Applied Statistical Methods

Project 1: Data Manipulation with USHCN Data

Project Overview

In this project we aimed to study if and how the climate has changed over the past century or so. The focus of my analysis is specifically on changes in precipitation as captured by weather stations across the US.

Areas in the US with the highest annual rainfall tend to be along the coasts and in the southeast region, according to the map in Figure 1. With increasing severity of the effects of climate change there are reports that the frequency, quantity and spread of rainfall are all increasing across the US[1]. Although it is hypothesized that the Earth will become subject to a phenomenon known as weather wiplash, in which periods of extreme rain are followed by periods of extreme drought, it is predicted that when it does rain, this rainfall will be more severe[2].

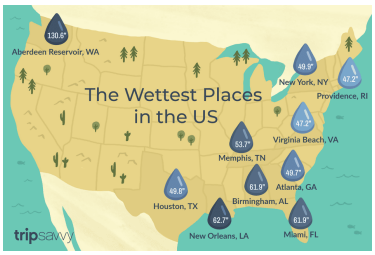


Figure 1. The Wettest Places in the US. (source: https://www.tripsavvy.com/wettest-places-in-the-usa-4135027)

In this analysis, I have set out to determine if and how precipitation has increased from the early 1900s to the 2000s. At the very least this analysis aims to determine if there have been any significant changes or apparent trends in rainfall over the past century.

The data are taken from the NOAA USHCN data archive. Overall, there were 1218 weather stations with precipitation data. These weather stations were spread across the entire US. All analyses were done in Python in a Jupyter Notebook[3], and the major analytical components are the following:

1. Comparison of Precipitation at Two Time Points (Task 1)
2. Precipitation Analysis via Permutation (Task 2)
3. Comparison of Average Rainfall Trends between US Regions

Comparison of Precipitation at Two Time Points (Task 1)

The first analysis carried out was a simple comparison of the average precipitation values (mm) in January 1910 and January 2010 across all weather stations in the dataset. These average precipitation values were split into and classified by seven approximately equally-sized intervals: [0, 98), [98, 196), [196, 294), [294, 392), [392, 490), [490, 588), [588, 687.1). Two separate map plots were then created (Figure 2), one for January 1910 and another for January 2010. Each point on each map plot corresponds to one weather station and the color of this point corresponds to the interval in which the average January precipitation amount for this weather station lies. The map plots were created with the matplotlib “scatter” function. If a weather station was missing either average

[1] USA Today. https://www.usatoday.com/story/news/weather/2019/08/04/climate-change-extreme-weather-getting-worse-in-these-20-places/39873609/

[2] Vox. https://www.vox.com/a/weather-climate-change-us-cities-global-warming

[3] https://github.com/shandu-m/applied-statistical-methods-project1

precipitation value, it was placed into a “missing” category and these weather stations are plotted on the map with a very faint white color.

