\boldsymbol{\theta_4}$	March offset	3.500	2.731
$\boldsymbol{\theta_5}$	April offset	2.880	2.140
$\boldsymbol{\theta_6}$	May offset	3.478	0.909
$\boldsymbol{\theta_7}$	June offset	4.928	0.118
$\boldsymbol{\theta_8}$	July offset	3.714	0.029

$\boldsymbol{\theta_9}$	August offset	3.907	0.074
$\boldsymbol{\theta_{10}}$	September offset	5.257	0.151
$\boldsymbol{\theta_{11}}$	October offset	2.163	2.006
$\boldsymbol{\theta_{12}}$	November offset	2.985	4.163
$\boldsymbol{\theta_{13}}$	December offset	3.278	11.146

Table 1: Least Squares Parameter Description Table: This table includes the parameters and their descriptions and values for the least squares model of Bergen County and Napa County's precipitation.

A way to see how these parameters work is to use the equation to predict one of the values in the dataset. When looking at Bergen County from January 2000 to December 2004, May 2002's precipitation can be predicted by using the May offset and the equation for $\hat{y}(n)$. In this case, $n = 29$, $\theta_{29 \, mod \, 12 + 1} = \theta_6 = 3.478$, and $\theta_1 = 0.024$. Therefore, $\hat{y}(n) = 4.174$ inches, which is close to the actual value of 4.92 inches. The following plots showcase the least squares model's precipitation predictions from January 2000 to December 2004.

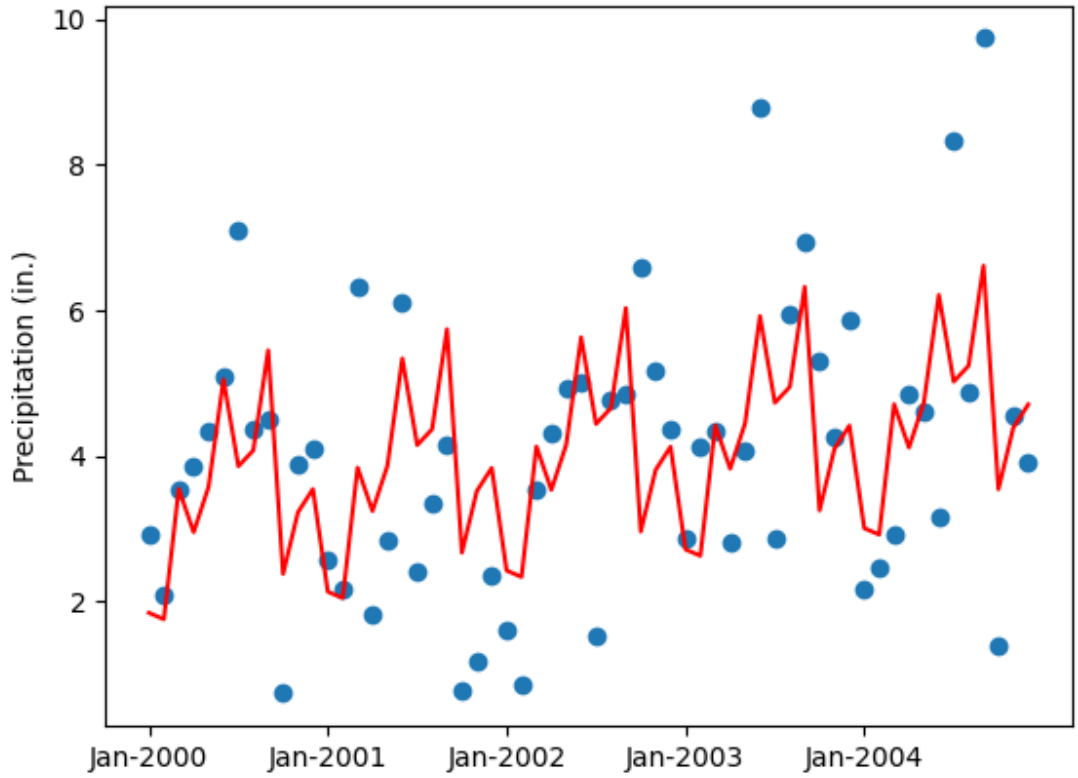

Figure 15: Least Squares Regression w/ Seasonal Component for Bergen County, NJ Precipitation. The red line is the least squares fit of the data and the blue dots represent the actual precipitation data. The fit exhibits a periodic trend to capture the seasonality of the data.

