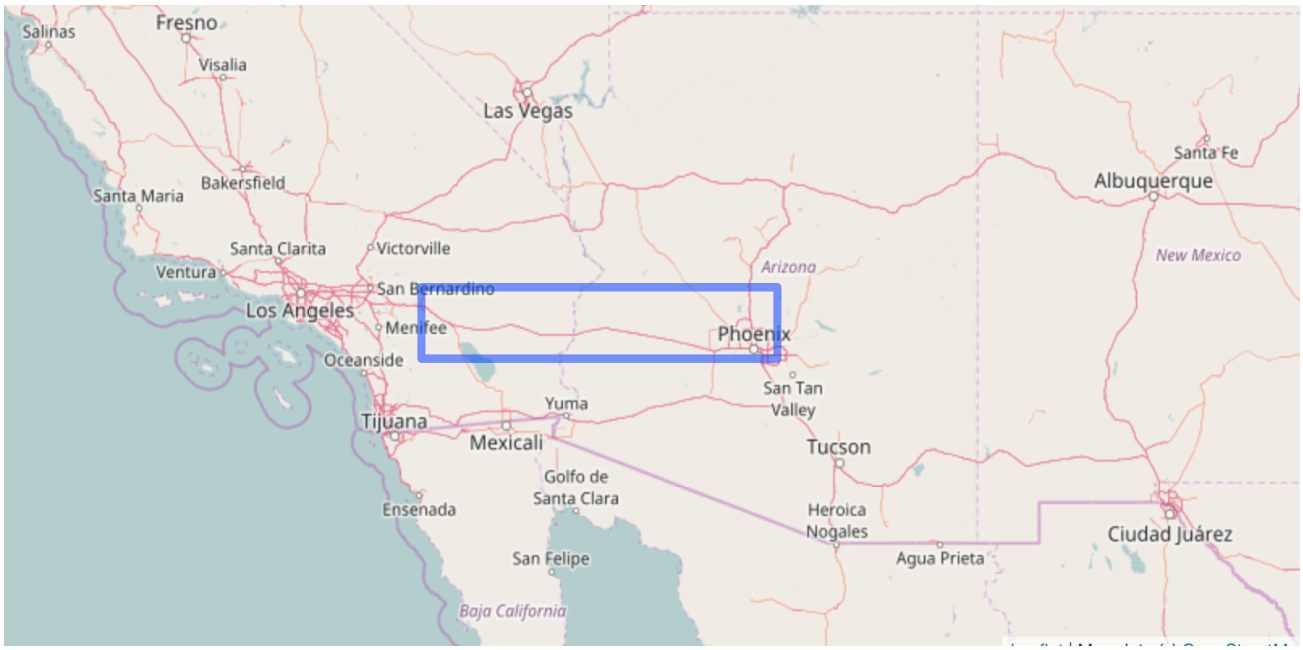
In [3]:
```
L=!ls p_figures/
for l in L:
    print "![%s](p_figures/%s)"%(l,l)
```

```
![square.png](p_figures/square.png)
```

In [4]:
```
!open p_figures/square.png
```

Arizona Weather Analysis

This is a report on the historical analysis of weather patterns in an area that approximately covers part of the eatern region of california and western region of Arizona(with Phoenix as the major hub) as shown by the figure below.



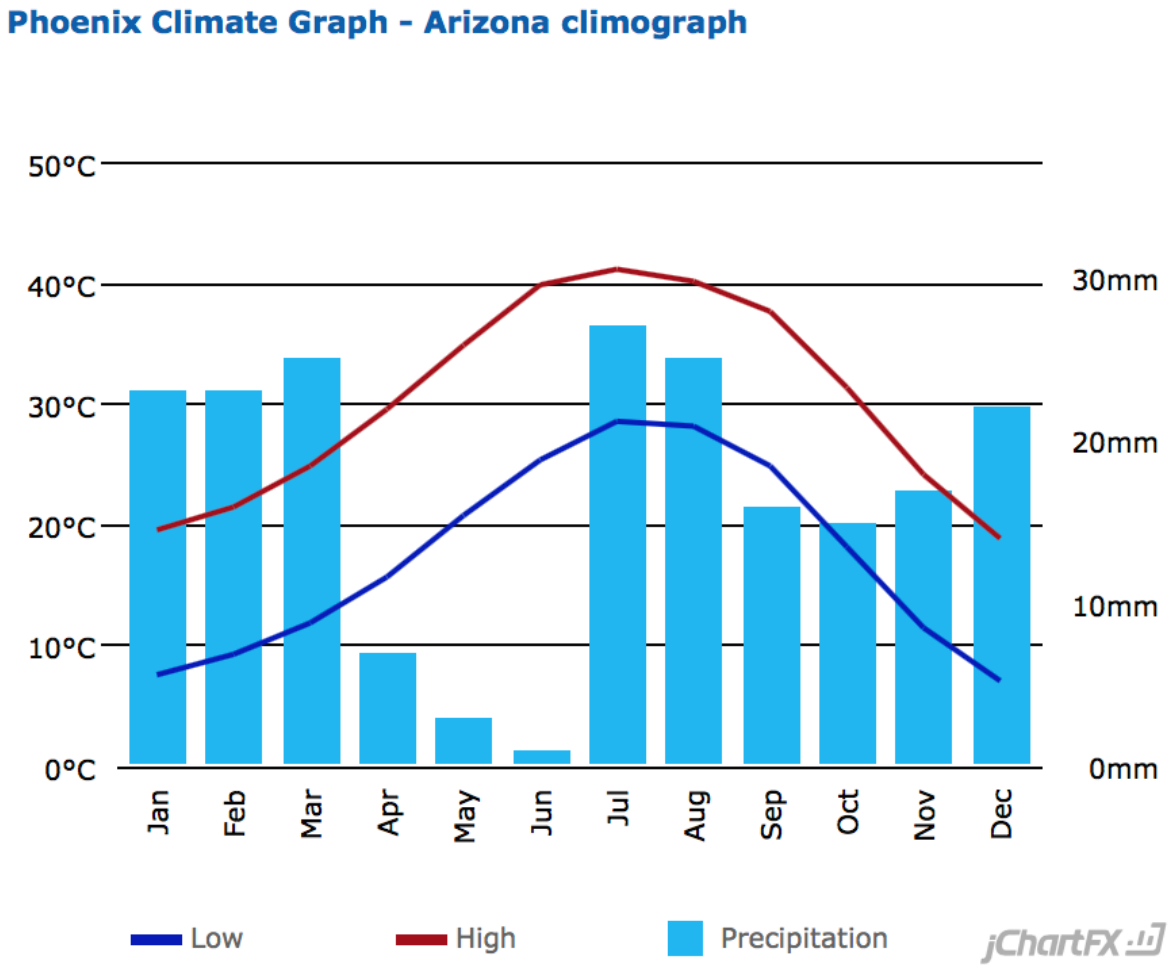
The data we will use here comes from NOAA (https://www.ncdc.noaa.gov/). Specifically, it was downloaded from This FTP site.

