In [1]:
```
L=!ls r_figures/
for l in L:
    print "![%s](r_figures/%s)"%(l,l)
```

```
![5__maps_using_iPyLeaflet.jpg](r_figures/5__maps_using_iPyLeaflet.jpg)
![bar.jpg](r_figures/bar.jpg)
![c1_r_a.png](r_figures/c1_r_a.png)
![c1_r_d.png](r_figures/c1_r_d.png)
![c2_r_a.png](r_figures/c2_r_a.png)
![c2_r_d.png](r_figures/c2_r_d.png)
![c3_r_a.png](r_figures/c3_r_a.png)
![c3_r_d.png](r_figures/c3_r_d.png)
![Climate_Boston_-_Massachusetts_and_Weather_averages_Boston.jpg](r_figures/Climate_Boston_-_Massachusetts_and_Weather_averages_Boston.jpg)
![location.png](r_figures/location.png)
![Map of Average snow depth.jpg](r_figures/Map of Average snow depth.jpg)
![northdakota.png](r_figures/northdakota.png)
![percipitation.png](r_figures/percipitation.png)
![PRCP_ND.png](r_figures/PRCP_ND.png)
![PRCP.png](r_figures/PRCP.png)
![sndepth.png](r_figures/sndepth.png)
![snownew.png](r_figures/snownew.png)
![SNOW.png](r_figures/SNOW.png)
![snowVE.png](r_figures/snowVE.png)
![SNWD_mean_eigs.png](r_figures/SNWD_mean_eigs.png)
![SNWD.png](r_figures/SNWD.png)
![SNWD_res_1_CDF.png](r_figures/SNWD_res_1_CDF.png)
![SNWD_res_2_CDF.png](r_figures/SNWD_res_2_CDF.png)
![SNWD_res_3_CDF.png](r_figures/SNWD_res_3_CDF.png)
![tmin,tmax.png](r_figures/tmin,tmax.png)
![TMIN,TMAX.png](r_figures/TMIN,TMAX.png)
![TminVE.png](r_figures/TminVE.png)
![TOBS.png](r_figures/TOBS.png)
![tobs_prcp.png](r_figures/tobs_prcp.png)
![VarExplained1.png](r_figures/VarExplained1.png)
![VarExplained2.png](r_figures/VarExplained2.png)
```

In [2]:
```
!open r_figures/TMIN,TMAX.png
```

```
Couldn't get a file descriptor referring to the console
```

North Dakota Weather Analysis

This is a report on the historical analysis of weather patterns in an area that approximately overlaps the area of the state of North Dakota.

The data we will use here comes from NOAA (https://www.ncdc.noaa.gov/). Specifically, it was downloaded from This FTP site (ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/).

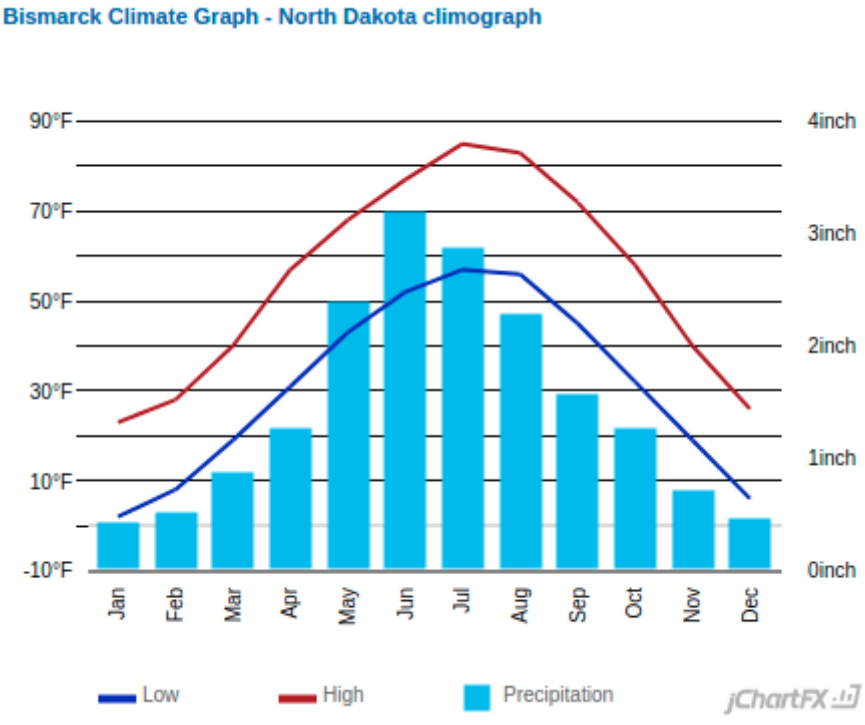
We focused on six measurements:

- **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.

