# Weather Analysis of South Eastern Part of USA

INTRODUCTION:

This report consists of the analysis of the Dataset: 'BSBSSSSS' which was given to me. This dataset primarily consists of the weather data accumulated from weather stations in the states of Florida, Atlanta and Georgia. The total number of weather stations under consideration are 120.
The geographical locations of the weather stations are shown in Fig. 1 and Fig. 2.

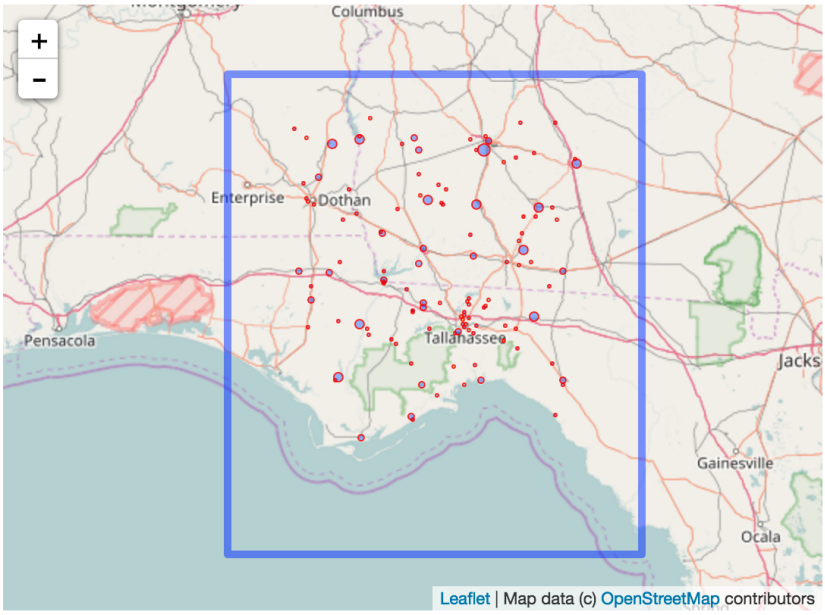


Figure 1: Zoomed in view of the weather stations on the US Map

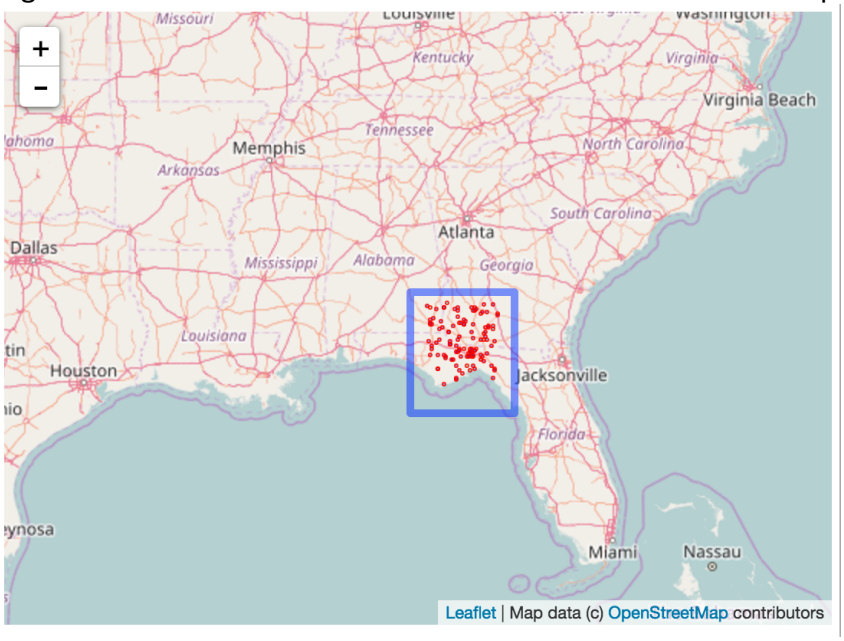


Figure 2: Zoomed out view of the weather stations on the US Map

## UNDERSTANDING THE DATASET:

The locations of the weather stations lie above the tropic of cancer and lie in the sub-tropical zone. So, we can in general associate them with moderate to high rainfall, moderate to high temperatures and low to no snowfall. In order to get an idea of the region, the general monthly statistics for the weather stations around Tallahassee, Florida is shown in Fig. 3.

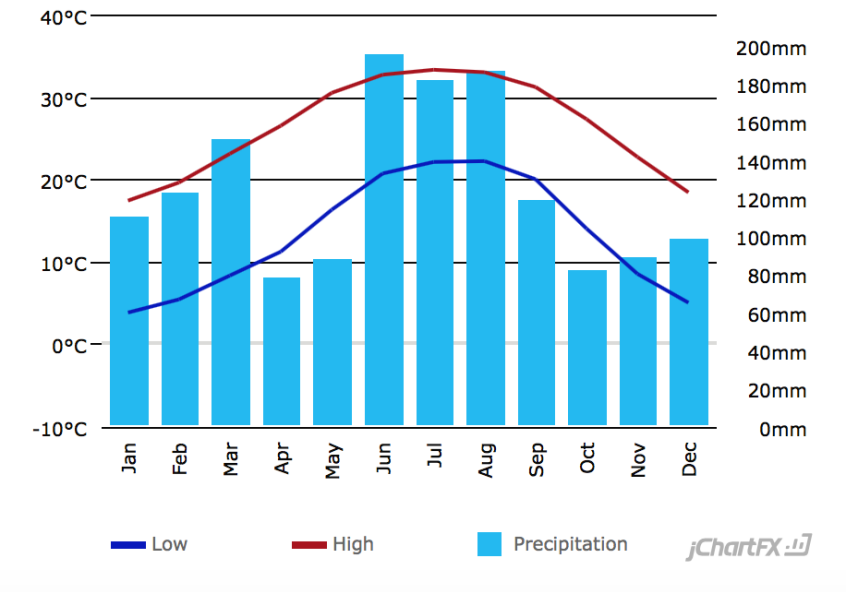


Figure 3: Temperature and Precipitation chart for Tallahassee, Florida

## STATISTICAL ANALYSIS ON THE DATASET:

The dataset used for analysis was obtained from the NOAA. We are primarily focused on performing analysis on six key variables:

- **TMIN –** The daily minimum temperature (in tenth of C)
- **TMAX –** The daily maximum temperature (in tenth of C)
- **TOBS –** The average temperature for each day (in tenth of C)
- **PRCP –** Daily Precipitation (in mm)
- **SNOW –** Daily snowfall (in mm)
- **SNWD –** The depth of accumulated snow

The TMIN and TMAX for the given dataset is shown in Fig. 4. The mean, mean-std and mean+std for TMIN and TMAX are plotted in Fig. 4. The mean TMAX and TMIN are similar to the temperatures obtained for Tallahassee, Florida as shown in Fig. 3.