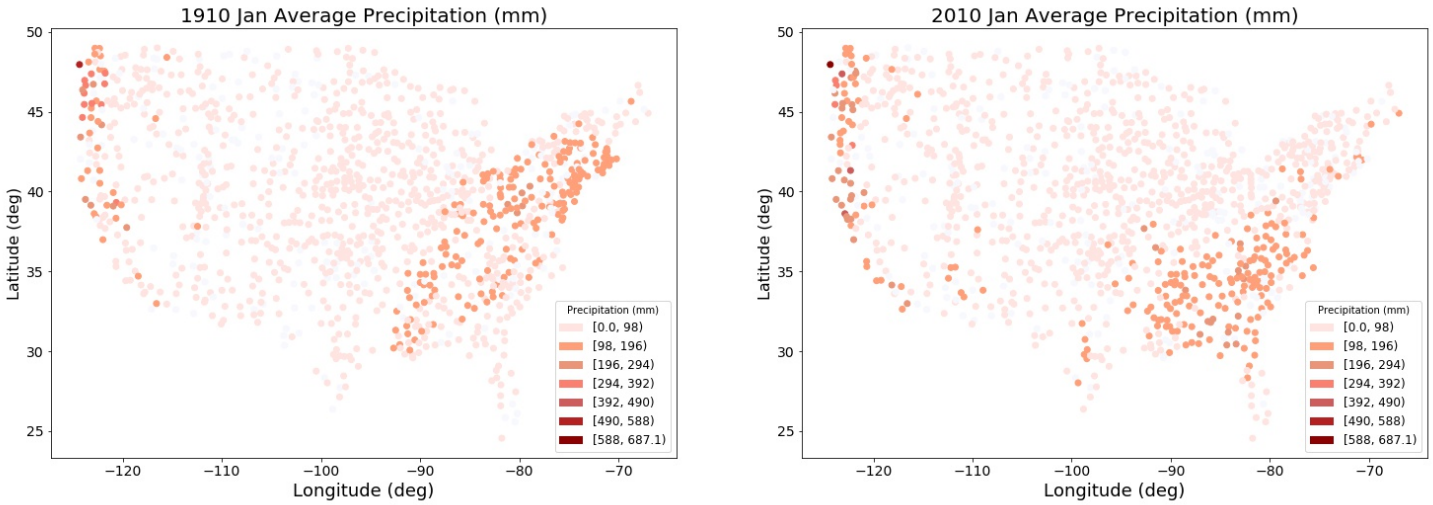


Figure 2. January 1910 and January 2010 average precipitation values (mm) reported by weather stations across the US.

From Figure 2, we see that the majority of weather stations reported average January precipitation values in the [0, 98) range in both 1910 and 2010. The next highest number of weather stations reported average January precipitation values in the [98, 196) range in both 1910 and 2010. Far fewer weather stations have average precipitation values in the higher precipitation categories.

As expected, regions along the coast experienced more rainfall in both January 1910 and January 2010 than inland areas did. We do not, however, see any major increase in volume of precipitation from 1910 to 2010 if only these January precipitation values are considered. Lastly, the eastern patch of heavier rainfall in 1910 appears to have moved more to the south in 2010, which is consistent with the usual changes in weather systems.

Most weather stations reported a ±25 precipitation change from Jan 1910 to Jan 2010

The conclusion that there was no significant change in average January precipitation for most weather stations from 1910 to 2010 can be confirmed by a histogram that displays the spread of precipitation differences between these two time points across all weather stations (Figure 3).

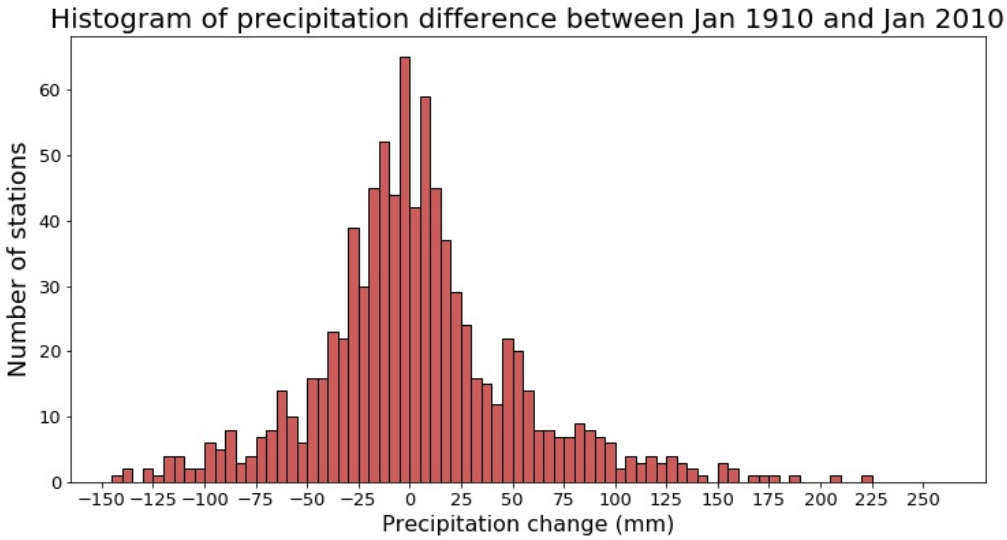


Figure 3. Histogram showing precipitation difference (mm) between January 1910 and January 2010.

The above distribution is approximately normally distributed and centered around 0mm. This indicates that the majority of weather stations reported no significant difference between January 1910 and January 2010 rainfall. Most weather stations reported a ±25 precipitation change from January 1910 to January 2010.

