Cedar City, UT Weather Analysis

This is a report on the historical analysis of weather patterns in an area that approximately overlaps the area of the Cedar City, Utah.

The data we will use here comes from NOAA.

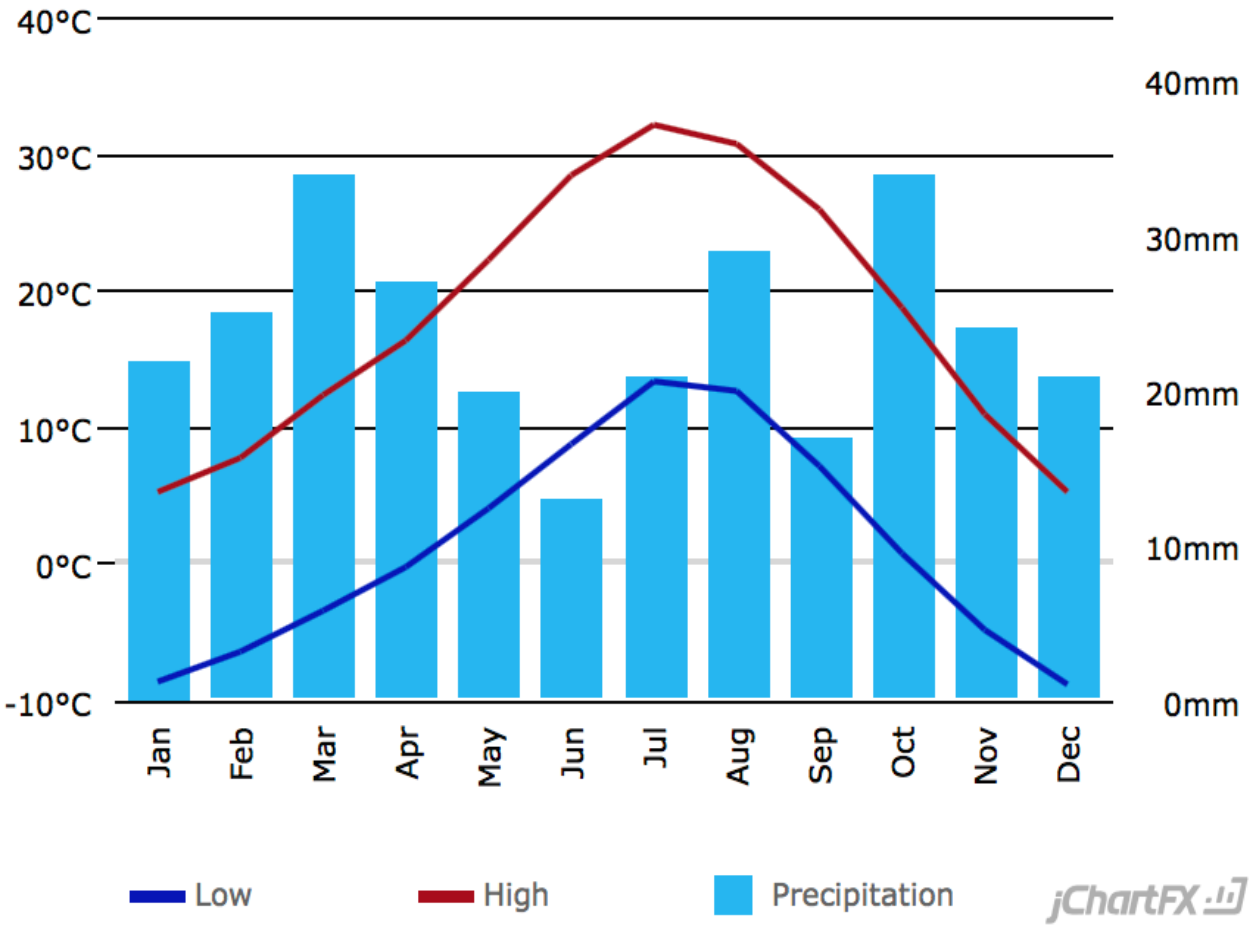
We mainly focused on six measurements:

- **TMIN, TMAX:** the daily minimum and maximum temperature.
- **TOBS:** The average temperature for each day.
- **PRCP:** Daily Percipitation (in mm)
- **SNOW:** Daily snowfall (in mm)
- **SNWD:** The depth of accumulated snow.

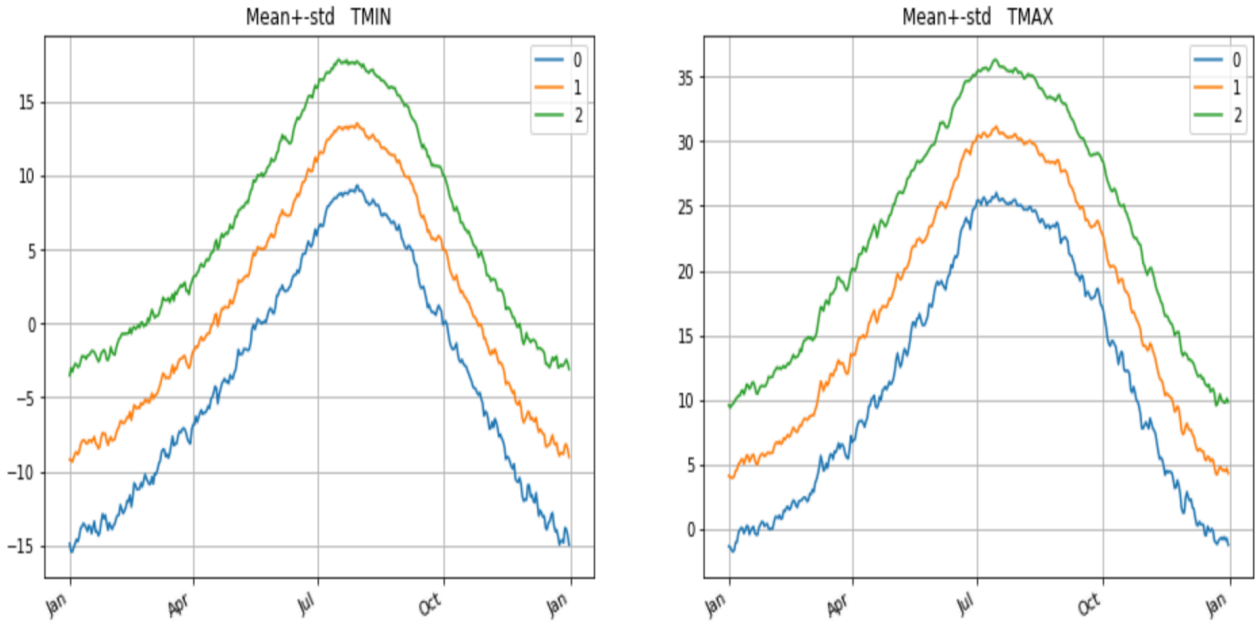
1. Sanity-check: comparison with outside sources

We start by comparing some of the general statistics with graphs that we obtained from a site called US Climate Data.

