CSE 255 - HW5 (file_index=SSSBSSBB)

Name: Balachander Padmanabha (A53202177)

Yosemite/Nevada Weather Analysis

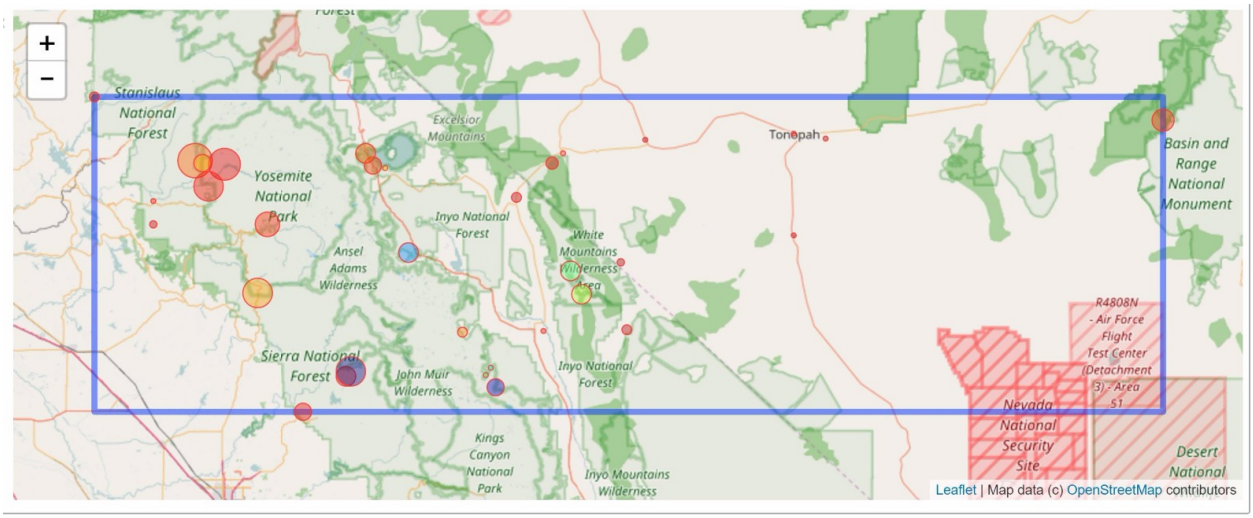
This is a report on the historical analysis of weather patterns in an area that approximately overlaps the area of the state of California and Nevada.

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:

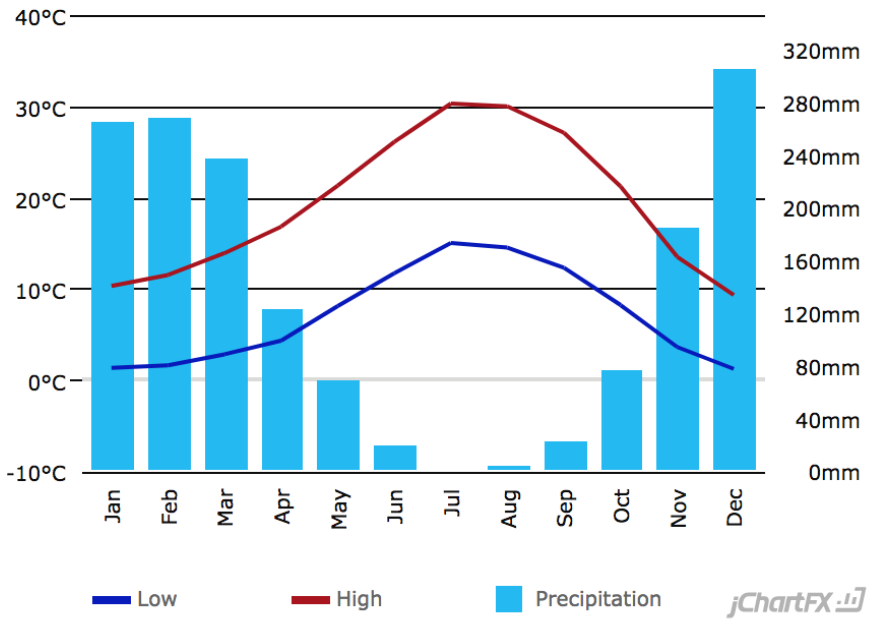
- **TMIN, TMAX:** the daily minimum and maximum temperature.
- **TOBS:** The average temperature for each day.
- **PRCP:** Daily Percipitation (in mm)
- **SNOW:** Daily snowfall (in mm)
- **SNWD:** The depth of accumulated snow.

Map corresponding to area in the dataset.