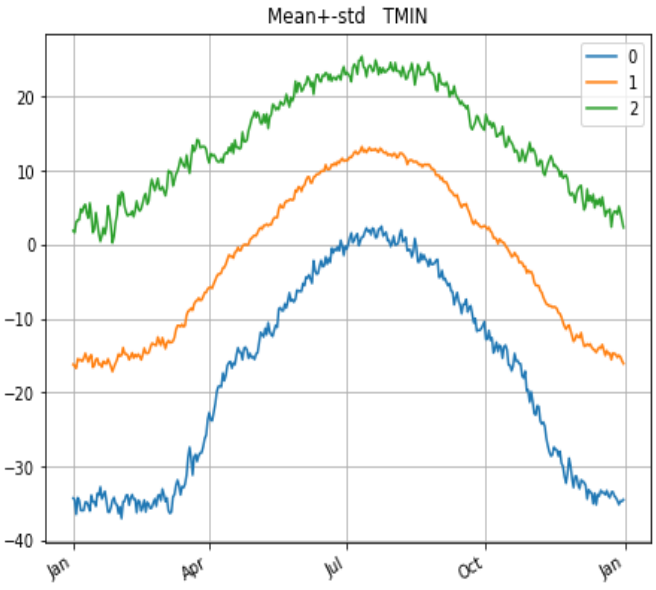
Sanity-check: comparison with outside sources

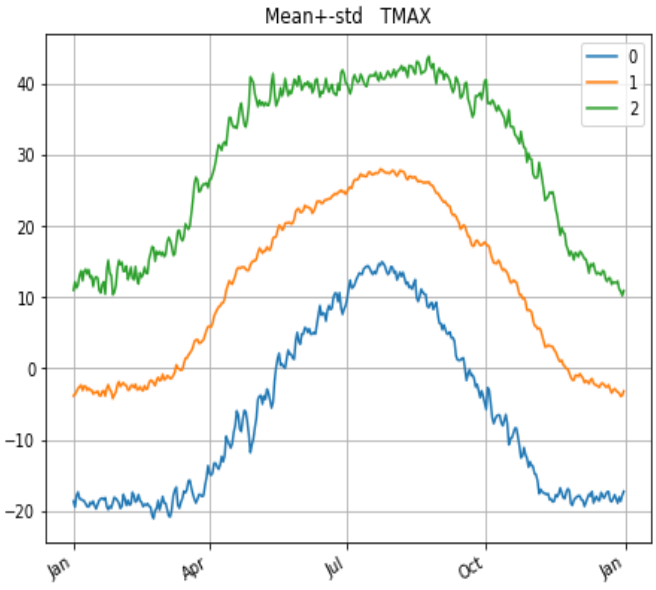
We start by comparing some of the general statistics with graphs that we obtained from a site called US Climate Data (http://www.usclimatedata.com/climate/boston/massachusetts/united-states/usma0046) The graph below shows the daily minimum and maximum temperatures for each month, as well as the total precipitation for each month.

An interesting point to note here is that almost half of the stations used, are located in Canada. In addition to the weather of North Dakota, we are also assessing the weather of a small part of Canada.

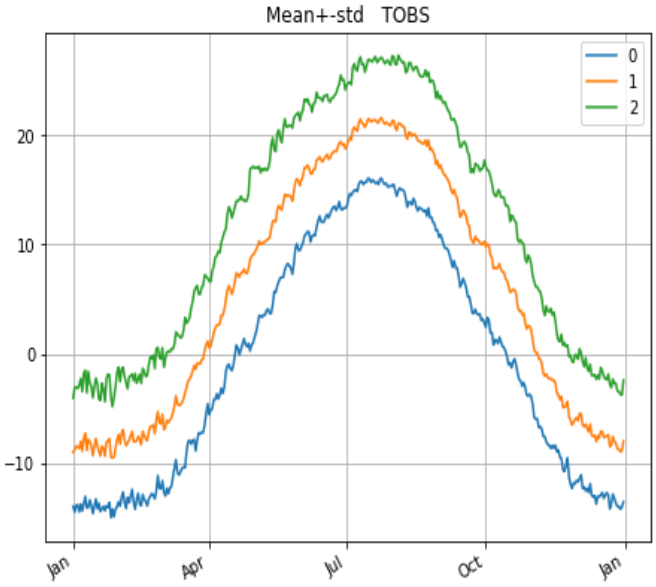


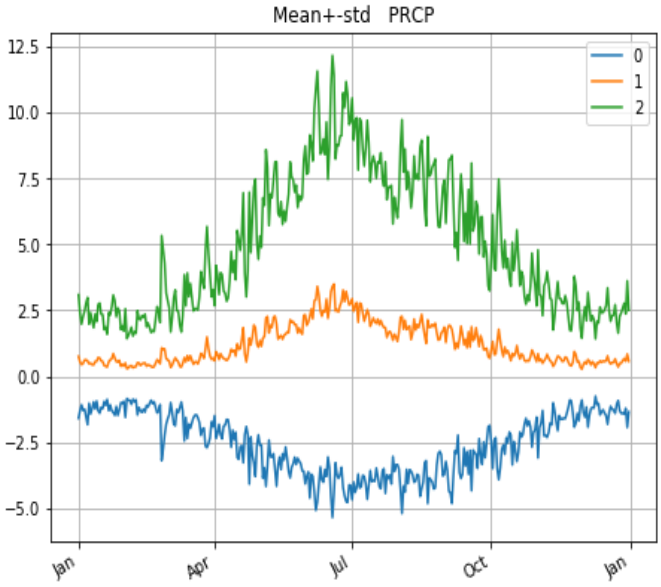
We see that the min and max daily temperature almost agree with the ones we got from our data, once we translate Fahrenheit to Centigrade. Thus proving the authenticity of our results.