The graph below shows the daily minimum and maximum temperatures for each month, as well as the total precipitation for each month.

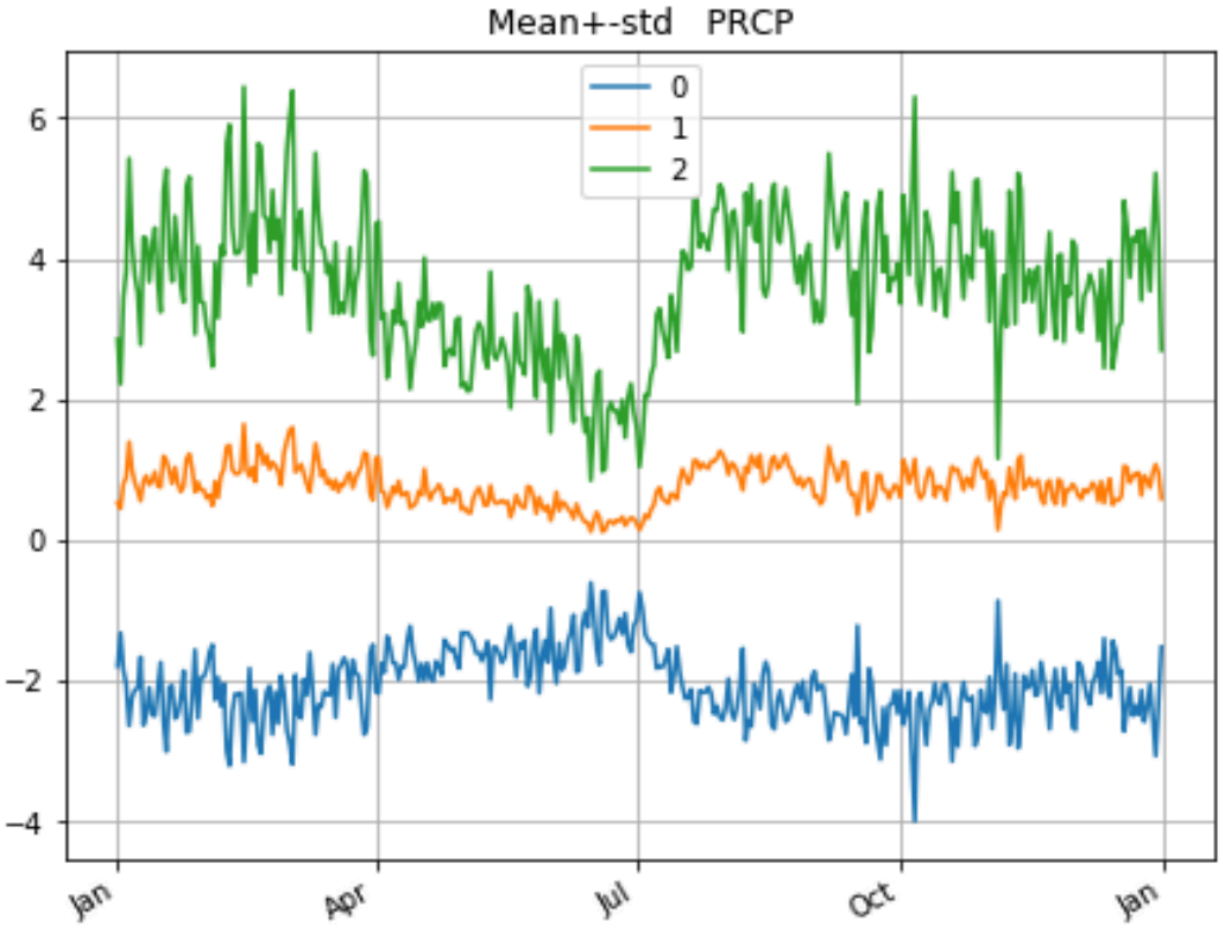


We see that the min and max daily temperature agree with the ones we got from our data, once we translate Fahrenheit to Centigrade.



To compare the precipitation we need to translate millimeter/day to millimeter/month.