We focused on six measurements:

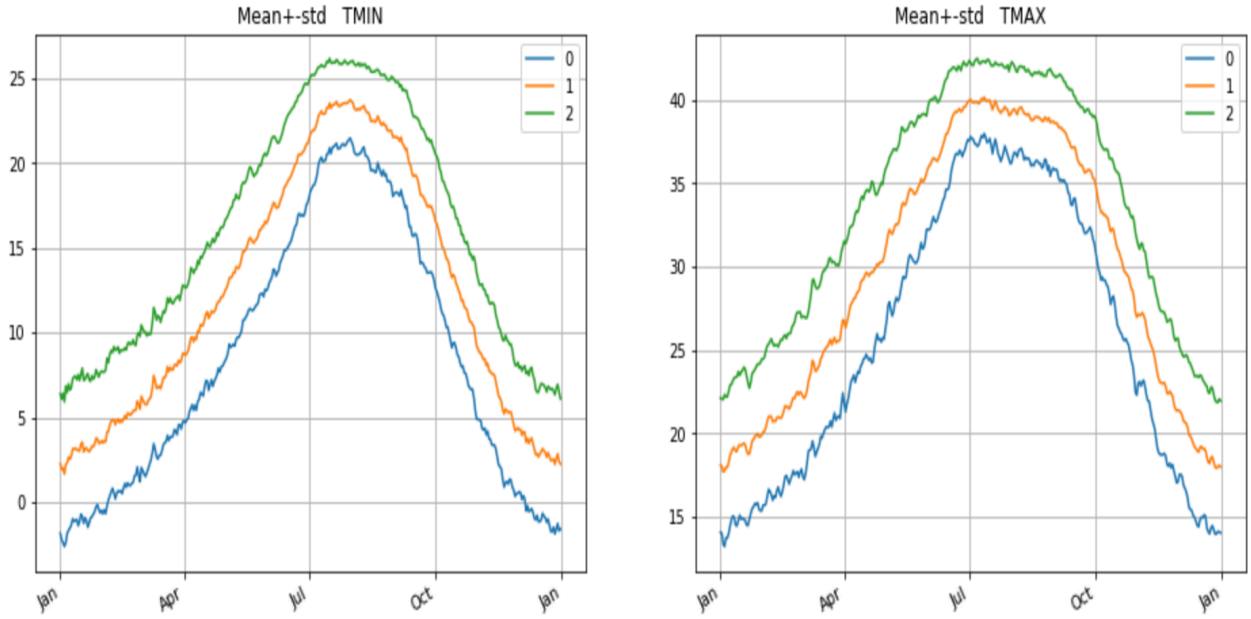
- **TMIN, TMAX:** the daily minimum and maximum temperature.
- **TOBS:** The average temperature for each day.
- **PRCP:** Daily Percipitation (in mm)
- **SNOW:** Daily snowfall (in mm)
- **SNWD:** The depth of accumulated snow.

Sanity-check: comparison with outside sources

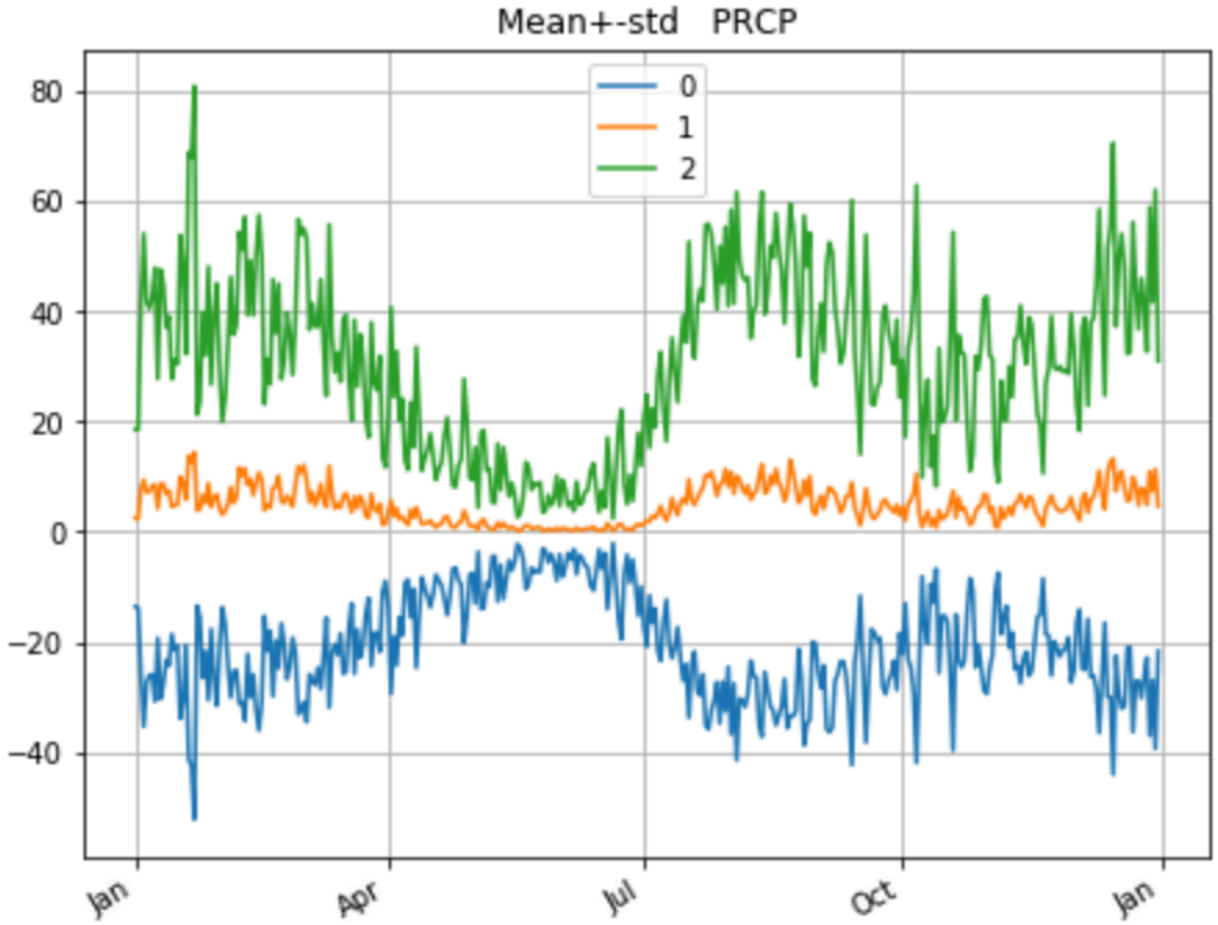
We start by comparing some of the general statistics with graphs that we obtained from a site called US Climate Data (http://www.usclimatedata.com/climate/boston/massachusetts/united-states/usma0046) The graph below shows the daily minimum and maximum temperatures for each month, as well as the total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