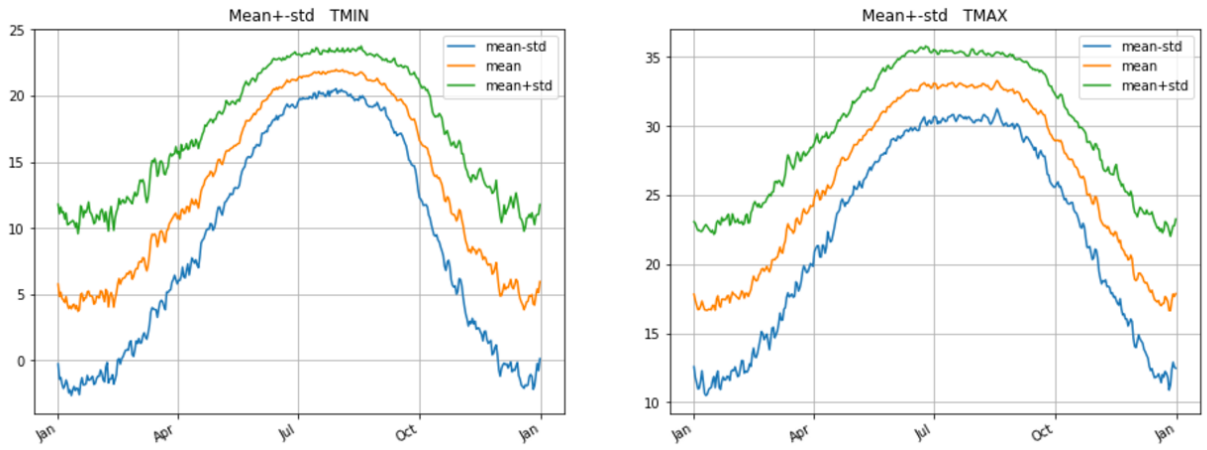


Figure 4: The Mean+-std of the TMIN and TMAX (Temperature in Celsius).

The average precipitation levels for the dataset are shown in Fig. 5.

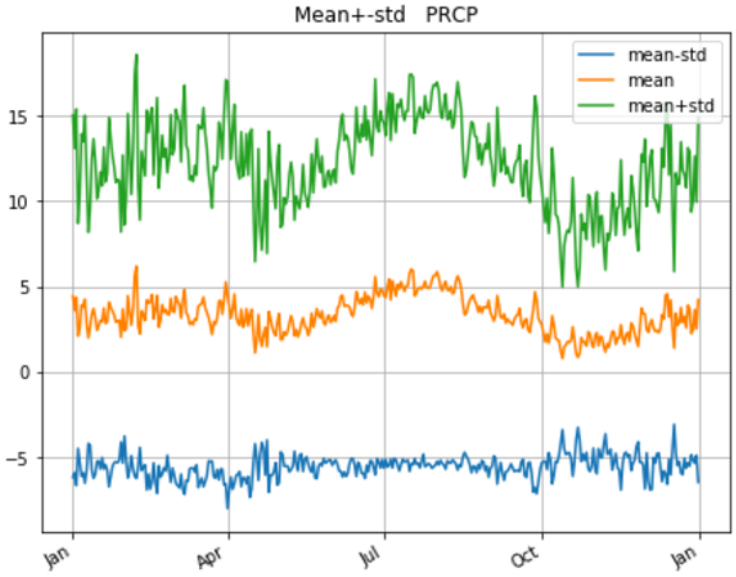


Figure 5: Mean values of precipitation (in inches)

The average precipitation for the Tallahassee region is also close to 5.9 inches, and the mean precipitation obtained from Fig. 5 is close to 5.5 inches.

Thus, the dataset under consideration is in agreement with the data obtained from external sources.

## PCA ANALYSIS:

The six variables under consideration are taken and we compute the percentage of variance explained as a function of the eigen-vectors used. The percentage of variance explained for TMIN, TOBS , TMAX, SNOW, SNWD and PRCP are given in Fig. 6 and Fig. 7 for their first 5 eigen vectors.

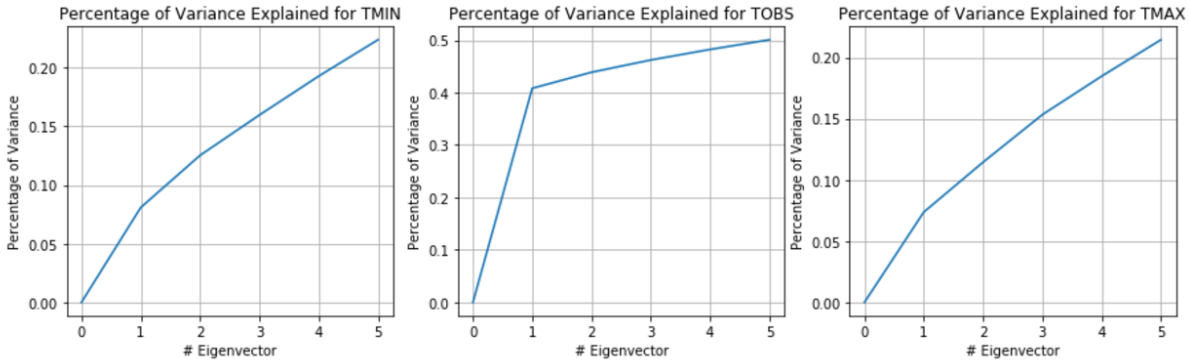


Figure 6: Percentage of Variance explained for TMIN, TOBS and TMAX w.r.t their eigenvectors.

As we can see from Fig.6, the top 5 eigen vectors explains 23% of the variance for TMIN, 50% of the variance for TOBS and 22% of the variance for TMAX. Thus, we can see that TOBS is best explained by the top 5 eigenvectors. This is particularly true for the first eigen vector of TOBS, which explains close to 41% of the variance.

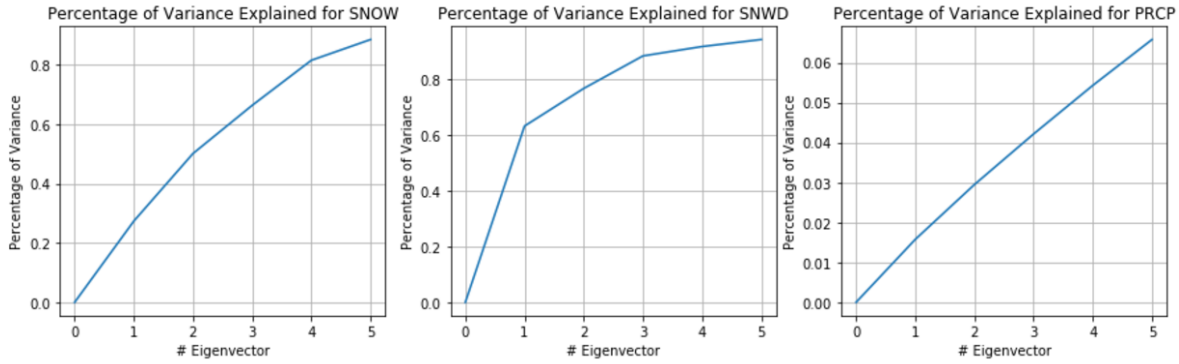


Figure 7: Percentage of Variance explained for SNOW, SNWD and PRCP w.r.t their eigenvectors.