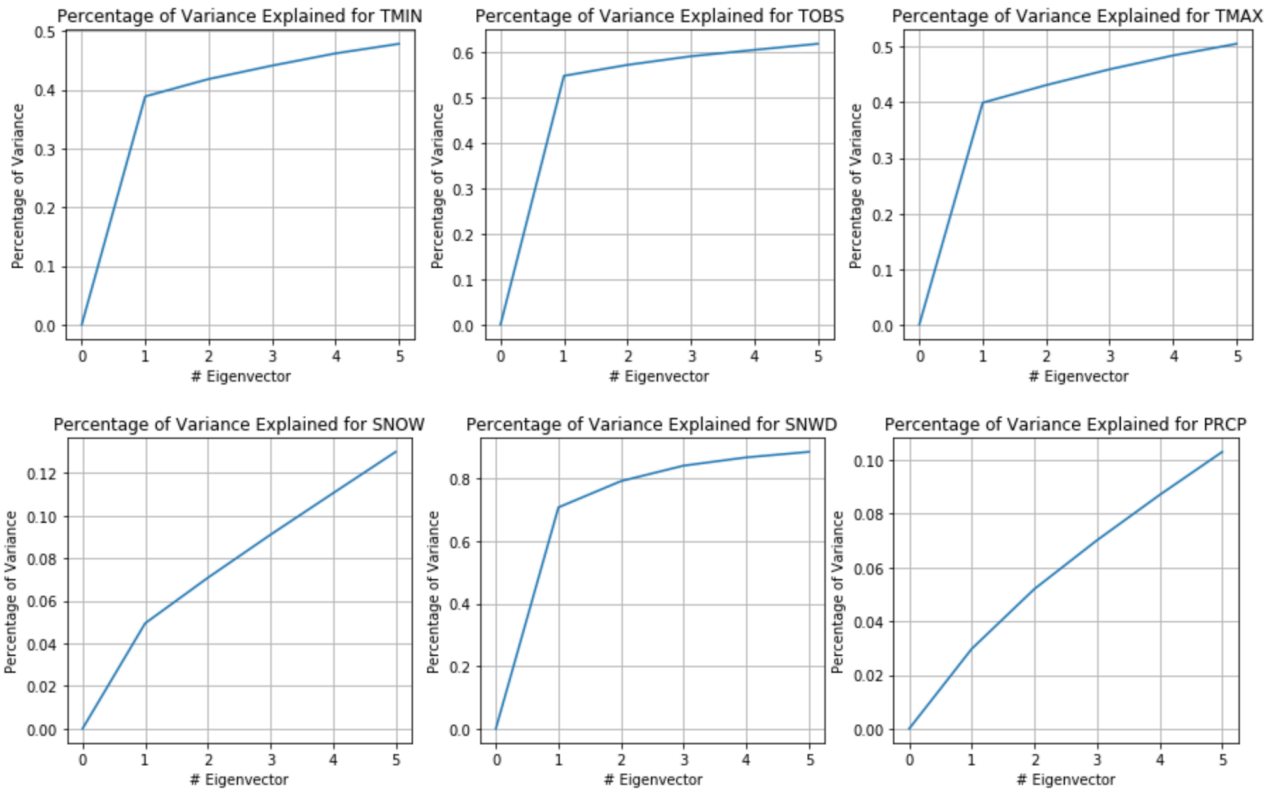
According to our analysis, the average rainfall is approximately 1.00 mm/day which translates to about 30 mm per month. According to US-Climate-Data the average rainfall is closer to 25 mm per month. However, there is clear agreement that average precipitation is close to a constant throughout the year.



2. PCA analysis

For each of the six measurements, we compute the percentage of the variance explained as a function of the number of eigen-vectors used.

2.1 Percentage of variance explained



measurement	top 5 eigen-vector variances explained	first eigen-vector variance explained
TMIN	48%	40%
TBOS	62%	55%
TMAX	51%	40%
SNOW	13%	5%
SNWD	90%	70%
PRCP	10%	3%

We see that SNWD is best explained by its top 5 eigen-vectors. The first 5 eigen-vectors almost explain all the useful daily-snowfall information in Cedar City. It is especially true for the first eigen-vector which, by itself, explains approximately 70% of the variance.

The following 3 measurements with similar variance explaining performance are TBOS, TMAX and TMIN. Their percentages of variance explained are 62%, 51% and 48%, and the corresponding first eigen-vector variance explained are 55%, 40% and 40%. For these 3 temperature measurements, the top 5 eigen-vectors explain about half of the total variance, and their first eigen-vector basically explains over 80% variance of that the top 5 eigen-vectors explained.

Measurements with the worst performance are SNOW and PRCP, which only explain about 10% variance. That is to say the top 5 eigen-vectors do not contain much useful information. In other words, these two measurements have much more noises than that of other measurements.

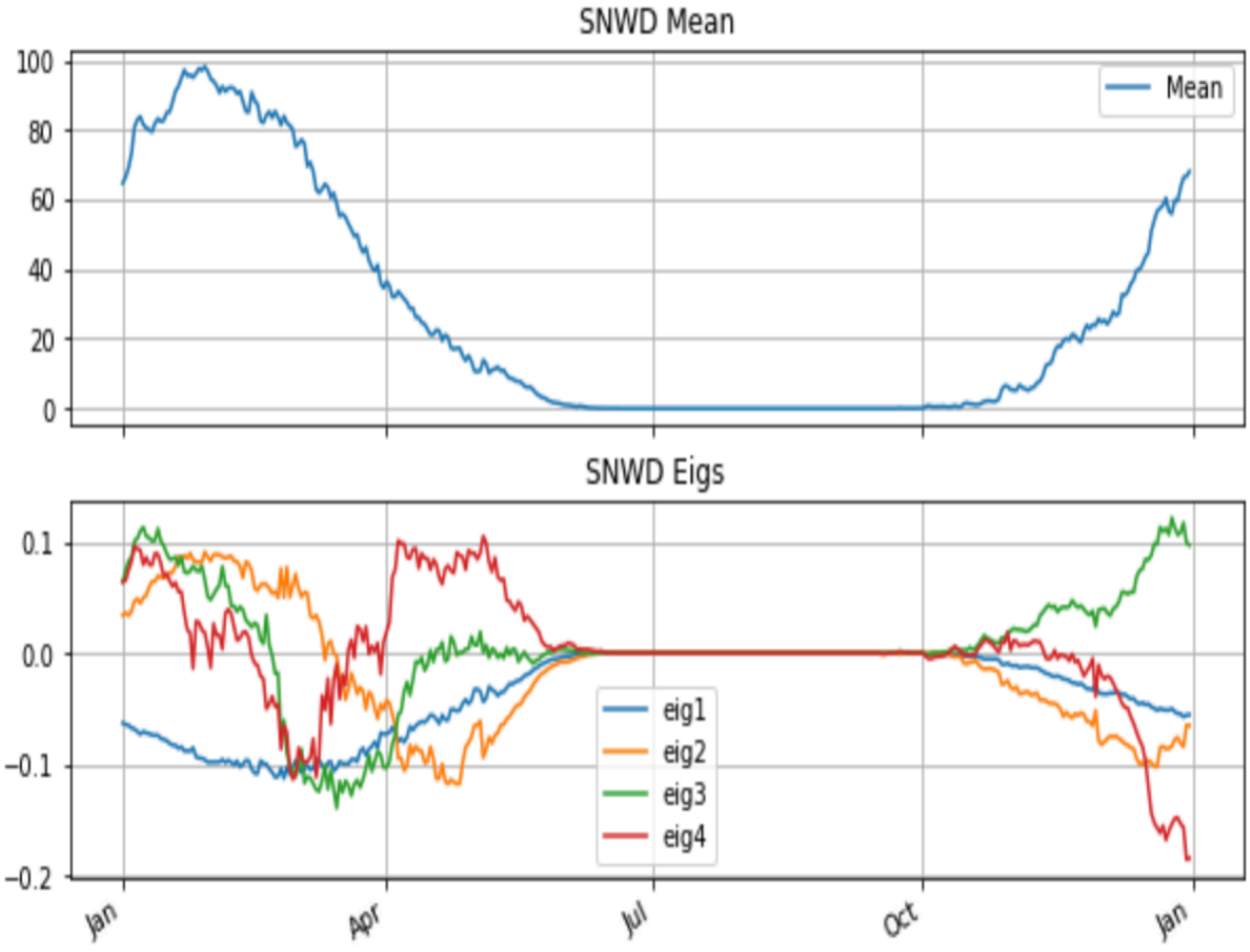
3. Analysis of snow depth

Based on the PCA analysis above, we choose to analyze the eigen-decomposition for snow-depth because even the first 4 eigen-vectors explain 85% of the variance. In other words, first 4 eigen-vectors for SNWD make sense.