To compare the precipitation we need to translate millimeter/day to inches/month. According to our analysis, the peak time of rainfall is in between April to October. The average rainfall during the month of June-July is around 2.5 mm/day which translates to about 3 Inches per month. According to US-Climate-Data the average rainfall is closer to 3 inch per month.

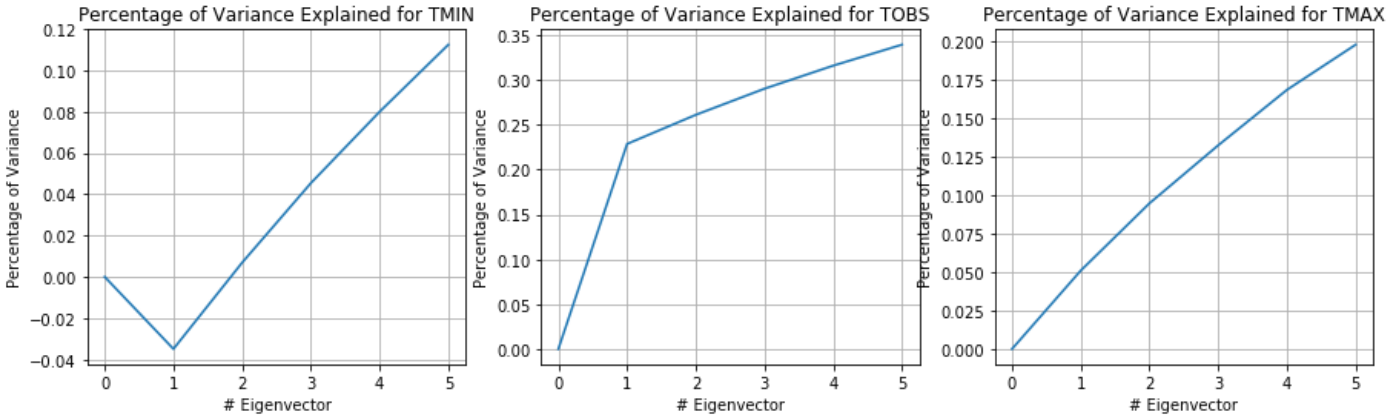




PCA analysis

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

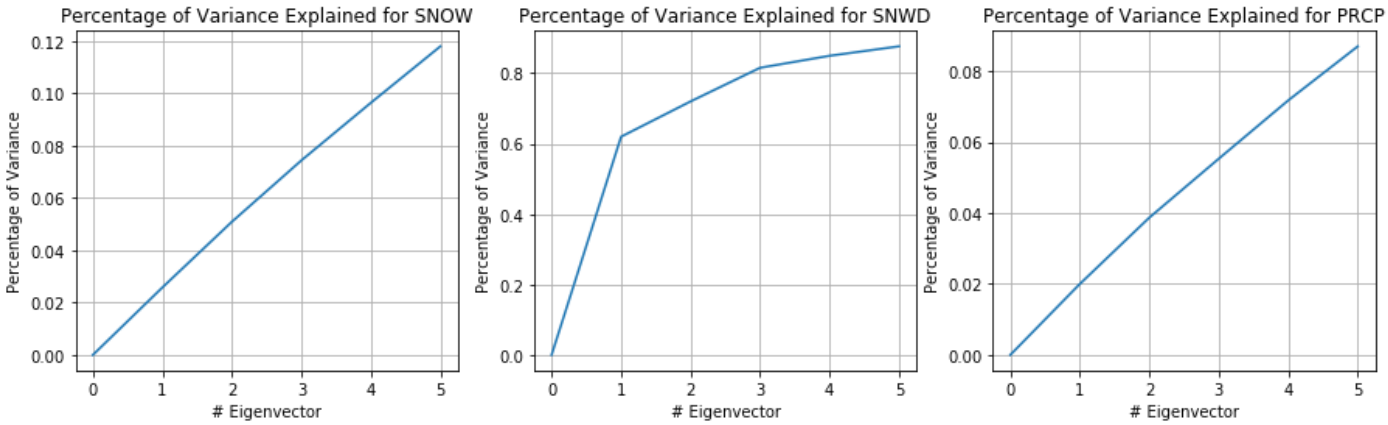
Percentage of variance explained.



We see that the top 5 eigen-vectors explain 11% of variance for TMIN, 35% for TOBS and 20% for TMAX.

We conclude that of the three, TOBS is best explained by the top 5 eigenvectors. This is especially true for the first eigen-vector which, by itself, explains 23% of the variance.

Here in TMIN we see a negative eigenvalues. In most cases, a covariance matrix is symmetric and positive semidefinite, which means all its eigenvalues are non-negative. But if somehow we get a non-psd result when calculating the covariance matrix, we can obtain a nearest psd matrix.



