1 Massacusets 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 Massachusets.

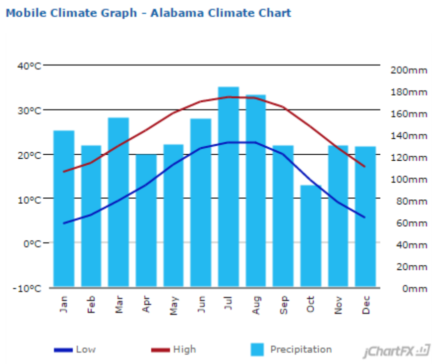
The data we will use here comes from NOAA (https://www.ncdc.noaa.gov/). Specifically, it was downloaded from This FTP site.

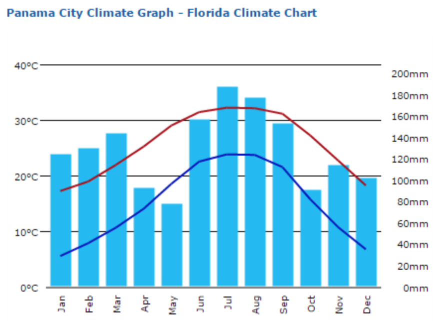
We focused on six measurements:

- **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.

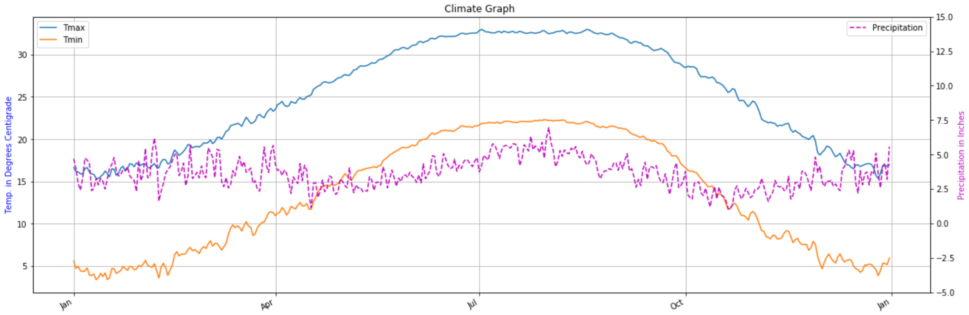
1.1 Sanity-check: comparison with outside sources

We start by comparing some of the general statistics with graphs that we obtained from a site called US Climate Data (http://www.usclimatedata.com/climate/boston/massachusetts/united-states/usma0046). **Mobile, AL** and **PanamaCity, FL** are the two major cities in the give region(of BSSSBSBS indexed data). The following graph below shows the daily minimum and maximum temperatures per month, as well as the total precipitation per month, for these two cities.



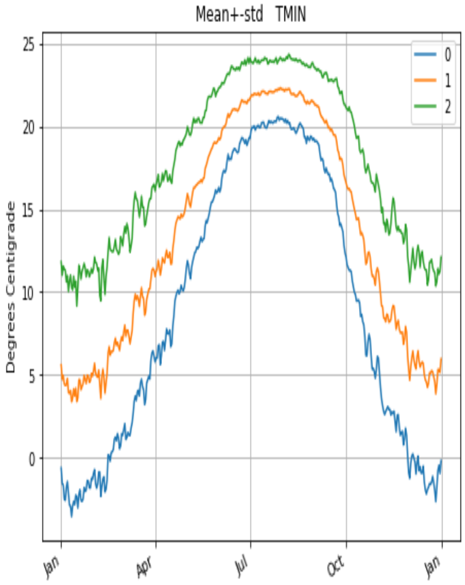


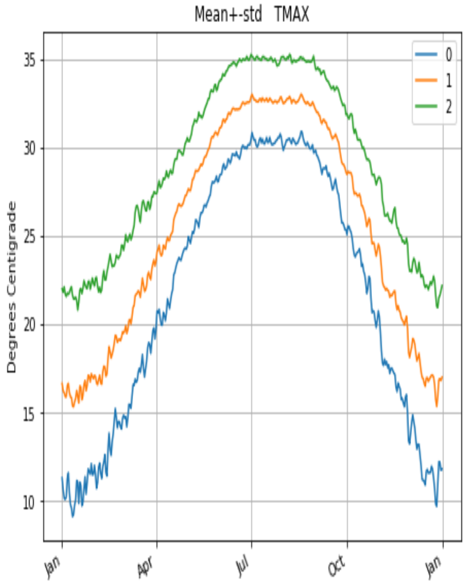
The following Climate graphs shows the mean of Tmax , Tmin and Precipitation for the current BSSSBSBS region. We see that the min and max daily temperature agree with the ones we got from our data as shown below. We can see the temperatures are ranging from 32°C to 5°C. Constant Precipitation through out the year ranging between 4.7 - 6 inches (120mm to 160 mm)