3.1 Eigenvector Interpretation

First, we graph the mean and the top 4 eigen-vectors.

We observe that the snow season is from November to the end of April, where the beginning of February marks the peak of the snow-depth.



Next we interpret the eigen-functions.
According to analysis above, the first eigen vector explains over 70% variance and the corresponding first eigen-function (eig1) has a shape very similar to the mean function just in the opposite direction. The interpretation of this shape is that eig1 represents the overall amount of snow below the mean.

eig2, eig3 and eig4 are similar in the following way. They all oscillate between positive and negative values. In other words, they correspond to changing the distribution of the snow depth over the winter months, but they don't change the total (much).

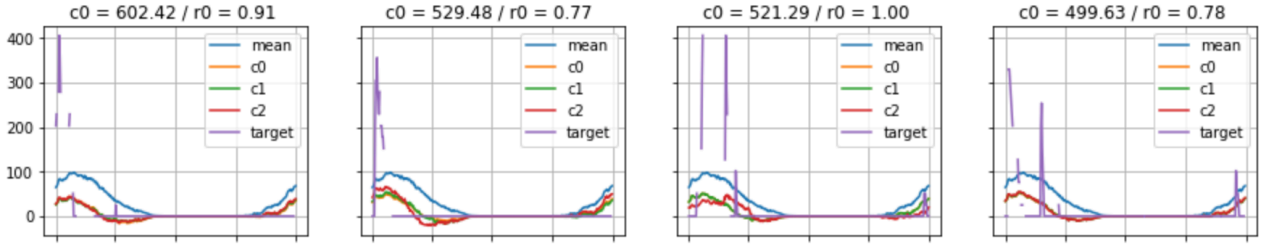
They can be interpreted as follows:

- **eig2:** more snow in Jan - Mar, less snow in Mar - mid Jun and Nov - Jan.
- **eig3:** more snow in Nov - Feb, less snow in Mar - May.
- **eig4:** more snow in Jan - Feb, Apr -Jun, less snow in Feb - mid Mar, Dec - Jan.

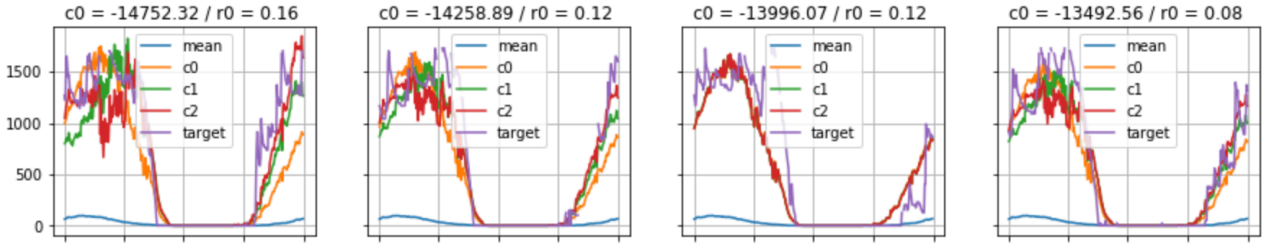
3.2 Reconstructions and Distribution of first 3 coefficients

3.2.1 Coeff1

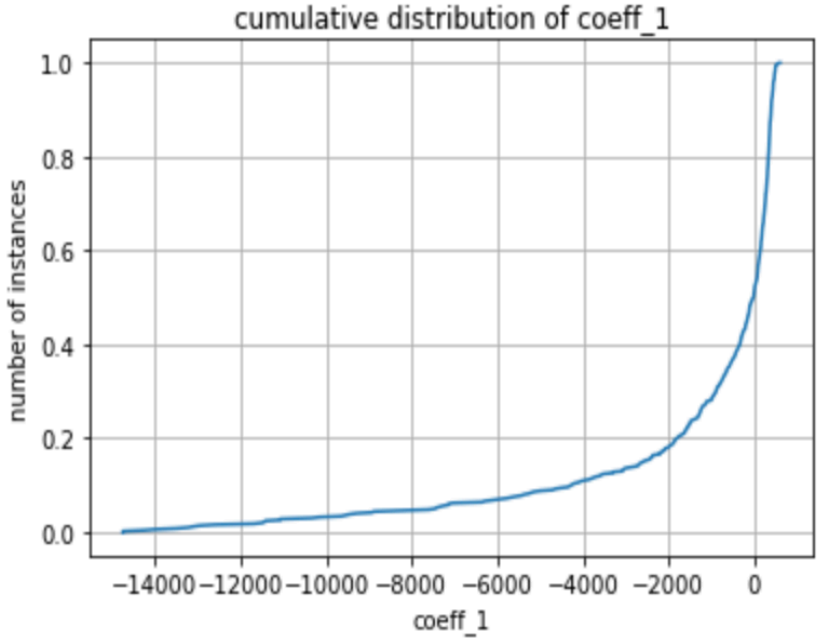
Coeff1: most positive



Coeff1: most negative

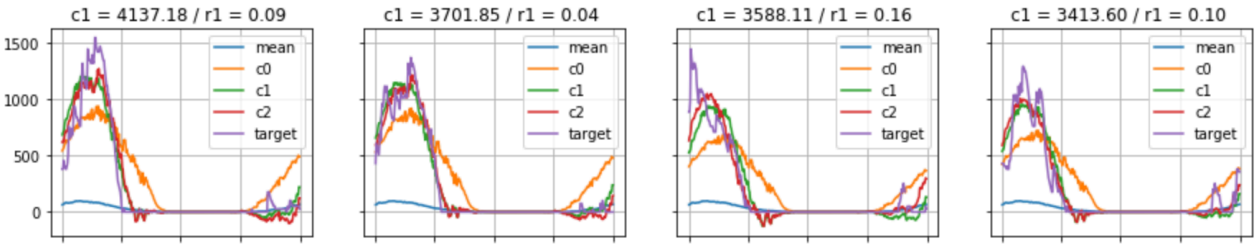


Large positive values of coeff1 correspond to less than average snow.
Low values correspond to more than average snow.