The top 5 eigenvectors explain 0.09% of the variance for PRCP and 12% for SNOW. Both are low values. On the other hand the top 5 eigenvectors explain %90 of the variance for SNWD. This means that these top 5 eigenvectors capture most of the variation in the snow signals. Based on that we will dig deeper into the PCA analysis for snow-depth.

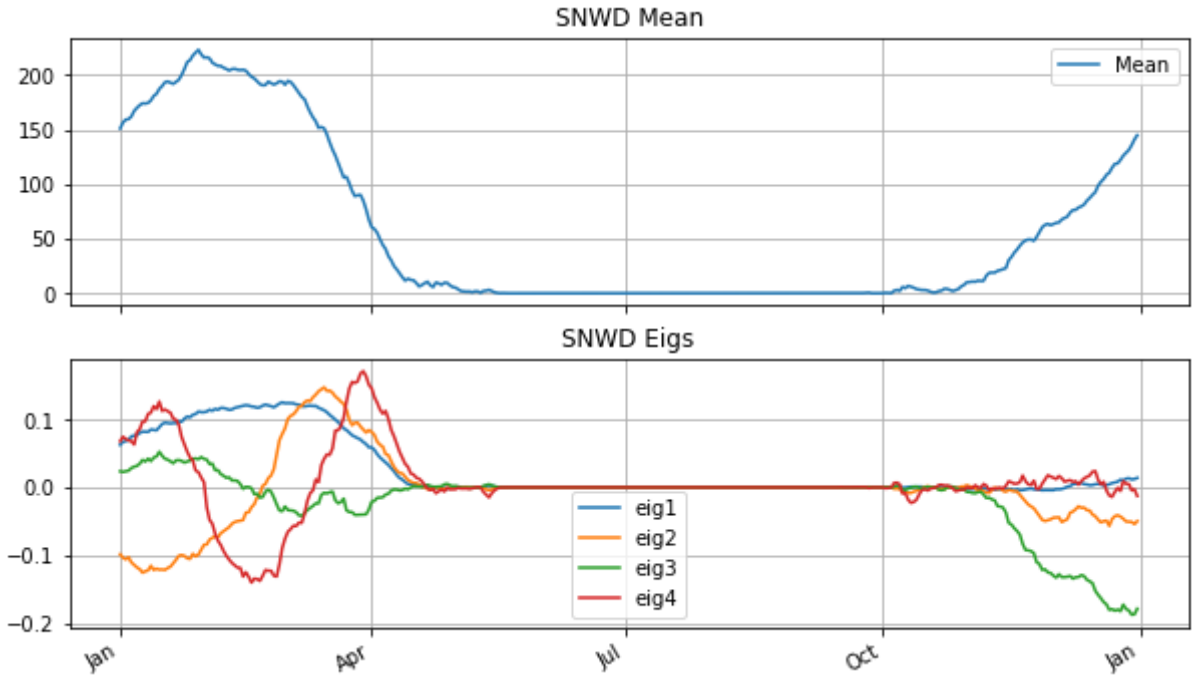
It makes sense that SNWD would be less noisy than SNOW. That is because SNWD is a decaying integral of SNOW and, as such, varies less between days and between the same date on different years.

Analysis of snow depth

We choose to analyze the eigen-decomposition for snow-depth because the first 3 eigen-vectors explain 80% of the variance. This makes the data rich.

First, we graph the mean and the top 3 eigen-vectors.

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



Next we interpret the eigen-functions. The first eigen-function (eig1) has a shape very similar to the mean function. The main difference is that the eigen-function is close to zero during october-december while the mean is not. The interpretation of this shape is that eig1 represents the overall amount of snow above/below the mean, but without changing the distribution over time.

eig2, eig3 and eig4 are similar in the following way. They all oscilate 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).

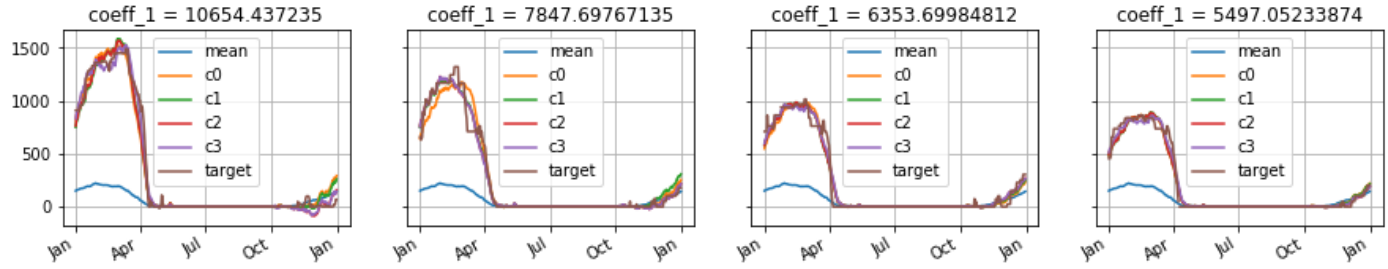
They can be interpreted as follows:

- **eig2:** less snow in jan - feb, more snow in march-april.
- **eig3:** more snow in jan - feb, less snow in march-april.
- **eig4:** more snow in jan and april (little amount of snow in nov-dec), less snow in feb - march.

