Massacussets Weather Analysis

This report is analysis the weather pattern overlapping with the region of Minnesota (Massacussets).

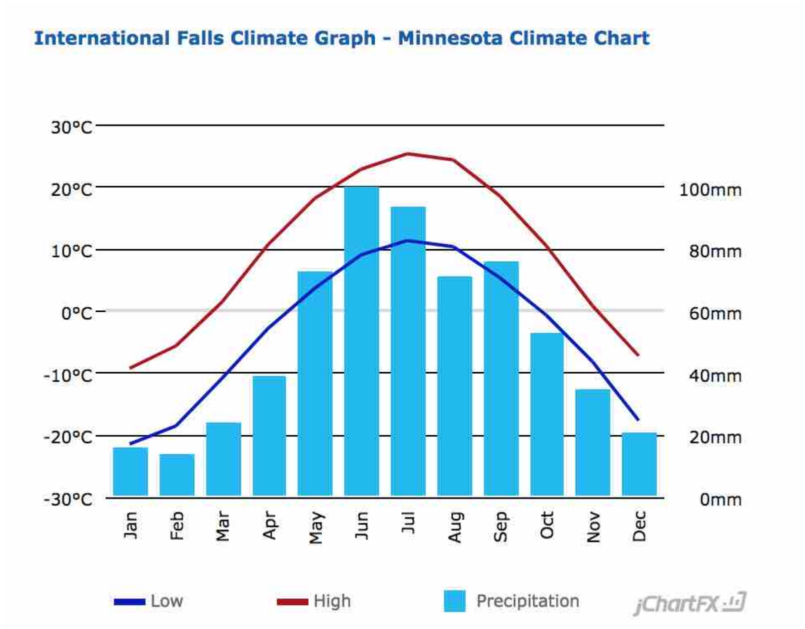
The data we will use here comes from NOAA (https://www.ncdc.noaa.gov/). Specifically, it was downloaded from This FTP site.

We focused on six measurements:

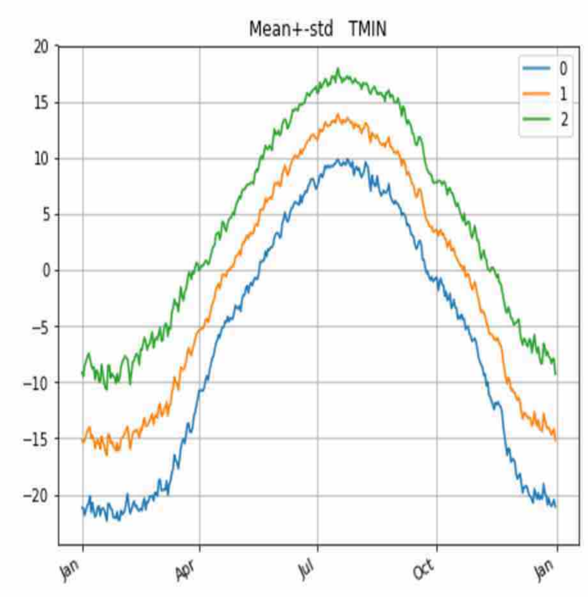
- **TMIN:** the daily minimum temperature (tenths of degrees C).
- **TMAX:** the daily maximum temperature (tenths of degrees C).
- **TOBS:** Temperature at the time of observation (tenths of degrees C).
- **PRCP:** Daily Percipitation (in mm).
- **SNOW:** Daily snowfall (in mm).
- **SNWD:** The depth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

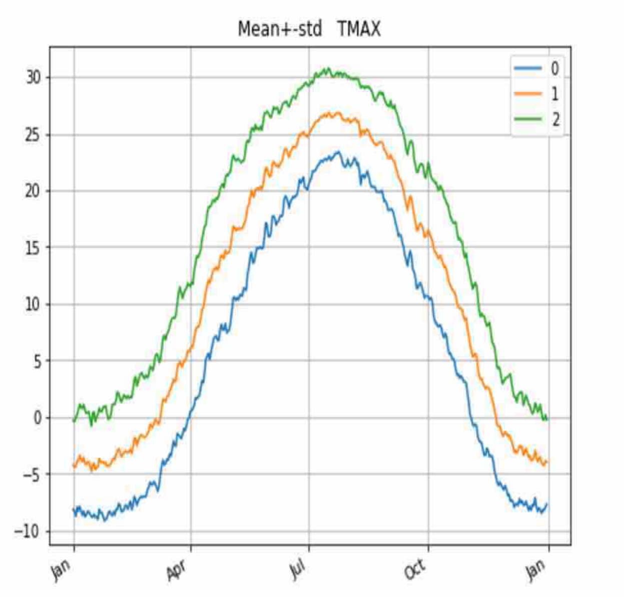
Sanity-check: comparison with outside sources

We sanity-checked the data from US Climate Data (http://www.usclimatedata.com/climate/international-falls/minnesota/united-states/usmn0376) The graph below shows the daily minimum and maximum temperatures for each month, as well as the total precipitation for each month.

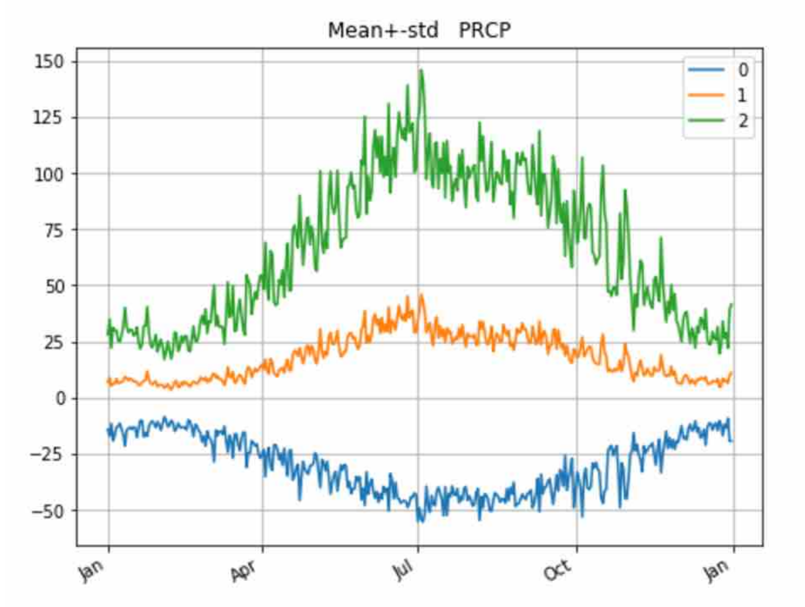


As can be seen below, min and max daily temperature approximately matches the ones we got from our data.



