CSE 255, Spring 2017: Homework 5
Weather analysis of North Dakota & Minnesota

Jahnavi Singhal (PID: A53205623)

May 16, 2017

1 Introduction

This report is about the analysis of the weather in a few cities of USA, particularly parts of North Dakota and Minnesota. The data was made available to us from NOAA (National Centers for Environmental Information).

I have focused on six types of measurement in the beginning and then tapered the analysis towards one measurement:

- TMIN, TMAX: the daily minimum and maximum temperature (tenths of degree C)
- TOBS: Temperature at the time of observation
- PRCP: Daily Precipitation (in tenths of mm)
- SNOW: Daily snowfall (in mm)
- SNWD: The depth of accumulated snow.

Every observation that we take or plot on graph is a vector of 365 dimensions (which represents the info for 365 days of an year).

Various experiments have been performed to analyse the historical data of different stations. Weather maps are created by plotting or tracing the values of relevant quantities mentioned above onto a geographical map.

2 Comparison with outside sources

I have performed a sanity check by comparing some of the general statistics with the patterns obtained from online sources like US Climate Data and Weather.com to confirm that our analysis matches with the statistics obtained by the experts.

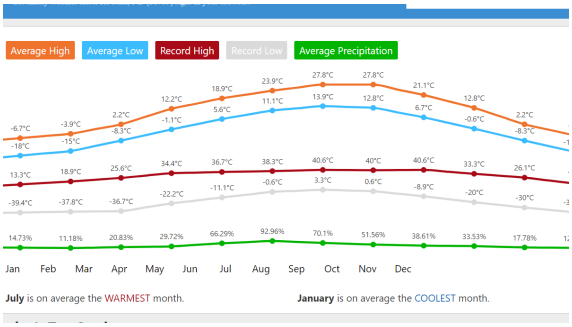


Figure 1: Plot showing the temperature trend record for the year for North Dakota

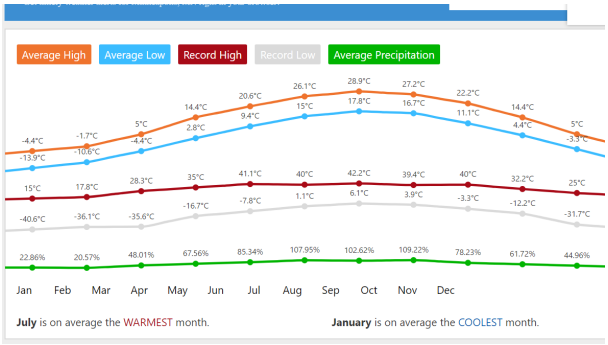


Figure 2: Plot showing the temperature min/-max trend record for the year for Minnesota

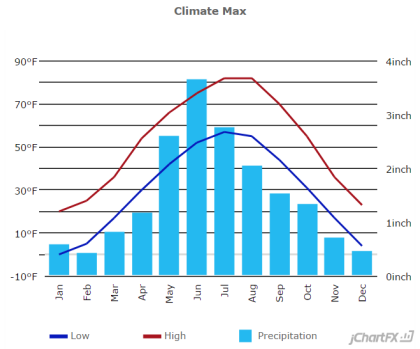


Figure 3: Histogram showing trends for precipitation, high low temperature for North Dakota

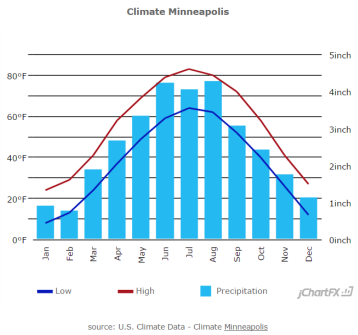


Figure 4: Histogram showing trends for precipitation, high low temperature for Minnesota

The histogram and line plots shown above are obtained from US climate data for North Dakota and Minnesota where the high/low temperature trends are quite similar; January being the coldest month and July being the hottest and the precipitation happens the most in the month of June.

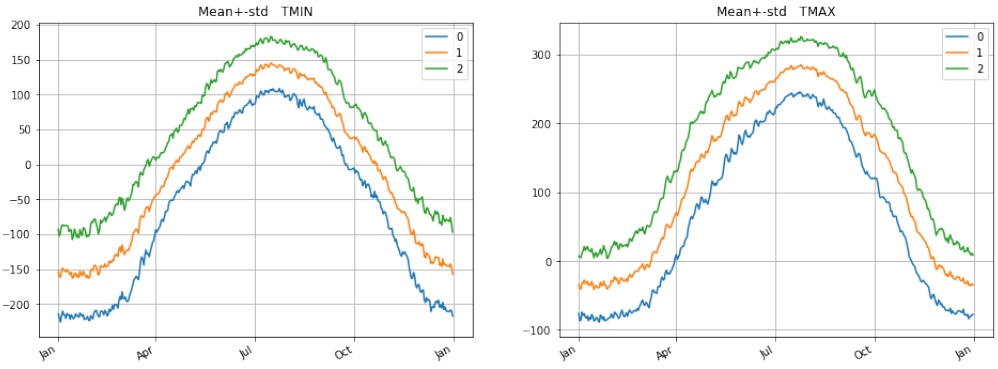


Figure 5: Plots showing the minimum and maximum temperatures of the region across the months

The above plot is drawn using the data of various years of stations under analysis. This is in sync with what is expected. The high and low temperatures vary from -10 degree C in cold months (January) to 20 - 30 degree C in hottest months (July).

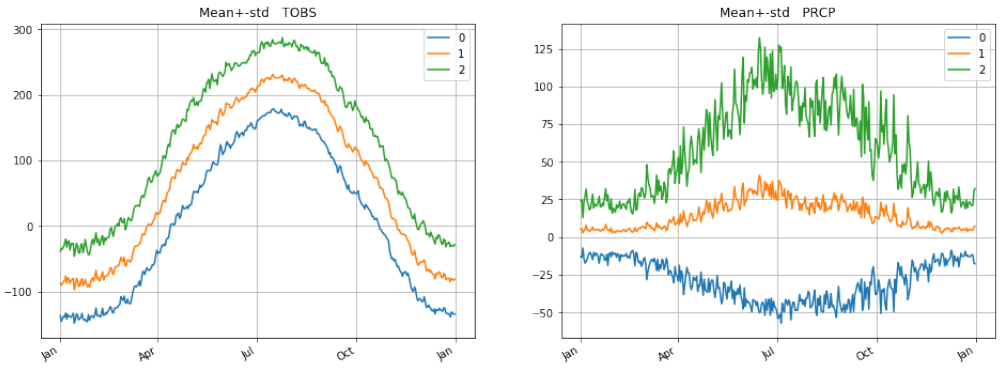


Figure 6: Plots showing the temperature at the time of observation and precipitation of the region

From the plots above, we can clearly see that the observations match with the graphs obtained from the websites. June receives most precipitation (125 in tenths of mm which is approx 5 inches) which is same as in figure 3.

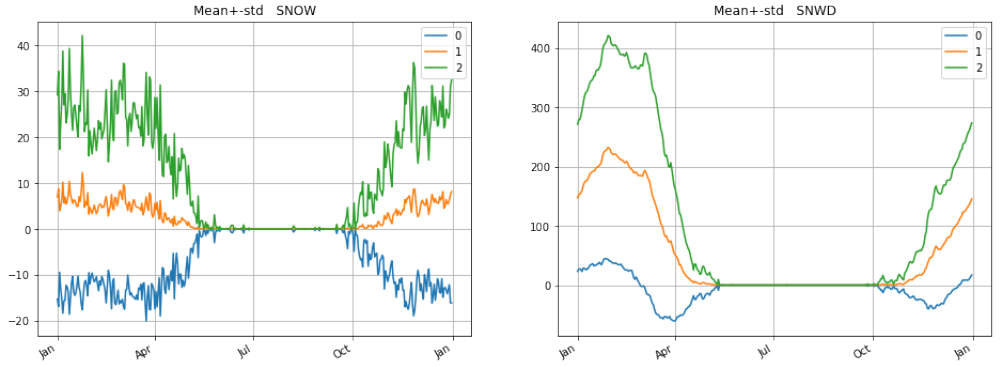


Figure 7: Plots showing the snowfall and snow depth in the region for the different months

From this figure, we can see that months from May to October receive almost no snowfall and have zero snow depth.

3 PCA analysis

Principal component analysis (PCA) is a statistical tool / process which we can use to convert a bunch of observations of correlated variables into a set of values of linearly uncorrelated variables. This way we reduce the dimensions of the data and thus, is very useful for data analysis. The first principal component has the largest possible variance (that is, accounts for as much of the variance in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components.