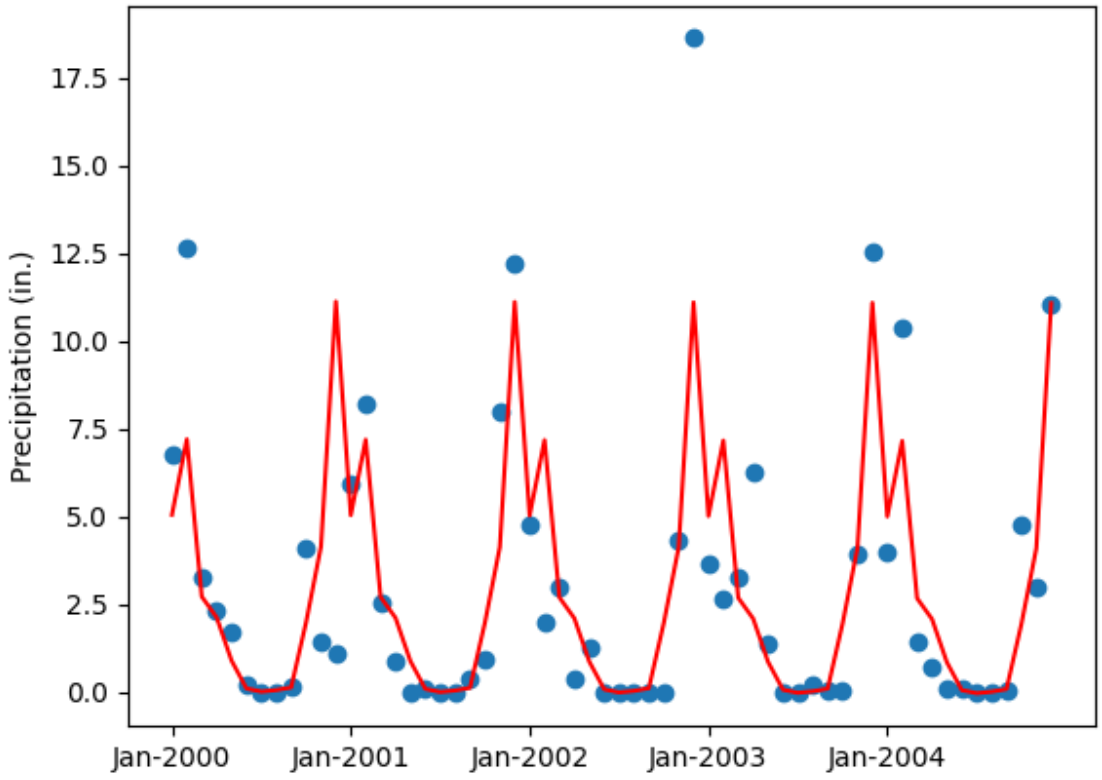

Figure 16: Least Squares Regression w/ Seasonal Component for Napa County, CA Precipitation. The blue dots are the actual data points, and the red line is the prediction that the model is making. The least squares fit tightly follows the seasonal pattern of the precipitation.

We can see how Napa's distinct seasonal pattern lends itself well to the least squares model's ability to forecast precipitation. Since Bergen County does not have as prominent of a seasonality to its dataset, its least squares model struggles to create a cohesive trend that accurately shows off its precipitation trends.

The least squares regression does incorporate the data's overall trend into its equation. Since the seasonal pattern is stronger than any potential trend with the data, we decided to look at

another model that would play to those strengths. Therefore, in order to attempt a more accurate prediction model, we decided to use a Seasonal Autoregressive Integrated Moving Average model, known as SARIMA.