As we can see from Fig.7, the top 5 eigen vectors explains 83% of the variance for SNOW, 84% of the variance for SNWD and 6.5% of the variance for PRCP. Here, we can see that most of the variance of the SNOW and SNWD is explained by the first 5 eigen vectors, this is intuitive as there isn't a lot of change in the SNOW measurements for these regions as they are in a sub-tropical region.

## ANALYSIS OF PRECIPITATION:

The amount of variation in the PRCP variable is the highest and it very difficult to reconstruct it based on performing PCA and combining it with the mean and the first k eigen values. The mean and the first 3 eigen vectors of PRCP are shown in Fig. 8.

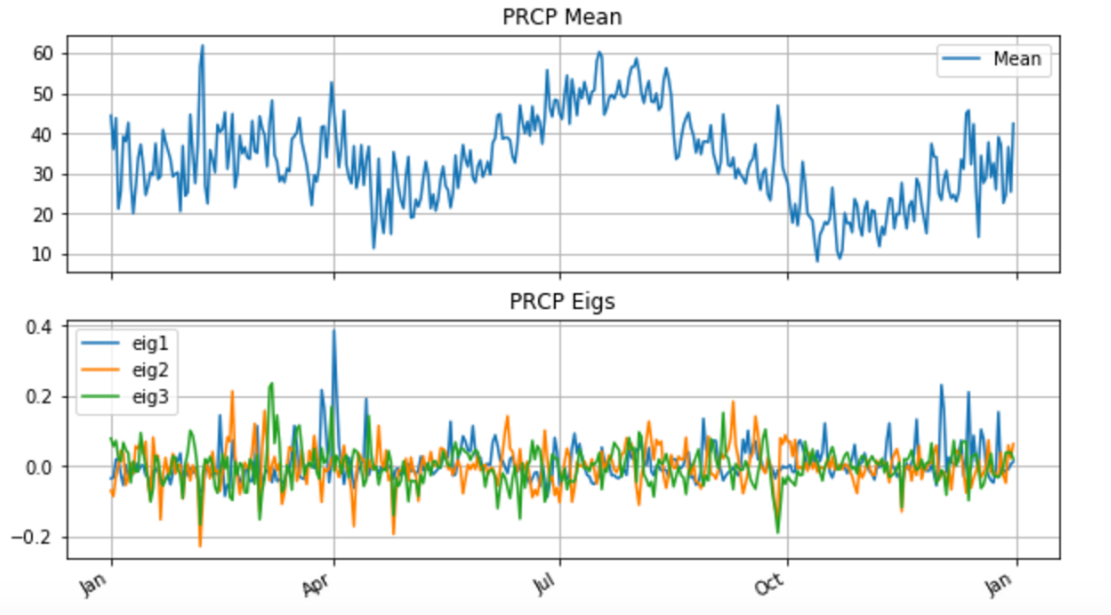


Figure 8: Mean and the first three eigen vectors of PRCP

From Fig. 8, we can observe that there is a lot of variation in the mean and it looks very difficult to try to reconstruct the PRCP variable using a small number of eigen vectors.

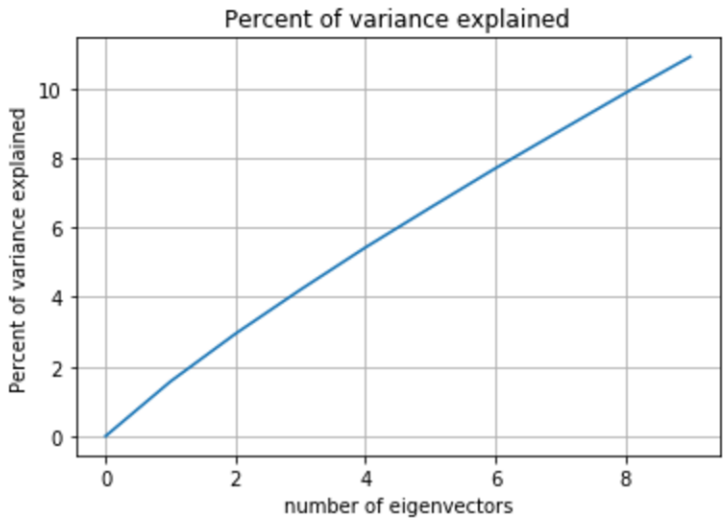


Figure 9: Percentage of variance explained by the number of eigen vectors

From Fig. 9, we can see that there is a nearly linear relation to increasing the number of eigenvectors and the percentage of variance explained.

Next, we try to compute the residual values by subtracting the mean, projection on the first eigen vector, projection on the second eigen vector and projection on the third eigen vector. After that, we try to filter the residual mean values which are less than 1 (i.e.) it gives a worse approximation of mean than zero.

That results in 2597 stations, which are having a mean residual better explained than zero. Out of these stations we display the twelve best entries(station and the corresponding year) which were best explained by the mean and the first three eigen vectors. These entries have low values of residual sum. This is given in Fig. 10.