There is however, a slight right-skew to the distribution, which may indicate that a greater proportion of weather stations saw a decrease in precipitation from January 1910 to January 2010.

Most weather stations reported an increase in average precipitation from 1910 to 2010

January 1910 and January 2010 are two *very* specific time points. As a result, we also analyzed the difference in average rainfall over all months from 1910 to 2010 instead, to get a clearer picture of any changes in precipitation trends between these two years at opposite ends of a century.

We analyzed this change in average precipitation from 1910 to 2010 through a histogram which shows the spread of precipitation differences between the 1910 and 2010 average precipitation across all weather stations (Figure 4).

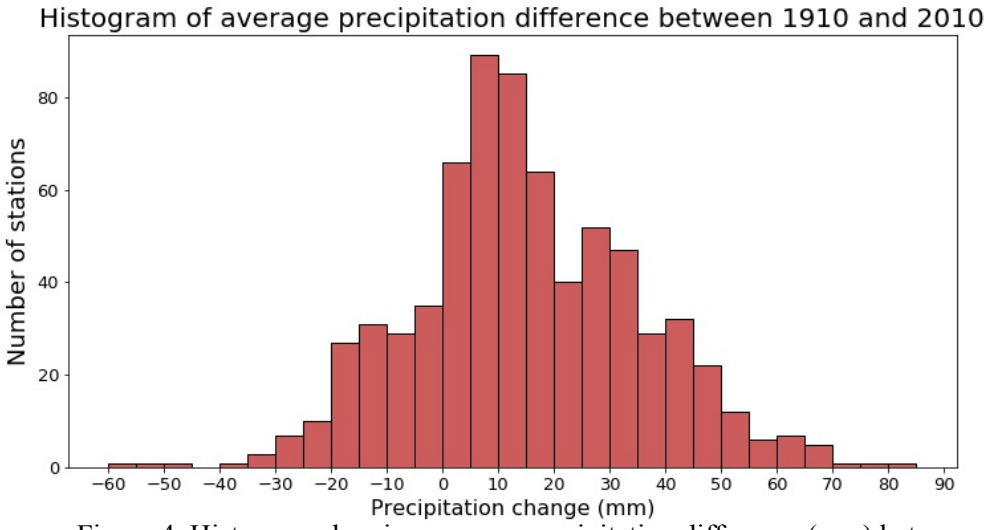


Figure 4. Histogram showing average precipitation difference (mm) between 1910 and 2010.

Like the distribution in Figure 3, this distribution is also quite normally distributed and is less right skewed than the previous distribution. This distribution, however, is centered around 10mm (not 0mm). This indicates that on average, most weather stations saw an increase in average precipitation from 1910 to 2010.

This observation that more weather stations appear to have reported more rainfall in 2010 than in 1910 can be confirmed through another map plot (Figure 5). Each point on each map plot corresponds to one weather station and the color of this point corresponds to the interval in which the average precipitation amount for this weather station lies, for both 1910 and 2010.