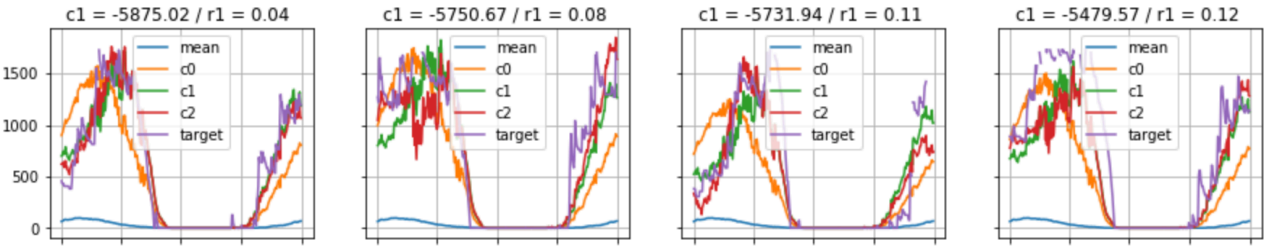


3.2.2 Coeff2

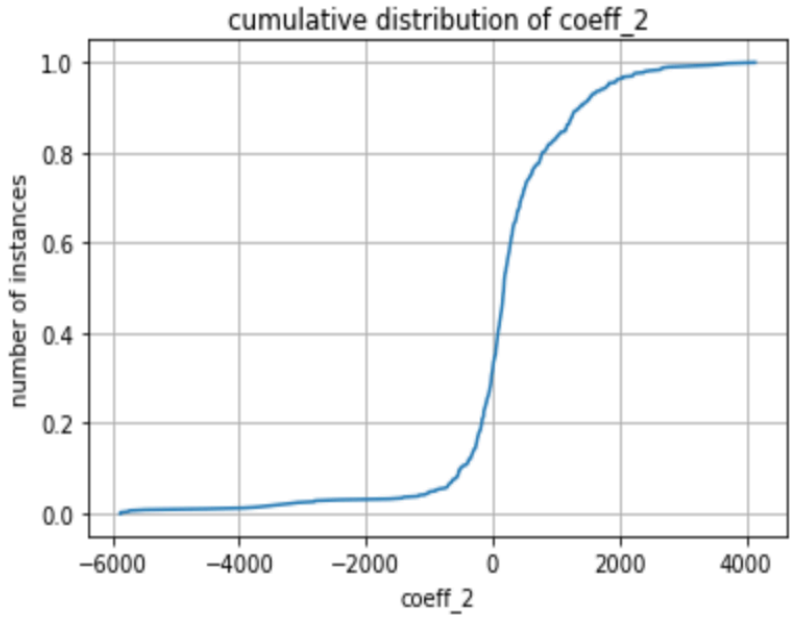
Coeff2: most positive



Coeff2: most negative

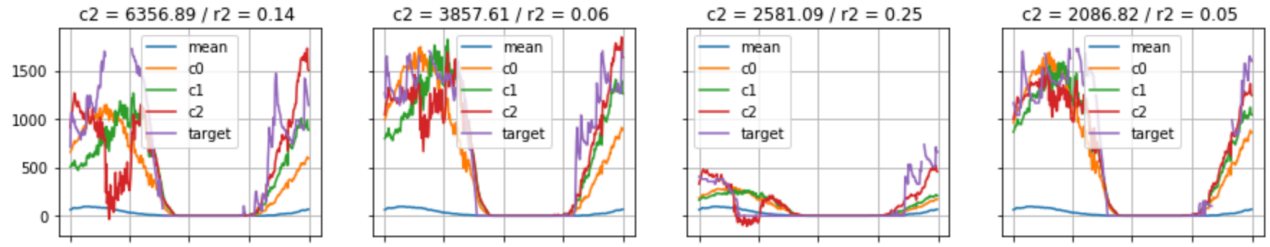


Large positive values of coeff2 correspond to an early snow season (most of the snowfall is before Feb).
Negative values for coeff2 correspond to an late snow season (most of the snow is after Apr).

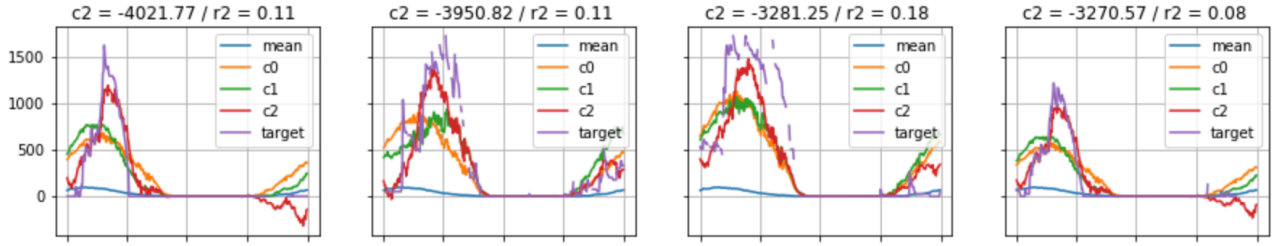


3.2.3 Coeff3

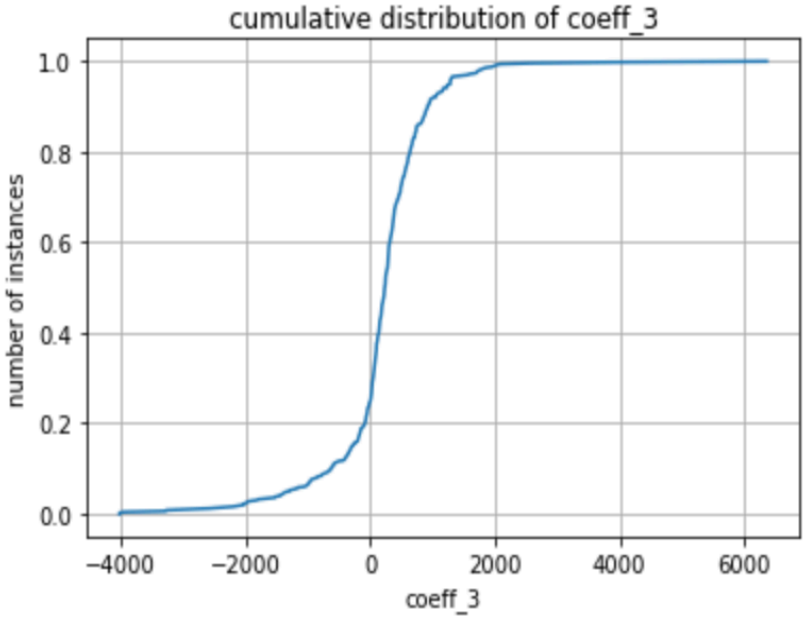
Coeff3: most positive



Coeff3: most negative



Large positive values of coeff3 correspond to several small peaks which indicate a long duration of heavy snow.
Large negative values of coeff3 correspond to a late peak (most of the snow is after beginning of Mat),which indicates a late snow season.