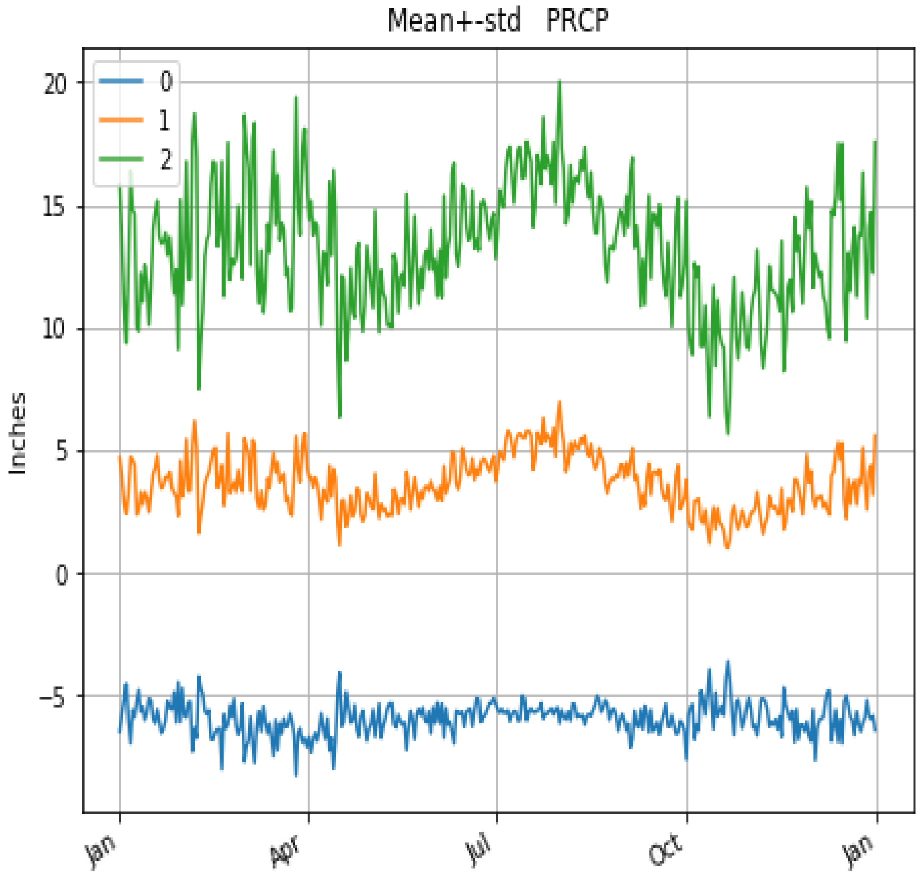
So the data from our analysis is in complete agreement with standard outside analysis.

A closer look at the distribution of TMIN and TMAX, reaffirms that months of June , July and August have the highest temperatures (Summer). And the months of Jan, Feb , Nov and Dec have the lowest temperatures (Winter). The standard deviation is so low and constant in Summers which means the whole region have equal high temperatures through out the years.





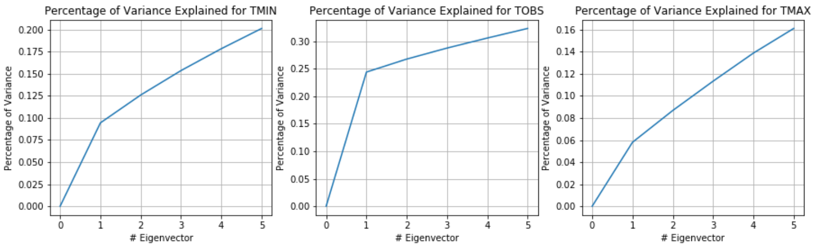
To compare the precipitation we need to translate millimeter/day to inches/month. According to our analysis the average rainfall is 3.00 mm/day which translates to about 3.55 Inches per month. According to US-Climate-Data the average rainfall is closer to 4 inch per month. However, there is clear agreement that average precipitation is close to a constant throughout the year



1.2 PCA analysis

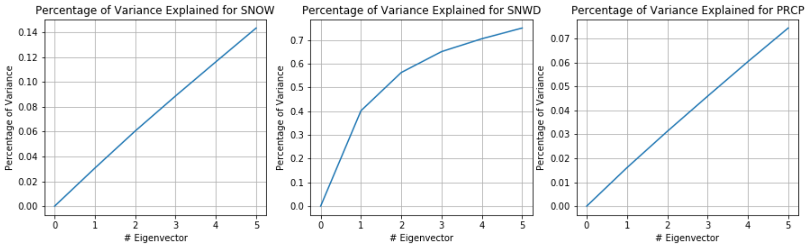
For each of the six measurement, we compute the percentate of the variance explained as a function of the number of eigen-vectors used.

