Middle California & Nevada Weather Analysis

0. Introduction

This is a report for analyzing the weather data for the district of middle California and middle Nevada.

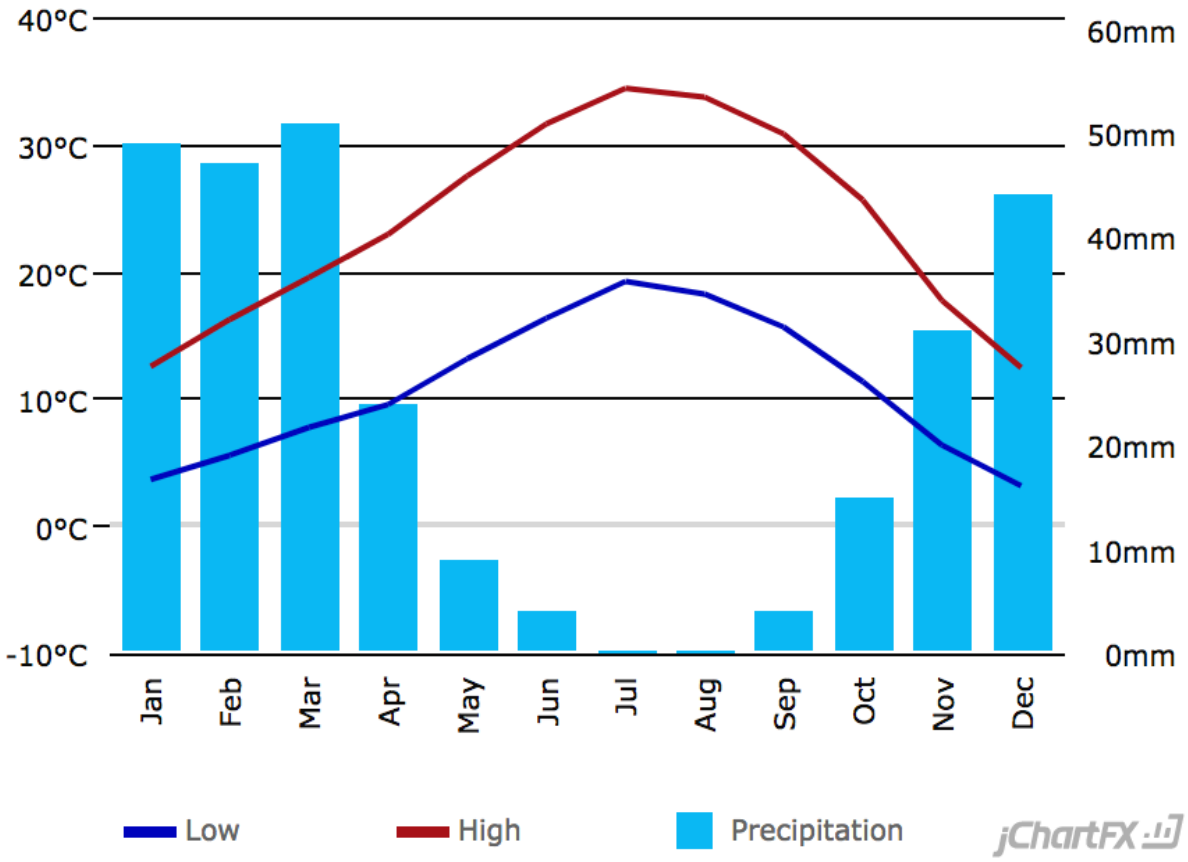
In the experiment, I totally used 12423 entries of data from NOAA.

In our experiment, we mainly focused on six measurements:

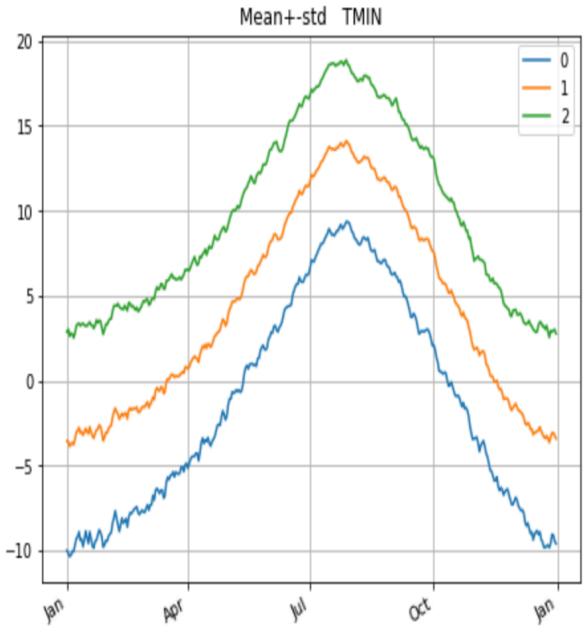
- **TMIN, TMAX:** the daily minimum and maximum temperature.
- **TOBS:** The average temperature for each day.
- **PRCP:** Daily Precipitation (in mm)
- **SNOW:** Daily snowfall (in mm)
- **SNWD:** The depth of accumulated snow.