According to the graphs above, we can conclude that distributions of first 3 coefficients are all normal like distributions with zero mean.

In specific, distribution of coeff1 has the largest variance, then is coeff2, and coeff3 has the lowest variance among these three coefficients.

3.3 Estimating the effect of the year vs the effect of the station

According to the analysis above, we can conclude that the measurement of daily snowfall changes from season to season.

Next, we are going to estimate whether variation of SNWD is more spatially or temporally influenced.

For **spatial** influence, we measure the daily snowfall variation from different stations.

For **temperal** influence, we measure the daily snowfall variation from year to year.

3.3.1 Visualized Distribution of years and stations

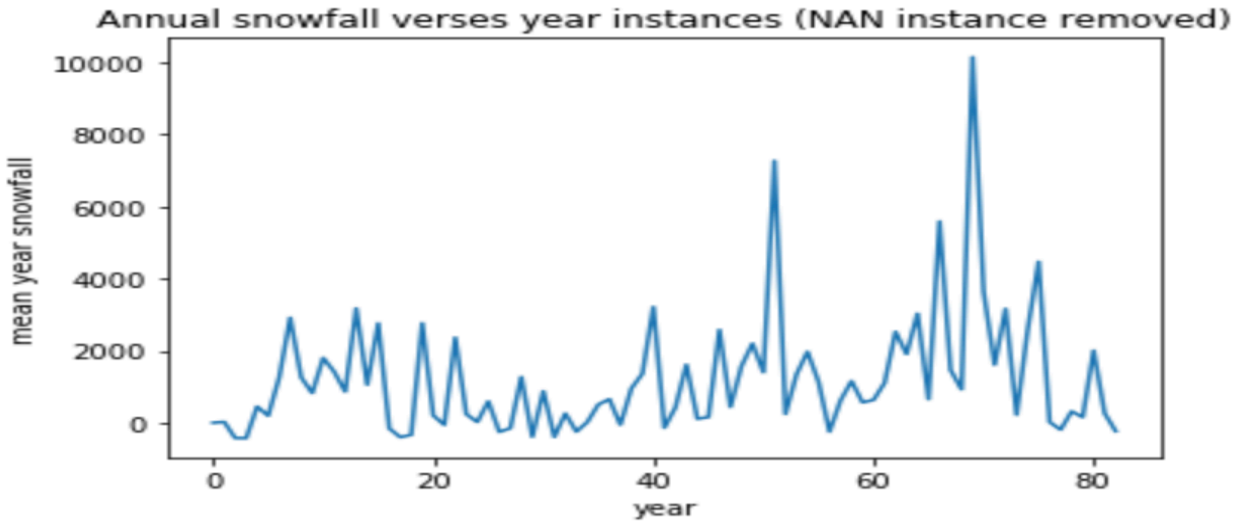


Figure above shows the annual snowfall of different year records.

It is obvious that the annual snowfall has some relations with the variation of year.

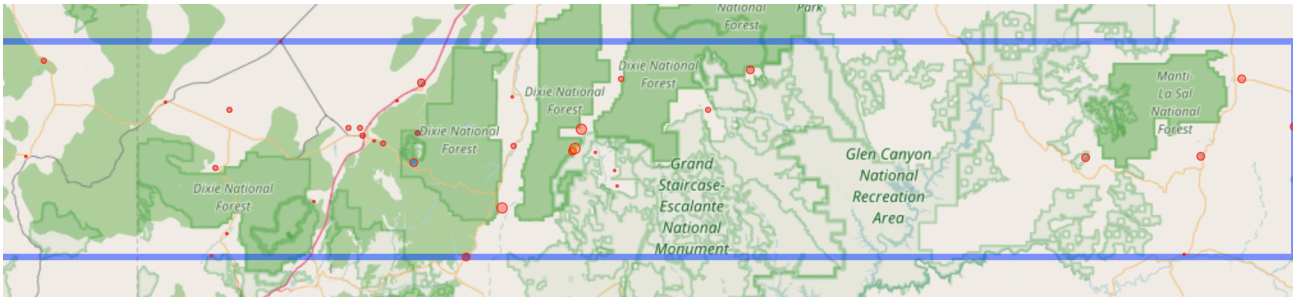


Figure above shows the distribution of observation stations on a map.

From the map, we can see that stations are distributed spatial dispersedly. Besides, topography varies from station to station, and different elevations for different stations can also significantly influence the snowfall.

3.3.2 Statistical Analysis

Since the first 4 eigen-vectors of SNWD contains more than 85% information, here we choose these 4 coeffs to represent SNWD.

Then we build tables with year as row label, station as column label and coeff as item in the table.

Next, we calculate the mean square of the table as an initial base.

We then compute the MS before and after subtracting either the row average (station mean) or the column average (year mean).

Situations	MS	Fraction explained
coeff_1 Initial	9109409.61	
coeff_1 Removing mean by station	2063631.97	77.3%
coeff_1 Removing mean by year	6202867.92	31.9%
coeff_2 Initial	1213821.00	
coeff_2 Removing mean by station	720944.12	40.6%
coeff_2 Removing mean by year	679848.79	44.0%
coeff_3 Initial	726612.86	
coeff_3 Removing mean by station	549688.04	24.3%
coeff_3 Removing mean by year	437566.47	39.8%
coeff_4 Initial	406484.22	
coeff_4 Removing mean by station	344546.79	15.2%
coeff_4 Removing mean by year	242599.54	40.3%

According to the table above, for coeff_1, we can conclude that locations of observation stations have more effects on snowfall than that of different years.