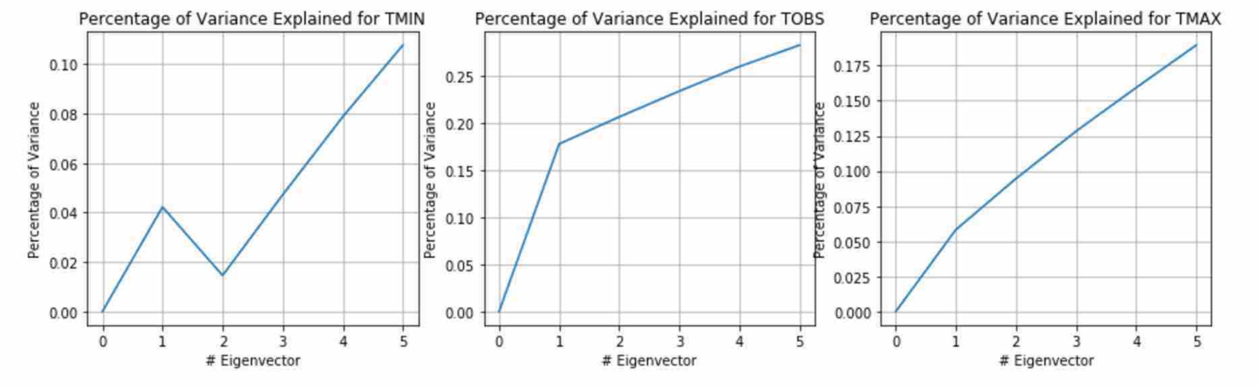


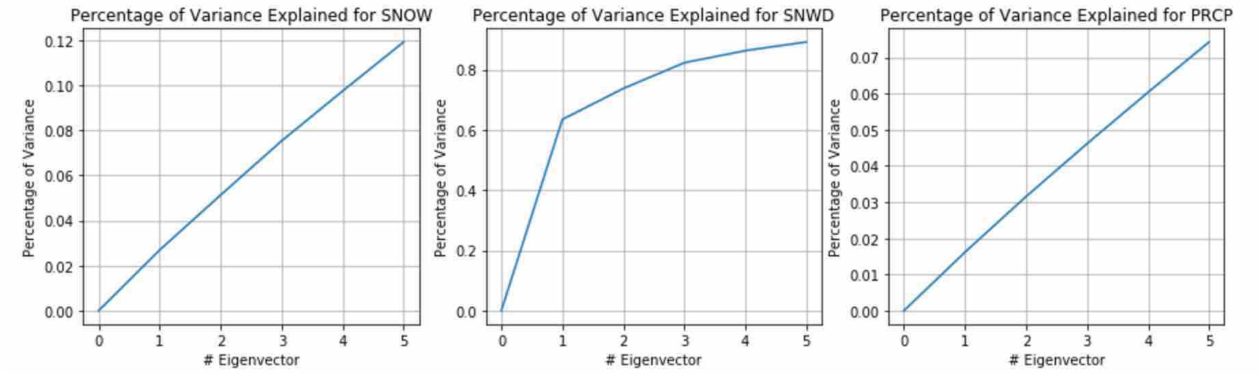
The precipitation figures also match to some approximation as can be seen the figure below:



PCA analysis

Next we checked the percent of variances explained using the top 5 eigen-values for each type of measurement:





Note that the dip in case of TMIN should not have been there as more the number of eigen values used, more should be the variance explained.

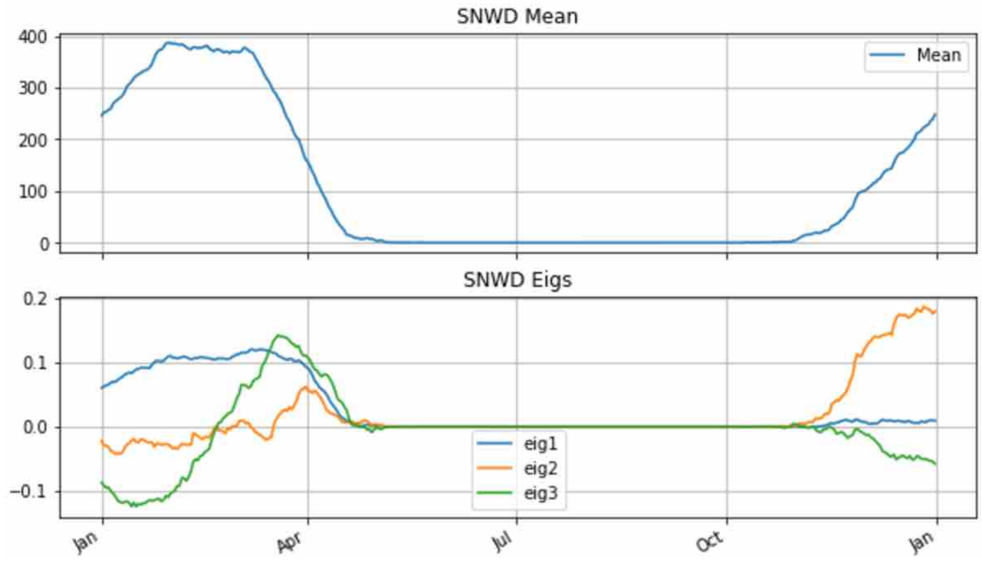
The graphs clearly show that SNWD is best explained using top 5 eigen values while PRCP is least explained using its corresponding top eigen values. So, we will next try to analyze these two measurements further.

Snow Depth Analysis

We choose to analyze the eigen-decomposition for snow-depth because the first 4 eigen-vectors explain 80% of the variance.

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

We observe that the snow season is from mid-november to the end of april, where the Feb-March marks the peak of the snow-depth.



Possible eigen-function interpretations:

Although, the first eigen-function (eig1) appears very similar to the mean function, the eigen-function is close to zero during november-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 and eig3 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). This can be verified from the 'variance explained' figure for 'SNWD' as eig1 explains more than 60% of variance in data.

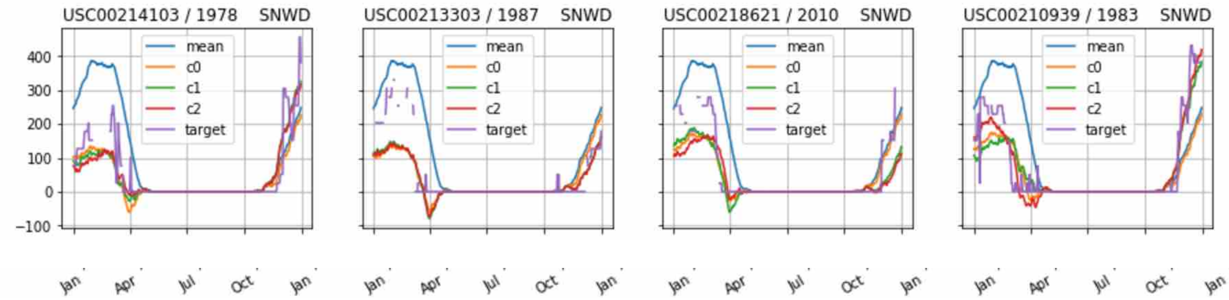
They can be interpreted as follows:

- **eig2:** less snow in dev - mid feb, more snow in mid feb-march.
- **eig3:** less snow in jan, more snow in feb, slightly more snow in march.

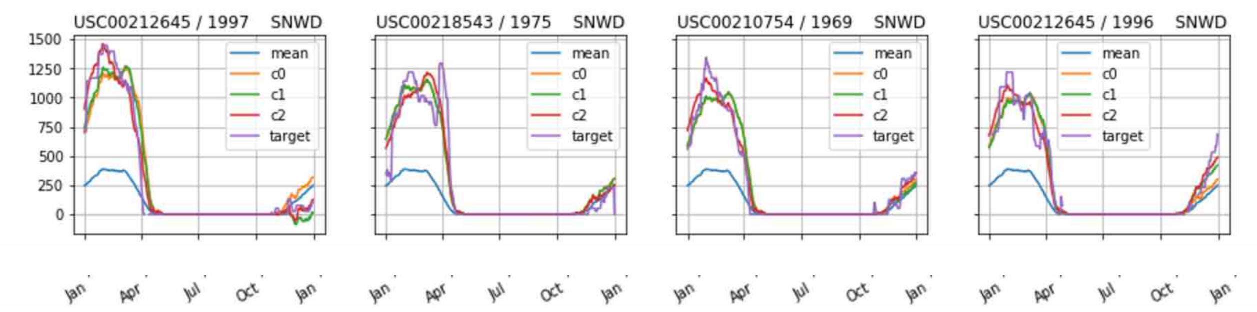
In the figures below we analyze the extreme values of coefficients of the top three eigen-values when performing reconstructions:

Examples of reconstructions

```
Coeff_1 most negative:
```

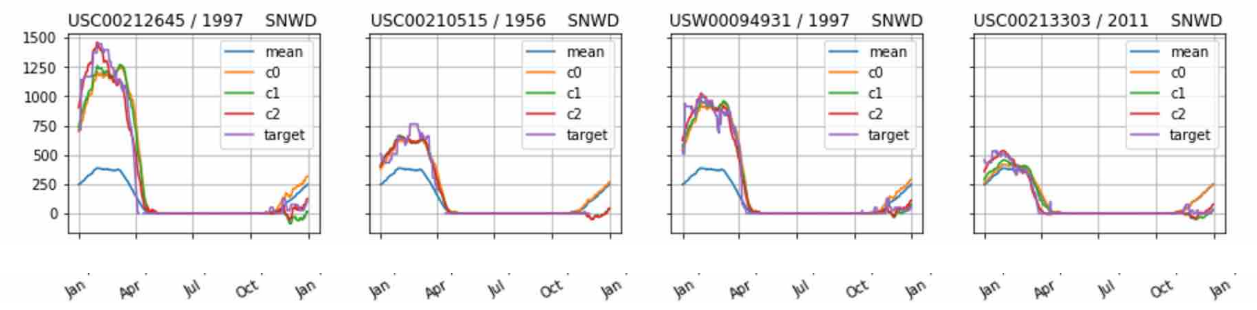


```
Coeff_1 most positive:
```

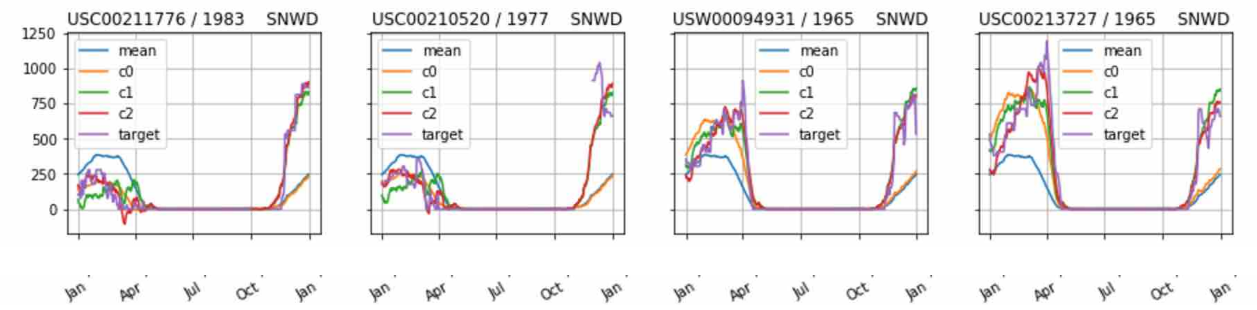


Large positive values of coeff1 correspond to more than average snow. Low values correspond to less than average snow.

```
Coeff_2 most negative:
```

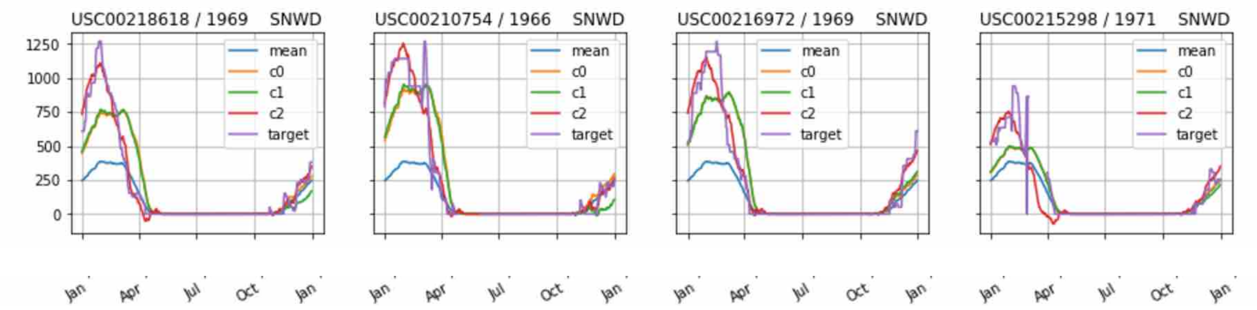


```
Coeff_2 most positive:
```

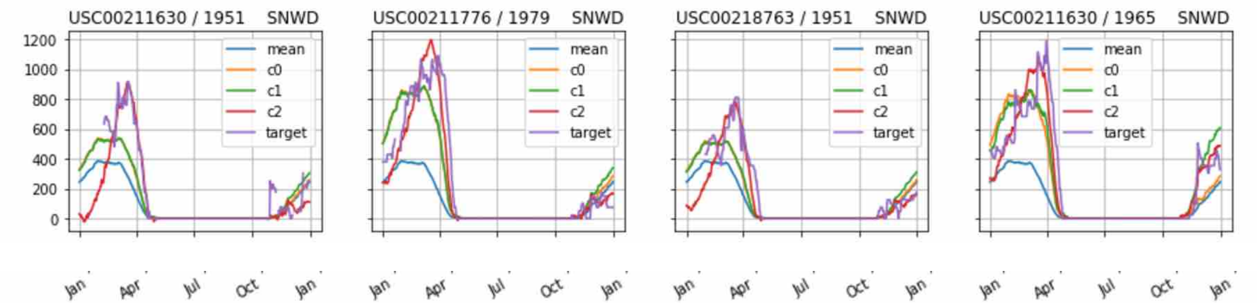


Large positive values of coeff2 also correspond to the amount of snow that is not explained by coeff_1.

```
Coeff_3 most negative:
```



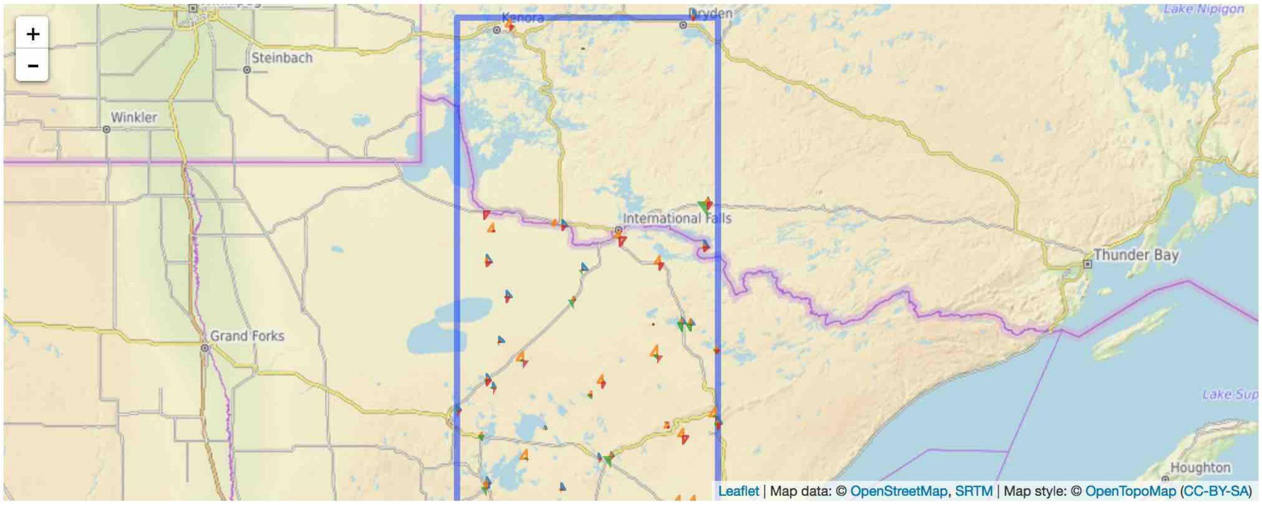
```
Coeff_3 most positive:
```



Large positive values of coeff3 correspond to a snow season which has a peak in february. Negative values of coeff3 correspond to a snow season which has a peak in january.

Geographical distribution of first 4 coefficients.