For all the six type of measurements, we compute the % variance explained as a function of the number of eigen-vectors used.

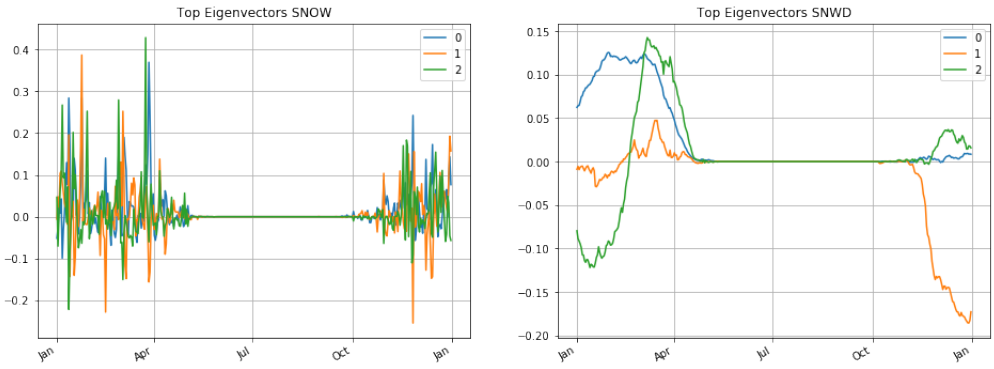


Figure 8: Plot showing top 3 eigen vectors for the measurement SNOW and SNWD

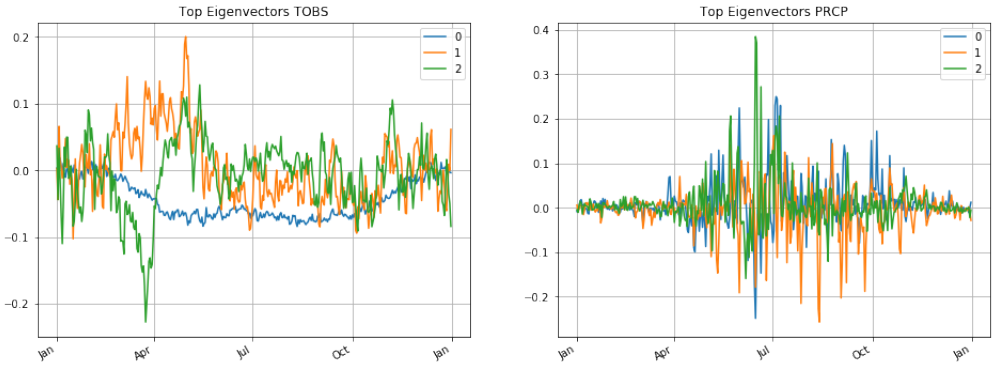
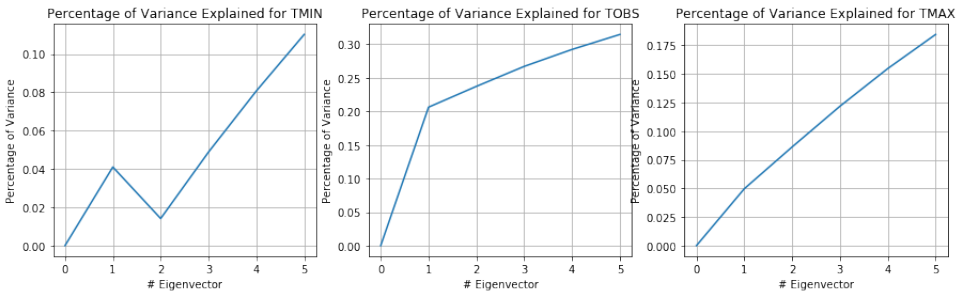


Figure 9: Plot showing top 3 eigen vectors for the measurement TOBS and PRCP

The above plots suggest that the trend of snow depth during winters using the top 3 eigenvectors is quite similar to the one in figure 7 and the snow depth is almost zero in the months of May to October as it is the summer season. However, the plot for snowfall using top 3 eigen vectors is not a very clear representation of the trend of snowfall across the months (as in figure 7).

Talking about the precipitation, the plot using top 3 eigen vectors seems to be somewhat random and doesn't reflect the actual trend. This can be further explained by the graphs shown below which capture the percentage variance covered by a top few eigen vectors.



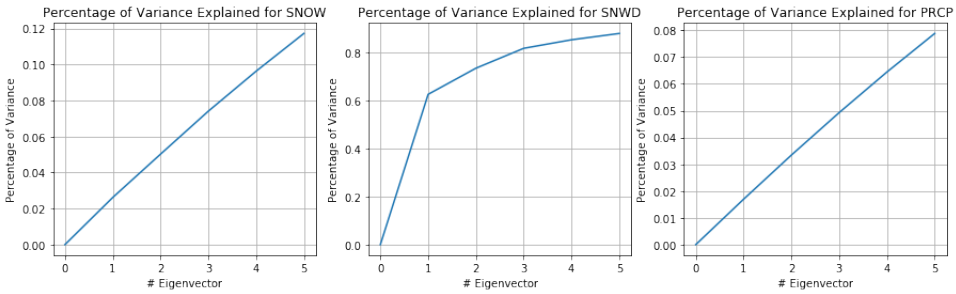


Figure 10: Plots showing percentage of variance explained

Observation from the plots from figure above:-

- Around 15% of variance is explained by top 5 eigen vectors for TMIN
- Around 35% of variance is explained by top 5 eigen vectors for TOBS
- Around 18% of variance is explained by top 5 eigen vectors for TMAX
- Around 12% of variance is explained by top 5 eigen vectors for SNOW
- Around **85%** of variance is explained by top 5 eigen vectors for SNWD, more than 60% is explained by the first eigen vector itself
- Around 8% of variance is explained by top 5 eigen vectors for PRCP

Based on the observations above, I chose to analyze SNWD further as the representation using eigen vectors which reduces the dimensions of the data and still represent the data quite well.

4 Analysis of Snow Depth

I have plot the data for the different months for the observation - snow depth. The graphs below tell the mean of the observations and the top 5 eigen vectors. Looking at the graph carefully, we figure out that the first eigen vector is quite similar to the plot of the mean and thus, represents the most of the variance. The snow season is from November to April with a peak in February.

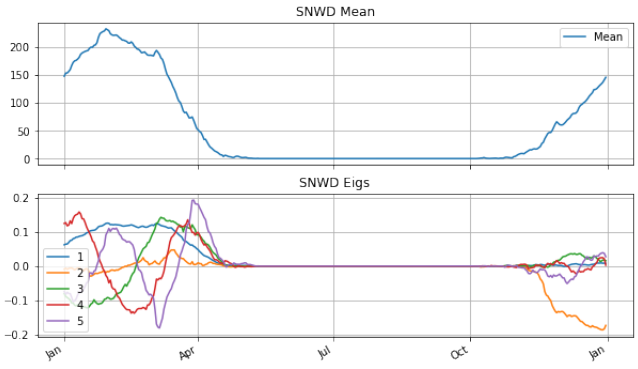


Figure 11: Plot of mean and top 5 eigen vectors for SNWD

4.1 Residuals & reconstructions

In general, the difference between the observed value of a variable and the predicted value is called the residual. Each data point has one residual. Residual = Observed value - Predicted value.
Thus, for our case, residual is the value obtained after subtracting the mean and the projection of the first 5 eigen vectors from the actual value of the observation.

I did experiments on the data to analyze the snow depth through out the year and included some of the results here. The 6 figures below include 8 plots each which show the reconstruction of the snow depth for a particular station and a particular year using top 5 eigen vectors. We can clearly see, how well the reconstruction is close enough to the target.