This conclusion is reasonable for coeff_1 because it represents the overall snowfall situations.
And station locations contain information about elevation, topography, vegetation etc, which are key influential factors on snowfall.
On the contrary, the annual snowfall among stations are not likely to vary a lot in different year.

For coeff*2,3,4, the effects are weaken, which means variation of year-by-year explains more than variation of station-to-station. And this is possibly because that coeff*2,3,4 do more about timing factors on snowfall instead of location factors.

3.4 Analyzing Residuals to find out relations between stations

From 3.3, we found that the spatial factors like observation stations have some kind of impacts on the snowfall.

In this section, we are going to analyze the relationship between different stations.

Instead of finding correlations between the amount of snow on the same day in different stations, we choose to compare whether or not it snowed on the same day in different stations.

3.4.1 Statistical Analysis

Null Hypothesis: the snowfall in two stations is independent.

Alternative Hypothesis: the snowfall in two stations is not independent.

To test the null hypothesis, we compute the probability associated with the number of overlaps under the null hypothesis.

We first calculate the probability that the number of overlap days (both stations snowed) equals to l which is a random variable.

By calculating probability of every two stations, we build a probability matrix.

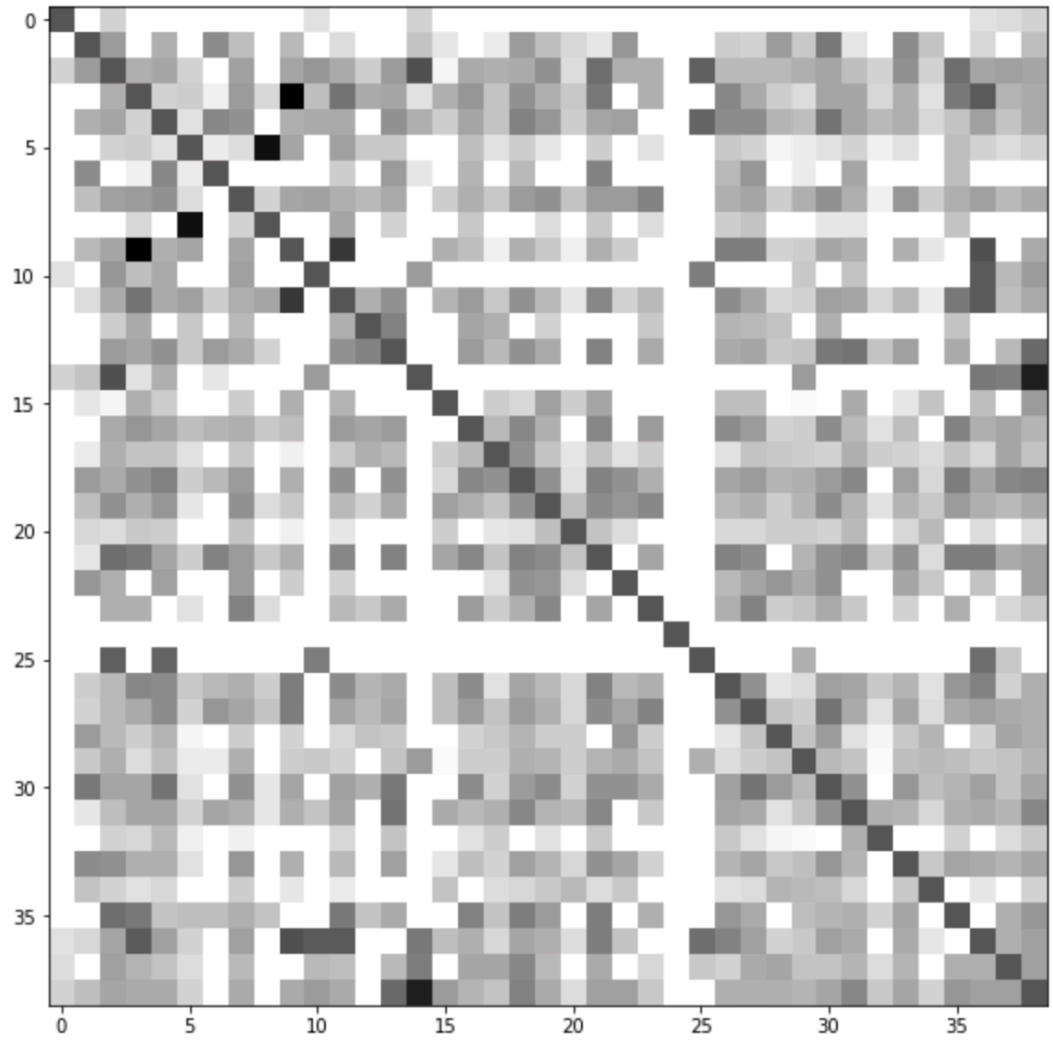
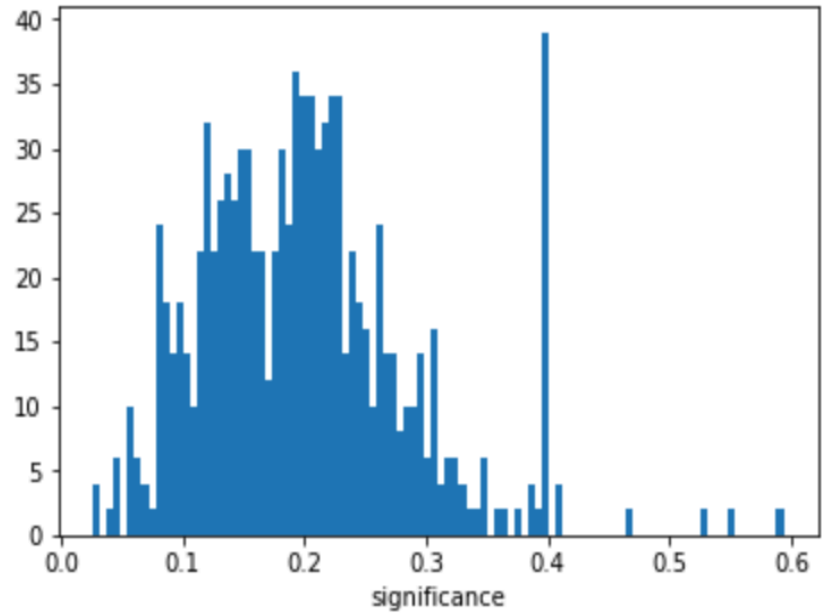


Figure above shows the visualization of the probability matrix.