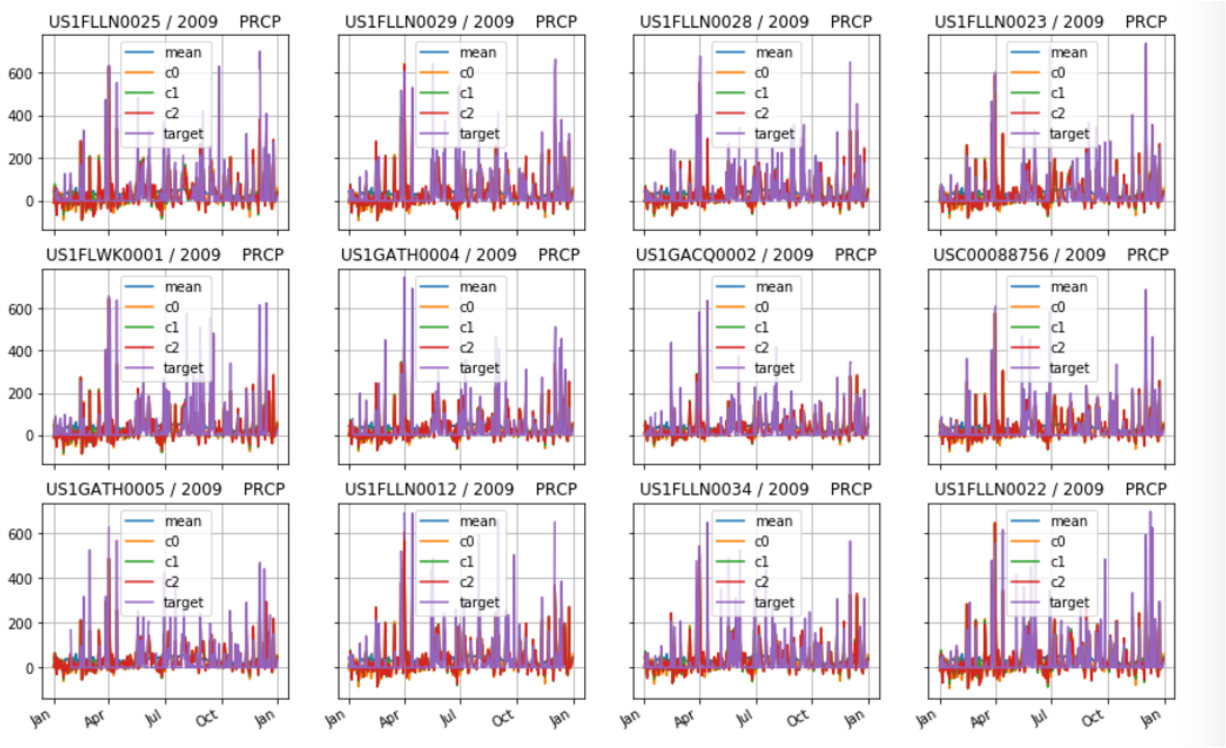


Figure 10: Reconstruction the PRCP target vector using mean, eig vector 1,2,3 for the respective stations (best 12).

Out of the stations we display the twelve worst entries(station and the corresponding year) which were worst explained by the mean and the first three eigen vectors. These entries have high values of residual sum. This is given in Fig. 11.

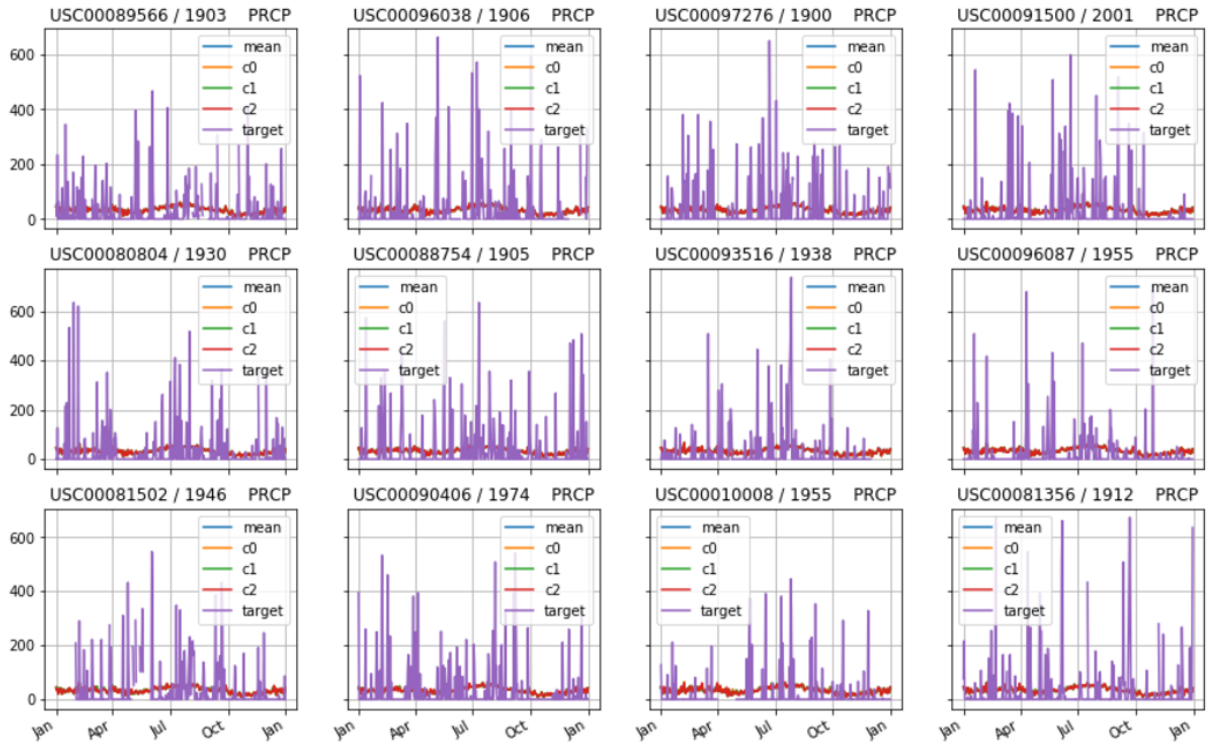


Figure 11: Reconstruction the PRCP target vector using mean, eig vector 1,2,3 for the respective stations (worst 12).

## ANALYSIS OF SNOW DEPTH:

On performing analysis of the SNWD variable for the given dataset, we find that there is not much variation with respect to the different months. This is primarily because the region we are dealing with is a sub-tropical region with no snowfall. The occurrences of snow can be due to high elevation which will be explored later.

The mean of the snow depth and the first three eigen vectors are given in Fig. 12.

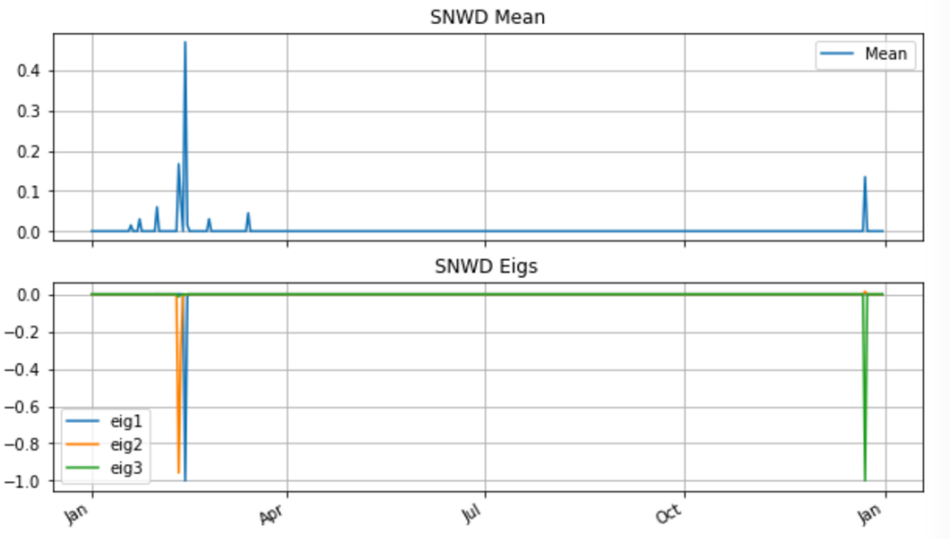


Figure 12: Mean and the first three eigen vectors of PRCP

We can observe that, because of the small amount of snowfall in the region, there isn't much variance in the mean of SNWD. The mean has three prominent peaks in February, March and in December and the three eigen vectors explain the occurances of snow in these months. The first eigen vector explains the most snowdepth.