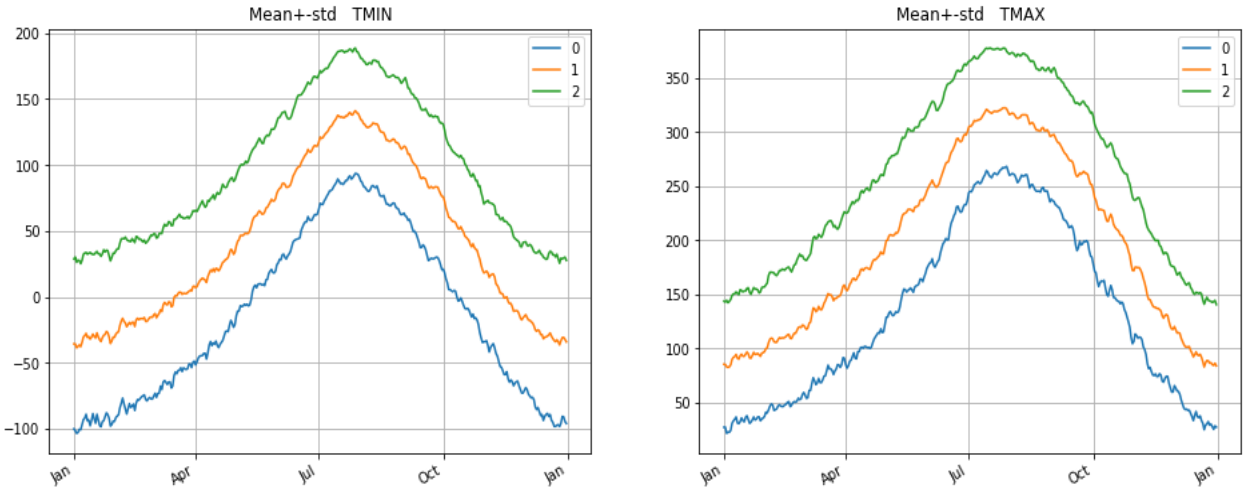


Sanity-check: comparison with outside sources

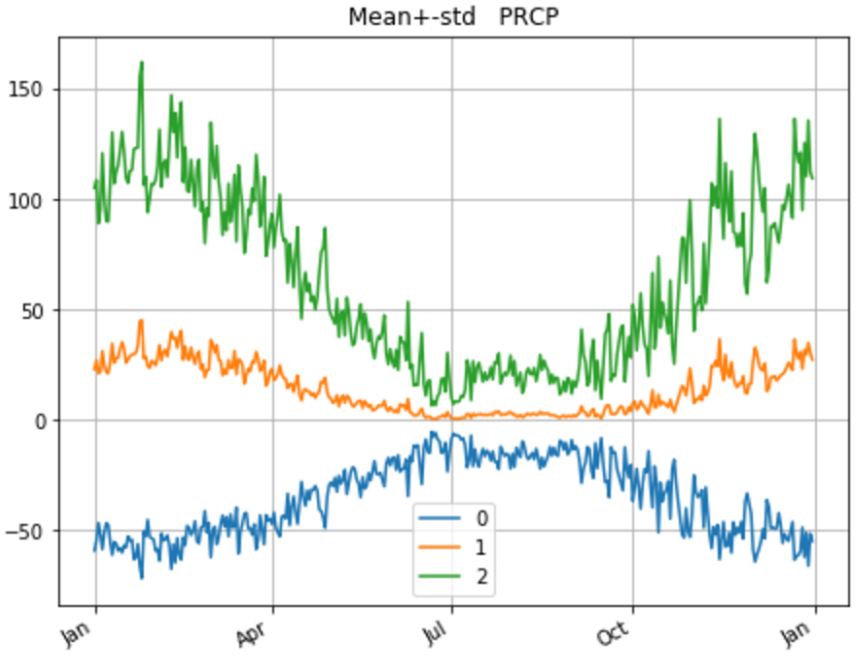
We start by comparing some of the general statistics with graphs that we obtained from a site called US Climate Data (http://www.usclimatedata.com/climate/boston/massachusetts/united-states/usma0046) The graph below shows the daily minimum and maximum temperatures for each month, as well as the total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



We see that the min and max daily temperature agree with the ones we got from our data, once we translate Fahrenheit to Centigrade. Below is plot from dataset of TMIN and TMAX with 0: mean-std 1: mean and 2: mean+std



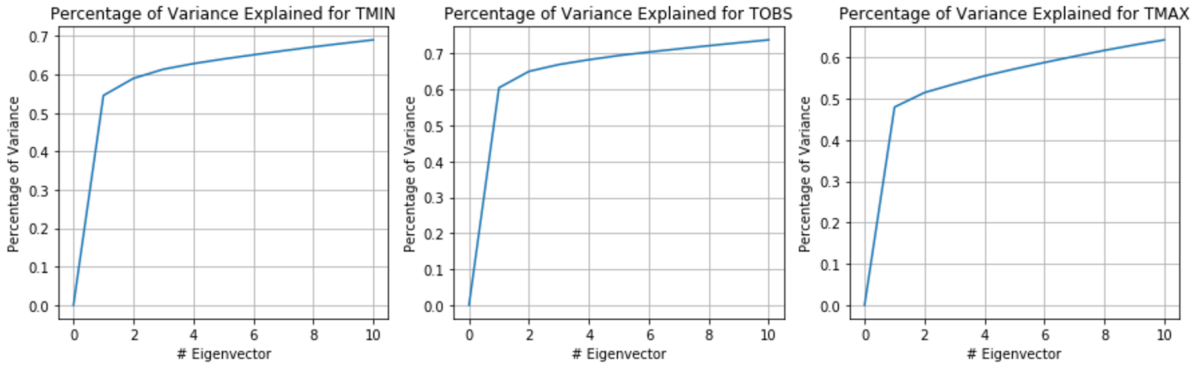
To compare the precipitation we need to translate millimeter/day to inches/month. Below we have 0: mean-std 1: mean and 2: mean+std for PRCP values from the dataset