SARIMA

Seasonality of a time series represents a repeating pattern that repeats over a certain amount of time. For monthly data, that time period tends to be 12 months. Based on Figure 10, we have a dataset that has a 12-month repeating pattern. SARIMA models have 7 parameters that are shown in the format of SARIMA$(p, d, q)(P, D, Q)s$, where p is the order of non-seasonal autoregression, d is the non-seasonal differences, and q is the order of non-seasonal moving average. P, D, Q are the seasonal counterparts to the previously stated parameters and s is the length of the seasonal pattern [8]. In order to create a SARIMA model, we need to assign values to each of these parameters. From earlier, we already have the seasonal length of 12. For non-seasonal differencing, d, there is the Augmented Dickey-Fuller test which is a hypothesis test to see if the data set is stationary. We can run the Augmented Dickey-Fuller test in python on both time series and we get p-values equal to zero for both. We can reject the null hypotheses and assume both data sets are stationary. Since it is stationary, we do not need to use differencing for the non-seasonal aspect of the time series and we can set d to 0. We can use the autoregression plots from earlier to help with some of the other parameters. We can look at how far and high the initial spikes of the autocorrelation plots go. We can see in the Bergen County plot, they only go to 1 lag and in the Napa plot, they go to the first 3 lags. Therefore, we can choose 1 through 3 for the autoregression and moving average parameters.

We did trial and error with various combinations of parameters to try and find the best fit. We ended up with ARIMA (2, 0, 0) x (2, 1, 0)12. Both time series had their lowest Akaike Information Criterion (AIC) scores with those parameters.

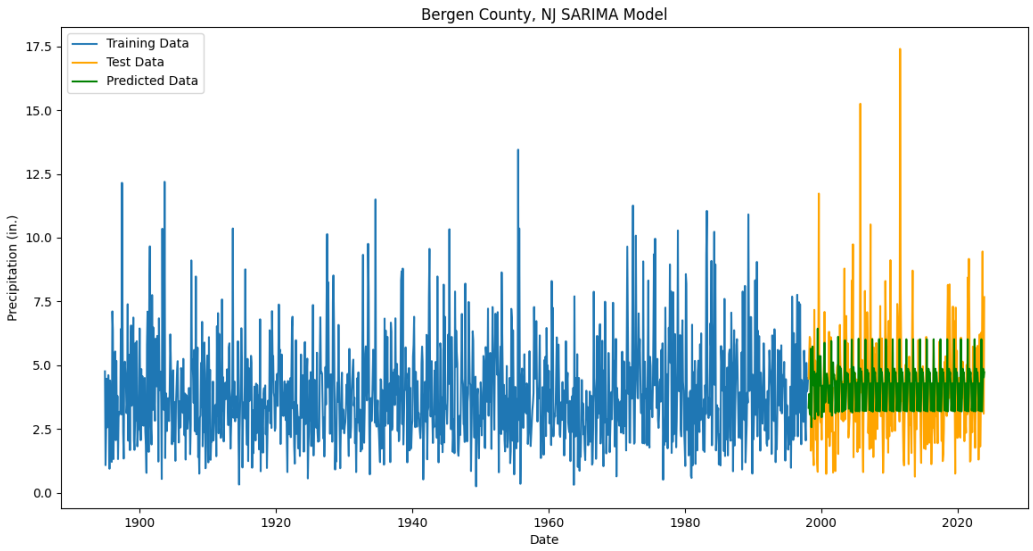

Figure 17: SARIMA Model of Precipitation in Bergen County, NJ. The above plot shows the SARIMA model's prediction for Bergen County. The green represents the forecast from the model and the yellow is the test data. The forecasted data mimics the up and down nature of the original data.