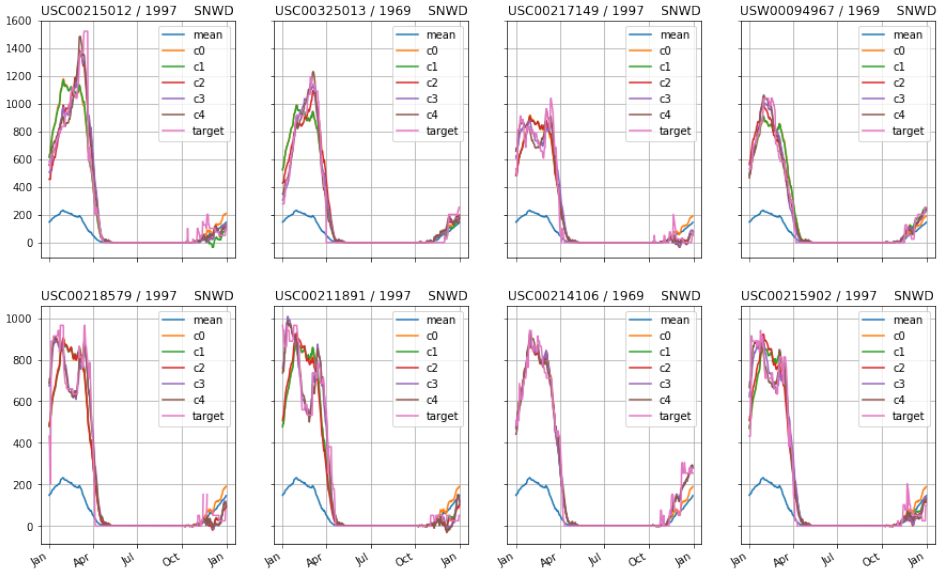


Figure 12: Reconstructions - coeff1 most positive

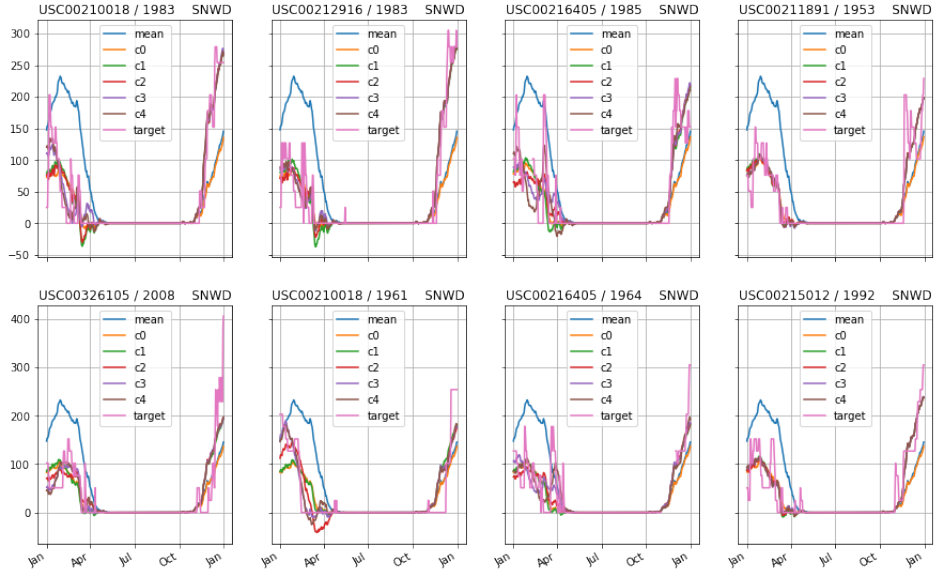


Figure 13: Reconstructions - coeff1 most negative

As the first eigen-function has a similar trend to mean; only it is close to zero during the end of the year while the mean is not. We can say that the 1st eigen vector represents the overall amount of snow above and below the mean. Large and positive values of coefficient 1 correspond to the parts of snow which are above average while low values denote the parts which are less than average.

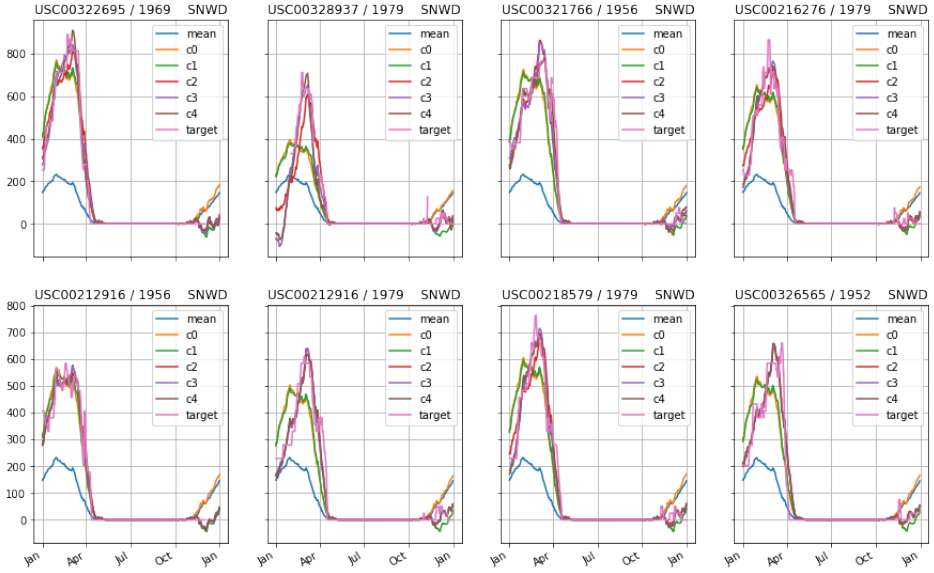


Figure 14: Reconstructions - coeff2 most positive

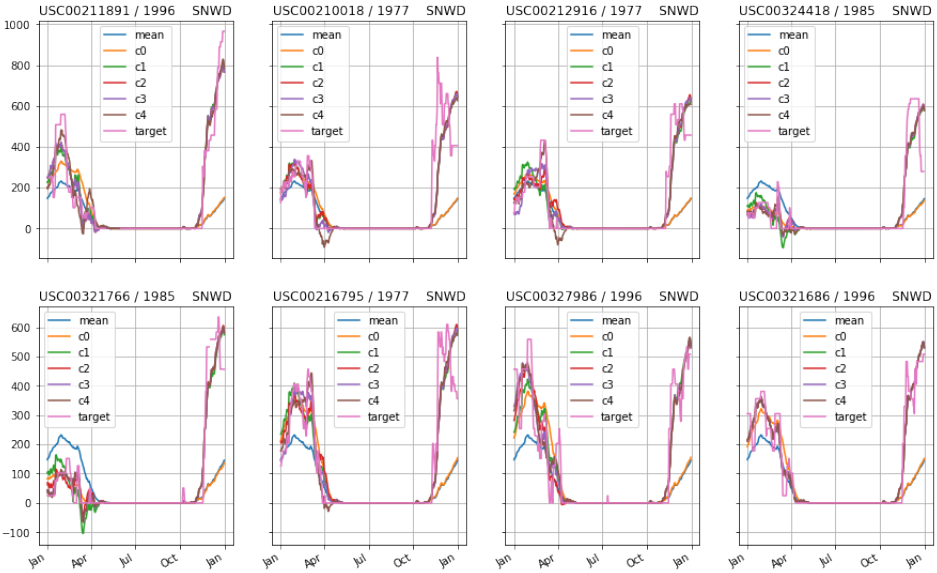


Figure 15: Reconstructions - coeff2 most negative

The second eigen vector curve tends to rise up during March and shows less snow during January. Large and positive values of coefficient 2 correspond to a late snow season (which is in March). Negative values for coefficient 2 denote slightly early snow season (which is January) and late in December.