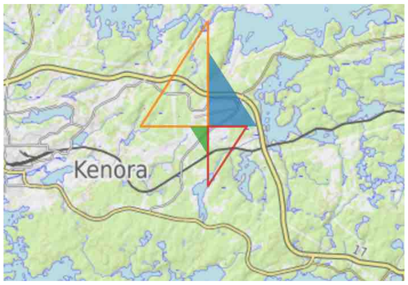
The figure below shows the stations with there corresponding coefficients (shown as triangles).



Where the notation is:

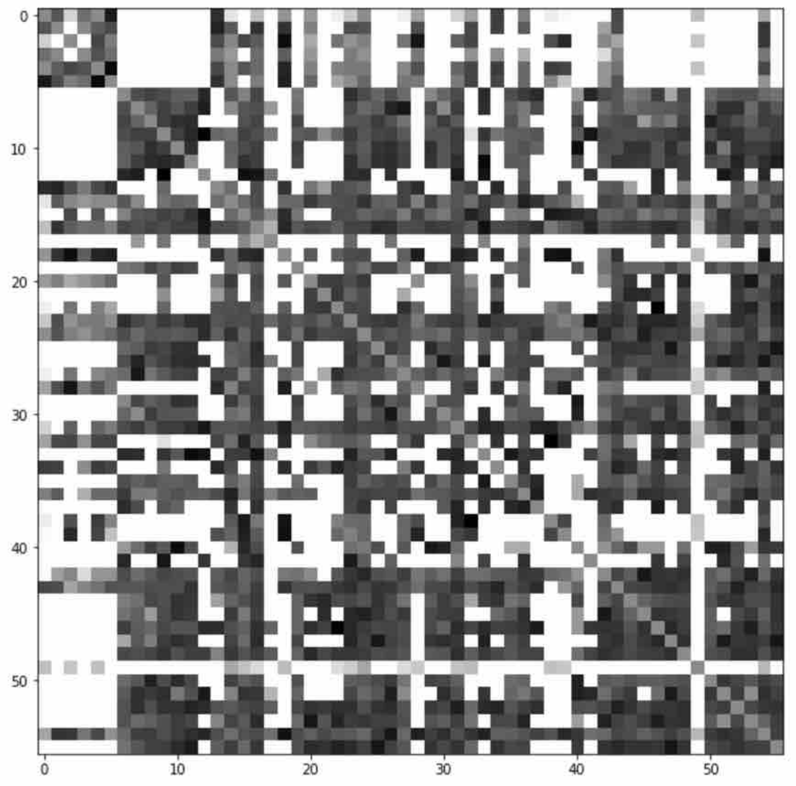
1. coeff_1 - (Blue Color)
2. coeff_2 - (Orange Color)
3. coeff_3 - (Green Color)
4. coeff_4 - (Red Color)
5. Size of triangle corresponds to the magnitude of the coefficient value.
6. Filled triangles corresponds to negative values.

Zoomed in image of one random station near Kenora:

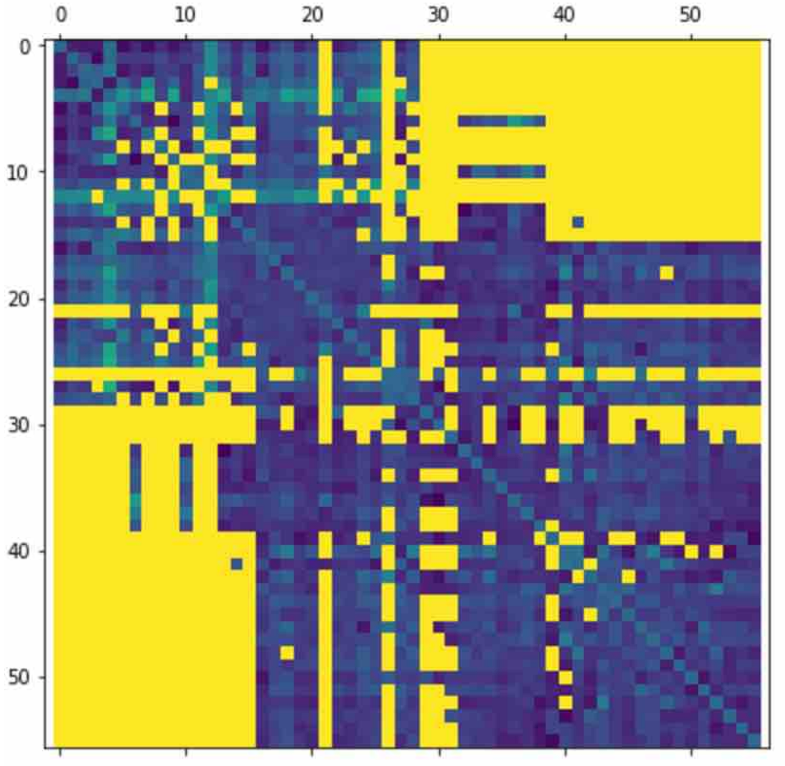


It is difficult to infer good meaningful inferences using just the maps shown above. So, we use the correlation matrix to find similarity between different stations.

Correlation matrix for 56 stations for SNWD measurements:



Correlation matrix after reordering on the basis of most common dimension (characteristic):



Darker values in the matrix represent smaller p-values which means strong evidence against the null-hyposthesis.

Clearly, only the stations from 29 to 56 appear to differ somehow from one another while most of the other pairs of stations appear to have common features (darker shades of color signifies more similarity).

The variation in the snow depth is mostly due to year-to-year variation

Below 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 RMS = 1607.08350635 RMS removing mean-by-station= 1475.61439358, fraction explained= 8.18060245454 RMS removing mean-by-year = 852.75670797, fraction explained= 46.9376230543

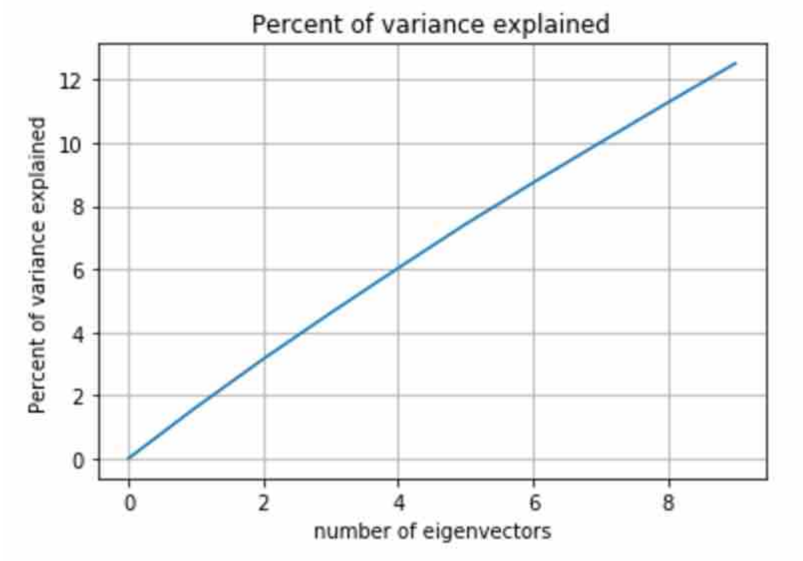
coeff_2 total RMS = 750.23614434 RMS removing mean-by-station= 732.635844818, fraction explained= 2.34596795361 RMS removing mean-by-year = 396.294511574, fraction explained= 47.177363479

coeff_3 total RMS = 710.360614167 RMS removing mean-by-station= 672.647841391, fraction explained= 5.30896167717 RMS removing mean-by-year = 457.688315369, fraction explained= 35.5695816686

We see that the variation by year explains more than the variation by station. We see that for coeff_1,2,3 the stations explain 2-9% of the variance while the year explaines 35-50%.