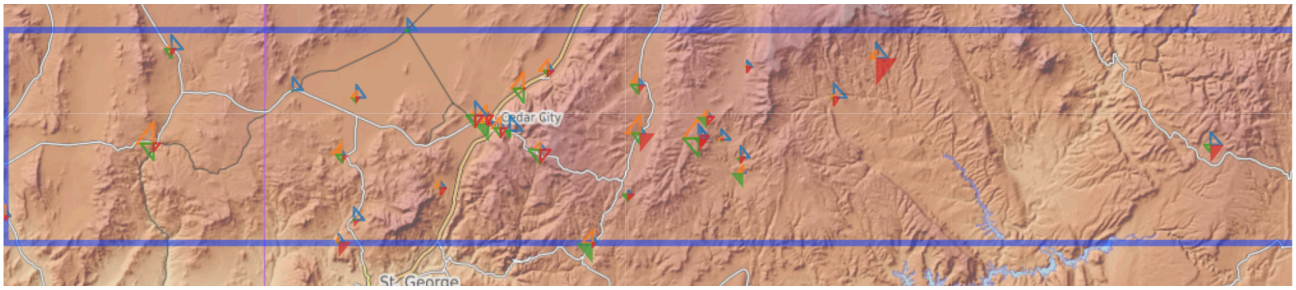
From the figure we can find out some boxes with dark color which means the corresponding two stations are strongly correlated, and the boxes with white color indicate that two stations are independent to some degree.



At 0.05 significance, for about 11% stations the null hypothesis are not rejected and for the remaining 89% stations, the null hypothesis is rejected. In other words, only a small number of stations are independent in 0.05 significance, and the majority of stations are correlated with each other.

3.4.2 Group of correlated stations

Since we have concluded from the above analysis that the first 4 eigen-vectors of SNWD explain over 85% variance, we can using these 4 vectors to group correlated stations. The following map is topography with 4 coeffs on every station.



We first utilize PCA to find out the first 4 eigen-vectors, and then using these vectors to calculate the probability matrix of every two stations overlapping snow days.

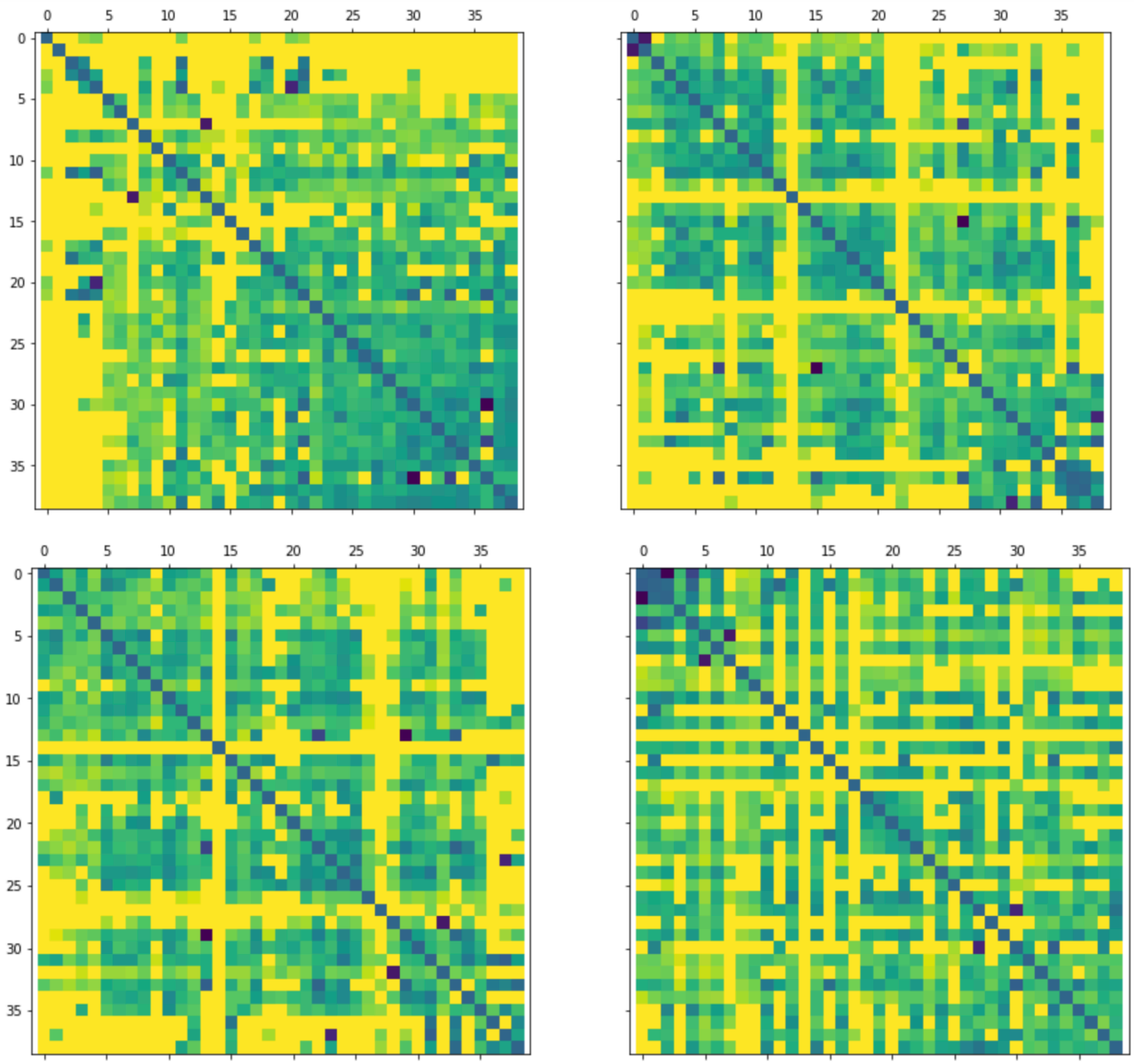


Figure above shows the visualization of the grouped probability matrix.

After we reordered the rows and columns using one of the eigen-vectors, we can find some grouping patterns. For example, the bottom right corner of the first matrix. Stations at position 20-38 are clearly strongly correlated with each other. And in the second matrix, there are clearly 9 regions in which the stations are correlated with each other.