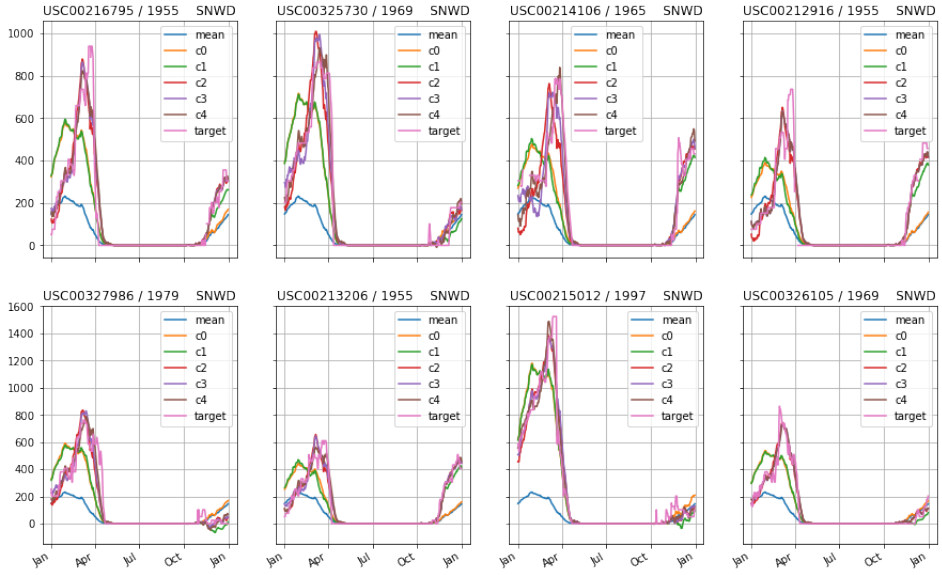


Figure 16: Reconstructions - coeff3 most positive

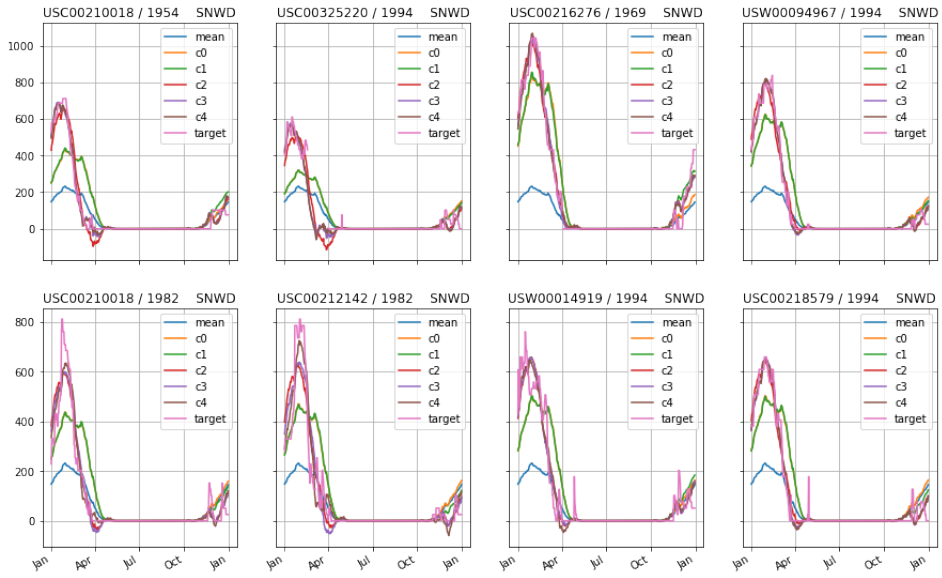


Figure 17: Reconstructions - coeff3 most negative

The third eigen vector corresponds to a peak in March and small peak in December. Large and positive values of coefficient 3 correspond to a snow season with two peaks. One is in February / March and other is in December. Negative values of coefficient 3 correspond to a season with a single peak in January.

Referring to figure 11, the fourth eigen vector corresponds to a peak in April and some snowfall in December. Large and positive values of coefficient 3 correspond to a snow season in April mostly and one peak in start of January. Negative values of coefficient 3 correspond to a some snow in February and December.

4.2 Best reconstruction

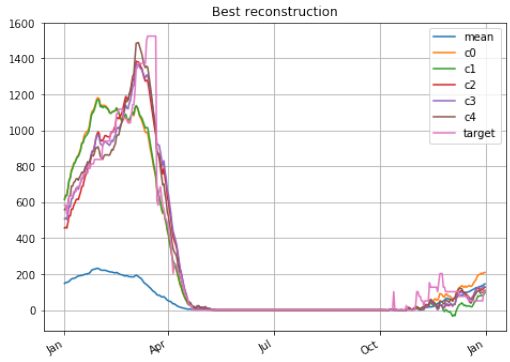


Figure 18: Best reconstruction for SNWD

The observation for the station $USC00215012$ and year 1997 got reconstructed best with the residual values very low as mentioned below:

- Residual normalized norm after mean: 0.700526
- Residual normalized norm after mean + top 1 eigs: 0.1110144
- Residual normalized norm after mean + top 2 eigs: 0.104294
- Residual normalized norm after mean + top 3 eigs: 0.043745
- Residual normalized norm after mean + top 4 eigs: 0.041141
- Residual normalized norm after mean + top 5 eigs: 0.03396
- Eigen vector Coefficients: 7570.823, 658.252, 1975.825, 409.758, -679.981

4.3 Some more analysis

The graphs below show the cumulative distribution of residuals and coefficients. It first increases and then becomes stable as the residuals are sorted which means the residuals add up quickly in the beginning and then stabilize due to smaller values.

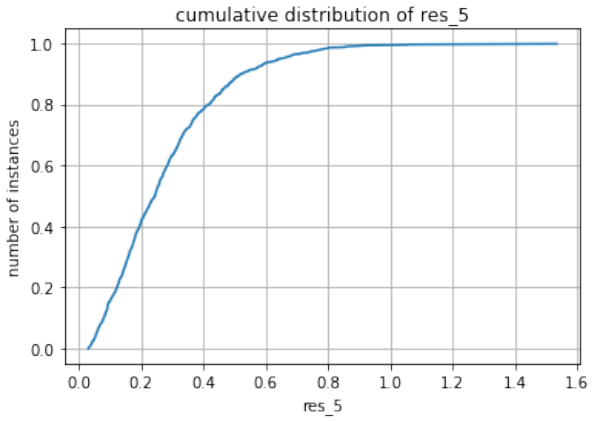


Figure 19: Cumulative distribution of residual 5 for all instances of observations

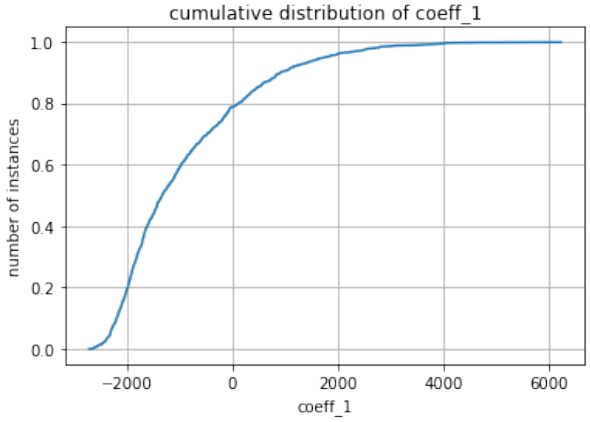


Figure 20: Cumulative distribution of coefficient 1 for all instances of observations

5 Analysis of PRCP

I have plotted the data for the different months for the observation - precipitation to compare with the results of snow depth. The graphs below tell the mean of the observations and the top 5 eigen vectors. Looking at the graph carefully, we see that we can figure out which eigen vector corresponds to which season of precipitation. There could be 2 reasons for this. First, our region is one of the driest regions of United States. Second, each eigen vector corresponds to the entire bandwidth which means there is no real variance which is being captured by the eigen vectors.

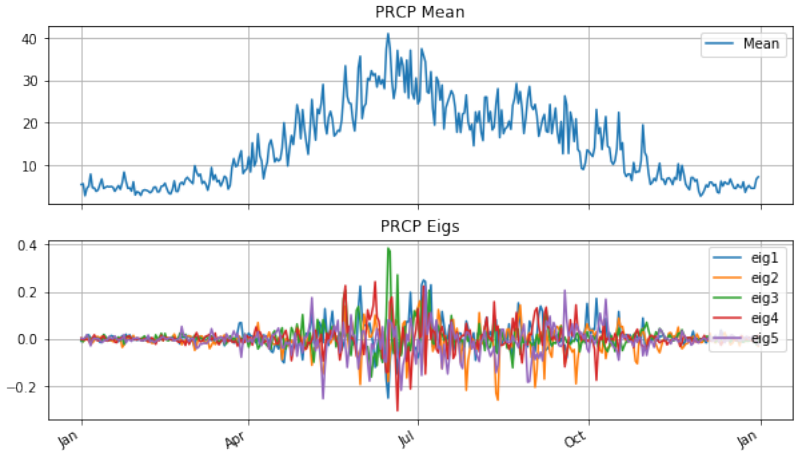


Figure 21: Plot of mean and top 5 eigen vectors for SNWD

6 Analysis of TMIN

I have plot the data for the different months for the observation - TMIN (minimum temperature of the day) to compare with the results of snow depth. The graphs below tell the mean of the observations and the top 5 eigen vectors. Looking at the graph carefully, we figure out that the first eigen vector does not represent the data well. The minimum temperatures are low in the month of January-February and December.

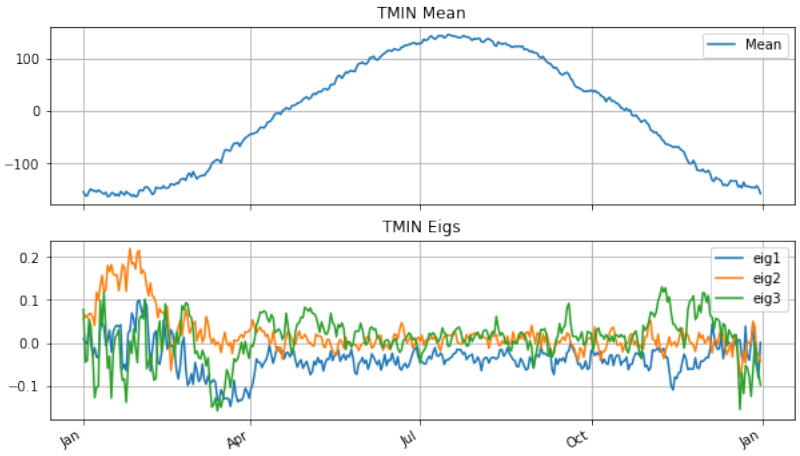


Figure 22: Plot of mean and top 5 eigen vectors for SNWD

6.1 Residuals & reconstructions

I did experiments on the data to analyze the minimum temperature throughout the year and included some of the results here. The 3 figures below include 6 plots each which show the reconstruction of the snow depth for a particular station and a particular year using top 5 eigen vectors. We can clearly see, in the reconstruction plot that each eigen vector is contributing to the minimum temperature curve and only represents 17% of the variance. Thus, reducing the dimensions doesn't make much sense for this feature.