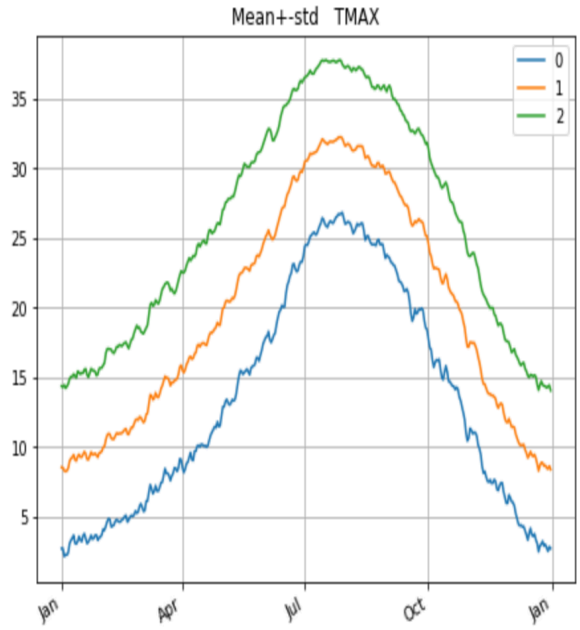
1. Sanity-check: comparison with outside sources

We first did the sanity-check, where we compared some general statistics with graphs that we obtained from the site called US Climate Data The graph below shows the daily minimum and maximum temperatures for each month, as well as the total precipitation for each month.



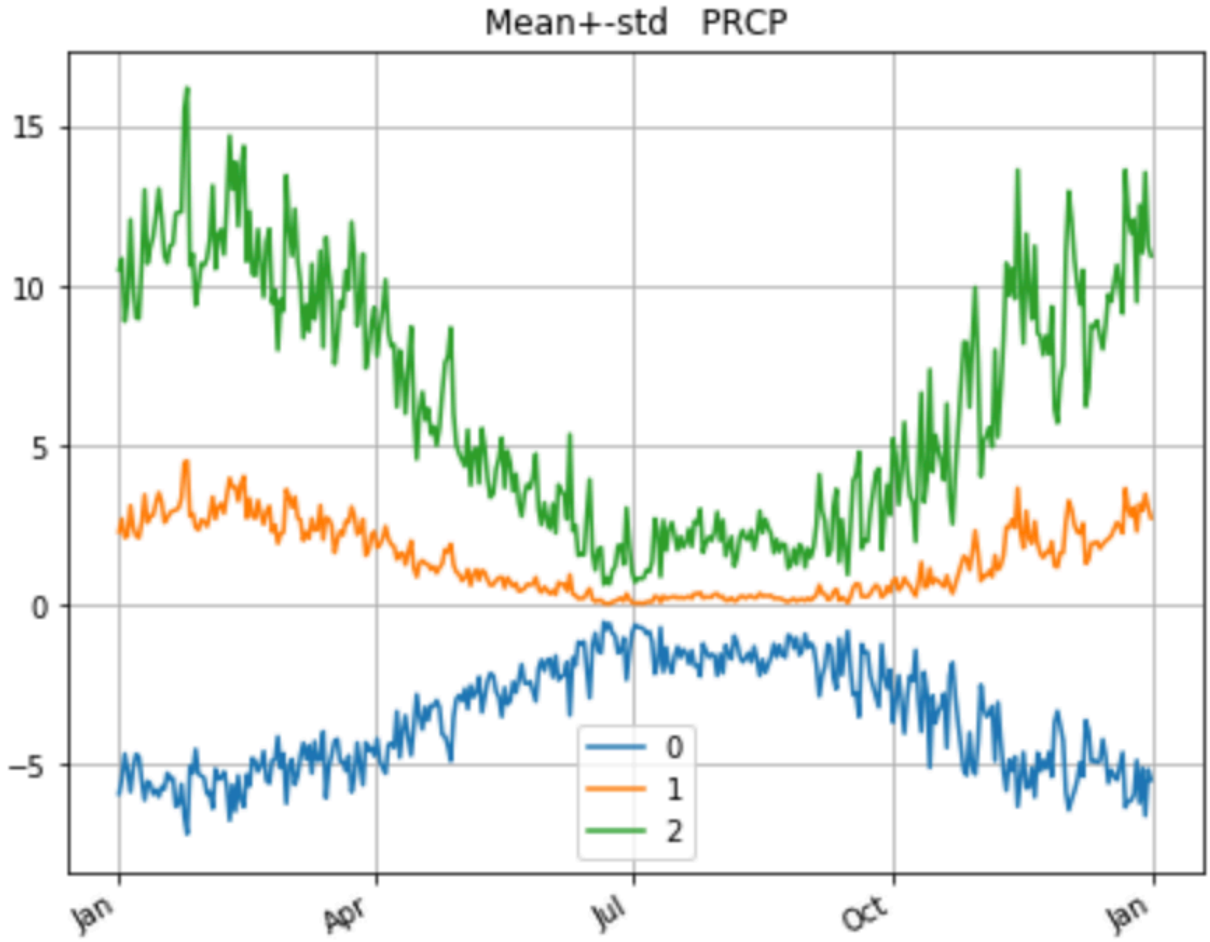
The graph below shows the min and max daily temperature agrees with the ones we got from our data, once we translate Fahrenheit to Centigrade.