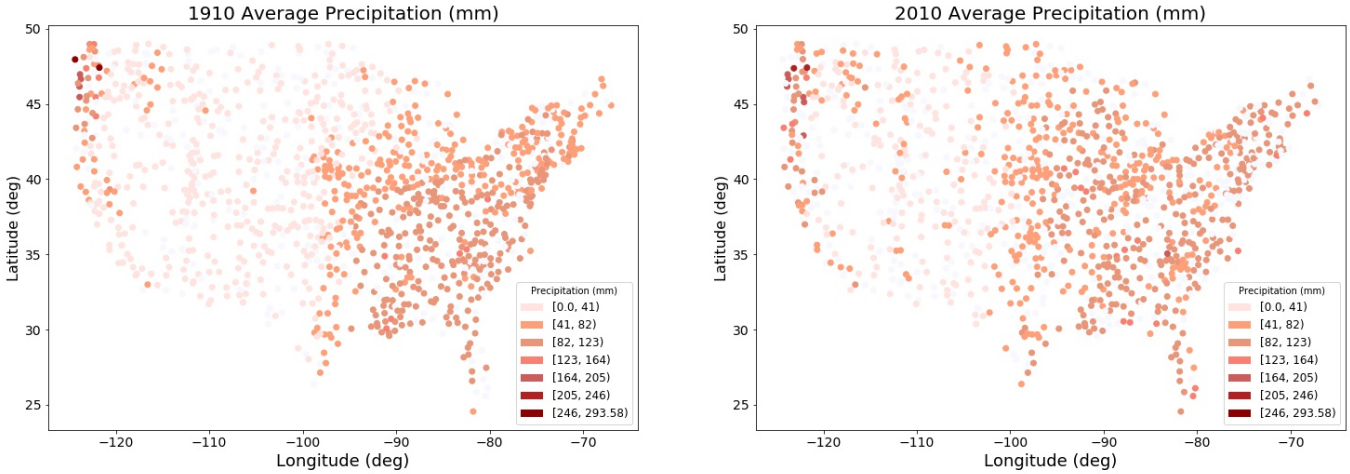


Figure 5. 1910 and 2010 average precipitation values (mm) reported by weather stations across the US.

By studying precipitation in 1910 and 2010 more holistically (looking at average over the year instead of in one month) we see a more observable change. In the 2010 map, the area with average rainfall values that fall in the second interval is now spread out over a much larger number of states. There also appears to be more rainfall inland in 2010 than there was in 1910.

Overall, it appears that in 2010 a greater proportion of weather stations reported higher average rainfall than in 1910. This supports the hypothesis that rainfall is increasing in severity.

Precipitation Analysis via Permutation (Task 2)

To determine if this change in rainfall over the past century or so is significant, we performed a permutation test. The time periods of focus were 1905-1935 and 1985-2015. My focus was on average summer rainfall, that is the average of rainfall that occurs between June and September. Again, we predicted that there would be more rainfall in the later time period than in the earlier one. The p-values from this permutation test are plotted in the following map plot (Figure 6). Each point on the map plot corresponds to the p-value from the permutation test for one weather station and the color of this point corresponds to the interval in which the p-value lies.

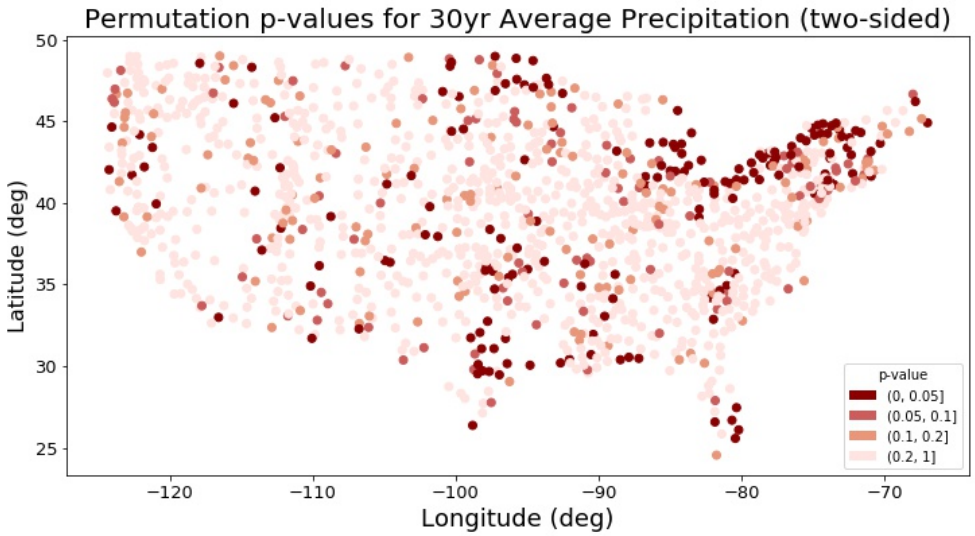


Figure 6. Permutation p-values for 30-year Average Precipitation across the time intervals 1905-1935 and 1985-2015.

From this map plot, it appears that there is a statistically significant difference in average summer rainfall from the 1905-1935 time period to the 1985-2015 time period. The most significant changes in this average summer rainfall occur across the north-easterly states and a few states in the middle of the US.

So, this suggests that there is a statistically significant difference in average summer rainfall between these two time periods across some states, but we do not yet know whether this is an increase or decrease in average summer rainfall. To answer this question, we first plotted a histogram to show the spread of average summer precipitation differences from the 1905-1935 to the 1985-2015 time period (Figure 7). Essentially we plotted the test statistic from our permutation test.