First, these graphs are the reconstruction of snow depth using mean and top 5 eigen vectors. The top 6 reconstructions are shown. The observation for the station $USC0015562$ and year 1987 got reconstructed best. This data was sorted according to the coefficient 1 (in ascending order).

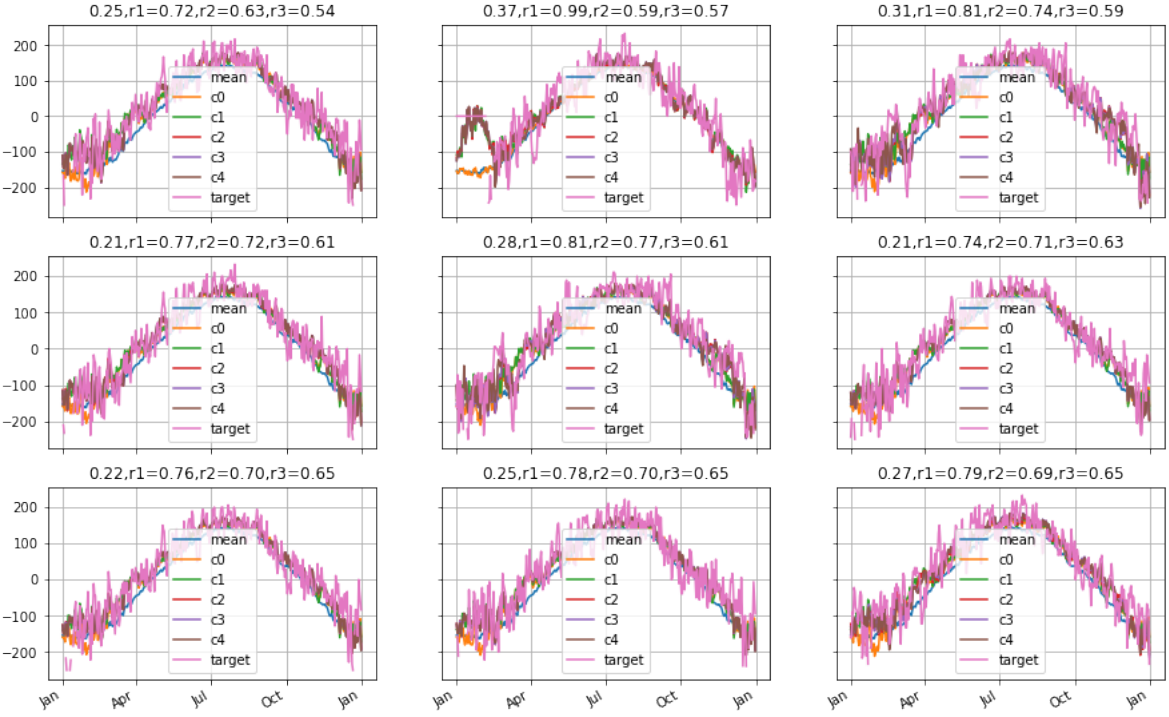


Figure 23: Reconstructions of the TMIN observations

6.2 Best reconstruction

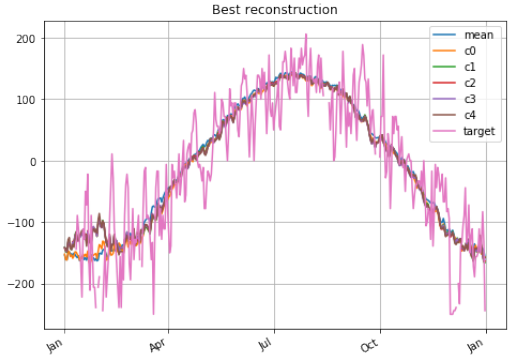


Figure 24: Best reconstruction for SNWD

The observation for the station $USC00325220$ and year 1914 got reconstructed best with the residual values very low as mentioned below:

- Residual normalized norm after mean: 1.6725
- Residual normalized norm after mean + top 1 eigs: 1.22017
- Residual normalized norm after mean + top 2 eigs: 1.00108
- Residual normalized norm after mean + top 3 eigs: 0.998276
- Residual normalized norm after mean + top 4 eigs: 0.97827
- Residual normalized norm after mean + top 5 eigs: 0.976628
- Eigen vector Coefficients: 283.008, 244.788, 70.30, 0.0063, 20.884

7 Visualizing the distribution across the locations

7.1 Analysis for snow depth

I further analyzed the observation of snow depth and plotted the distribution of coefficients for the stations on a map of United States. Different color and size of markers show the variation in snow depth and the number of observations for the stations.

The blue rectangle in the maps below mark the region under analysis.

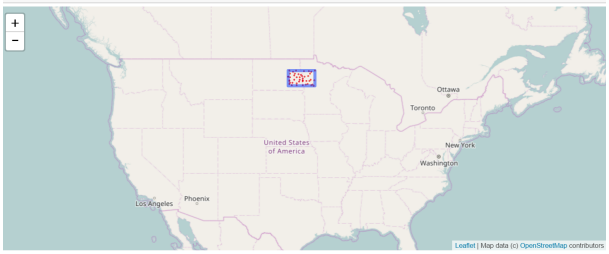


Figure 25: Map of United States showing the region of analysis

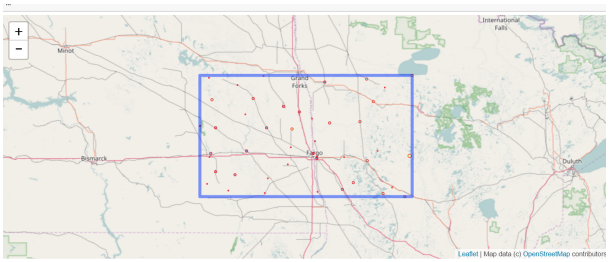


Figure 26: Zoomed version of the map showing the distribution in North Dakota & Minnesota

The **color scheme** for the circles is in RGB format, explained as follows:

- Red color denotes that the coefficient is high
- Blue color denotes that the coefficient is low (can be negative as well).
- The coefficient for the stations marked green and yellow color lie in the middle of the range.

We can see that there are some clusters of stations showing similar snow depth coefficients. Eg. There is one cluster in the middle (near Fargo) which has less coefficient 1. In the next section, I will analyze the distribution across years as well.