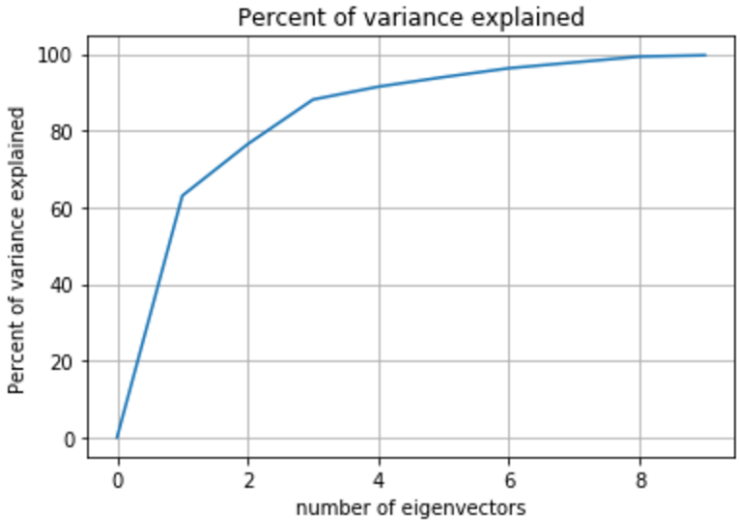


Figure 13: Percentage of variance explained by the number of eigen vectors

From Fig. 13, we can see that the SNWD variable is explained completely by just 8 eigen vectors. The first eigen vector explains close to 62% of the variance.

On further analysis of the SNWD information, we find that only 19 stations in the dataset have Snow Depth recorded. Out of these, only two stations have had correlated snow occurances and these stations are 'USC00095585' and 'USC00080804'. So, the occurrences of snow are rare and mostly for the years 1958 and1973.

The analysis of SNWD to find if it is a spatial variation or a temporal variation, we observe that it is primarily as temporal variation where most of the 19 stations, experienced snow in the year 1958 and also coupled with the fact that this sub tropical region doesn't experience much snow, so the logical explanation would be that SNWD is a temporal variable.

The occurrences of snow cannot be attributed to the elevation of the stations are all the stations that recorded snow had elevation less than 500 meters.

## VISUALIZING THE STATIONS ON A MAP:

The locations of the stations from which the data for the PRCP variable was collected is shown in Fig. 14.

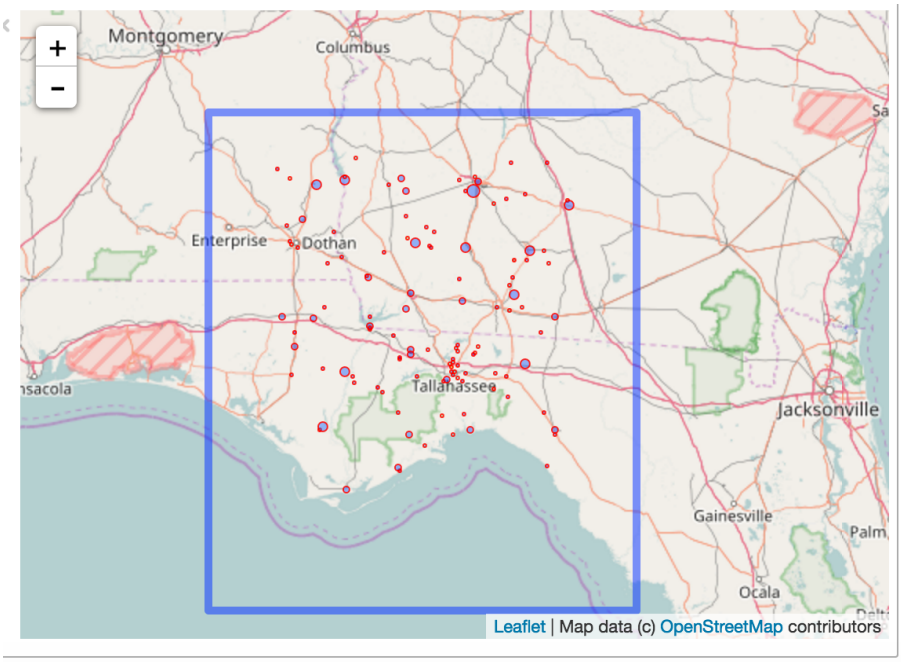


Figure 14: Locations of the stations where PRCP data was collected

The radius of the bubble indicates that higher the count of the stations in a particular region then larger the radius.

The locations of the stations from which the SNWD data was observed is shown in Fig. 15.

We can see that just 19 stations in the map from which the snow depth data was observed.

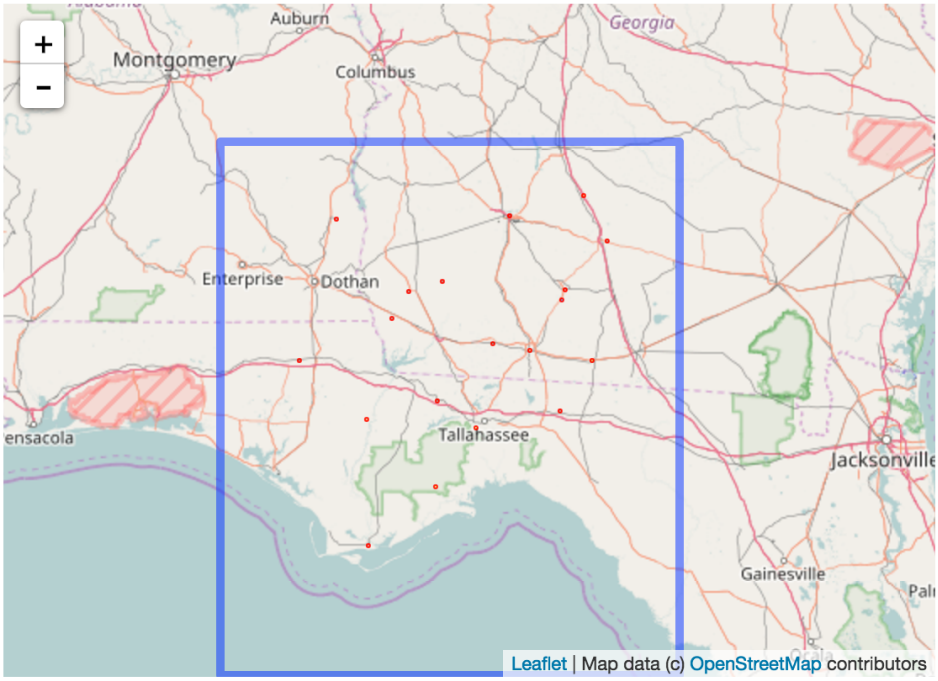


Figure 15: Locations of the stations where SNWD data was collected

# ANALYSIZING THE PRECIPITATION PATTERN USING EIGEN VECTORS:

We try to unpack the station, year and the PRCP values for the corresponding year. We find out that the maximum recorded length possible is 44530 days. Now, we fill in the precipitation values collected for all the stations and the for the years for which data is available and fill in zeros for all other entries.