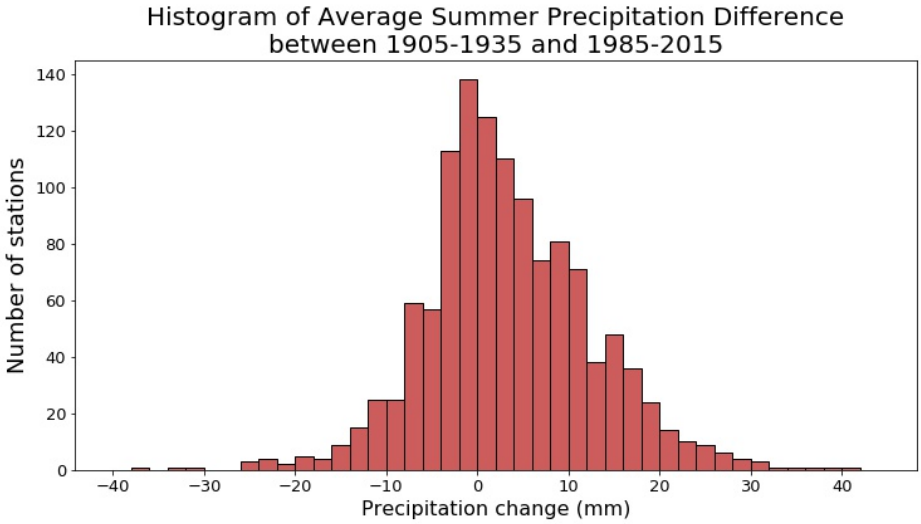


Figure 7. Histogram of average summer precipitation difference from the 1905-1935 time period to the 1985-2015 time period.

This distribution is also relatively normally distributed but has a slight left-skew. This indicates that a slightly greater number of weather stations reported a higher average summer rainfall over the 1985-2015 time period than they reported for the 1905-1935 time period. The majority of weather stations experienced a change in ±10mm of average summer rainfall between these two periods. Thus it does appear that more summer rainfall is being experienced when we compare the averages across these two time periods.

Additionally upper-tail and lower-tail permutation tests were carried out with the same data and in a similar manner to the two-sided permutation test. The p-values from these tests are shown in Figure 8.

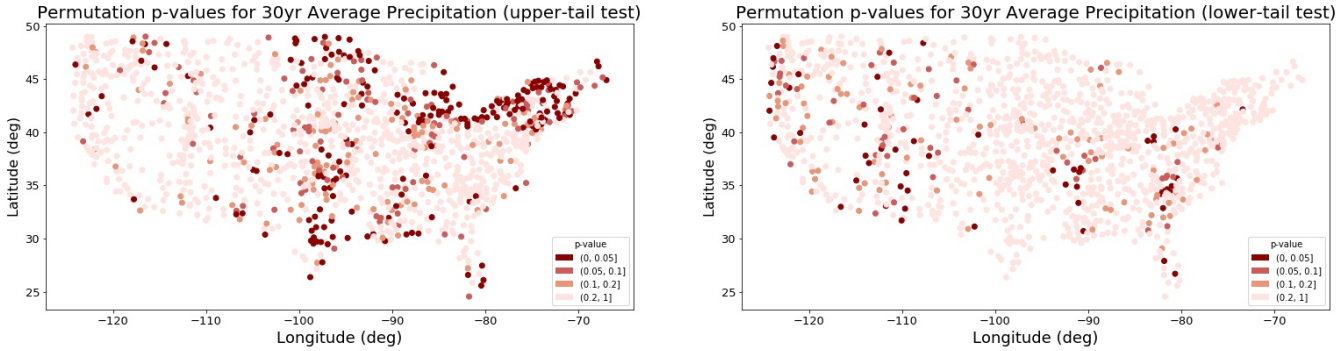
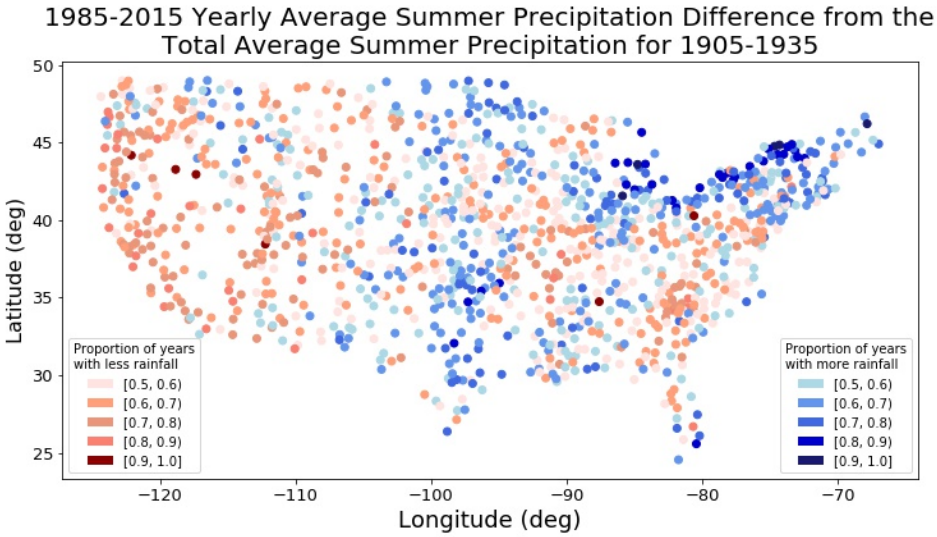