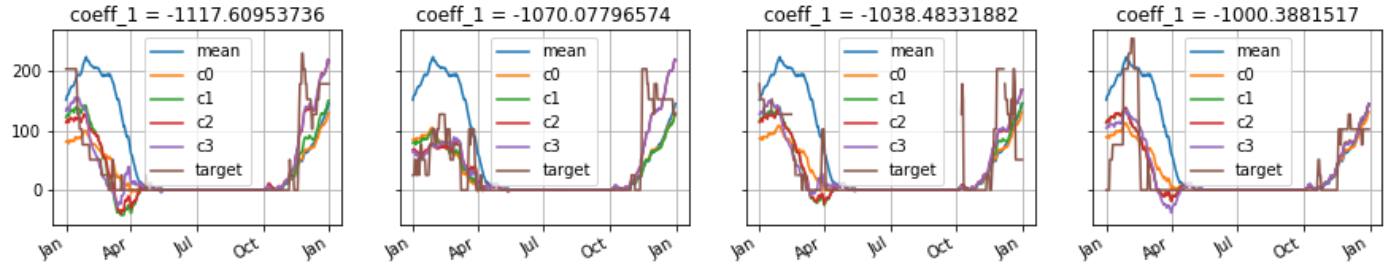
Examples of reconstructions

Coeff1

Coeff1: most positive



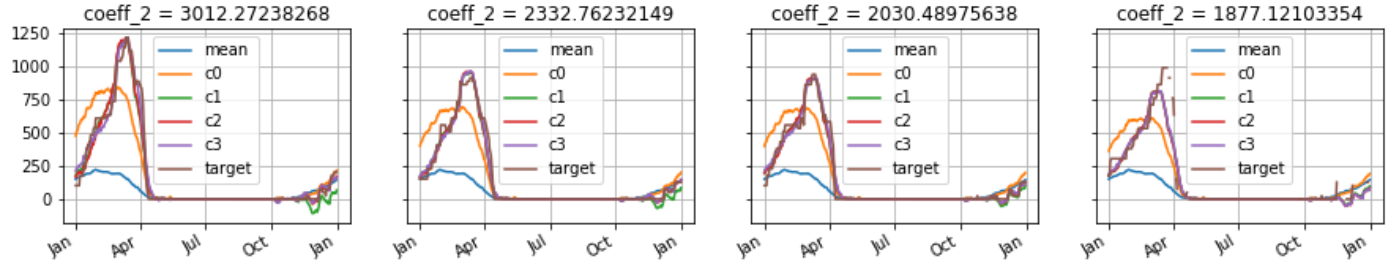
Coeff1: most negative



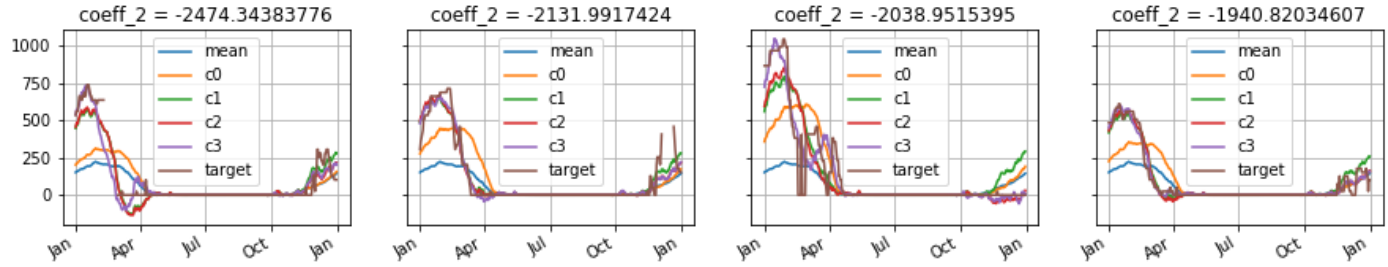
Large positive values of coeff1 correspond to more than average snow. Low values correspond to less than average snow.