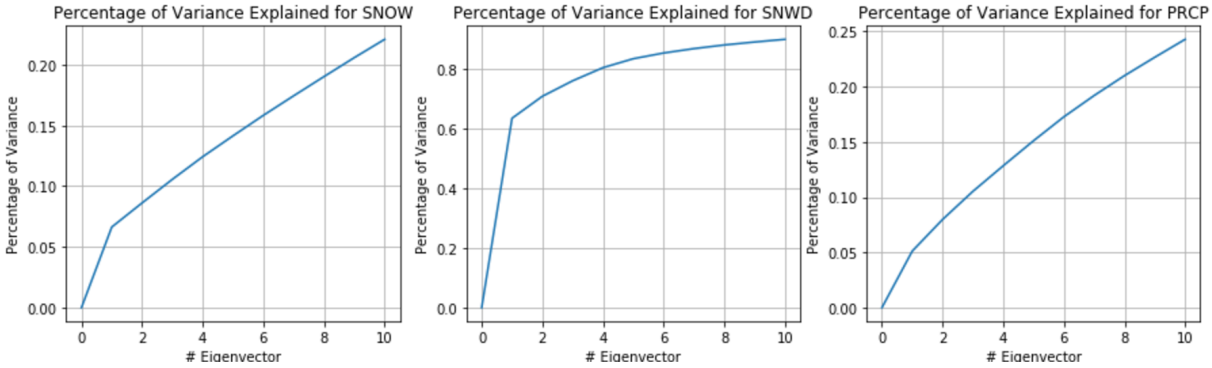
PCA analysis

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 65% of variance for TMIN, 70% for TOBS and 55% for TMAX. Based on these numbers we can 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 62% of the variance.



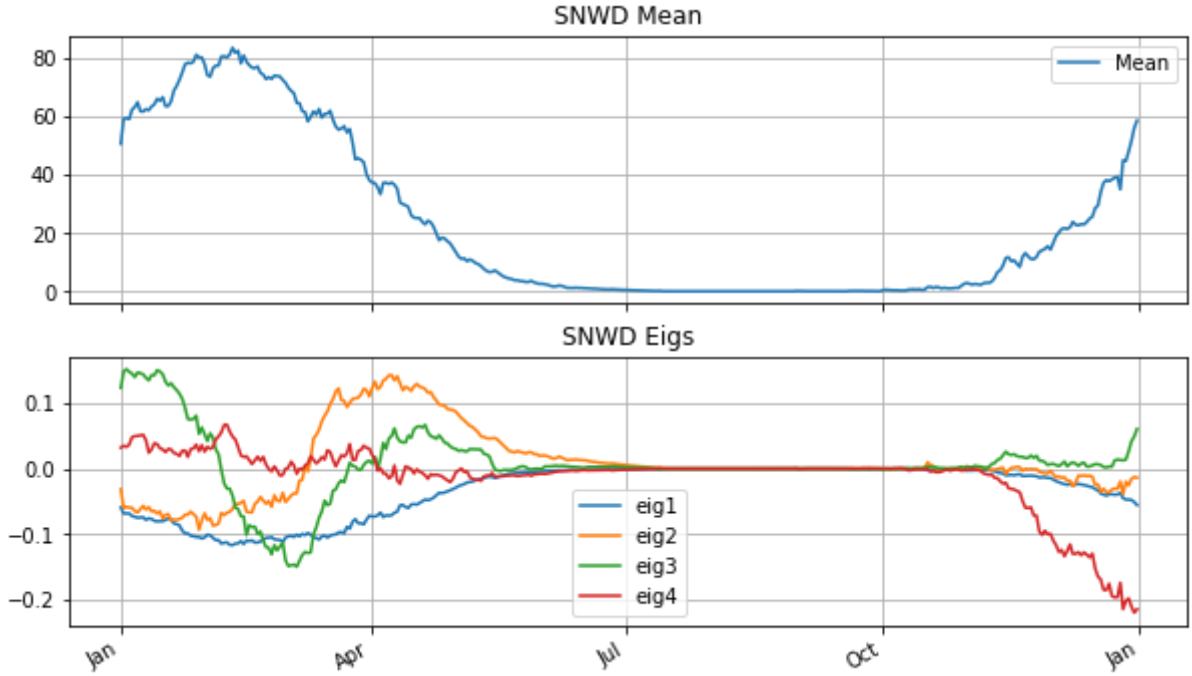
The top 5 eigenvectors explain 15% of the variance for PRCP and 13% for SNOW and both these are low values. On the other hand the top 5 eigenvectors explain 82% of the variance for SNWD. This means that these top 5 eigenvectors capture most of the variation in the snow signals.

It points to the fact that SNWD would be less noisy than SNOW as SNWD varies less between days and between the same date on diffferent years when compared to SNOW.

Based on these numbers I wish to further analyze SNWD and TOBS measurements from the dataset.