We find out that we have 120 stations which have PRCP values collected. We then select the top two stations (number of days the PRCP values were collected) and try to find if it rained in both locations on the same day. The stations 'USC00090979' and 'USC00097276' were taken since these have the maximum number of entries, 35357 and 34623 respectively. After removing the nan values we have 28351 entries for both the stations. The probability of occurrence of precipitation between the two stations is 0.0776.

We then go on to compute log probability for each pair of stations and try to plot a histogram for it, the value 0.4 is obtained when log probability is computed for the same station. The histogram obtained is shown in Fig. 16.

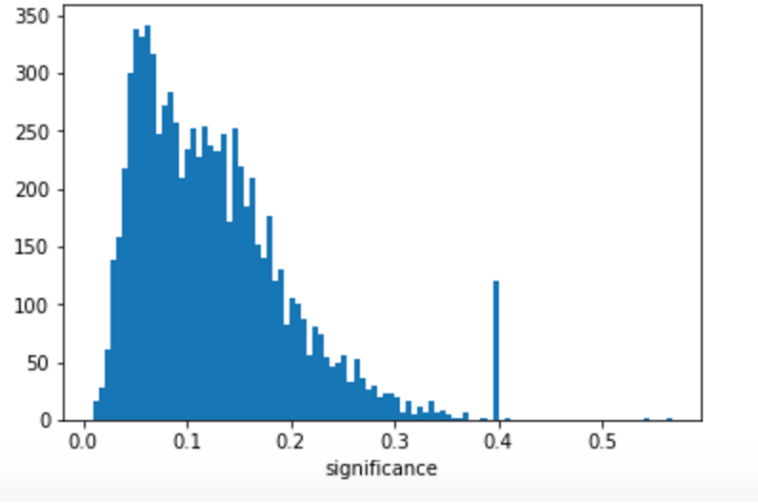


Figure 16: Histogram of the log probabilities observed and the number of occurrences

From the figure, we can observe that most of the stations have a low value of log probability for the pair less than 0.2.
This shows us that the our p value is very low and so we can reject our null hypothesis which states that no two stations are correlated.
This is further explained by our correlation matrix.
The matrix showing the likelihood between the pair of stations in shown in Fig. 17.

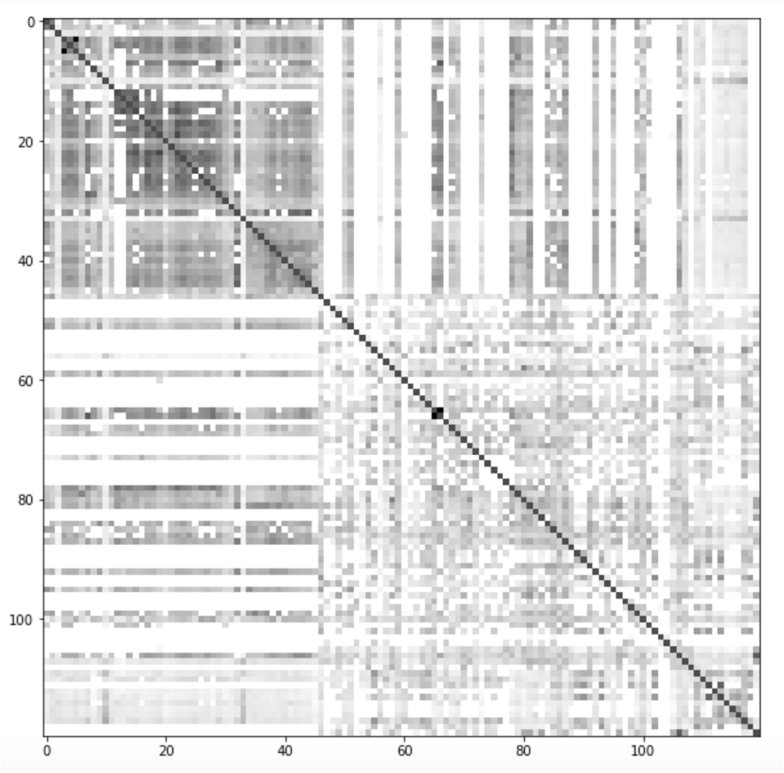


Figure 17: Correlation matrix between the 120 stations

From Fig. 17, we can see that some stations have much more correlated PRCP data than others, stations from 15 to 30 seem to have higher correlation than the others.

Now we try to find the structure in the dependency matrix, by performing SVD and then we can see that top 10 eigenvectors explain about 80% of the square magnitude in the matrix. This is shown in Fig. 18.

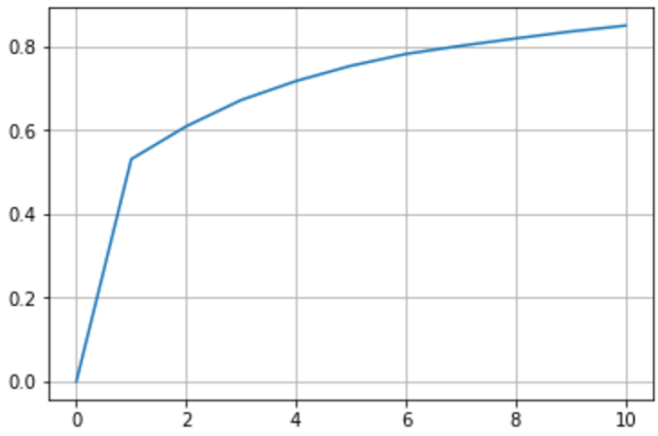


Figure 18: Number of the eigen vectors vs percentage of the square magnitude explained

We now consider the first four eigen vectors which explain close to 75% of the square magnitude. The dependency matrix obtained for the four eigen vectors is shown in Fig. 19.