Coeff2

Coeff2: most positive



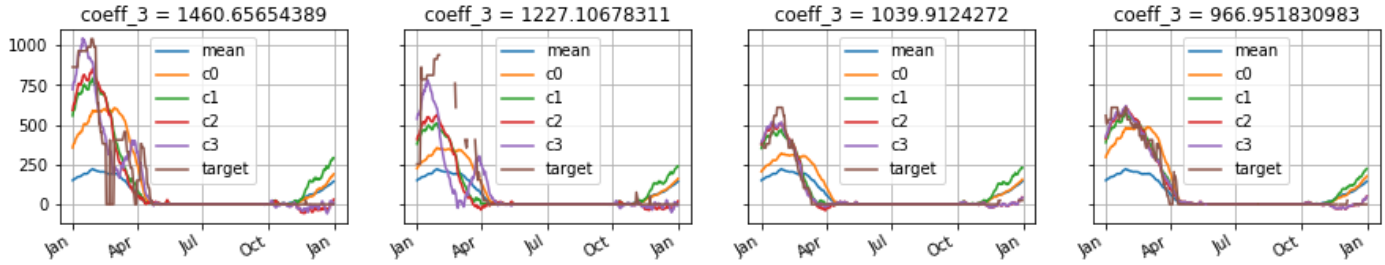
Coeff2: most negative



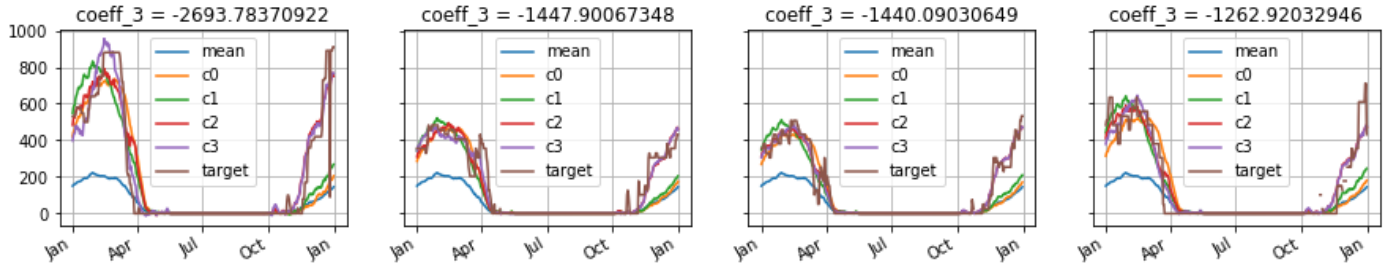
Large positive values of coeff2 correspond to a late snow season (most of the snowfall is after mid feb. Negative values for coeff2 correspond to an early snow season (most of the snow is before mid-feb.

Coeff3

Coeff3: most positive



Coeff3: most negative



Large positive values of coeff3 correspond to a snow season with two spikes: one in the start of january, the other at the end of february. Negative values of coeff3 correspond to a season with a single peak at the end of Jan.

The variation in the timing of snow is mostly due to year-to-year variation

In the previous section we see the variation of Coeff1, which corresponds to the total amount of snow, with respect to location. We now estimate the relative importance of location-to-location variation relative to year-by-year variation.

These are measured using the fraction by which the variance is reduced when we subtract from each station/year entry the average-per-year or the average-per-station respectively. Here are the results:

coeff_1
total MS = 1972884.43
MS removing mean-by-station= 1506065.51, fraction explained= 23.66
MS removing mean-by-year = 742903.49, fraction explained= 62.34

coeff_2
total MS = 417468.57
MS removing mean-by-station= 381541.42 , fraction explained= 8.60
MS removing mean-by-year = 151462.16 , fraction explained= 63.71

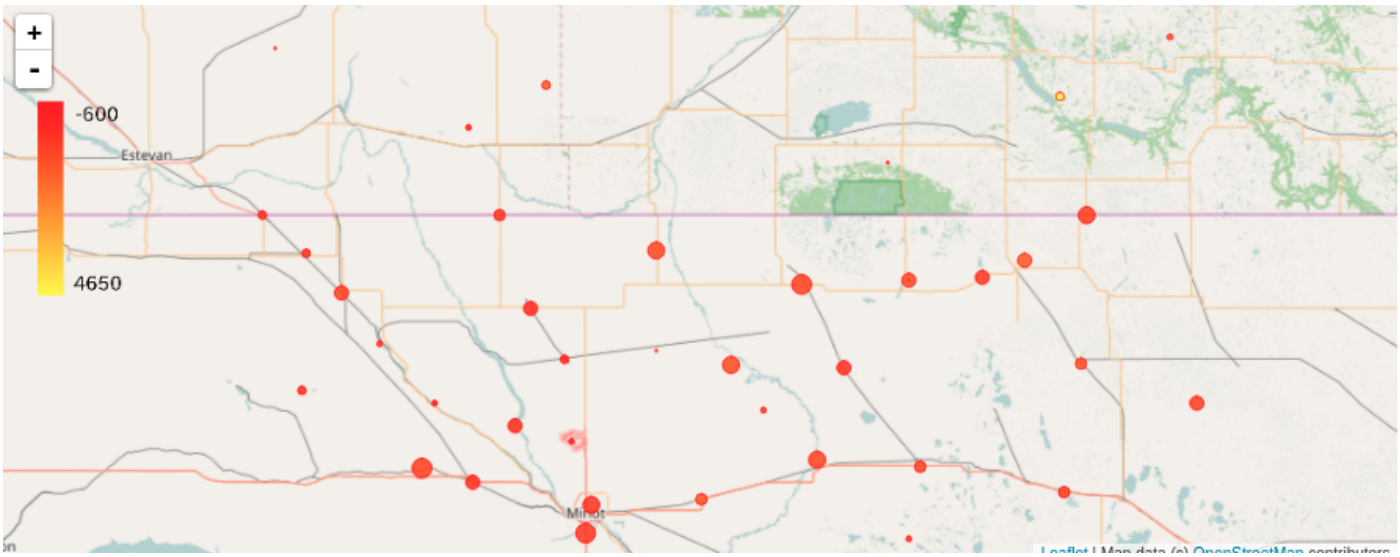
coeff_3
total MS = 351831.52
MS removing mean-by-station= 326485.51, fraction explained= 7.20
MS removing mean-by-year = 92637.88, fraction explained= 73.66

We see that the variation by year explains more than the variation by station. However this effect is a little weaker consider coeff_1, which has to do with the total snowfall, vs. coeff_2,3 which, as we saw above have to do with the timing of snowfall. We see that for coeff_2,3 the stations explain 7-8% of the variance while the year explaines 60-70%.