We see that the min and max daily temperature agree with the ones we got from our data in Centigrate.



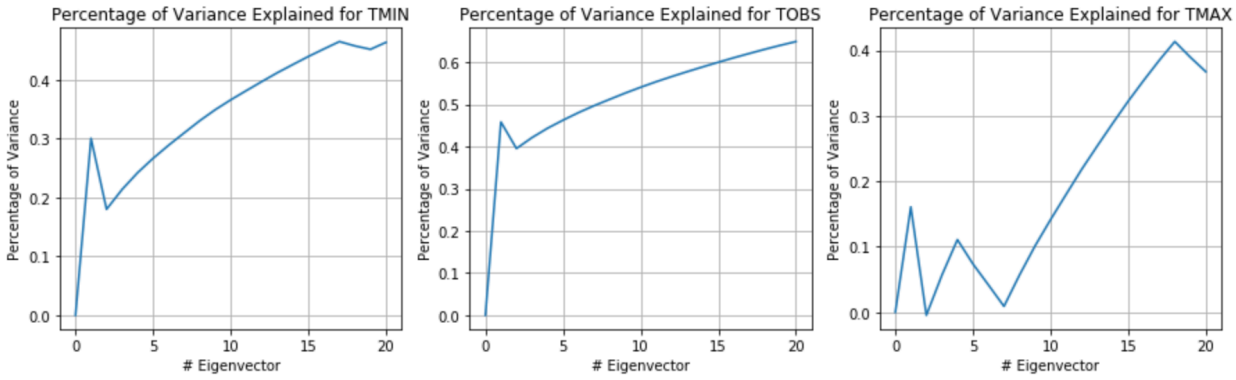
To compare the precipitation we look at the distribution of the rainfall over all the months in general. We see that the shape of the curve matched exactly with that obtained with our data. The highest amount of precipitation is in the end of July to beginning of August time frame. Comparing the recorded values for rainfall, we observe that our data records around 10 mm less rainfall uniformly. This could be strongly because the data also includes parts of Arizona that do not experience much rainfall and parts of California which are comparitively dry, thus lowering the mean precipitation.



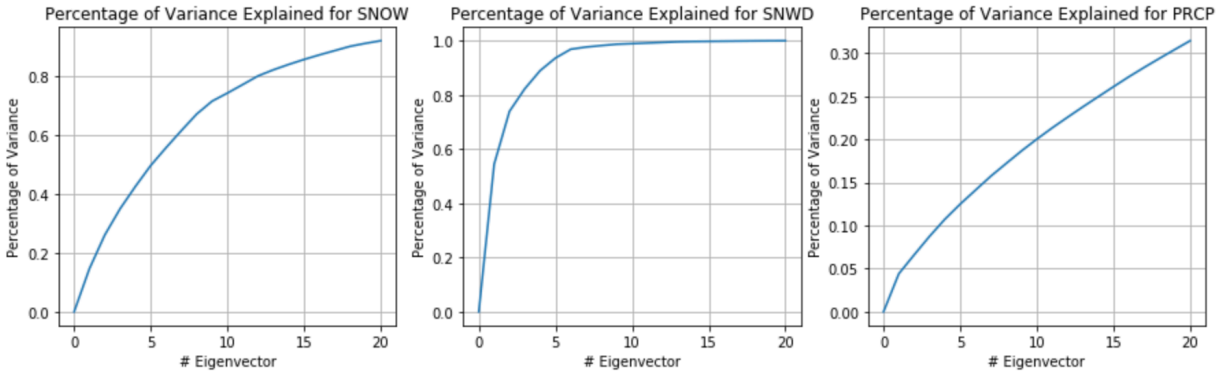
PCA analysis

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 observe an uneven rise and drop in the graphs for TMIN, TOBS and TMAX. This is due to the presence of negative eigen values for these measurements, which result when the covariance matrix is non-positive semi-definite. We defer the analysis of this to further sections.



The top 5 eigenvectors explain 13% of the variance for PRCP, which is a low value. The top 5 eigen values for SNOW explian 50% of the variance. However, the top 5 eigenvectors explain %95 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.