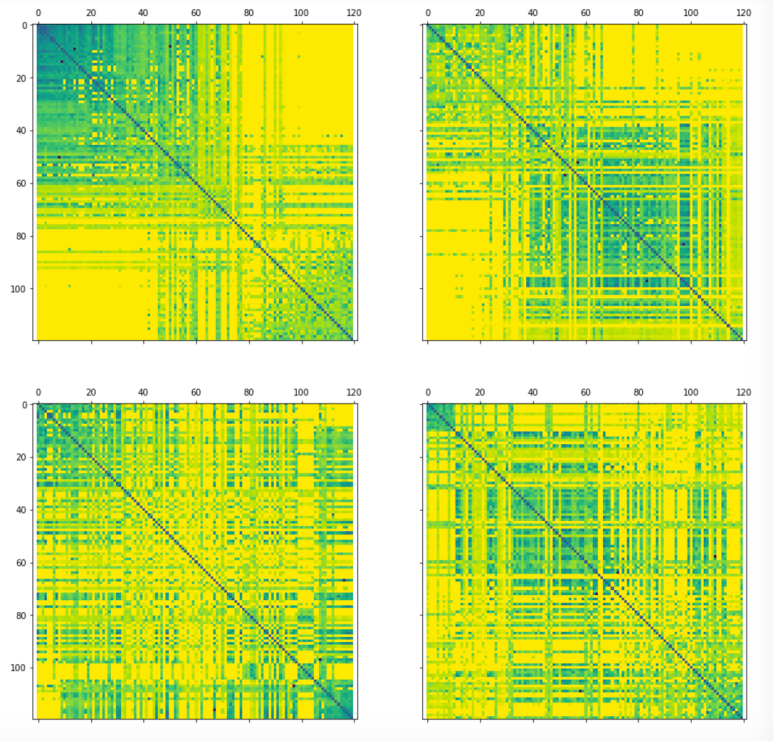

Figure 19: Dependency matrices for the first 4 eigen vectors

From Fig. 19, we can see that this type of Block Diagonal organization, helps us identify the patterns in data much better. From the first eigen vector, we can see that stations in the first 25 positions are strongly correlated with each other and from the second eigen vector we find a grouping of strongly correlated stations from indices 65-100.

So, we tried to visualize the first 25 stations from the first eigen vector and we the observed the grouping as shown in Fig. 20.

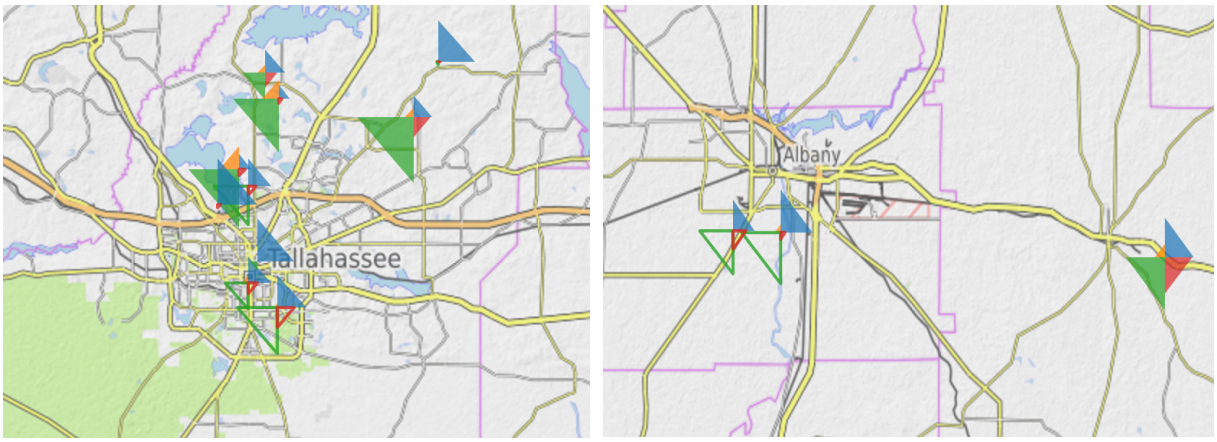

Figure 20: The locations of the majority of the first 25 stations whose indices were obtained from the first eigen vector.

From Fig. 20, we can see that the pattern of the grouping is very similar for stations that are close together. Few stations near Tallahassee, Florida and Albany, Georgia show this strong correlation. So, the PRCP can be viewed as a temporal variable rather than a spatial variable. The opaque polygons represent a positive coefficient and the transparent polygon represents the negative coefficient.

## ANALYSIZING THE TEMP OBSERVED PATTERN USING EIGEN VECTORS:

To clearly understand the effective utilization of PCA to analyse the contribution of the eigen vectors to a variable, we analyze the TOBS variable.

From Fig. 21, we can see that mean of the TOBS variable provides the general shape to the variable and eigen vectors try to reshape it to fit the target curve better.

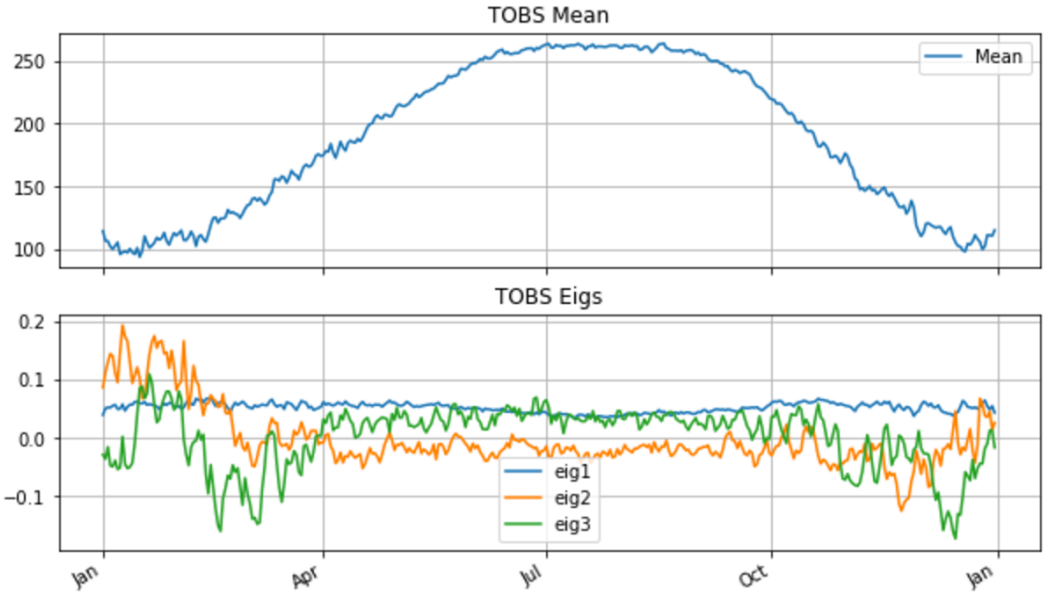


Figure 21: Mean and the first three eigen vectors of TOBS

The percentage of variance explained vs the number of eigen vectors is shown in Fig. 22, we can see that 6 eigen vectors explain close to 50% of the variance.