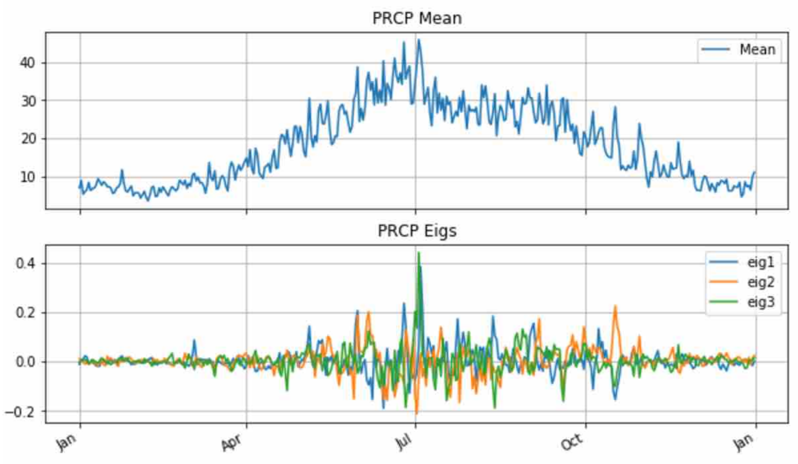
PRCP Analysis

We choose to analyze the eigen-decomposition for PRCP because the first 4 eigen-vectors explain just 6% of the variance. Hence, we want to show that it is very difficult to find meaningful patterns for PRCP measurements using eigen-decomposition.



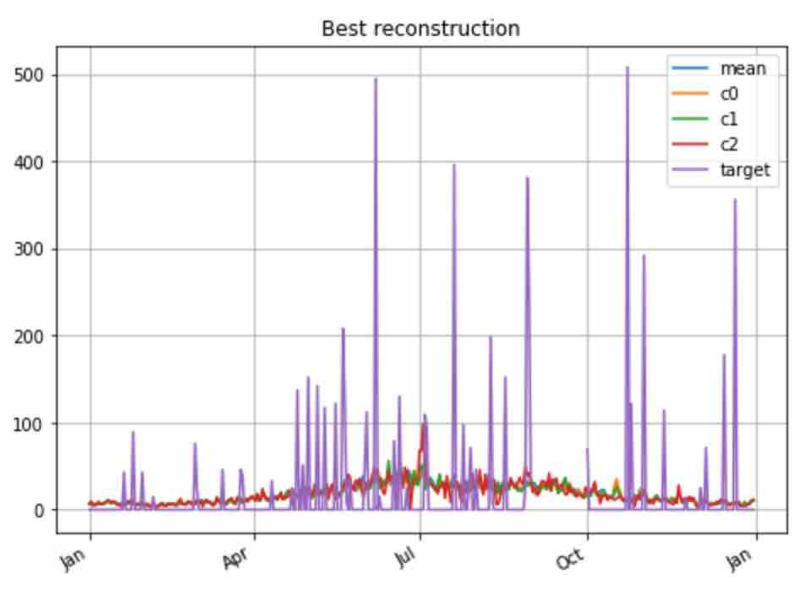
Next, we graph the mean and the top 3 eigen-vectors.

We observe that unlike SNWD measurements, PRCP measurements appear to be very noisy and hence all eigen-values keep on oscillating up and down.



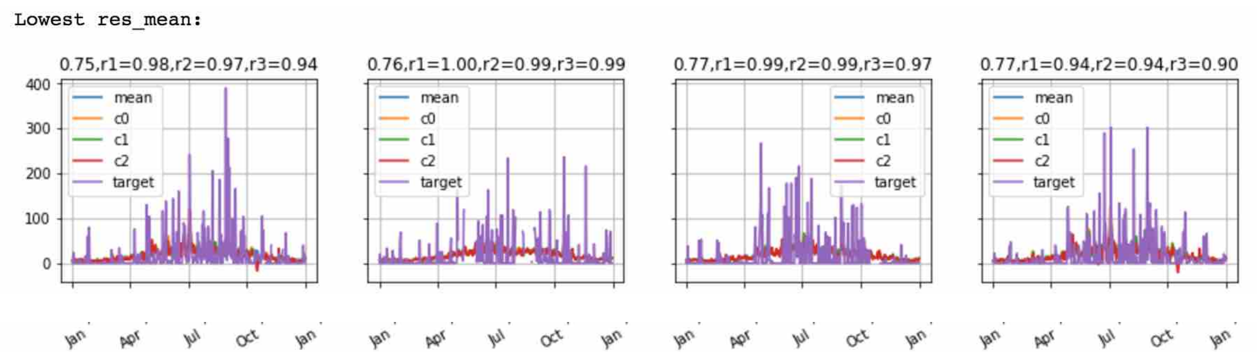
Examples of reconstructions

Best reconstruction



Plotting reconstructions with

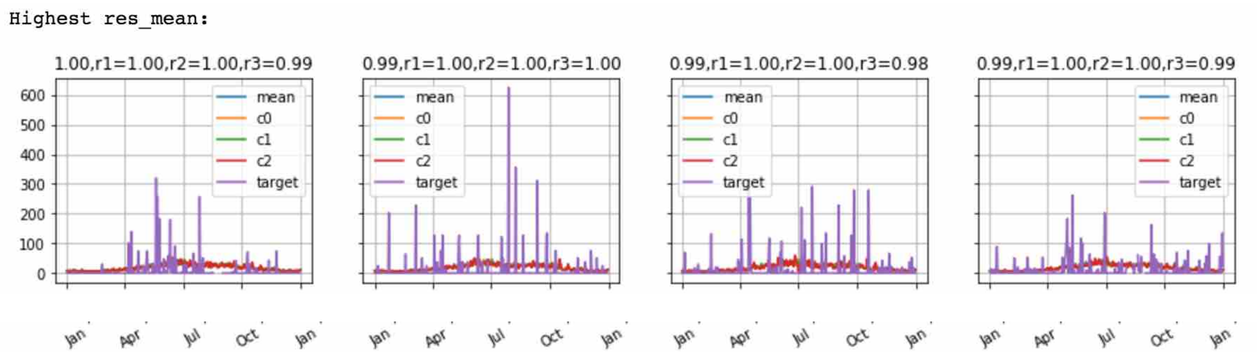
Graphs when mean explains the most variance:



```
Lowest res_mean:
```

0.75,r1=0.98,r2=0.97,r3=0.94
0.76,r1=1.00,r2=0.99,r3=0.99
0.77,r1=0.99,r2=0.99,r3=0.97
0.77,r1=0.94,r2=0.94,r3=0.90
400
300
200
100
0
mean
c0
c1
c2
target
Jan
Apr
Jul
Oct
Jan

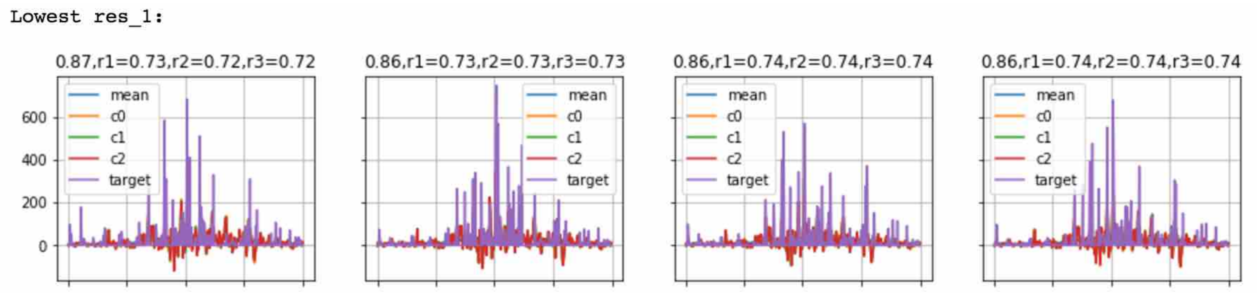
Graphs when mean explains the least variance:



```
Highest res_mean:
```