Analysis of Snow Depth (SNWD)

I choose to analyze the eigen-decomposition for snow-depth because the first 4 eigen-vectors explain 80% of the variance by using the graph of mean and top 4 eigen-vectors. It is observed that the snow season is from mid-November to the end of June, where the middle of February marks the peak of the snow-depth.



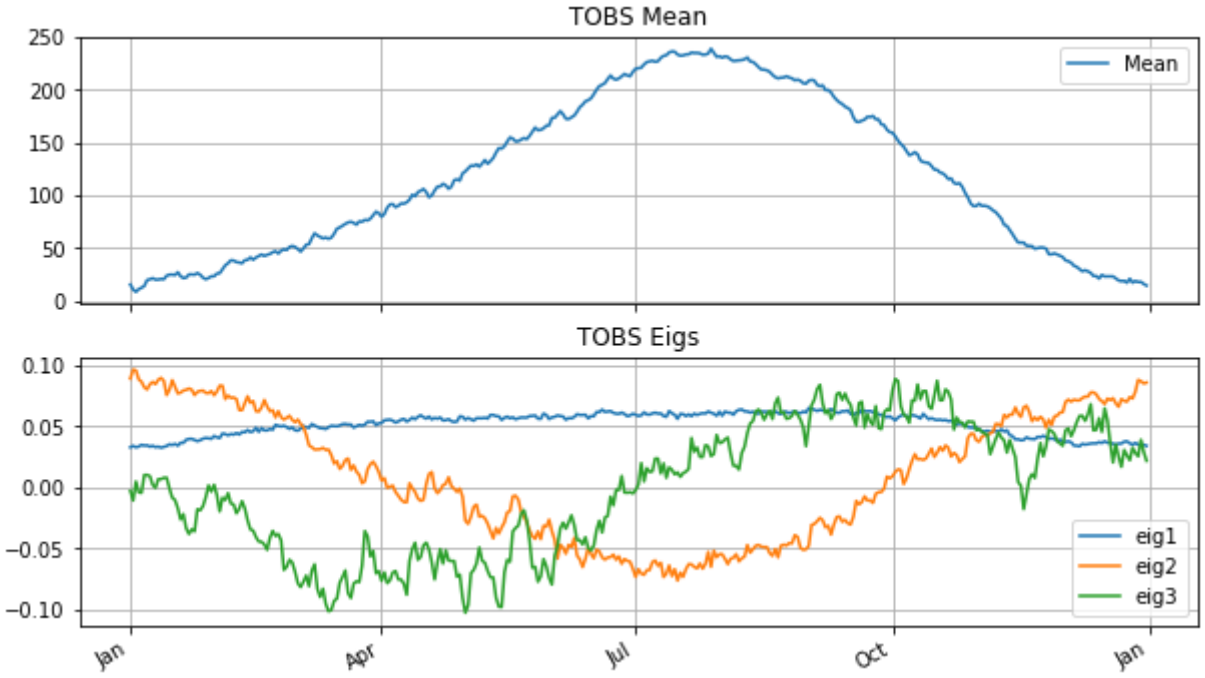
Interpreting the eigen-functions we see that all 4 eigen-functions have certain portions similar to the mean SNWD plot which helps in the overall description of snow-depth. **eig1, eig2, eig3 and eig4** are similar in the following way. They all oscilate between positive and negative values. In other words, they correspond to changing the distribution of the snow depth over the winter months.

They can be interpreted as follows:

- **eig1:** less snow in January - mid June, less snow in mid November-December.
- **eig2:** less snow in January - mid March, more snow in mid March-June, less snow in mid November-December.
- **eig3:** slightly more snow in January - mid March, less snow in mid March-April, more snow in May-June, more snow in mid November-December.
- **eig4:** little snow in January - mid March, almost no snow in April-June, less snow in November-December

Analysis of Average Temparature (TOBS)

I choose to analyze the eigen-decomposition for snow-depth because the first 3 eigen-vectors explain 66% of the variance by using the graph of mean and top 3 eigen-vectors.



Interpreting the eigen-functions we see that all 3 eigen-functions have certain portions similar to the mean TOBS plot which helps in the overall description of snow-depth. Eig1 to some extent mimics the overall average temparature similar to the mean plot.