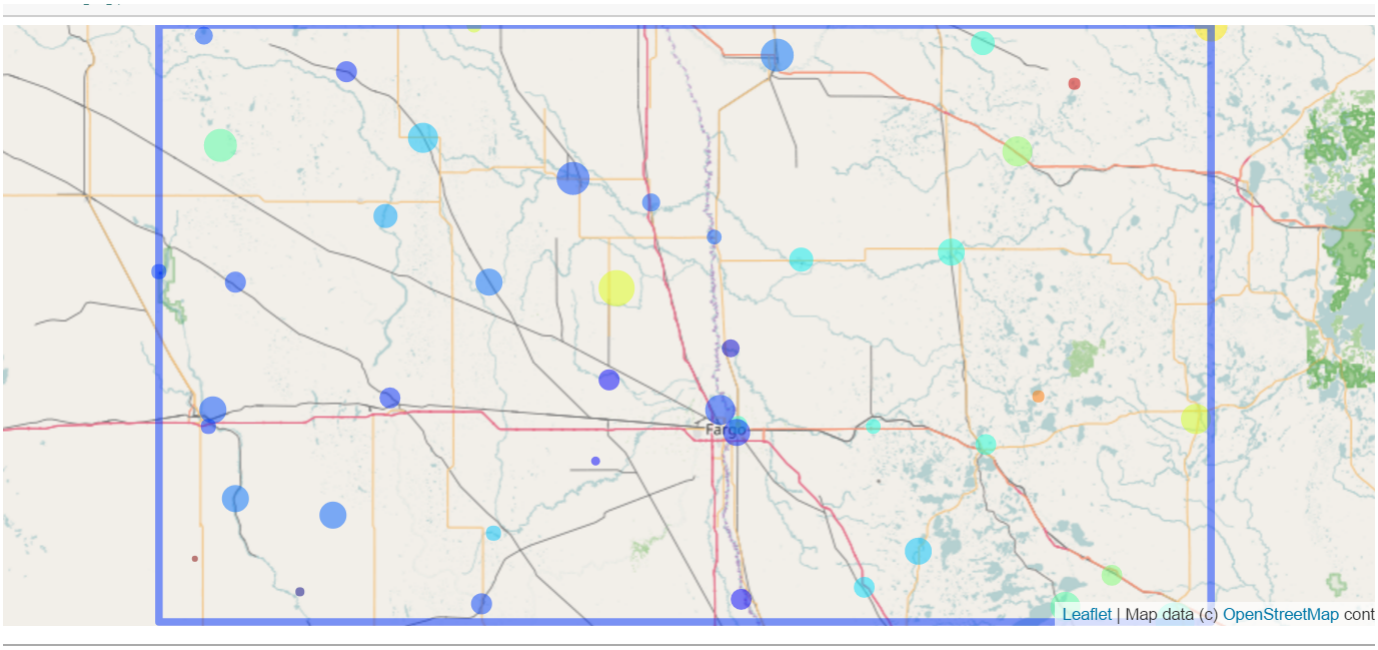


Figure 27: Map showing distribution of coefficients in the region

I have also plotted the stations on a topographic map to take the topographic features of the region into account. But since the region is limited to a small area, the topographic differences are not huge. The fact that this region is mostly cold, and close to water bodies buttresses the analysis of snow depth for the region.

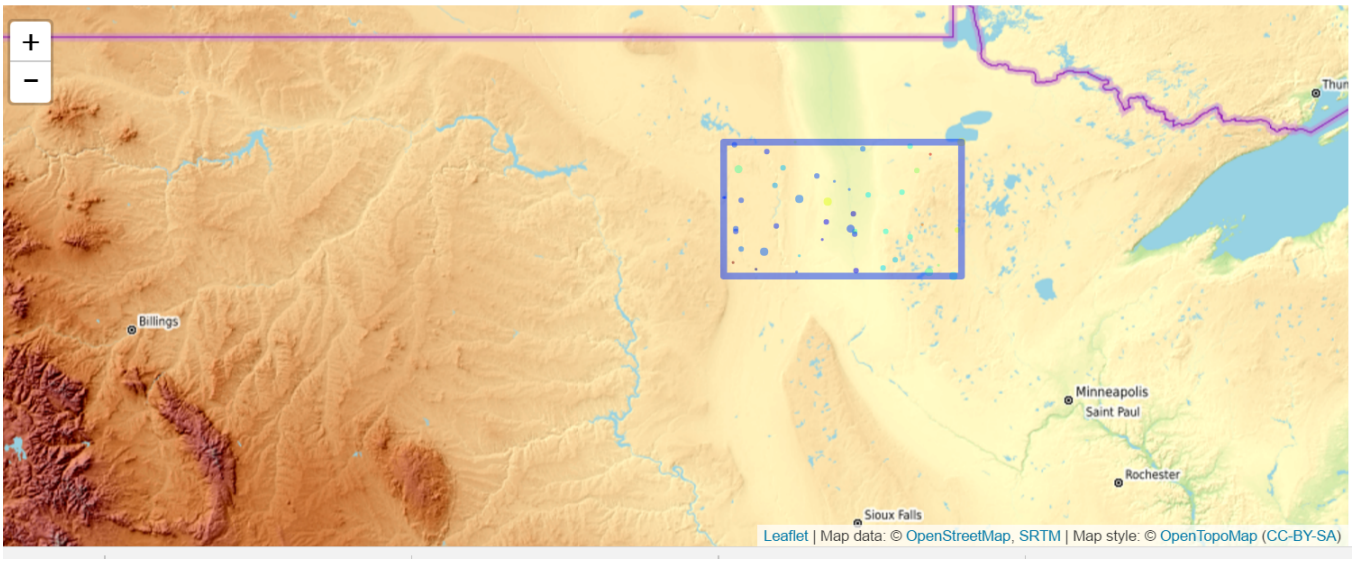


Figure 28: Topographic map showing distribution of coefficients in the region

Observations:

- Highest coefficient is for the station : M Heart Ranch Airport $(USC003265651)$ - small red circled station
- Lowest coefficient is for the station : James River $(USC003236654)$ - small blue circled station

The scatter plots below show the relation between the elevation and longitude of the stations and the coefficient 1 for the linear combination of the eigen vectors.

I did not find any useful covariance between the latitude and the coefficient 1. This could be because we are limited to a small region where the latitude doesn't affect much. However, there seems to be a good and positive correlation between elevation and longitude of the stations with the coefficients which means the more elevation the station, the better coefficient it has, which strongly supports the snow depth observation.

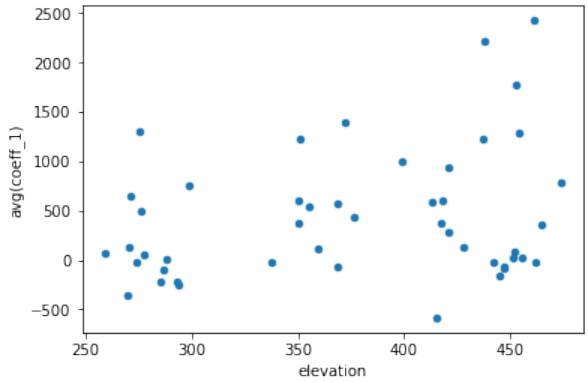


Figure 29: Plot showing avg(coeff_1) vs elevation of stations under analysis

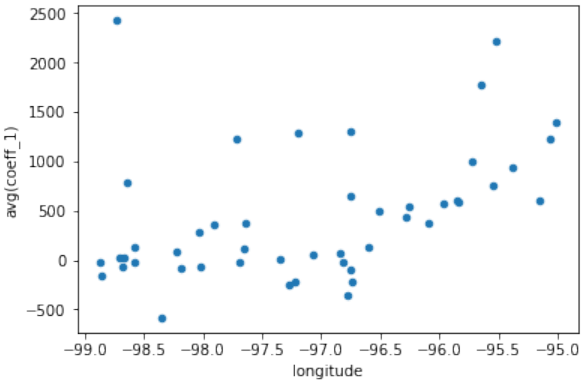


Figure 30: Plot showing avg(coeff_1) vs longitude of stations under analysis

7.2 Analysis for precipitation

I have also analyzed precipitation and plotted the coefficients for the stations on the map to support the facts in section 4 and 5.

The blue rectangle in the maps below mark the region under analysis.

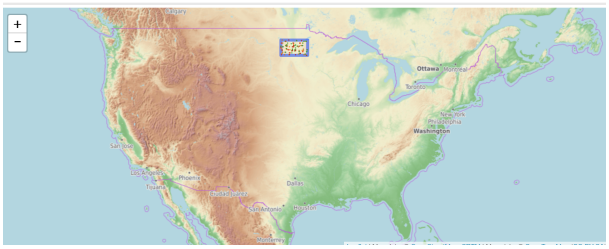


Figure 31: Map of United States showing the region of analysis

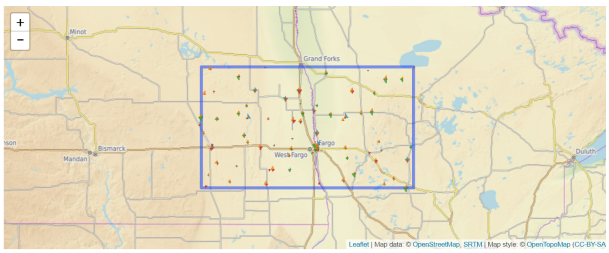


Figure 32: Map showing the distribution of PRCP in North Dakota & Minnesota

The **color scheme** for the triangles is in RGB format, explained as follows:

- Blue color denotes the 1st coefficient
- Orange color denotes the 2nd coefficient
- Green color denotes the 3rd coefficient
- Red color denotes the 4th coefficient
- Filled triangle denotes that the coefficient is negative