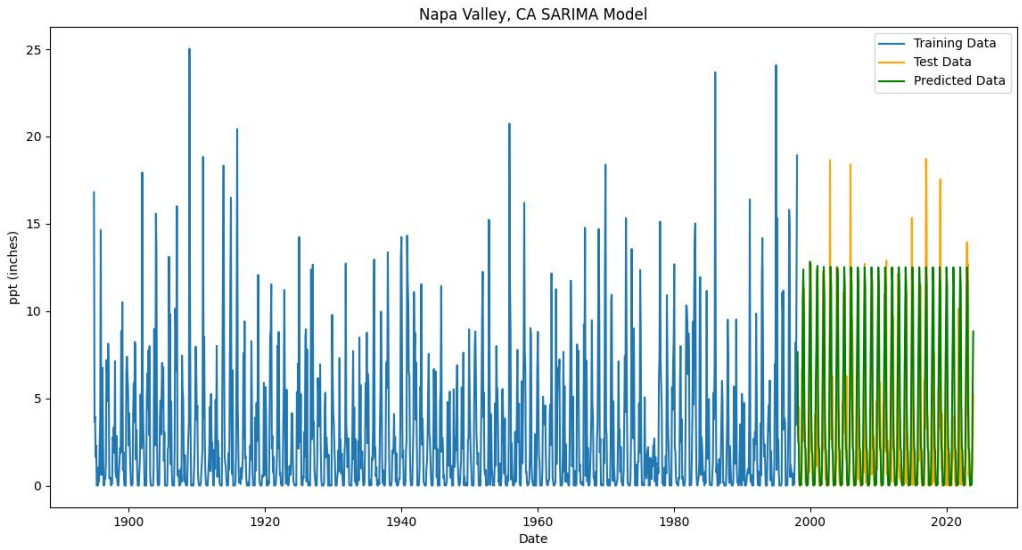

Figure 18: SARIMA Model of Precipitation in Napa County, CA. The Napa County predictions closely depicts the actual test data. It covers a larger range of the data than the Bergen County model.

The Napa SARIMA model can accurately represent the big swings between winter and summer precipitation. The Napa model had a root mean square error of 3.6 inches. The Bergen County SARIMA model stays in the small range of values that cover a typical New Jersey year, but it also swings back and forth to account for the sudden changes in precipitation that occur year to year as shown in Figures 3 and 4. The Bergen County model had a root mean square error of 2.1 inches.

Bergen County's shorter range of typical precipitation allows its predictions to have minimal distance from the actual data. Even though Bergen's SARIMA model did not follow the wavelength pattern as well as Napa's predictor did, it was able to stay tightly within the typical precipitation numbers to make an adequate prediction. However, Napa's larger fluctuation in

precipitation between seasons lent itself well to a model that will more sharply replicate that fluctuating seasonal pattern. The Napa SARIMA model succeeds in painting a better picture of what the Napa seasons look like than the SARIMA model of Bergen County was able to.

Analysis and Discussion

We were able to make out prominent strengths and weaknesses of two of the most highly used drought indices. Although SPI was made to be able to compare values across different climate locations, it fails to properly identify droughts when precipitation is near 0, historically. It is superior in identifying drought on a shorter time scale, but when trying to focus on consistent drought it fails. No matter how many months are in drought, once summer begins, SPI will no longer consider Napa County in drought. Even though those months typically receive little to no precipitation, it does not mean they are not affected by drought. If you were only to look at SPI values as an indication of drought for a location like Napa County, you would be negligent to drought preparation during the summer.

In this regard, PDSI does a much better job. We can see how its reliance on factors other than historic precipitation data helps it identify droughts that occur in vastly different climates. PDSI is also the opposite of SPI in its ability to identify long term droughts. PDSI did not identify drought in 66 years of Napa County's data and 70 years of Bergen County's data. It mostly starts to identify droughts when it can also identify droughts in several months close to each other. It is important to be mindful of the drought index you are using depending on the location you are using it on. With each having different strengths, it becomes important to consider what kind of drought you are looking for.