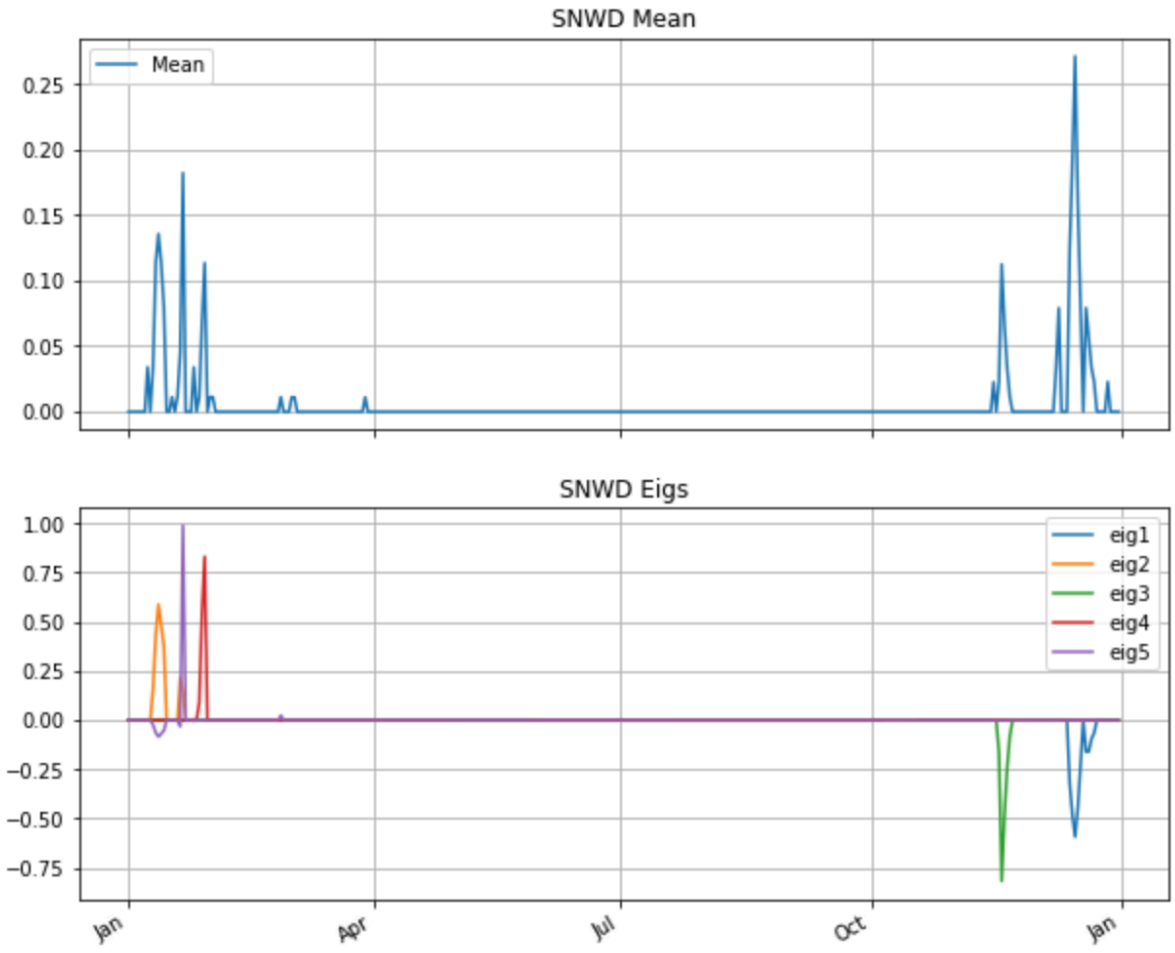
It makes sense that SNWD would be less noisy than SNOW. That is because SNWD is a decaying integral of SNOW and, as such, varies less between days and between the same date on diffferent years.

Analysis of snow depth

We choose to analyze the eigen-decomposition for snow-depth because the first 5 eigen-vectors explain over 95% of the variance.

First, we graph the mean and the top 5 eigen-vectors.

We observe that the region does't have snow over the major portions of the year. There is intermittent snow recorded in the mid-November to end-of-January period.

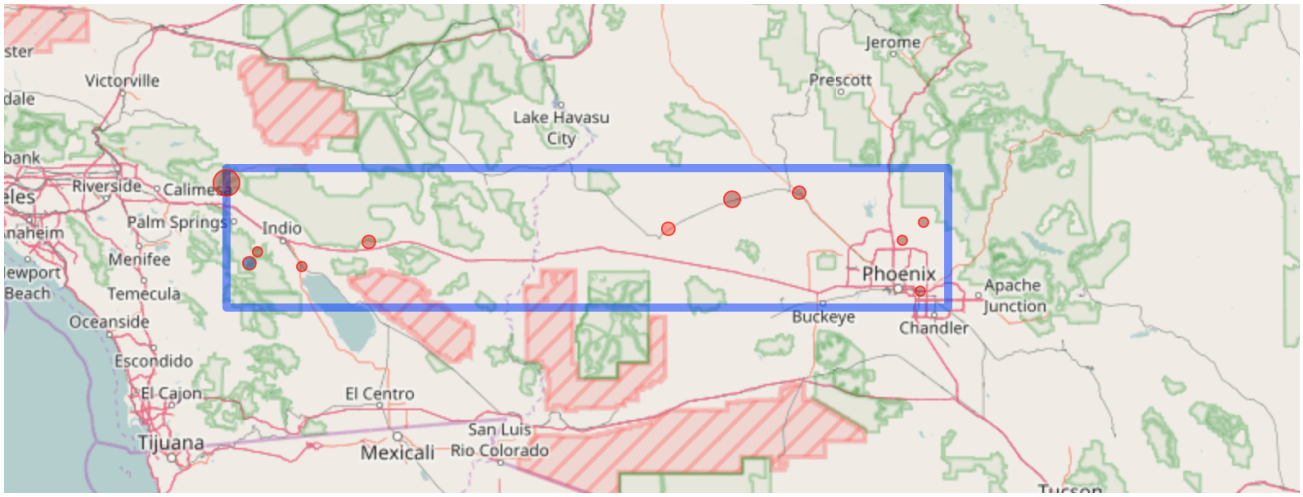


Next we interpret the eigen-functions. We observe that none of the them stand out in terms of explaining the ditribution of SNWD in terms of its mean. **eig1** mostly agrees with the time period over which there is 0 snowfall (all the colours all overlapping at the 0 mark due to which we cannot clearly differentiate between them) and says that the snow is lesser than its mean in the mid-Decemeber period. **eig 2** agrees with the mean in terms of the snow in mid-January, **eig 3** says that the snow in mid-November is lesser than the mean, **eig 4** seems to exactly coincide with the mean snowfall in beginning of February and **eig 5** exactly coicides with the mean snowfall at the end of January.