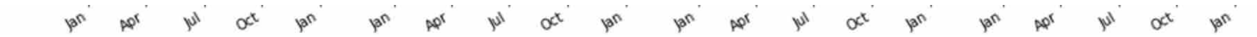
1.00,r1=1.00,r2=1.00,r3=0.99
0.99,r1=1.00,r2=1.00,r3=1.00
0.99,r1=1.00,r2=1.00,r3=0.98
0.99,r1=1.00,r2=1.00,r3=0.99
600
500
400
300
200
100
0
mean
c0
c1
c2
target
Jan
Apr
Jul
Oct
Jan

Graphs when mean and first eigen value explains the most variance:

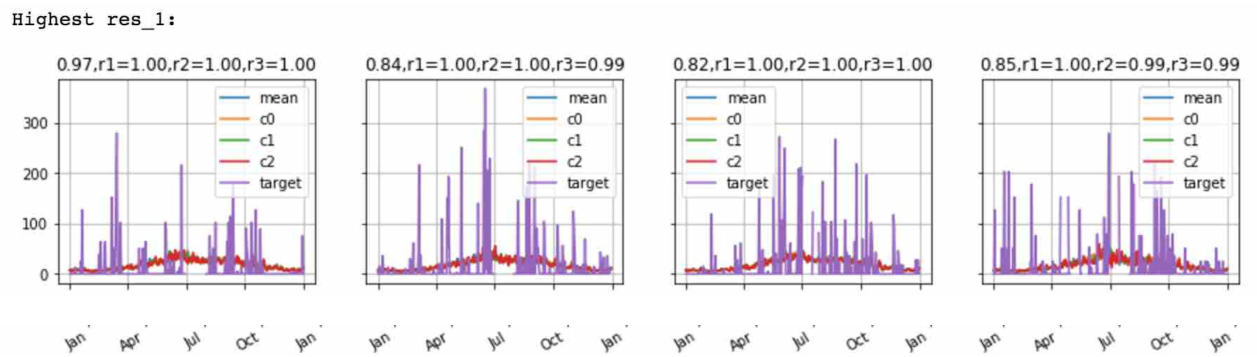


```
Lowest res_1:
```

0.87,r1=0.73,r2=0.72,r3=0.72
0.86,r1=0.73,r2=0.73,r3=0.73
0.86,r1=0.74,r2=0.74,r3=0.74
0.86,r1=0.74,r2=0.74,r3=0.74
600
400
200
0
mean
c0
c1
c2
target

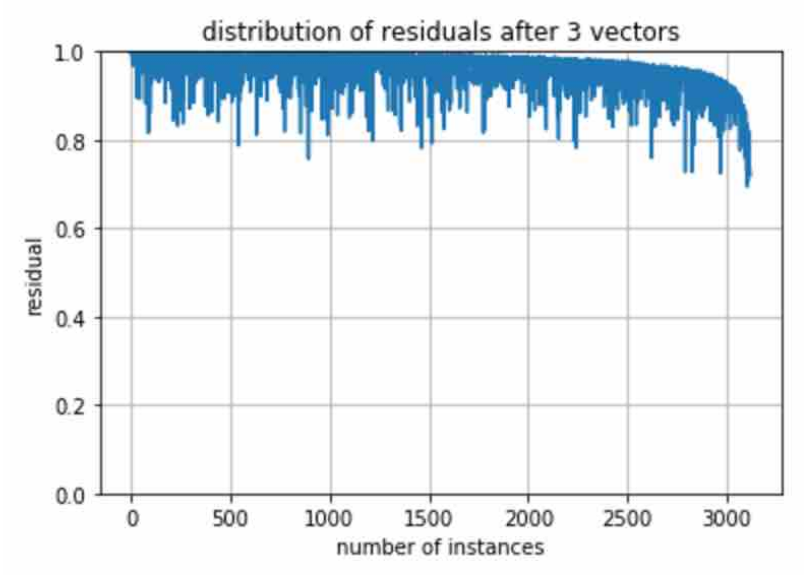


Graphs when mean and first eigen value explains the least variance:



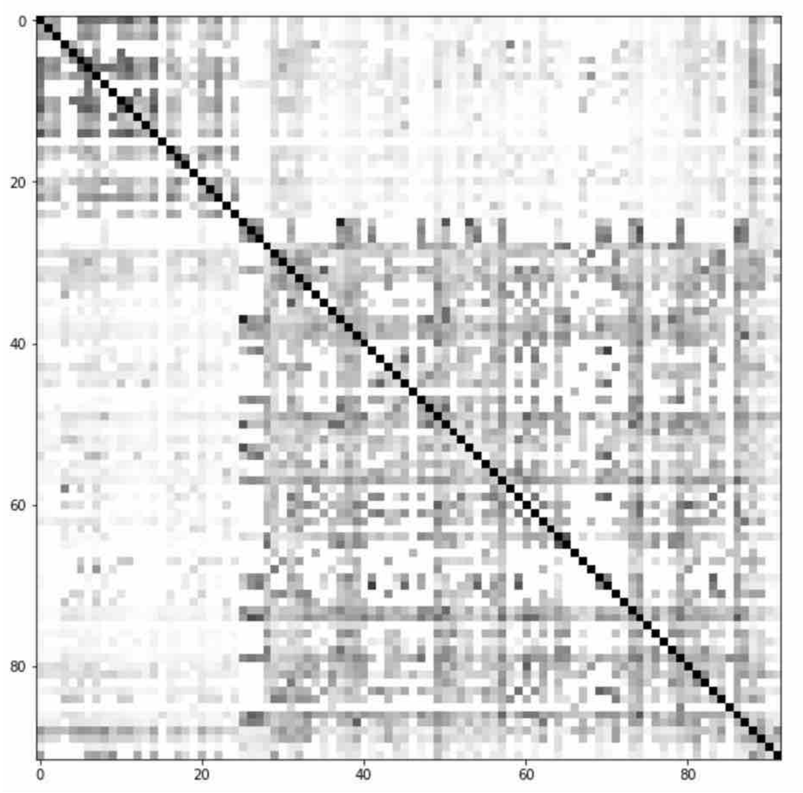
```
Highest res_1:
```

Even the lowest residual values after considering the variance explained by mean and first eigenvalue is more than 80 percent. The graph below verifies our claim that it is nearly impossible to make sense of PRCP measurements using small number of dimensions.

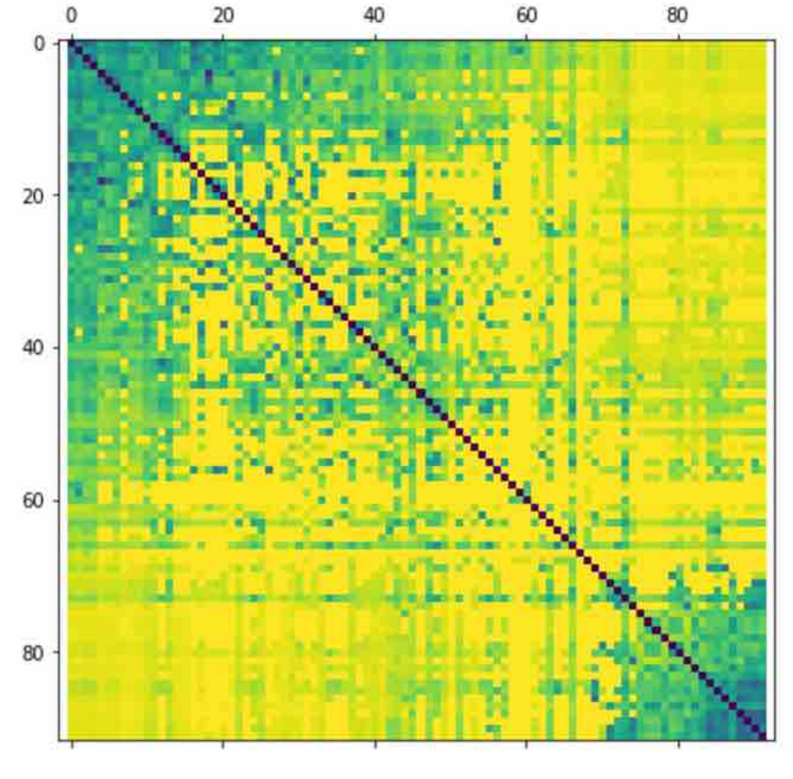


Plotting correlation matrix

Correlation matrix for 92 stations for PRCP measurements:



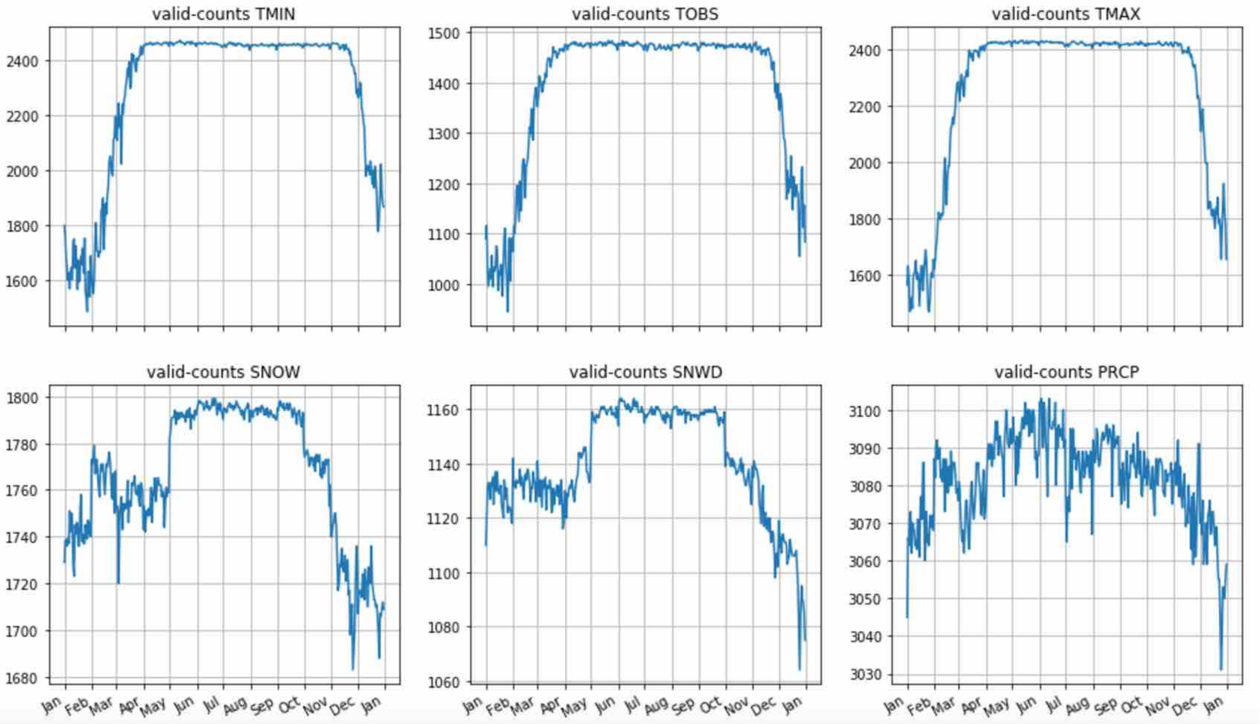
Correlation matrix after reordering on the basis most similar characteristic:



Again, the light shades of the matrix favor the null-hypothesis, verifying our initial claim that it is nearly impossible to find patterns in PRCP data using a small number of eigen-values.

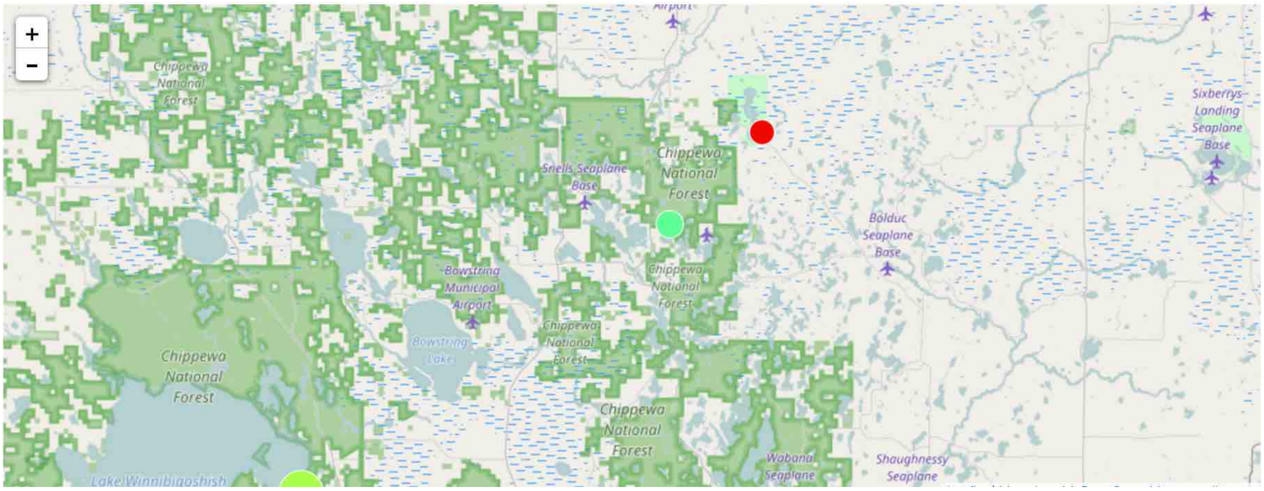
A different kind of Analysis 1

Is the data skewed?

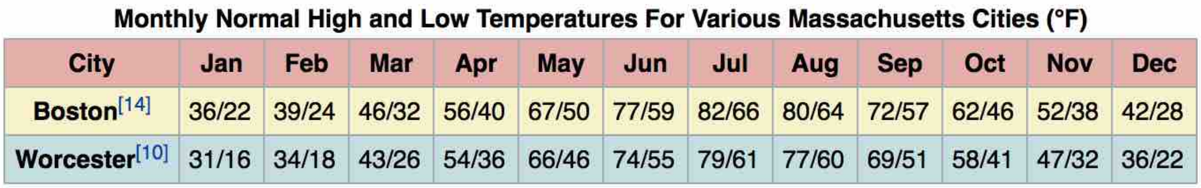


We can see that for all types of observations the number of valid observations reduce near the start and end of an year.

Below is a snapshot of a small region of massacussets:



This part of the united states of America is known to be very cold and filled with snow during the winter season. Below is a snapshot of an image taken from Wikipedia (https://en.wikipedia.org/wiki/Climate_of_Massachusetts) outlining the monthly normal min and max temperature in some parts of Massacusets: *(Note the temperature values in dec, jan and feb).*



Monthly Normal High and Low Temperatures For Various Massachusetts Cities (°F)

City	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Boston[14]	36/22	39/24	46/32	56/40	67/50	77/59	82/66	80/64	72/57	62/46	52/38	42/28
Worcester[10]	31/16	34/18	43/26	54/36	66/46	74/55	79/61	77/60	69/51	58/41	47/32	36/22