- Unfilled triangle denotes that the coefficient is positive

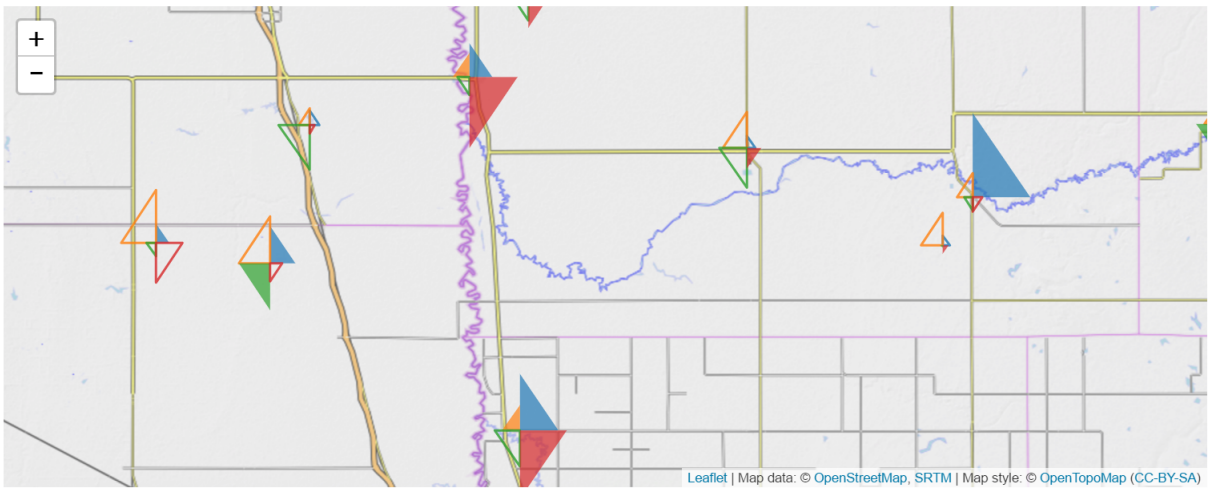


Figure 33: Map showing distribution of coefficients in the region

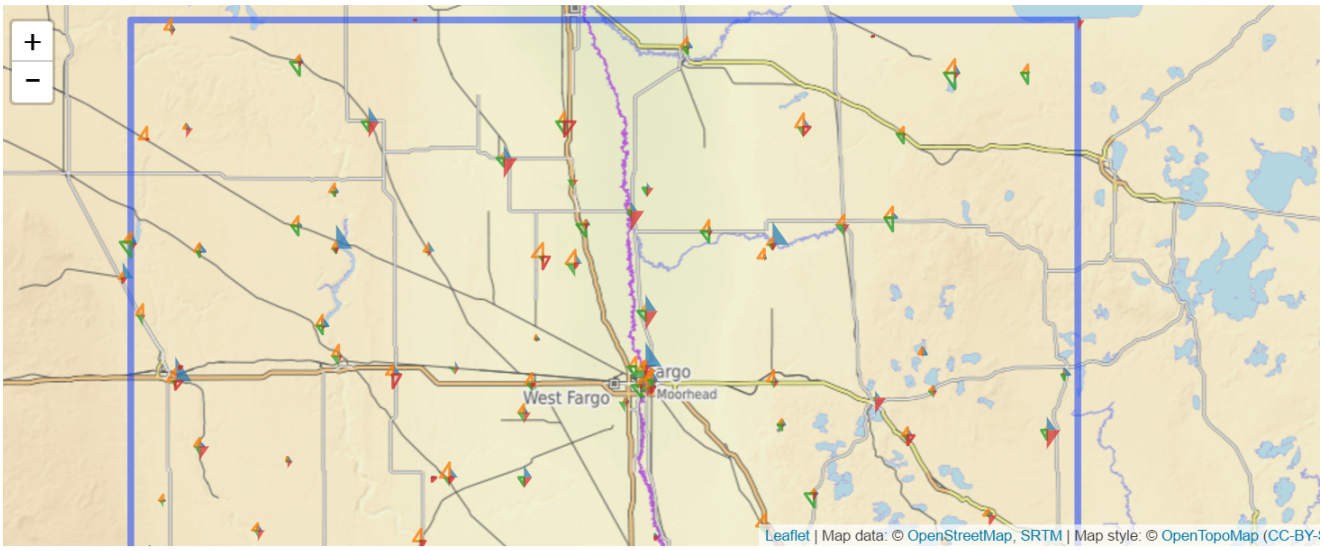


Figure 34: Topographic map showing distribution of coefficients in the region

The observations and analysis on this map about the values and variation of the different coefficients are further explained at the end of this report (in section 9).

8 Variation of year-to-year vs station-to-station

In the analysis of the snow depth, we have plotted the locations on the map. Let us now look from a different angle, i.e. the variation of the snow from year-to-year.

The mean square is calculated as the arithmetic mean of the squares of a set of numbers.

For this case, the values are as below:-

- **Coeff1**
  Total MS = 1949005.198
  MS removing mean-by-station = 1497844.5690
  Fraction explained = 23.12%
  RMS removing mean-by-year = 787664.1344
  Fraction explained = 59.73%

- **Coeff2**
  Total MS = 399370.0216
  MS removing mean-by-station = 381688.6106
  Fraction explained = 4.42%
  RMS removing mean-by-year = 108873.9813
  Fraction explained = 72.73%

- **Coeff3**
  Total MS = 346712.709319
  MS removing mean-by-station = 320630.376809
  Fraction explained = 7.522%
  RMS removing mean-by-year = 166467.30
  Fraction explained = 51.98%

The values shown above are for snow depth. We can clearly see that the variation by year explains more than the variation by station. Again, the coefficient 1 explains the overall snow for the year. The coefficient 2 is responsible for the peak in March. The coefficient 3 corresponds to a snow season in February and December.

9 Analyzing residuals for precipitation between stations per day

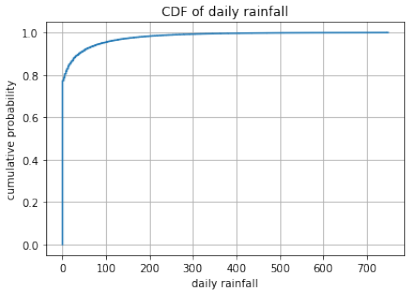


Figure 35: Cumulative Rainfall

This shows that it rains very less in the region under analysis. It rains on around **20% of the days**, which agrees with the overall rainfall statistics given on Wikipedia.

As we discussed in the class, studying the amount of rain on the same day in different stations is difficult, we will analyze on the number of common days it rained.

Lets consider the null hypothesis which says that the rainfall in two stations is independent. To discuss this further, lets compute the probability associated with the number of overlaps under the null hypothesis.

I considered two stations for analysis with the highest number of days it rained.

- $USC00211891$
- $USW00094967$

Total days under consideration = 44895
Number of days it rained in Station 1 = 43443
Number of days it rained in Station 2 = 42249

Results: The log probability (i.e. log of the probability that number of days it rained on both the stations is what we observed) = -0.0801
The number of overlap days = 40836

Then, I calculated the normalized log probability of each station to analyze if the stations close to each other are related in terms of daily rainfall statistics. To refer to the calculations, the in-class notebooks can be seen.

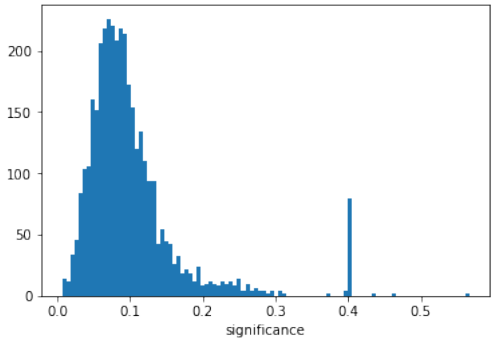


Figure 36: P-value distribution

The above histogram shows that the observations are skewed towards a lower pvalue (around 0.05) which makes us reject the null hypothesis and conclude that there has to be a correlation between the stations and their locations in terms of rainfall.

Thus, I analyzed further. There are 77 stations under observation. The matrix below shows the normalized log probability for each pair of stations.

Figure 37: Cumulative Rainfall

We can see that every station is coinciding with itself. Thus, we see a darker diagonal. In addition to this, there seems to be a strong correlation between stations 0-10 and station 25 & 76. These are the stations which are strongly co-related:- [u'USC00324418', u'USC00327986', u'USC00213104', u'USC00325660', u'USC00213463', u'US1MNNR0001', u'USC00323342', u'USC00327117', u'US1NDCS0030', u'USC00327270'] and [USC00213206-USC00325764].

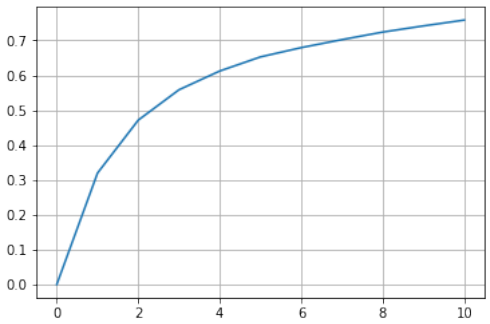


Figure 38: Percentage variance explained

The above graph shows that there first few eigen vectors explain around 75% of the variance and first 4 eigen vectors explain around 60%. Thus, I considered only top 4 eigen vectors for the further analysis.