There is an increasing trend in droughts in Napa Valley. Despite the differences between both indices, they each showed signs of that trend. Even though SPI can not identify droughts during Napa's summer, the index can still pick up overall trends in that area. It is important to

recognize when there is an increasing trend of droughts so that the location that is feeling the drought's effects can be prepared for potentially more frequent droughts in the future.

	Bergen County, NJ	**Napa County, CA**
Standard Deviation	1.99 inches	3.85 inches
Least Squares RMSE	1.52 inches	2.34 inches
SARIMA RMSE	2.29 inches	3.98 inches

Figure 21: Prediction Model Performance Table: This table displays the standard deviation and root mean square error of the prediction models for both Napa County and Bergen County.

The least squares models performed best out of the models tested in both locations. For Bergen County, the model had an RMSE of 1.52 inches and an R-squared of 0.37. The Napa County model had an RMSE of 2.34 inches and an R-squared of 0.66. The repeating seasonal aspect of the least squares model works best with data that has a consistent seasonal pattern of its own. The reason this model does not perform better on these data sets is because precipitation changes so drastically each year. Even Napa, where they get very consistently low amounts of water during the summer, has large fluctuations between years. As you can see in Figures 1 and 2, the range of precipitation some of these months get makes it extremely difficult for any tool to predict what will happen in the future. You would need a much more flexible model to improve on these ones.

The SARIMA model was able to forecast weather patterns for each location it was testing on, but it is not able to predict extreme weather events. The Bergen County models beat out the Napa models in root mean squared error. This is likely due to Bergen's precipitation having a smaller range in its total output than Napa. During the winter, some months in Napa can receive as much as 20 inches of precipitation. Fifteen inches of precipitation in December, January, or February would not be considered an outlier in Napa County. However, in Bergen, that would be an outlier for any month of the year. Ten inches of precipitation would be an outlier during any month in Bergen County. Therefore, a model like SARIMA has a smaller range of values to predict for Bergen County whereas Napa's values fluctuate a lot more. Still, we were able to see in the visualizations that both models created a clearer picture of Napa's seasonal pattern than that of Bergen County. Overall, the least squares model's predictions were closest to the actual dataset's monthly precipitation values. The SARIMA model performed the worse in predicting accurate precipitation values for Bergen County and Napa County. It seems we must accept that unpredictable precipitation phenomena will have to stay that way. However, there seems to be great benefit in understanding the weather patterns of locations with differing weather patterns. They can tell a lot about which tools may be most useful in its analysis.

Conclusions

Future work could make these same comparisons with a larger variety of climate distinct locations. It is important to gain a better understanding of how different kinds of weather patterns can be potentially predicted and how they can affect drought classifications. Future work could also include an analysis of large-scale climate patterns such as El Niño and their effect on locations with differing weather patterns. A good next step would be to look into the reasons for the increasing number of droughts in California and figure out ways to help mitigate their effects.

This study was able to show how different drought indices classify drought in distinct climate regions. Neither index served as a completely satisfying approach to identifying drought. They have their weaknesses that become apparent depending on the climate of the region that they are being used on. PDSI can predict drought in regions that have seasonal patterns that produce months which receive little to no precipitation. It is also stronger at identifying long-term drought. SPI completely fails to identify drought when precipitation numbers are historically and consistently close to 0. Droughts in Napa County have developed an increasing trend over the 128 years of the dataset.

This project showed how different climate regions' precipitation can be predicted using various time series models. Least squares models can incorporate the seasonal aspect of a location's climate data and map it to make future predictions. It cannot predict when extreme weather events will occur in any location. More complex models such as SARIMA can predict a variety of weather patterns as well. SARIMA can mimic a location's seasonal weather pattern when that location has strong seasonality. It will make a more confined and safer prediction when looking at a location that does not have a distinct seasonal pattern.

References

[1] “A Review of Drought Indices.” *International Journal of Constructive Research in Civil Engineering*, vol. 3, no. 4, 2017, https://doi.org/10.20431/2454-8693.0304005. Accessed 15 May 2020.