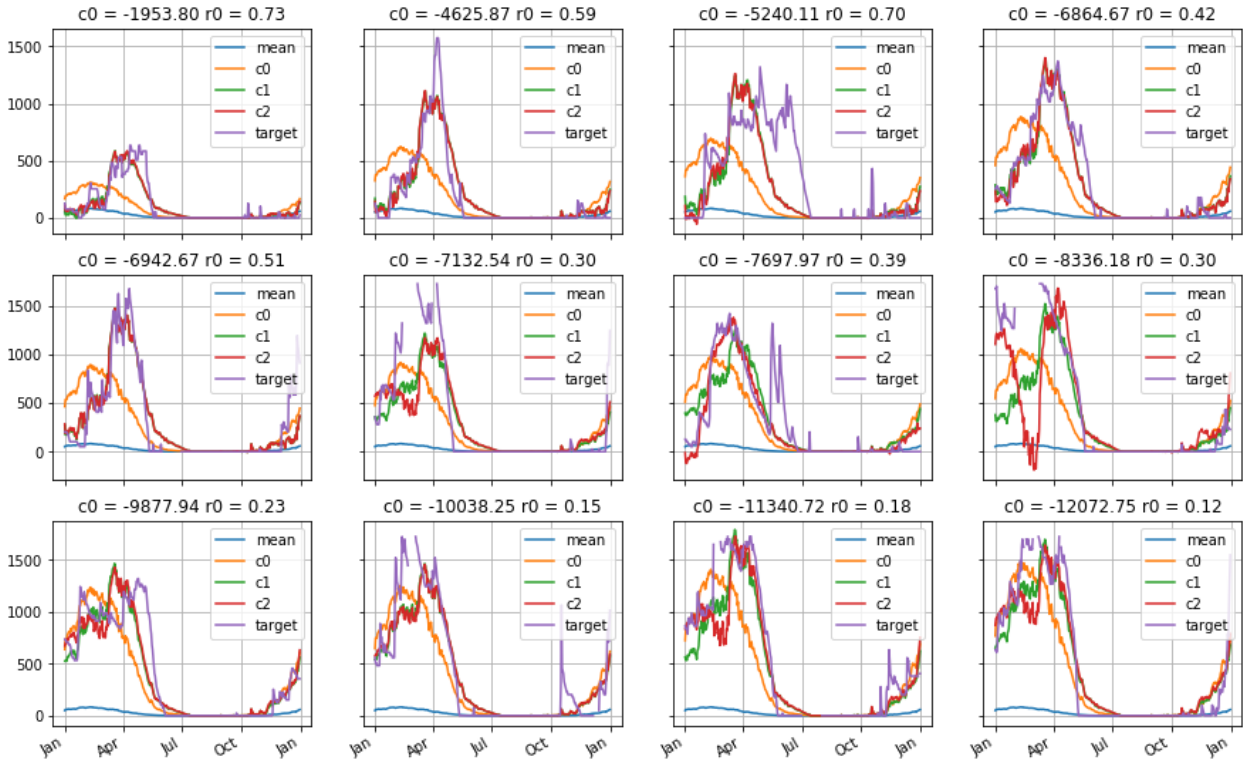
eig2 and eig3 are similar in the following way. They all oscilate between positive and negative values. In other words, they correspond to changing the distribution over the year.

They can be interpreted as follows:

- **eig1:** Indicates that average temparature is more in January-December.
- **eig2:** more average temparature in January - April, less average temparature in May - October and more average temparature in November - December.
- **eig3:** less average temparature in January - July, more average temparature in August - December with slight dip in mid November.

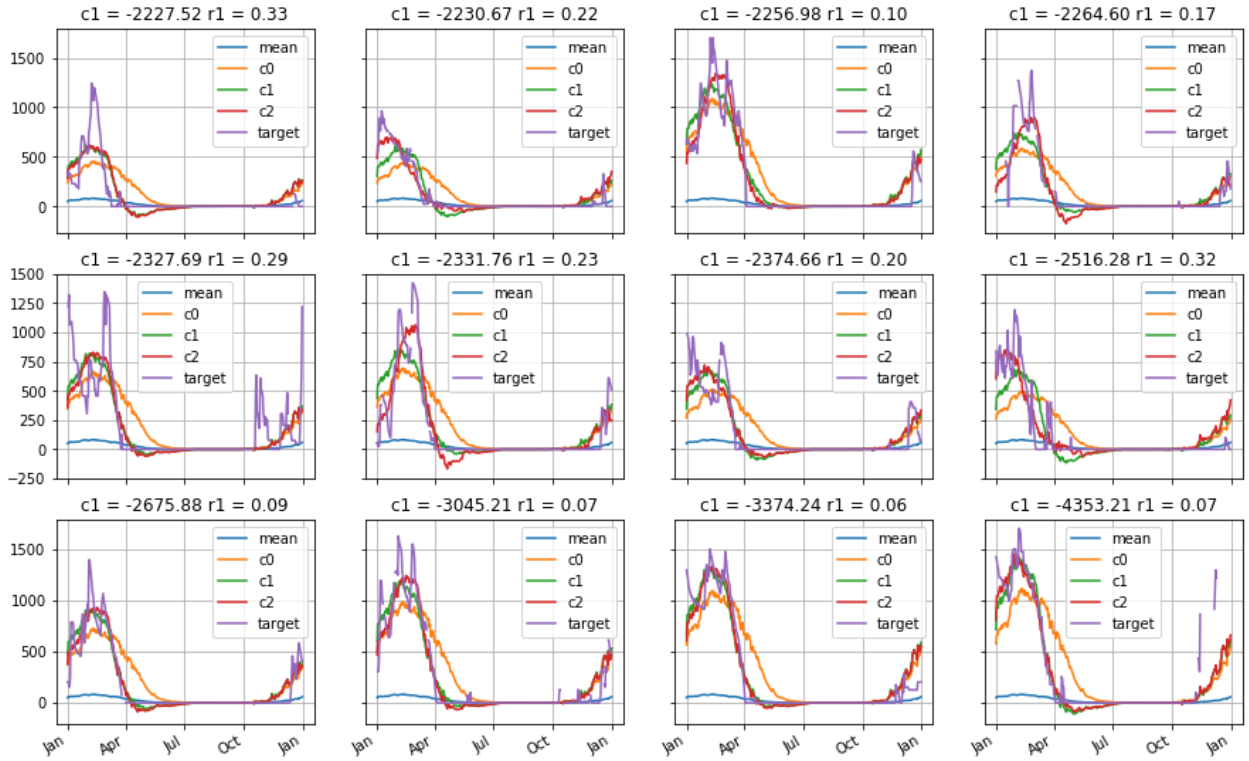
Examples of reconstructions (for SNWD)