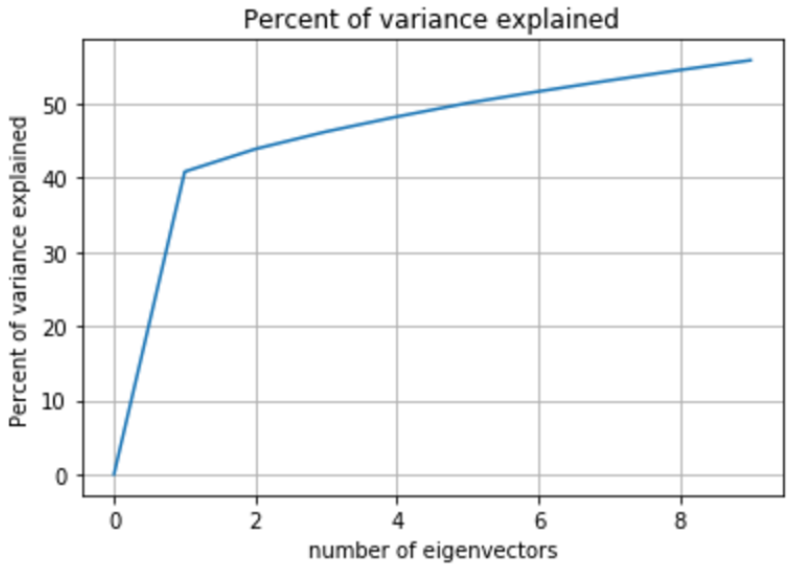


Figure 22: Number of eigen vectors of TOBS vs the percentage of variance explained

From Fig. 23, we can see that mean provides the necessary shape to attain the target variables, the eigen vector 1, the contribution of the eigen vector 1 is negative and higher its coefficient value it pulls the curve downwards.

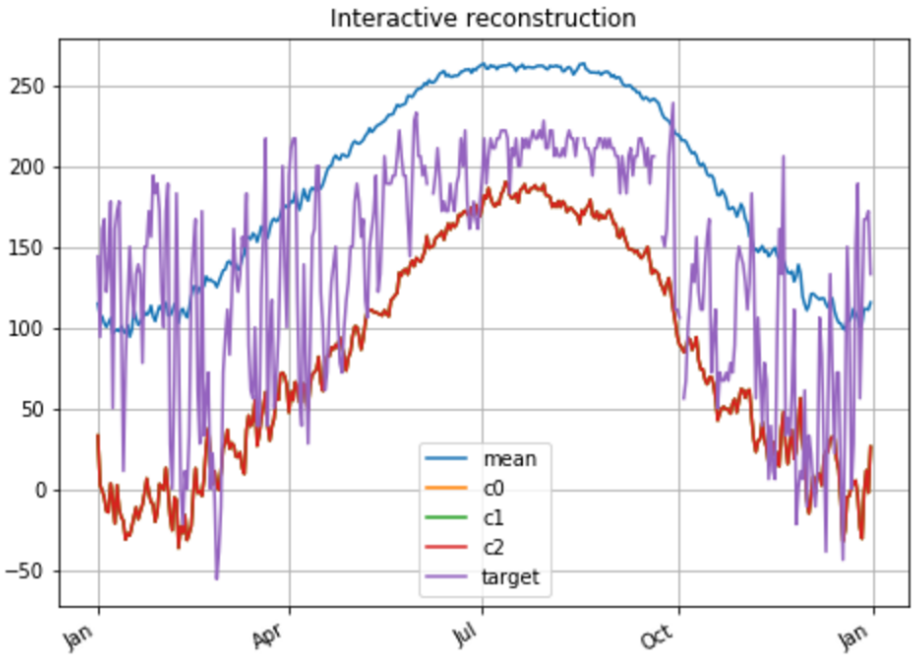


Figure 23: Target curve of TOBS and the contributions by mean and c0.

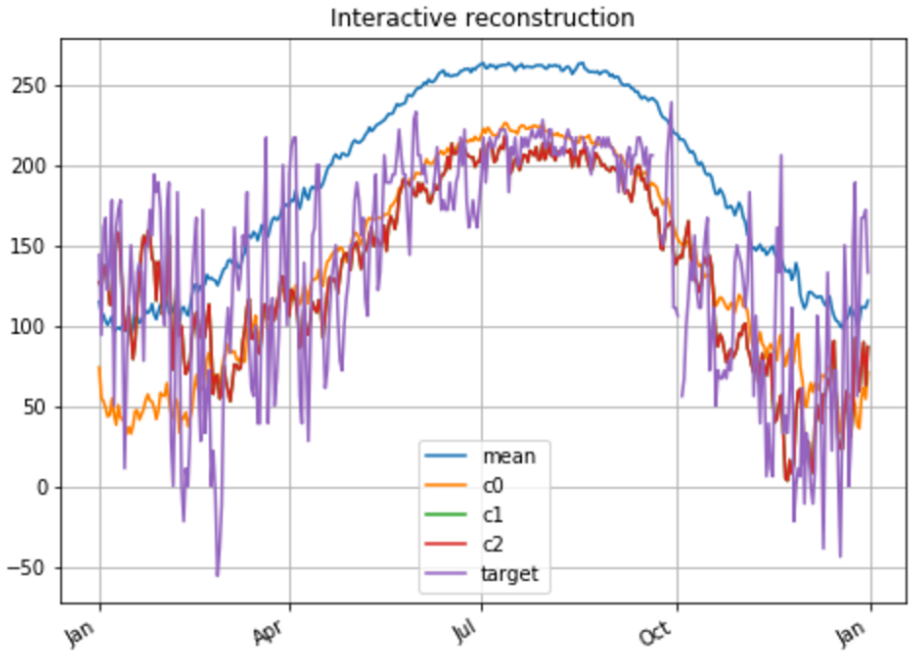


Figure 24: Target curve of TOBS and the contributions by mean and c0,c1.

From Fig. 24, we can see that giving the first two coefficients the optimal coefficient values, they model the target pretty well. So we can infer that the second eigen vector is crucial vector describing the target curve.

## CONCLUSION:

Thus, on performing statistical analysis on the 'BSBSSSSS' dataset, we can observe certain key characteristics and patterns which conform to the actual weather patterns observed in the south eastern region of the US. Firstly, the precipitation variable shows us that our null hypothesis that no two stations are related in terms of rain received is rejected. There are some stations near Tallahassee and Albany that have similar rain patterns, this shows that PRCP is a temporal variable rather than a spatial variable. Next, in terms of snow depth, the region in the dataset doesn't experience snow, in fact there was small amount of data collected in the year 1958 and 1977 which showed traces of snow and once again the SNWD can be concluded as a temporal variable. Next, on observing the temperature of the region, the stays the close the same almost throughout the year around the 15C-20C mark and dips below 15C during the months of November and December. The TOBS variable was explained better than the PRCP variable through the eigen vectors. We can conclude that the region given in the dataset is well and truly in the sub-tropical region.