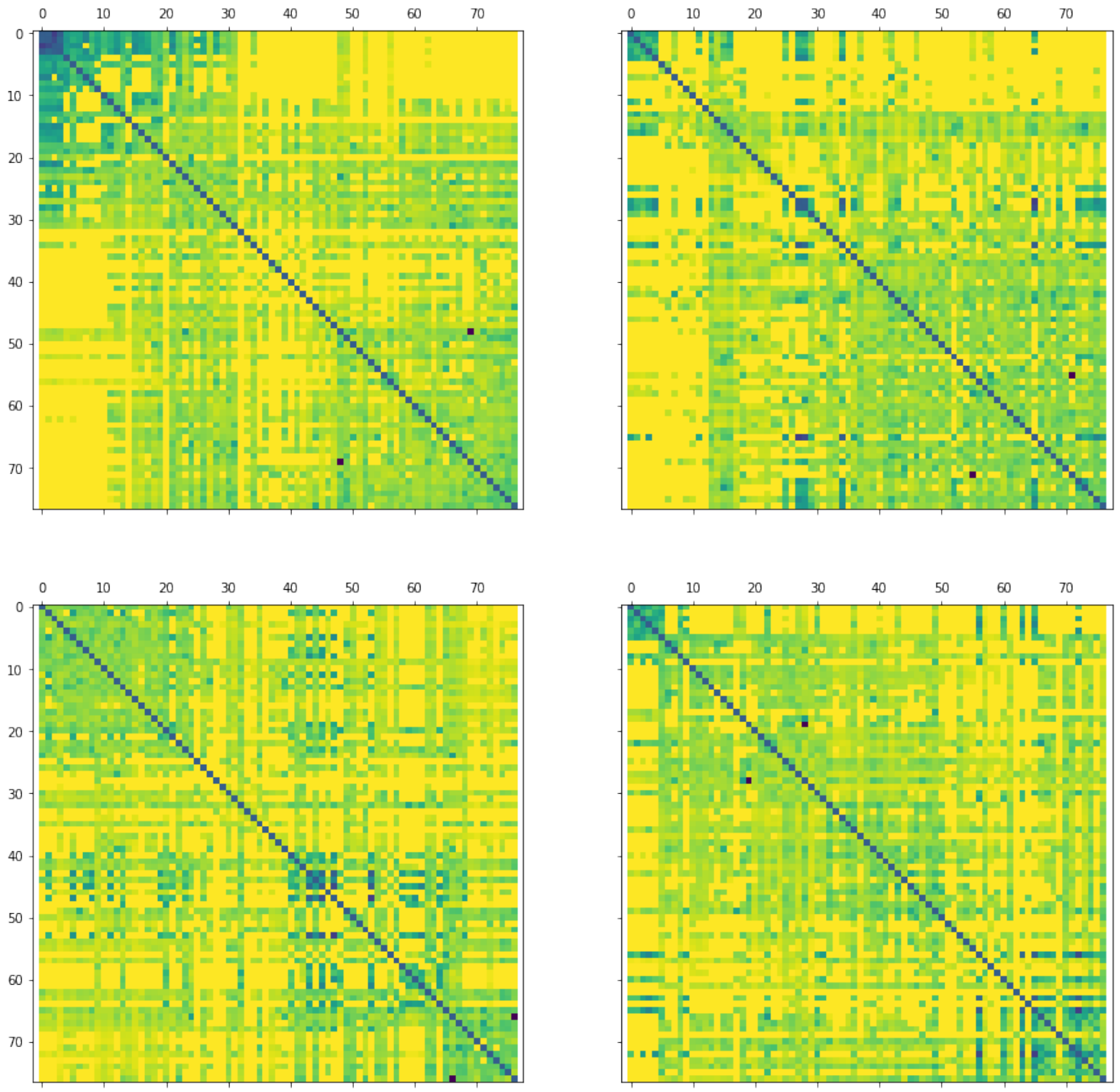


Figure 39: Percentage variance explained

I reordered the rows and columns of the matrix using the eigen vector sequence. Now, the clustering of the stations became more evident. We can clearly see two group of clusters more evident. Stations 0-5, 29-30, 65-25, 40-45 are very close to each other. Stations 5-10, 70-77 are close to each other too (may be not as strong as the first set).

In order to verify the result we obtained using eigen vectors and the rejecting the null hypothesis, I plotted some of the seemingly correlated stations on map.

The result was quite good. The stations are indeed close. I have plotted 3-4 clusters of 2-3 stations each in the map below.

- The stations colored in pink circles (25-65) are quite close to each and thus, are correlated to each other.
- The stations colored in blue circles (43-44) are close to each other and near water bodies, and therefore receive rainfall on similar days.
- The stations colored in orange circles (29-30) are close and high in latitude. They too are correlated by the eigen vector block diagonalization
- The stations in black circles are stations 1-4 and show very rich correlation between each other.

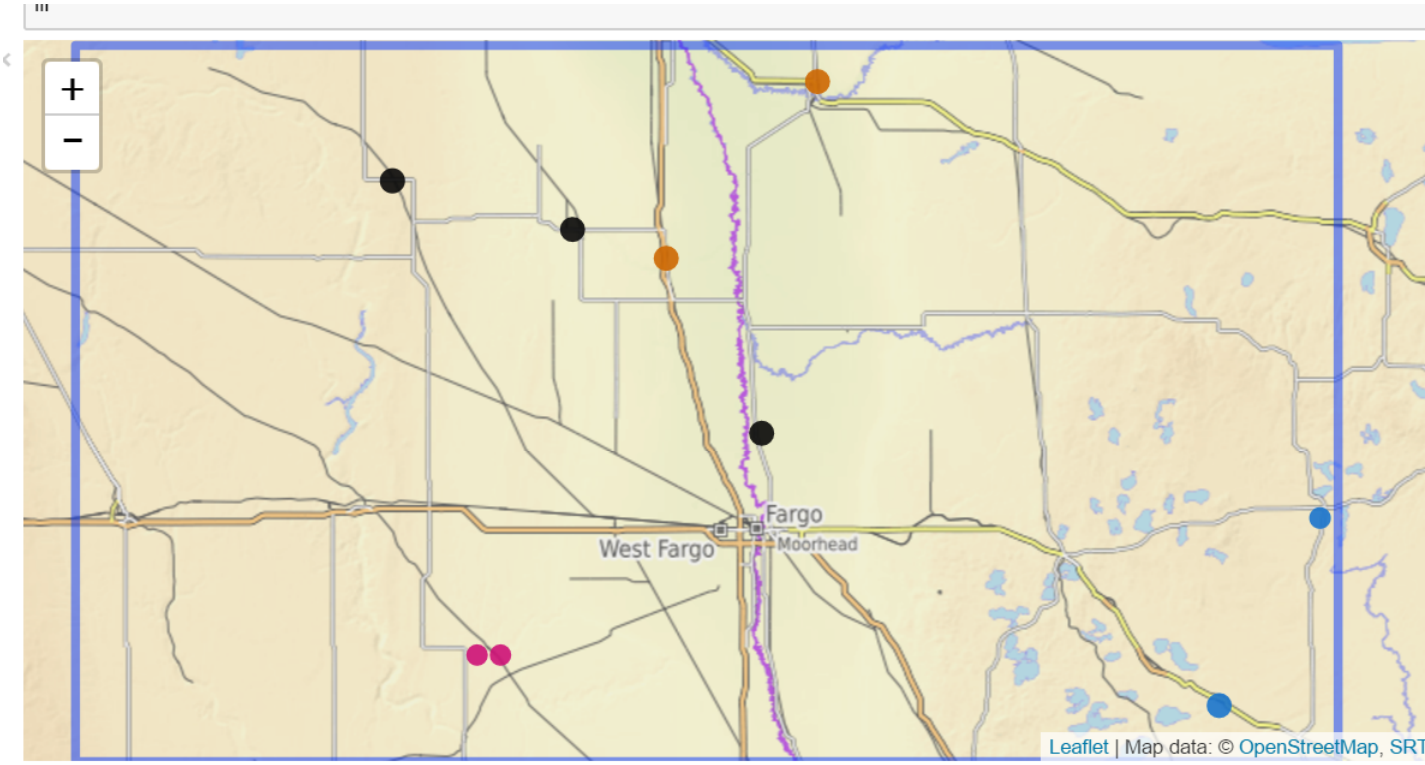


Figure 40: Map showing the correlated cities in terms of rainfall