Coeff0:



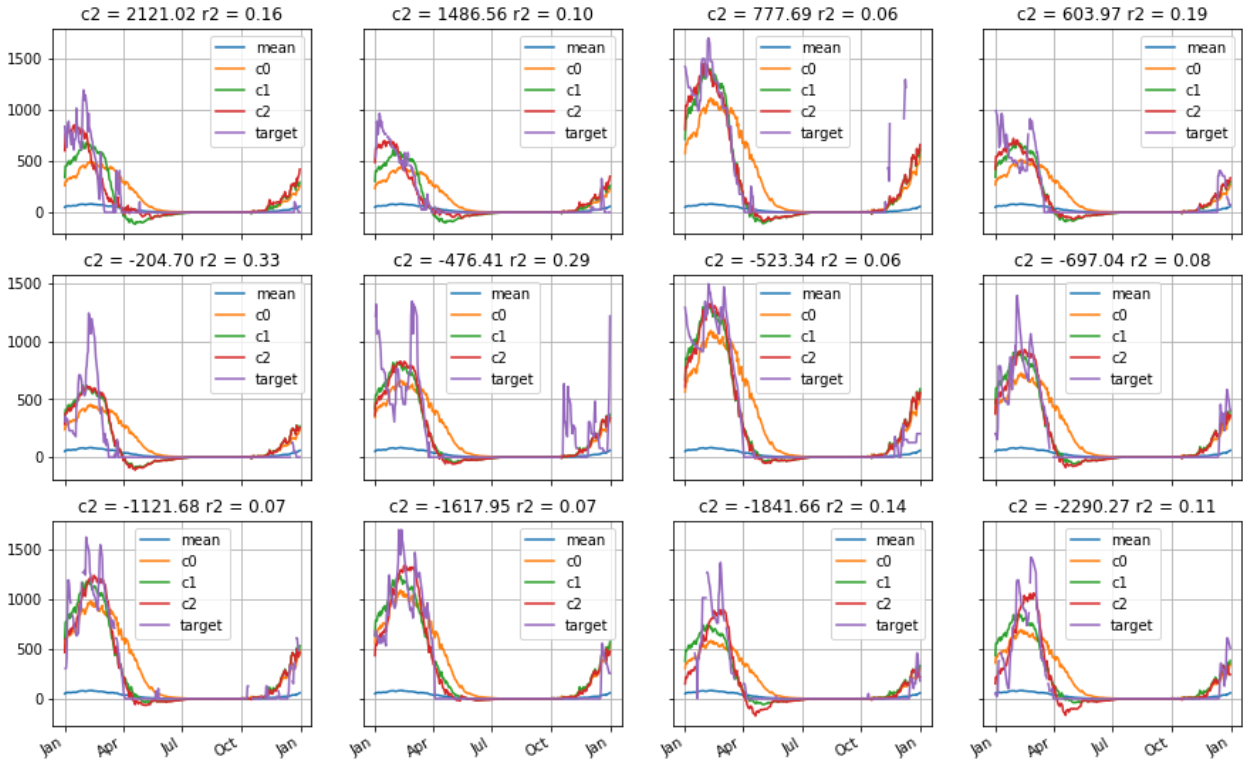
From the plots we can conclude the mean is consistent for all of them. As coefficient 0 becomes more negative it becomes closer to target for snow depth.

Coeff1:



From the plots we can conclude the mean is consistent for all of them. As Coefficient 1 becomes more negative it becomes closer to target for snow depth.

Coeff2:



From the plots we can conclude the mean is consistent for all of them. As coefficient becomes more positve it becomes closer to target for snow depth.

The variation in the timing of TOBS is mostly due to station-to-station variation

Coeff1:
total RMS = 872.516277667
RMS removing mean-by-station = 491.149555752, fraction explained = 43.7
RMS removing mean-by-year = 808.870716883, fraction explained = 7.3

Coeff2:
total RMS = 299.685060601
RMS removing mean-by-station = 169.89146285, fraction explained = 43.33
RMS removing mean-by-year = 267.8148107, fraction explained = 10.63

Coeff3:
total RMS = 198.688095552
RMS removing mean-by-station = 174.646531054, fraction explained = 12.1
RMS removing mean-by-year = 150.299247834, fraction explained = 24.3

We see that the variation by station explains more than the variation by year.