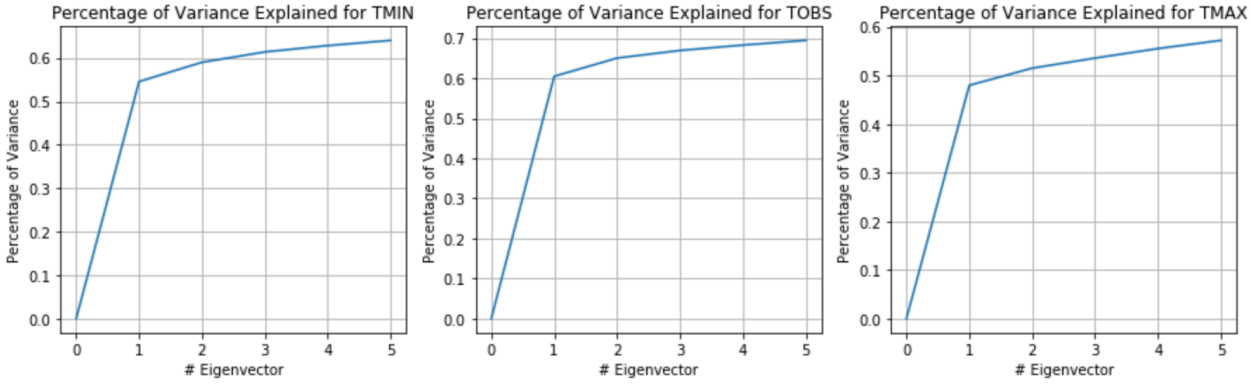
Additionally, to compare the precipitation we need to translate millimeter/day to millimeter/year.

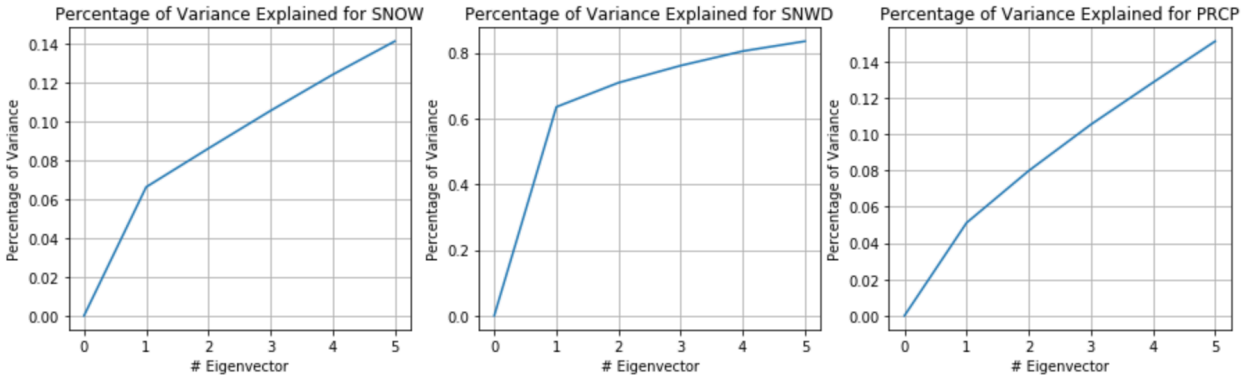
According to our analysis, the average rainfall is approximately 0.80 mm/day which translates to about 290 mm per year. According to the US-Climate-Data, the average rainfall is closer to 322 mm per year. However, there is clear agreement that average precipitation is close to a constant throughout the year.



2. PCA analysis

For each of the six measurements (TMIN, TMAX, TOBS, PRCP, SNOW, SNWD), we compute the portion of the variance explained as a function of the number of eigen-vectors used.





measurement	top 5 eigen-vectors explained variance
SNWD	85%
TMIN	63%
TBOS	70%
TMAX	58%
PRCP	15%
SNOW	14%

As the above graphs and table show, we got some interesting findings.