[2] Dai, Aiguo & National Center for Atmospheric Research Staff (Eds). Last modified 2023-08-19 "The Climate Data Guide: Palmer Drought Severity Index (PDSI).” Retrieved from https://climatedataguide.ucar.edu/climate-data/palmer-drought-severity-index-pdsi on 2024-04-02.

[3] Daly, Christopher & National Center for Atmospheric Research Staff (Eds). Last modified 2023-08-09 "The Climate Data Guide: PRISM High-Resolution Spatial Climate Data for the United States: Max/min temp, dewpoint, precipitation.” Retrieved From https://climatedataguide.ucar.edu/climate-data/prism-high-resolution-spatial-climate-data-united-states-maxmin-temp-dewpoint on 2024-04-02.

[4] National Drought Mitigation Center (2018). SPI Generator [software]. University of Nebraska–Lincoln. https://drought.unl.edu/Monitoring/SPI/SPIProgram.aspx

[5] Ravi Shah, Nitin Bharadiya, Vivek Manekar, Drought Index Computation Using Standardized Precipitation Index (SPI) Method For Surat District, Gujarat, Aquatic Procedia, Volume 4, 2015, Pages 1243-1249, ISSN 2214-241X, https://doi.org/10.1016/j.aqpro.2015.02.162.

[6] Buteikis, Andrius. “Time series with trend and seasonality components.” Vilnius University.

https://web.vu.lt/mif/a.buteikis/wp-content/uploads/2019/02/Lecture_03.pdf

[7] Boyd, Stephen, and Lieven Vandenberghe. *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares*. Cambridge University Press, 2018.

[8] Shengwei Wang, Juan Feng, Gang Liu, Application of seasonal time series model in the precipitation forecast, Mathematical and Computer Modelling, Volume 58, Issues 3–4, 2013, Pages 677-683, ISSN 0895-7177, https://doi.org/10.1016/j.mcm.2011.10.034.

[9] Martínez-Acosta, L.; Medrano-Barboza, J.P.; López-Ramos, Á.; Remolina López, J.F.; López-Lambraño, Á.A. SARIMA Approach to Generating Synthetic Monthly Rainfall in the Sinú River Watershed in Colombia. *Atmosphere* **2020**, *11*, 602. https://doi.org/10.3390/atmos11060602

[10] Oliveira Ewerton Cristhian Lima de, Nogueira Neto Antonio Vasconcelos, Santos Ana Paula Paes dos, da Costa Claudia Priscila Wanzeler, Freitas Julio Cezar Gonçalves de, Souza-Filho Pedro Walfir Martins, Rocha Rafael de Lima, Alves Ronnie Cley, Franco Vânia dos Santos, Carvalho Eduardo Costa de, Tedeschi Renata Gonçalves. Precipitation forecasting: from geophysical aspects to machine learning applications. Frontiers in Climate, Volume 5, 2023.

https://www.frontiersin.org/articles/10.3389/fclim.2023.1250201

[11] Stanke, C., Kerac, M., Prudhomme, C., Medlock, J., & Murray, V. (2013). Health effects of drought: a systematic review of the evidence. *PLoS currents*, *5*, ecurrents.dis.7a2cee9e980f91ad7697b570bcc4b004. https://doi.org/10.1371/currents.dis.7a2cee9e980f91ad7697b570bcc4b004

[12] PRISM Climate Group, Oregon State University, https://prism.oregonstate.edu, data created 4 Feb 2024, accessed 16 Feb 2024

[13] Vose, R.S., Applequist, S., Durre, I., Menne, M.J., Williams, C.N., Fenimore, C., Gleason, K., Arndt, D. 2014: Improved Historical Temperature and Precipitation Time Series For U.S. Climate Divisions Journal of Applied Meteorology and Climatology. DOI: http://dx.doi.org/10.1175/JAMC-D-13-0248.1

[14] Gilbert, Richard O. *Statistical Methods for Environmental Pollution Monitoring*. United States.