Figure 8. Permutation p-values for the upper-tail and lower-tail tests.

Figure 8 shows that a greater number of weather stations reported a statistically significantly greater average summer rainfall for the 1985-2015 time period than for the 1905-1935 time period. There are fewer weather stations that reported a statistically significantly lower average summer rainfall for the 1985-2015 time period than for the 1905-1935 time period. Overall, this supports our hypothesis that precipitation is becoming more severe.

As the final step in this part of the analysis, we wanted to look more closely at which weather stations experienced this higher level of rainfall in the later time period and to what extent. To do this, we computed the fraction of times the summer precipitation value for a year in the 1985-2015 time period was greater than or less than/equal to the average summer precipitation value for the 1905-1935 time period. These fractions were then plotted on a map plot (Figure 9). Each point in the plot corresponds to a weather station. Weather stations indicated by dots on the blue color spectrum are stations that, when compared to the total average summer precipitation over 1905-1935, reported a *higher* summer average rainfall for more than 50% of years; weather stations indicated by dots on the red color spectrum are stations that, when compared to the total average summer precipitation over 1905-1935, reported a *lower* summer average rainfall for more than 50% of years.



Number of stations at which more than 50% of years between 1985 and 2015 had a *higher* average summer precipitation than the total average summer precipitation for the 1905-1935 period	551
Number of stations at which more than 50% of years between 1985 and 2015 had a *lower* average summer precipitation than the total average summer precipitation for the 1905-1935 period	667

Figure 9. 1985-2015 yearly average summer precipitation difference (mm) from the total average summer precipitation (mm) for 1905-1935.

From this map plot and the above statistics, it seems that while there are slightly more weather stations at which more than 50% of years between 1985 and 2015 had a higher average summer precipitation than the total average summer precipitation for the 1905-1935 period, there are only approximately 100 (out of 1218) more stations at which this is the case. Therefore, there is still strong evidence that a many weather stations reported more summer rainfall in a high proportion of years in the more recent time period than in the earlier one.

Additionally, looking at the distribution of colors in this map plot, we see that there are more darker blue points than darker red points. This indicates that when a weather station reported *more* summer rainfall in a high proportion of years in the more recent time period than in the earlier one, this proportion of years with higher rainfall was typically greater than the proportion of years reported to have less rainfall by another weather station. In other words, if a weather station A had *x* proportion of years in the later time period with more rainfall than the first time period's average, and another

weather station B had *y* proportion of years in the later time period with less rainfall than the first time period's average, then *x* is likely to be higher than *y*.

Overall, this further supportes the claim that rainfall is increasing in severity, because when there is more rainfall there is *a lot more* rainfall, while when there is less rainfall, the difference is not as extreme.