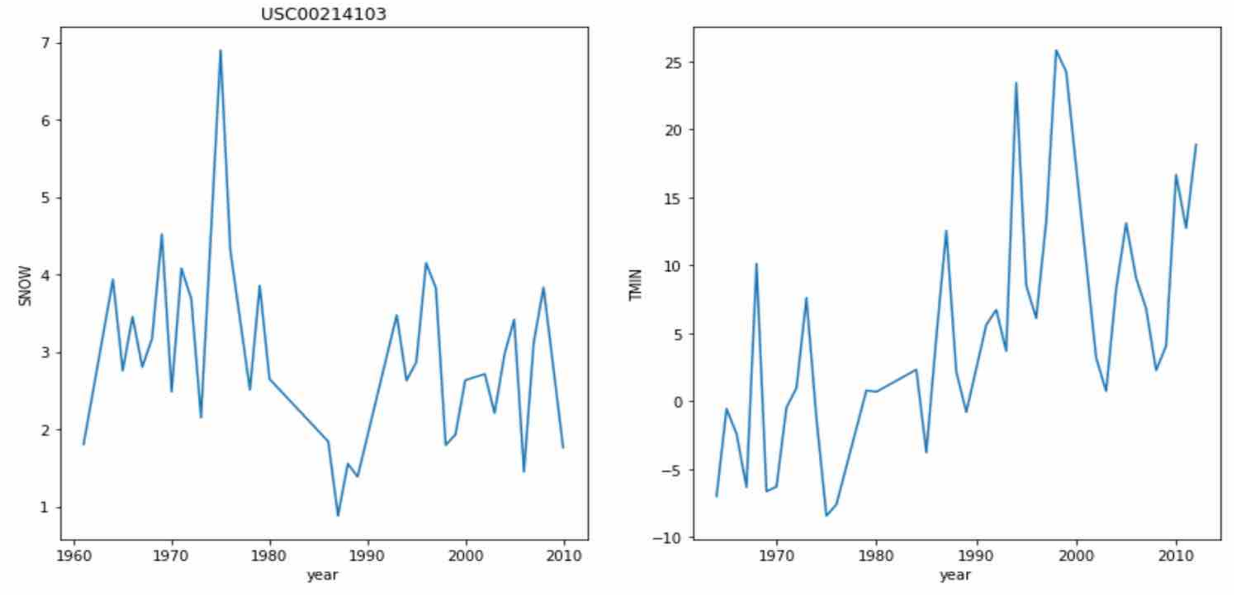
Possible reasons of dip in number of observations during the winter season

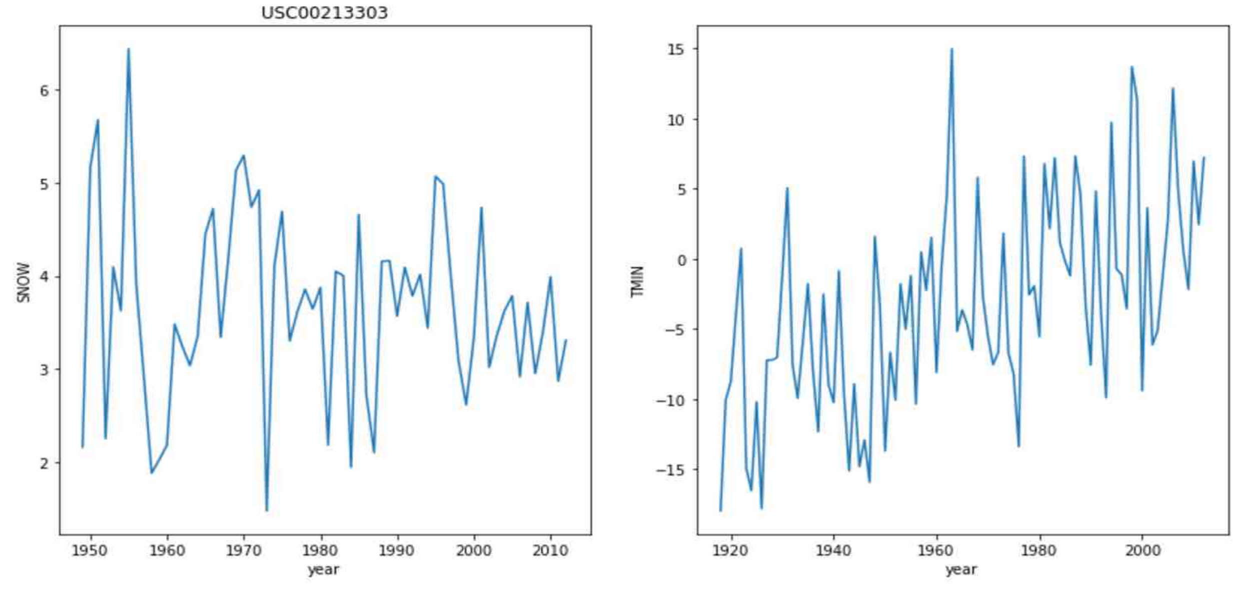
1. The extreme climate during the winter season makes it difficult to keep on making observations with the same frequency as that of non-winter seasons.
2. A subset of stations lie in extreme climate regions which are closed during winter seasons.
3. Since the data was recorded for many years, it is possible that for a subset of years it was not possible to make observations for some of the stations during winter season.

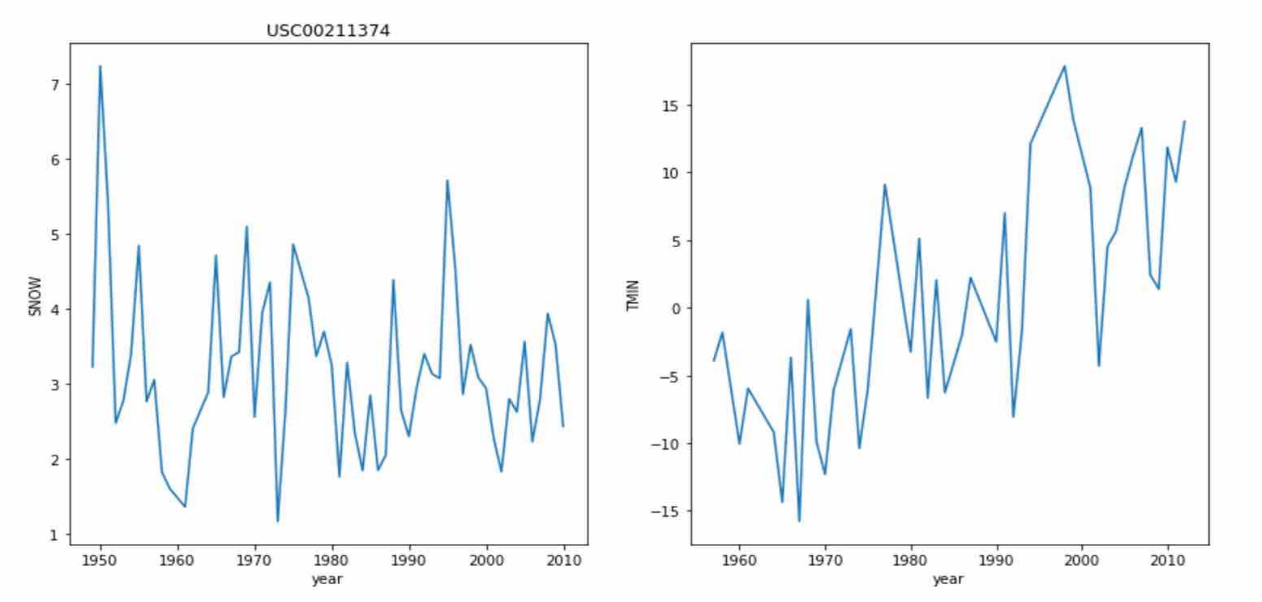
A different kind of Analysis 2

Is this Global Warming?

Below are some plots of stations which had atleast 20 years worth of data for both 'Snow' and 'TMIN' measurements.







The above graphs show that even though there are many fluctuations (change in weather) in the plots, the overall effect is increase in TMIN (or little decrease in snow) which can be attributed to climate change.

Note: We have only showed graphs for a subset of regions which showed increasing minimum temperatures or decreasing snow (or both) in the regions. There were staions where the minimum temperatures and snow did not change much with time but we did not find any evidence of **global cooling**.