Overall, we can say that the top eigen functions collectively agree to the fact that there is no snow in the March to November period, which is in agreement with the mean. Individually, they explain the increase or decrease of snow in certain time periods as mentioned above with respect to the mean.

Plotting stations with non-zero SNWD

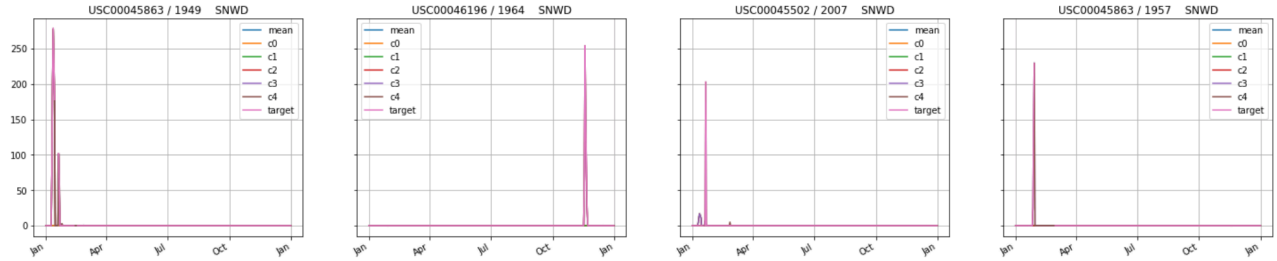
On plotting the stations which report a non-zero value for SNWD, we observe that only 11 stations feature in this list with a total of 23 measurements. These stations are distributed sparsely throughtout, with the station having 7 observations located in the top-left corner of the region (in California). Moreover, the colors of the colors correspond to different values of coeff_1.



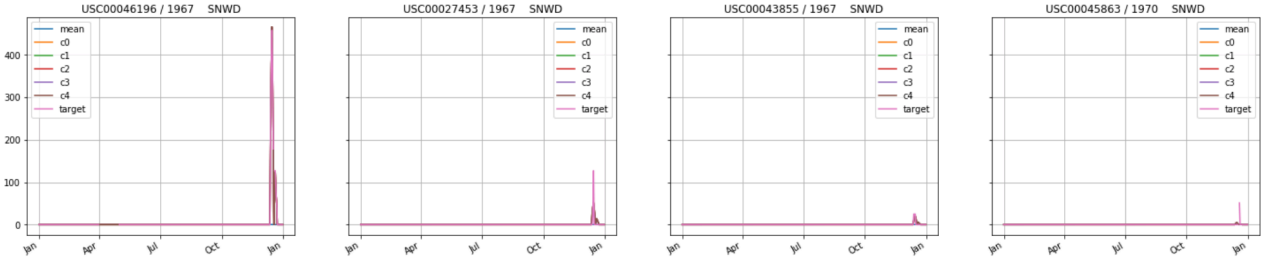
Examples of reconstructions

Coeff1

Coeff1: most positive



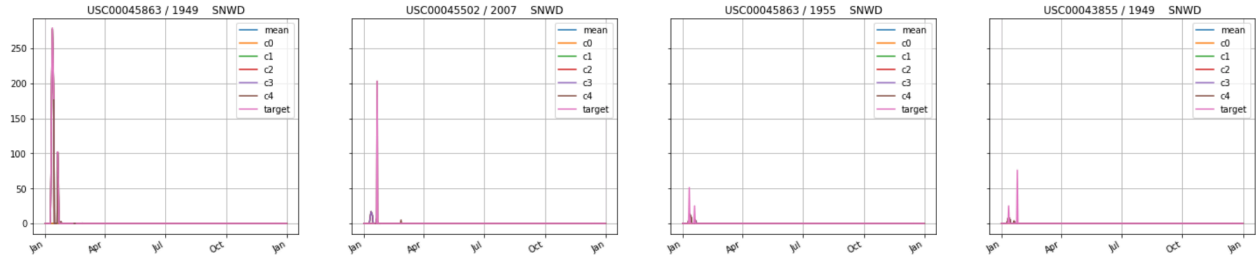
Coeff1: most negative



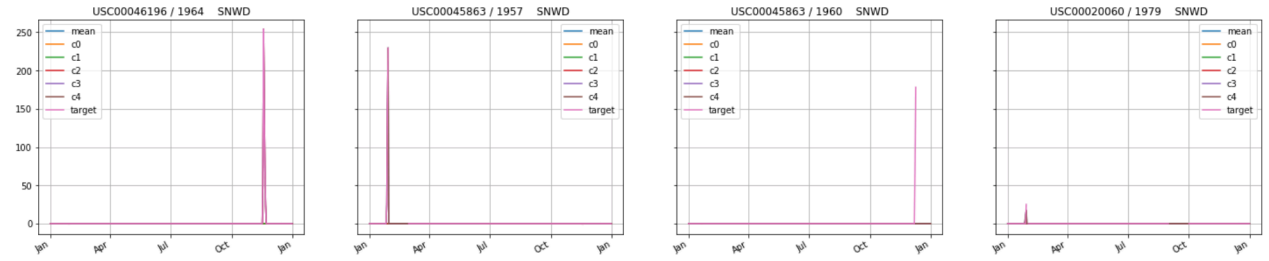
Large positive values of coeff1 correspond to high amounts of snow. Low values correspond to snow in Decemeber.