Comparison of Average Rainfall Trends between US Regions

Lastly, I was interested in any observable average rainfall trends between the Western, Midwestern, Northeastern and Southern regions in the US. To carry out this analysis, each weather station was placed into a region category based on the state in which it fell and based on the region segmentation of the US according to the US Census Bureau. For all weather stations in each region,

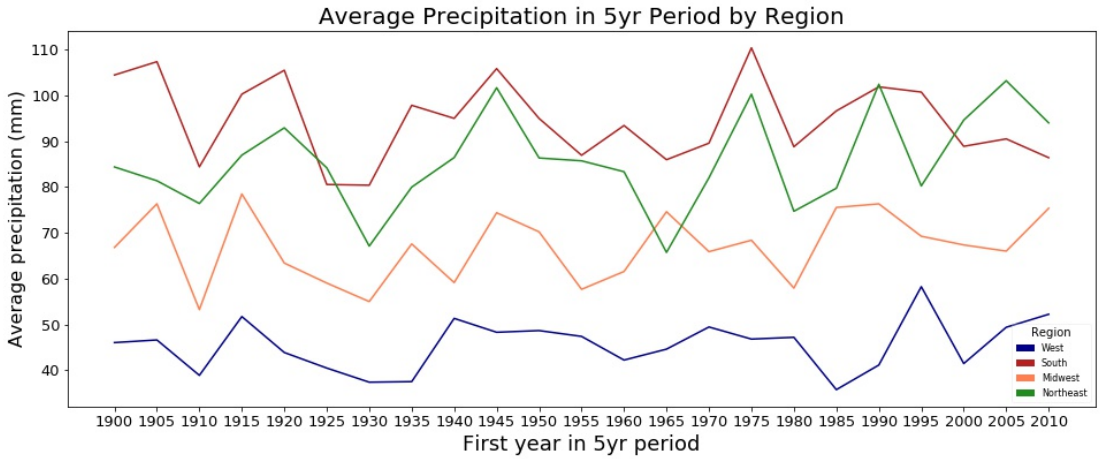


Figure 10. Average precipitation (mm) in 5-year period by region (West, South, Midwest, Northeast).

From this plot, it appears that each US region (west, northeast, midwest or south) has a very distinct range of average precipitation values. Yearly average precipitation in the South is consistently higher than in the Northeast, the yearly precipitation in the Northeast is consistently higher than in the Midwest, and the yearly average precipitation is consistently the lowest in the West.

This is interesting when compared with our permutation p-value plot (Figure 6), because the most significant changes in rainfall were not observed in the South. So it appears that while rainfall in the South is high, it is consistently high rather than having recently increased.

Also, the very regular peaks and troughs in each line in the plot above are somewhat relieving in light of the continued warnings about the immediate and severe effects of climate change. This is not to say that periods of more severe rainfall are not to come, but we can at least be assuaged by the fact that discrete periods of severe rainfall are common the history of the US.

Conclusion

So, is rainfall becoming more severe? From a simple comparison to two discrete time points (January 1910 and January 2010), there is not a significant amount of evidence to support the

prediction that rainfall is indeed becoming more severe. When we extend our analysis to include a wider range of time points, however, there is a somewhat overarching indication of increasing rainfall.

Firstly, the comparison of average precipitation for the entire 1910 and 2010 years indicated that most weather stations did indeed report a slight increase in precipitation from 1910 to 2010. Furthermore, looking specifically at summer rainfall, the permutation test analysis supports the idea that rainfall volume does change significantly from the early to the late 1900s when compared across the time periods 1905-1935 and 1985-2015. And through the upper- and lower-tail permutation tests we were able to confirm that this precipitation change was indeed most significantly an *increase*.

Moreover, in the 30-year time period from 1985-2015, around 45% (551) of weather stations experienced more than 15 years in which the yearly average summer rainfall exceeded the average for the 1905-1935 time period.

There is, therefore, substantial evidence that rainfall is becoming more severe. At the same time, however, the precipitation situation may not be too dire on the whole. Figure 10 shows that precipitation volume has had very regular ups and downs in all regions of the US over the past century. What these analyses do not show, though, and what climate change analysts worry about (and frankly what everyone should worry about) is whether or not weather patterns such as precipitation severity have changed more intensely over the past two decades and will continue to change with increasing intensity in the years to come.