1.2.1 Percentage of variance explained.



We see that the top 5 eigen-vectors explain 20% of variance for TMIN, 32% for TOBS and 16% 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 25% of the variance.



The top 5 eigenvectors explain 7% of the variance for PRCP and 14% for SNOW. Both are low values. On the other hand the top 5 eigenvectors explain %75 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.

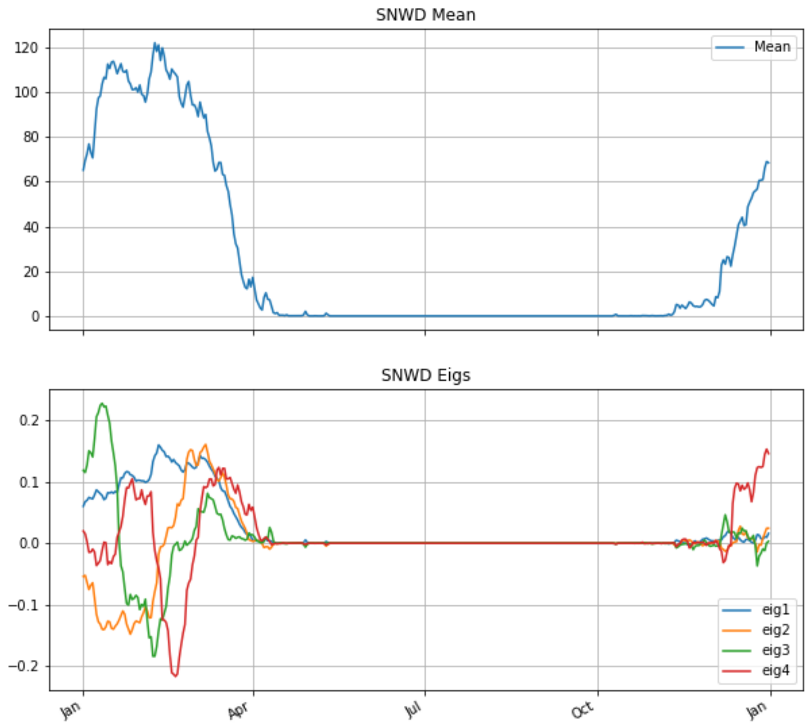
It makes sense that SNWD would be less noisy than SNOW. That is because SNWD is a decaying integral of SNOW and, as such, varies less between days and between the same date on diffferent years.

1.3 Analysis of snow depth

We choose to analyze the eigen-decomposition for snow-depth because the first 4 eigen-vectors explain 70% of the variance.

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

We observe that the snow season is from mid-november to the end of march, 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).

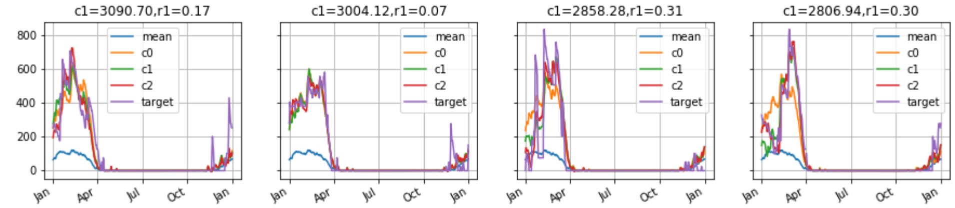
They can be interpreted as follows:

- **eig2:** less snow in jan - mid feb, more snow in mid feb-march.
- **eig3:** more snow in jan, less snow in feb, slightly more snow in march.
- **eig4:** more snow in dec, more snow in start feb, less snow in end of feb, more snow in march.