Coeff2

Coeff2: most positive



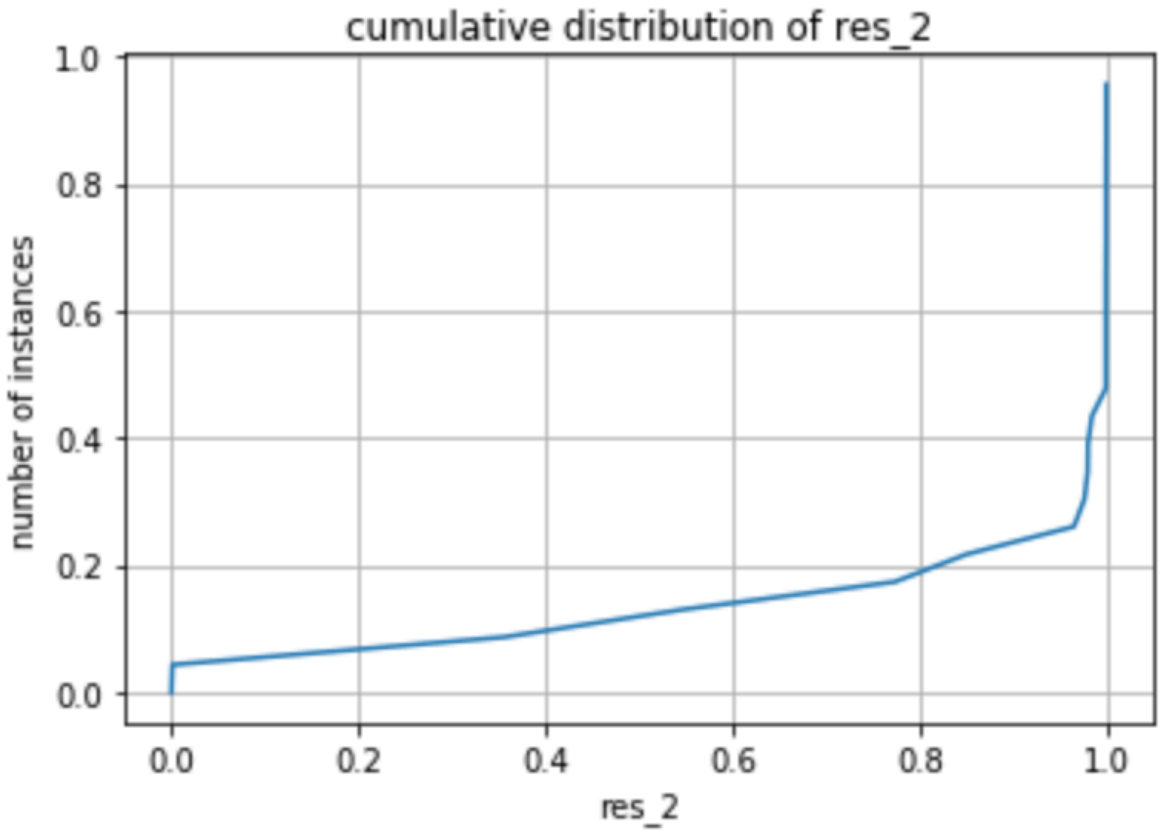
Coeff2: most negative



Large positive values of coeff2 correspond to a early snow season. Negative values for coeff2 do not seem to correspond to a pattern.

Moreover, the above data was taken after filtering out residuals < 1.0 as adding a smaller filter than 1 did not account for more than 1 station in a few cases and did not allow us to see a pattern. However, overall we can say that the high variance explained in case of SNWD is due to the fact that most of the stations (by a large margin) have 0 values recorded for this measurement and our eigen functions are largely predicting this 0 value.

Below graph shows that more than 50% is the data has a res_2 value of 1, which supports the above claim.

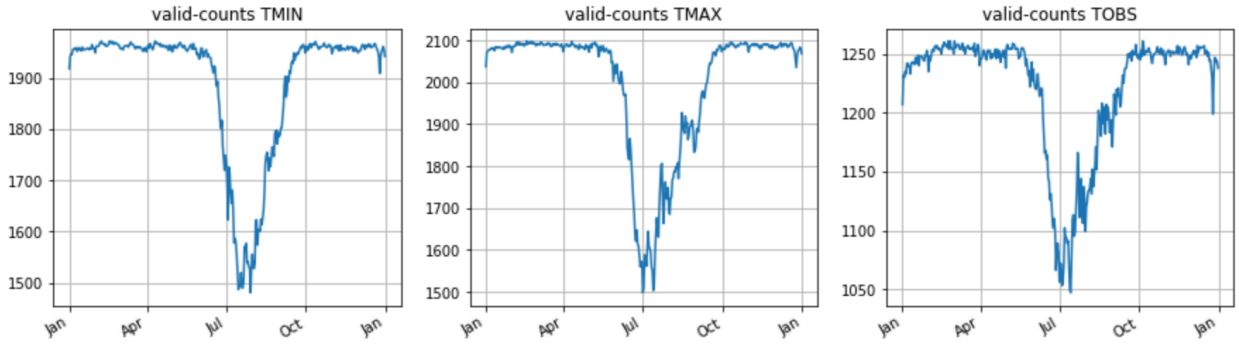


Analysis of Temperature (TMAX, TMIN, TOBS)