Geographical Visual Representation of Snow Depth

In the following map, the stations which provided the data for snow depth is marked with the circles. The size of the circle represents the amount of measurement data for snow depth received from the stations.

As the legend specifies, the shade of the color represents the average coefficient of the first eigen vector. Where the red color specifies less coefficient and yellow collow specifies more coefficient.

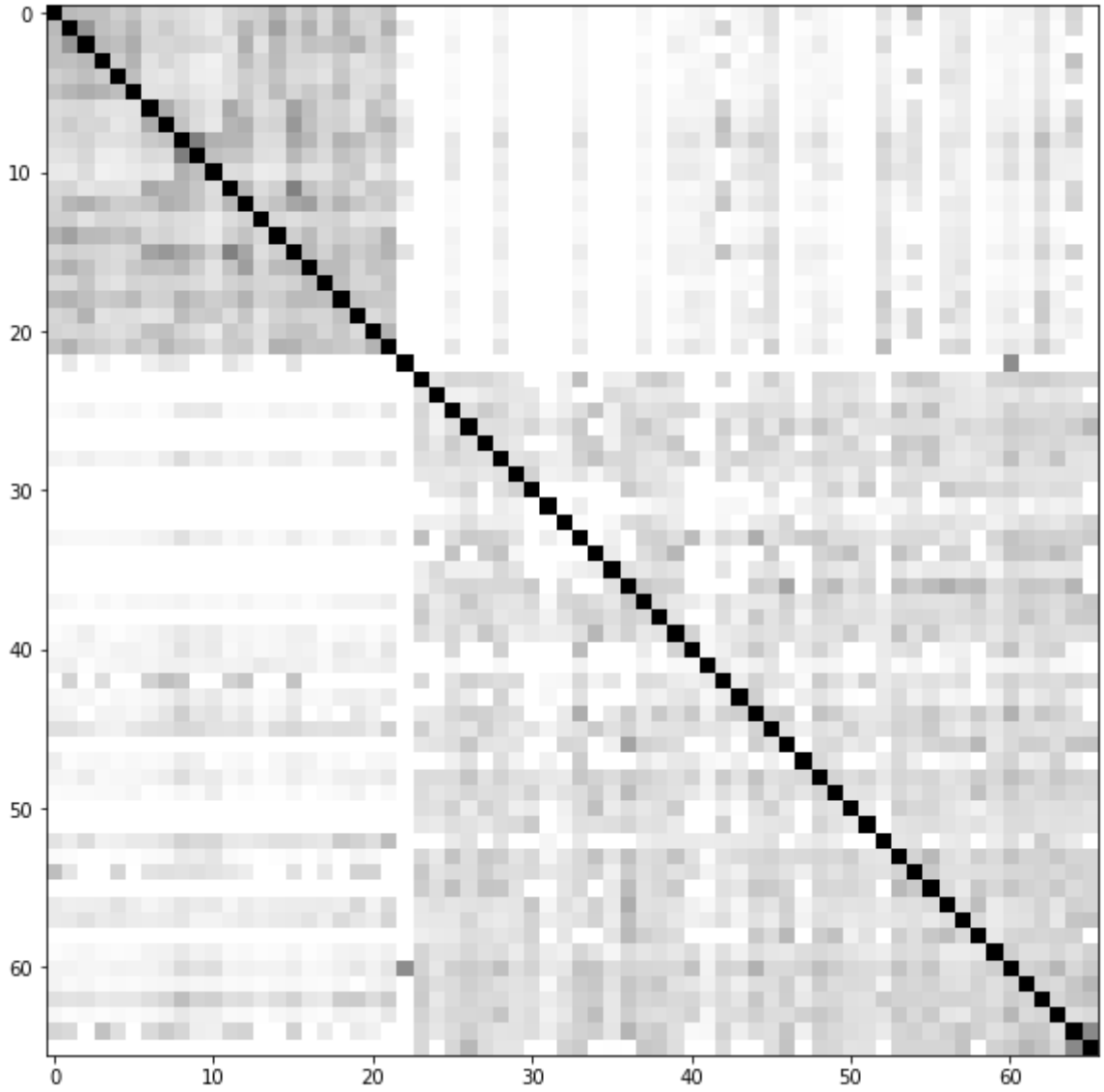


Measuring statistical Correlation for the Snow days in dependency matrix.

The matrix above shows, for each pair of stations, the normalized log probability that the overlap in snow days is particular and most of the stations are correlated.

We see immediately the first 21 stations are highly correlated with each other.