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

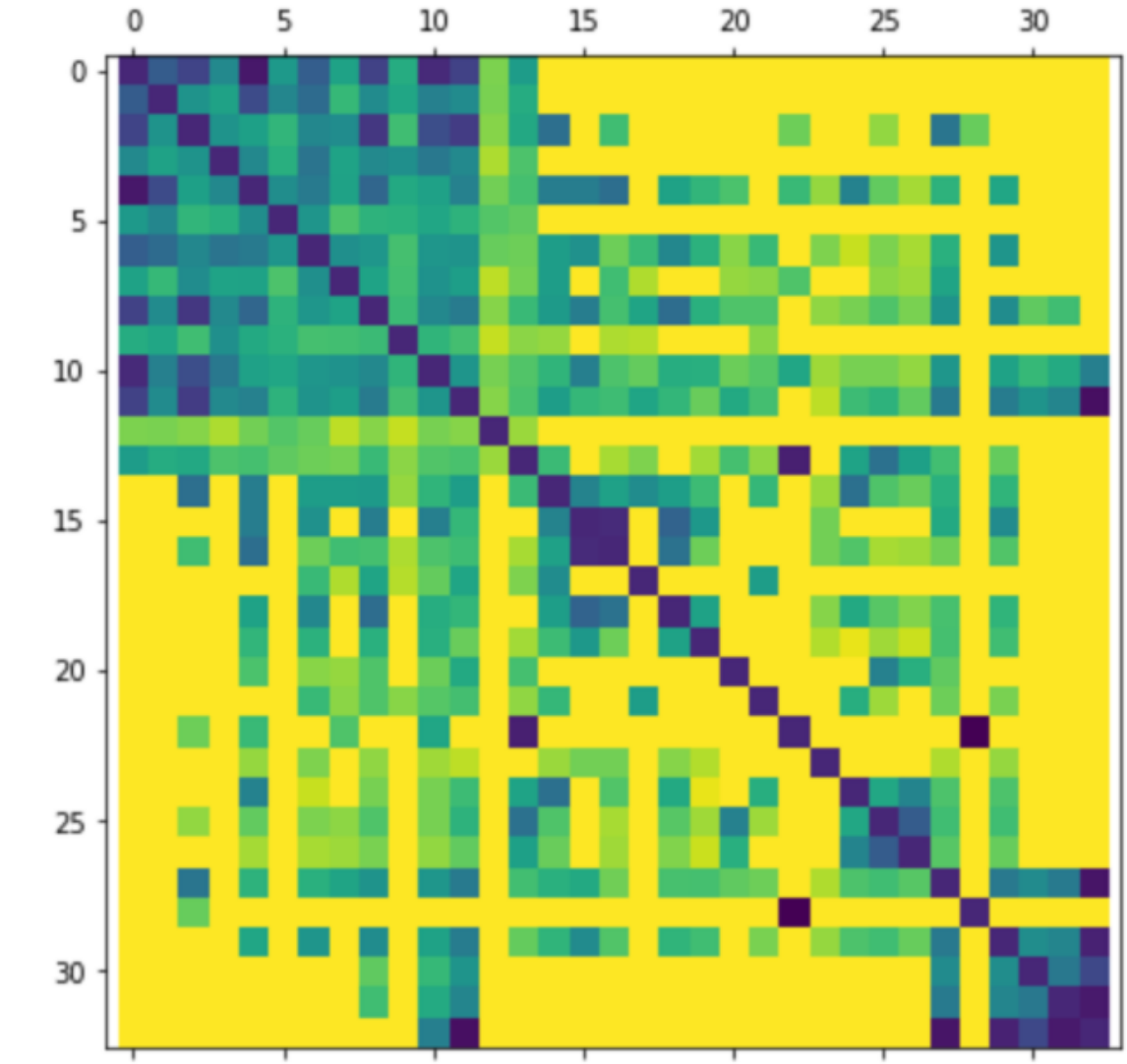
Coeff1:
total RMS = 3359.93064281
RMS removing mean-by-station = 1880.565768, fraction explained = 44
RMS removing mean-by-year = 2505.43652363, fraction explained = 25.43

Coeff2:
total RMS = 1229.5296533
RMS removing mean-by-station = 1090.23742062, fraction explained = 11.32
RMS removing mean-by-year = 859.293116903, fraction explained = 30.1

Coeff3:
total RMS = 1052.12813966
RMS removing mean-by-station = 1018.94773308, fraction explained = 3.1
RMS removing mean-by-year = 694.772844026, fraction explained = 33.96

We see that the variation by year explains more than the variation by station.

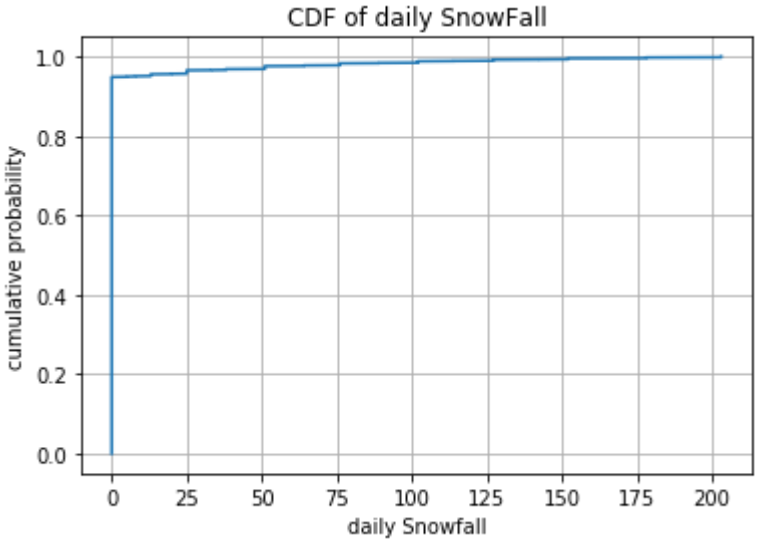
Residuals Analysis (for SWND)



Consider the upper left corner of the matrix. The stations at positions 0-11 are clearly strongly correlated with each other. This probably is due to the fact most of these stations are around the yosemite area and receive snow around the same time of the year (Spatial relationship among the stations).