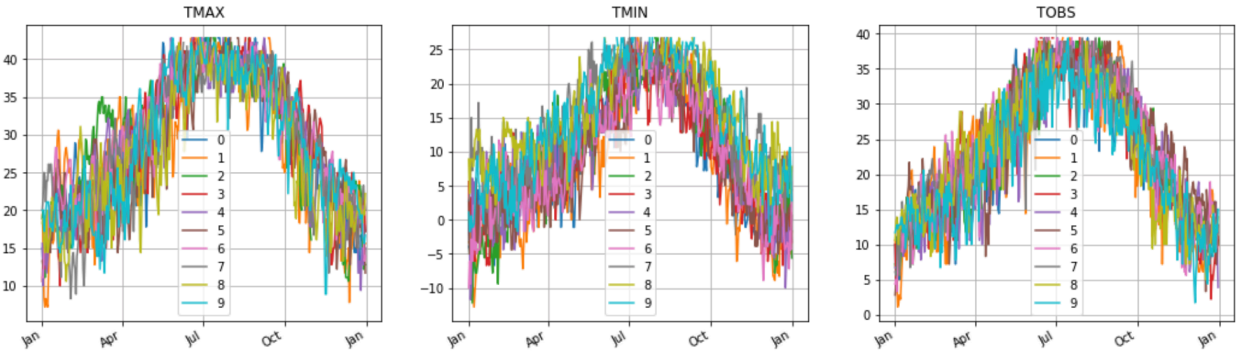
We choose to analyze and understand the rise and dip in the graphs for TMAX, TMIN and TOBS for the Percentage of variance explained. Upon further study it can be seen that the eigen values corresponding to the dip are negative, which resulted from the fact that the covariance matrices for these measuremets were not positive semi-definite.

Several reasons can cause the covariance matrix to be semi-definite such as, errors in calculations, linear dependency between the dimensions, too many NANs etc.

So we study the number of valid counts for each of these measurement as follows:



From the above plots we can clearly see a big drop in the number of valid counts in the data for the time period around July, when this region expriences maximum temeratures. To confirm this we plotted the time series data for 10 random years



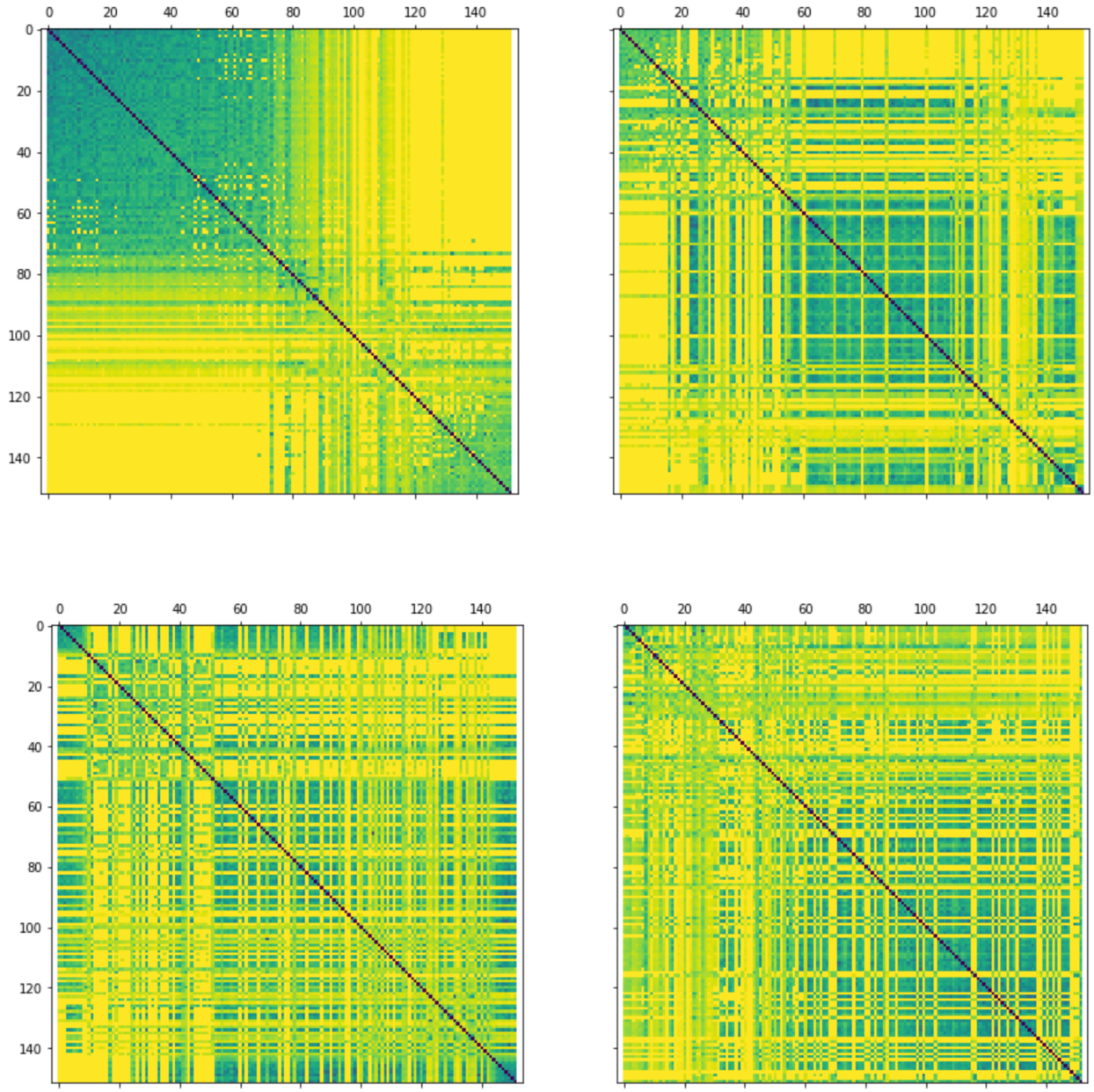
We can clearly see that during the data cleaning process, the values above temeratures of about 42 degree Celcius for TMAX, about 26 degree Celcius for TMIN and 39 degree Celcius for TOBS, which occur in the July time-frame, were marked as NAN. This lead to the negative eigen values for these measurements.

For further valid analysis on temperature, we may need to tune the data cleaning parameters specifically to this region as it has higher temperatures compared to majority of the regions in the USA. We also learn that these high temperatures were strong factors in accurately reconstructing our temperature measurments.

Analysis of Precipitation

We choose to analyze precipitation as the mean precipitation pattern matches very closely with the actual levels of precipitation and Arizona is expected to show good trends in terms of rainfall.