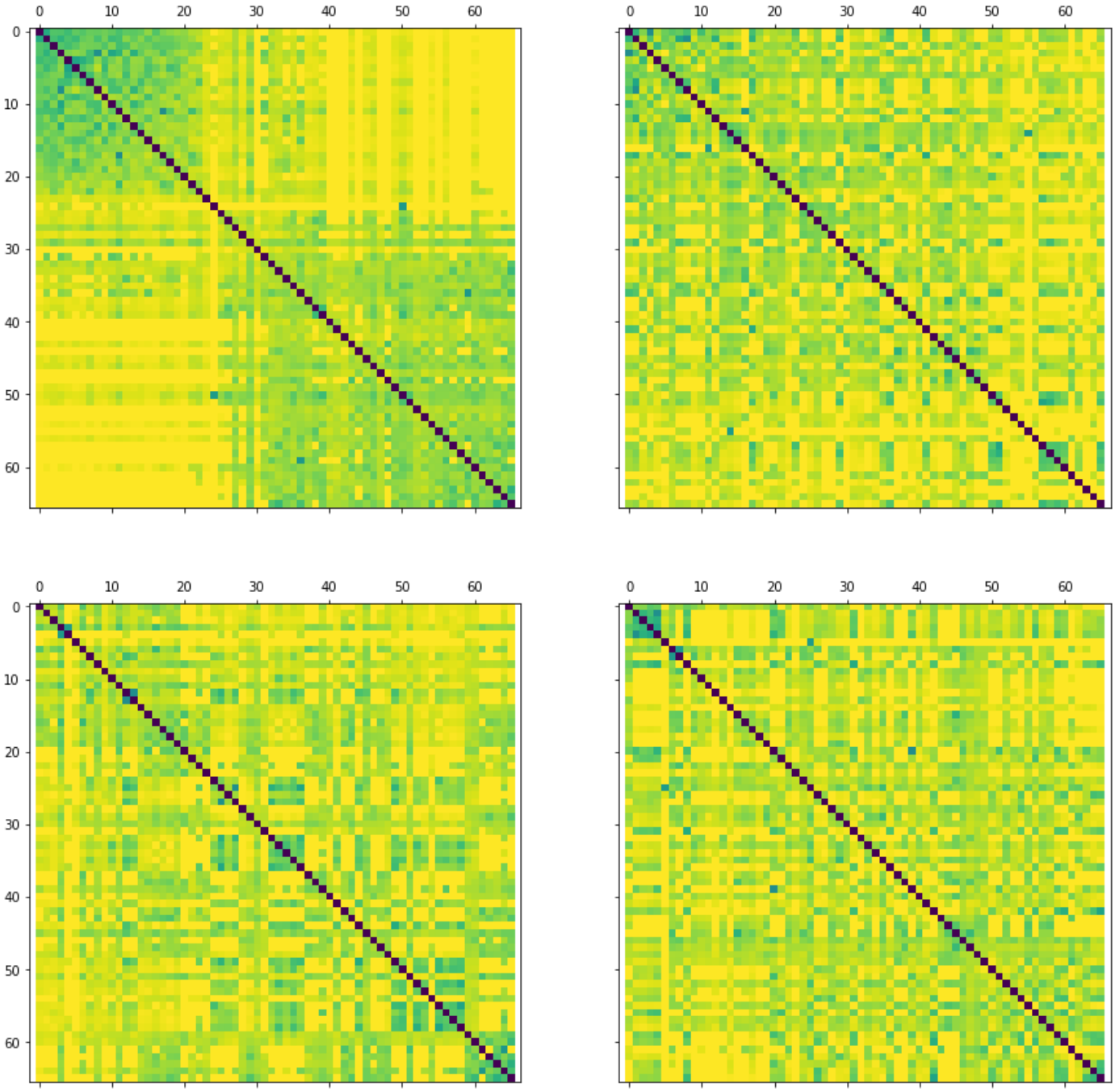
A group of very correlated stations is:
['CA005012545', 'USC00326025', 'CA004015045', 'USW00094011', 'CA005012960', 'CA005010QFQ', 'USC00323963', 'USC00322385', 'USC00327704', 'USC00328764', 'USC00323686', 'USC00328627', 'CA00501A7AR', 'USC00320729', 'USC00325993', 'USC00327655', 'CA005010480', 'USC00320626', 'CA005010485', 'CA005010547', 'USC00322304']



Explanation and possibe extensions

When we reorder the rows and columns of the matrix using one of the eigenvectors, the grouping of the stations becomes more evident. For example, consider the upper left corner of the scond matrix (The upper left one). The stations at positions 0-21 are clearly strongly correlated with each other considering snowfall.

This type of organization is called **Block Diagonal** and it typically reveals important structure such as grouping or clustering pf strongly related stations.



Geographical Representation of Correlation of Stations using 4 coefficients on Snow Data.

We represented the coefficients of eigenvectors using triangles. Size of the triangles on each side represents the magnitude of the coefficients.

From the following graph, since the stations are correlated, we can infer that the value of a coefficient almost follows the same pattern for all stations in that local region. Solid triangles are positive coefficients and Hollow triangles are negative coefficients.