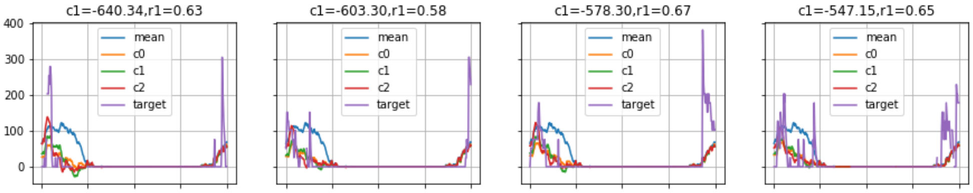
1.3.1 Examples of reconstructions

1.3.1.1 Coeff1

Coeff1: most positive


c1=3090.70,r1=0.17 c1=3004.12,r1=0.07 c1=2858.28,r1=0.31 c1=2806.94,r1=0.30

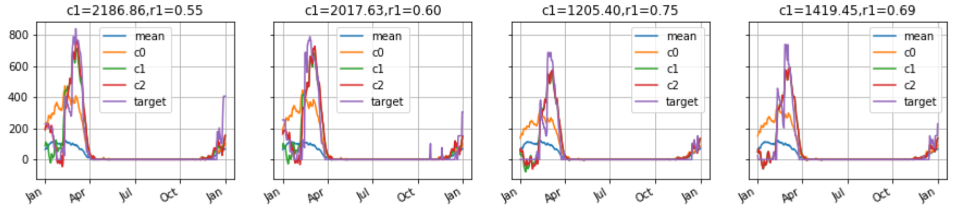
Coeff1: most negative


c1=-640.34,r1=0.63 c1=-603.30,r1=0.58 c1=-578.30,r1=0.67 c1=-547.15,r1=0.65

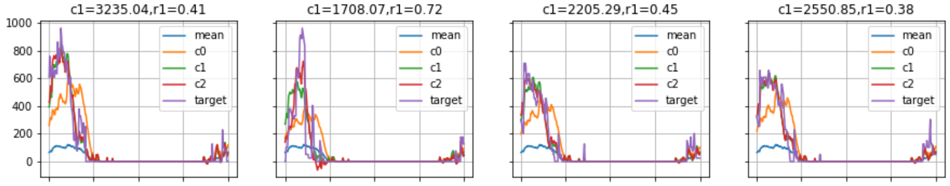
Large positive values of coeff1 correspond to more than average snow. Low values correspond to less than average snow.

1.3.1.2 Coeff2

Coeff2: most positive


c1=2186.86,r1=0.55 c1=2017.63,r1=0.60 c1=1205.40,r1=0.75 c1=1419.45,r1=0.69

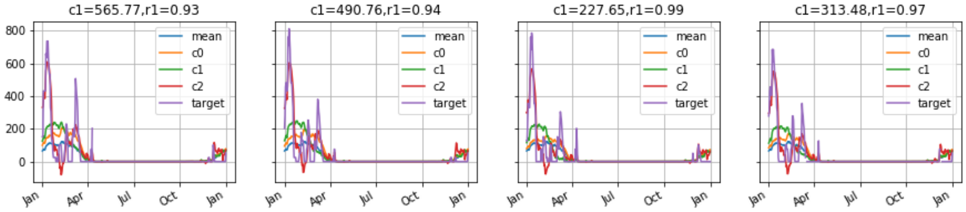
Coeff2: most negative


c1=3235.04,r1=0.41 c1=1708.07,r1=0.72 c1=2205.29,r1=0.45 c1=2550.85,r1=0.38

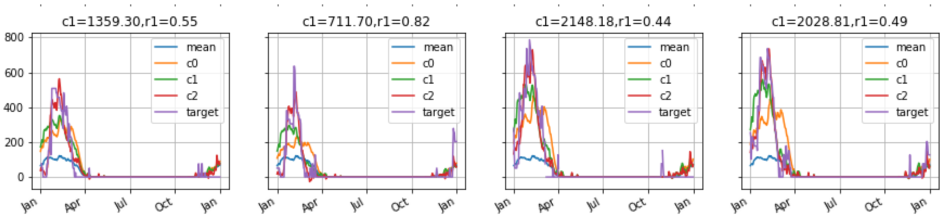
Large positive values of coeff2 correspond to a late snow season (most of the snowfall is after mid feb. Negative values for coeff2 correspond to an early snow season (most of the snow is before mid-feb.

1.3.1.3 Coeff3

Coeff3: most positive


c1=565.77,r1=0.93 c1=490.76,r1=0.94 c1=227.65,r1=0.99 c1=313.48,r1=0.97

Coeff3: most negative



Large positive values of coeff2 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.

1.4 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 = 822858.58
MS removing mean-by-station= 599862.22, fraction explained=27.1
MS removing mean-by-year = 284206.98, fraction explained=65.5

coeff_2
total MS = 425511.45
MS removing mean-by-station= 412035.01, fraction explained= 3.2
MS removing mean-by-year = 79354.27, fraction explained=81.4

coeff_3
total MS = 232267.77
MS removing mean-by-station= 223638.48, fraction explained= 3.7
MS removing mean-by-year = 26072.88, fraction explained=88.8

coeff_4
total MS = 140554.36
MS removing mean-by-station= 131569.68, fraction explained= 6.4
MS removing mean-by-year = 23618.12, fraction explained=83.2

We see that the variation by year explains more than the variation by station. However this effect is weaker consider coeff_1, which has to do with the total snowfall, vs. coeff_2,3,4 which, as we saw above have to do with the timining of snowfall. We see that for coeff_2,3,4 the stations explain 3-5% of the variance while the year explaines 80-90%.