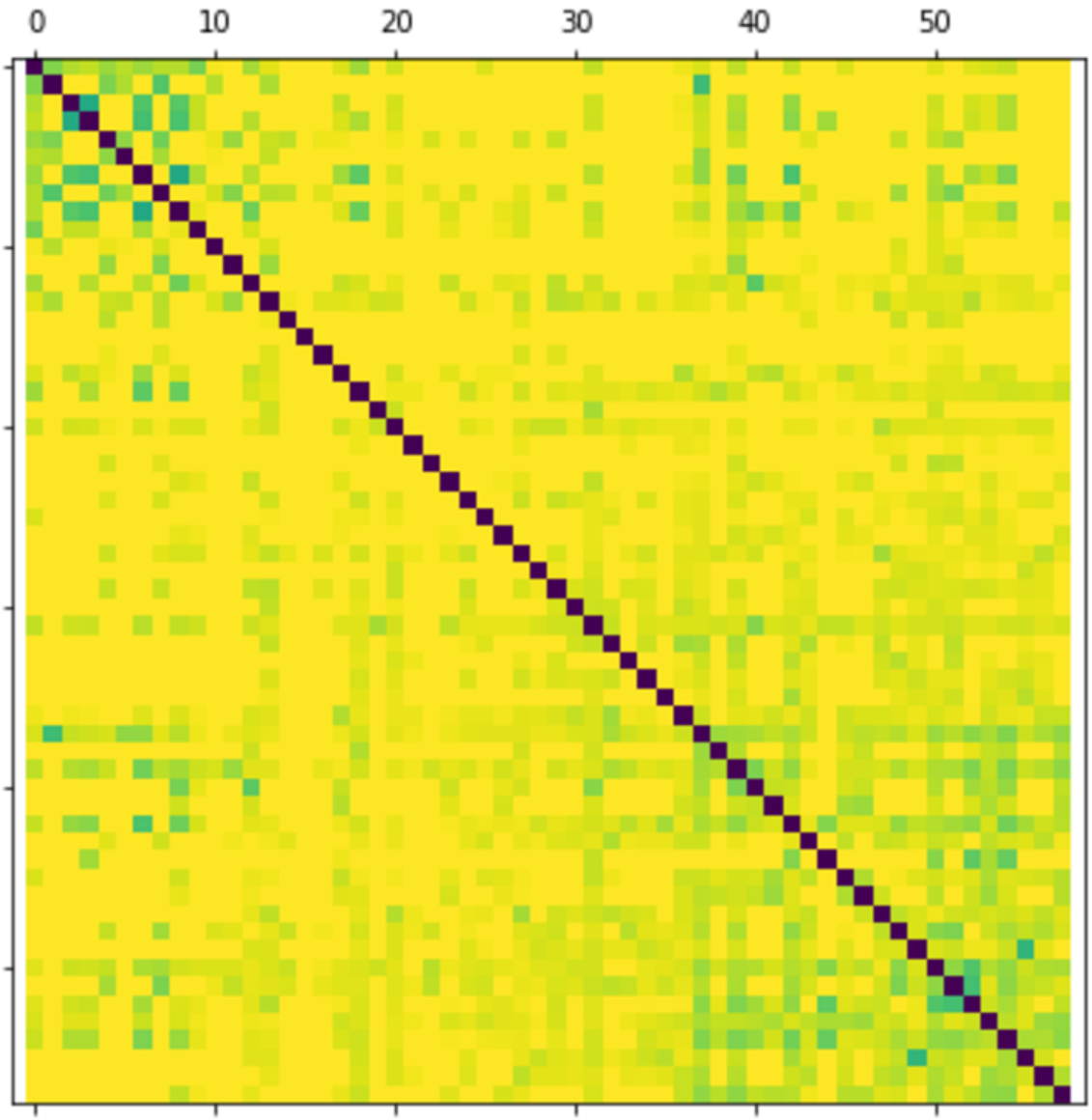
stations IDs: [u'USC00043939', u'USC00049855', u'USC00040755', u'USC00261755', u'USC00263285', u'USC00042331', u'USC00045400', u'USC00044679', u'USC00048406', u'USC00049632', u'USC00049633']

Snow Fall CDF:



Conclusion from the above plot would be is it's likely to be hard to find correlations between the amount of snow on the same day in different stations. Because amounts of snow vary a lot between even close locations and most of the time there is 0 snowfall.

This can also be concluded with the below correlation plot for SNOW which shows any existence of correlation among the stations.


0
10
20
30
40
50