We plot the correlation between the stations in terms of precipitation as shown below:

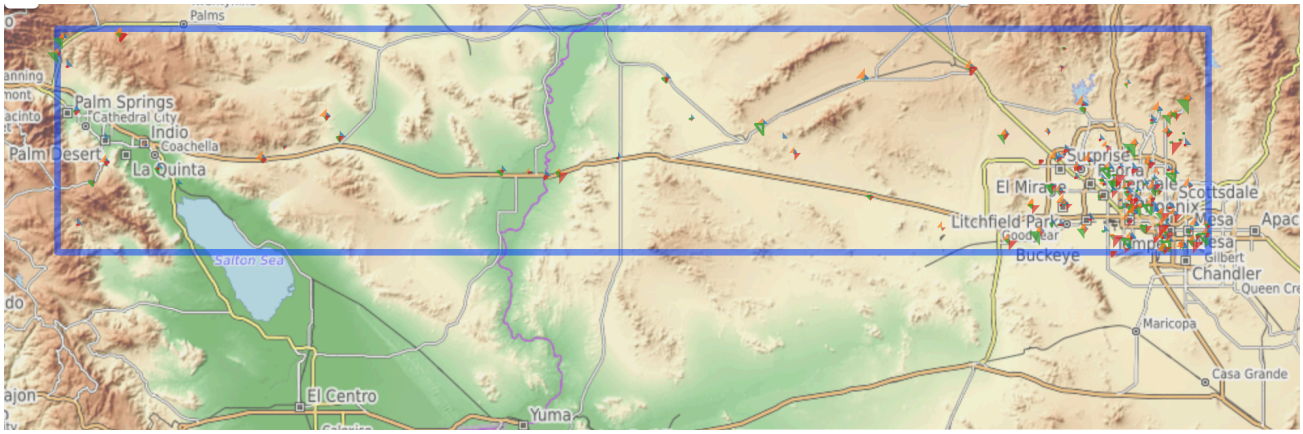


Consider the upper left square of the first matrix. The stations at positions 0-75 are clearly strongly correlated with each other. Within these, we see that there are some stations, in positions 0-45 or so, which are more related to each other than to the rest of this block.

The lower right region of the first matrix also shows some correlation, with the stations at positions 140-159 being more closely correlated with each other than the rest.

The same correlation is less visible in the other matrices, though such a trend is present in all of them. We clearly expect around 40% of the stations to be correlated with each other.

We plot these stations on a graph to see their spatial relationship with each other.



We can clearly see that the large number of stations are densely located in Phoenix. This prompts us to explore if precipitation is spacially or temporaly related in this region.

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

Here we 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 = 126.785315693
RMS removing mean-by-station= 105.654495455, fraction explained = 16.67%
RMS removing mean-by-year = 47.0170780143, fraction explained = 62.92%

coeff_2
total RMS = 87.1850363857
RMS removing mean-by-station= 84.3939920488, fraction explained = 3.2%
RMS removing mean-by-year = 45.3673615536, fraction explained = 47.96%

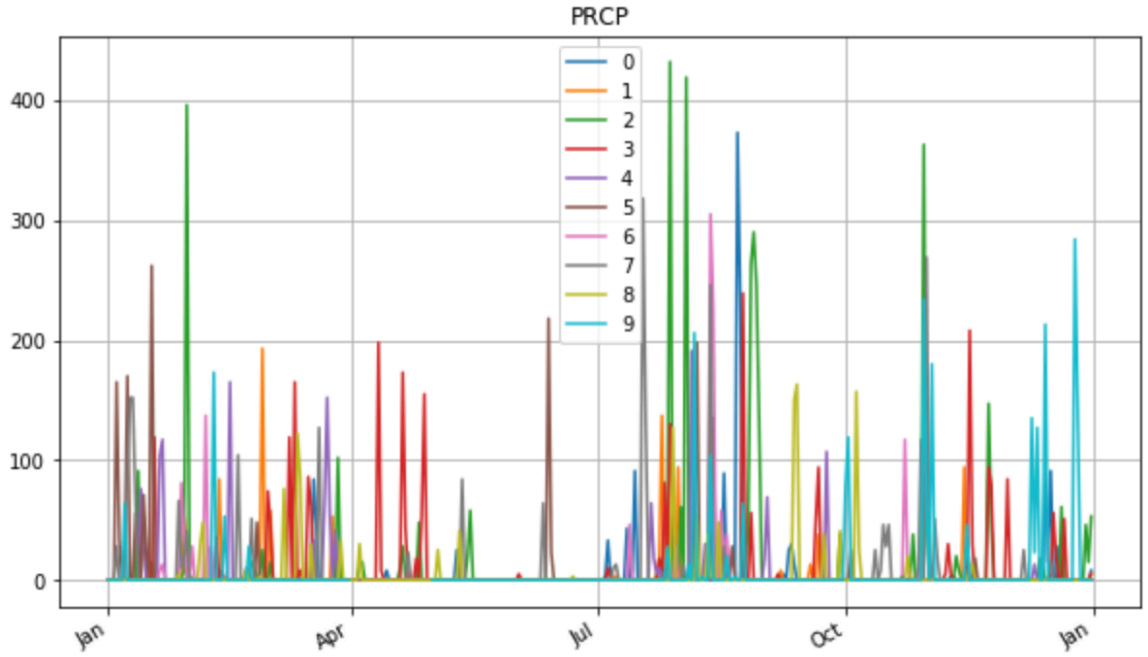
coeff_3
total RMS = 87.4634462146
RMS removing mean-by-station= 81.2733284255, fraction explained = 7.07%
RMS removing mean-by-year = 44.8067267246, fraction explained = 48.77%

We see that the variation by year explains more than the variation by station. Looking at the values for fraction explained, we can say that coeff_1 is more related to the timing of rainfall compared to coeff_2 and coeff_3. The explanantion provided by station-to-stations is very less, in the range 7%-17%, which clearl shows that precipitation is more of a temporal feature as opposed to a spatial one.

To visualize the observation made above, we plot the mean precipitation for 10 randomly chosen stations

stations.

Relationship between the stations:



We see very less overalap between the timings in precipation, which is due to the year-to-year variation.