We first found that the measurement SNWD was best explained by the top 5 eigenvectors. The first 5 eigenvectors almost explained all the useful information about daily snowfall. This is especially ture for the first eigen-vector which, by itself, explains approximately 65% of the variance.

Beyond the measurement SNWD, the measurements TBOS, TMIN and TMAX have the similar variance - their percentage of variance explained are 70%, 63% and 58%, and the corresponding first eigen-vector variance explained are 60%, 55% and 48%. For these three temperature measurements, the top 5 eigen-vectors explain over half variance.

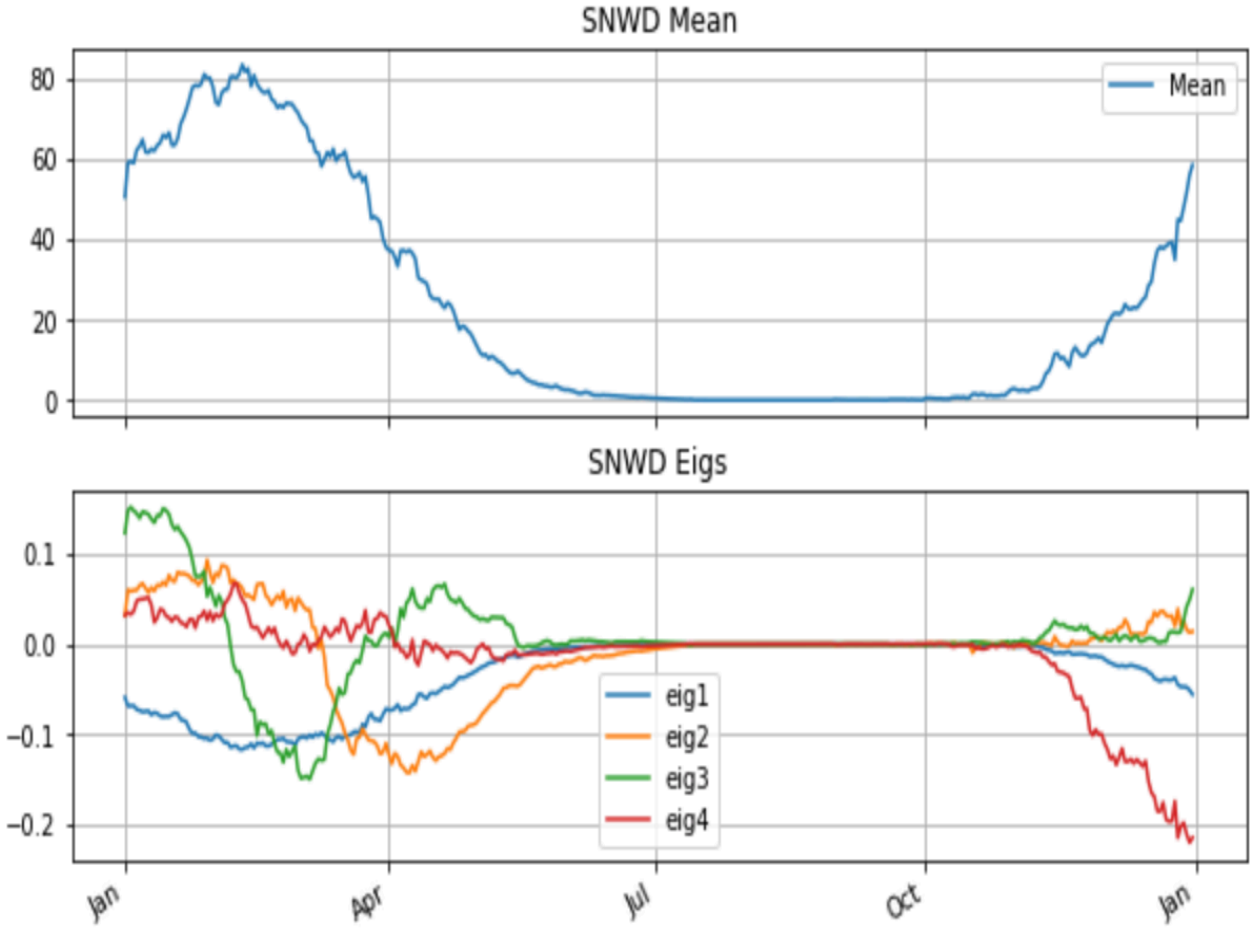
Also, we found that the measurements with the worst performance are SNOW and PRCP, which only explained about 14% and 15% variance. That is to say the top 5 eigen-vectors do not contain much useful information. In other words, these two measurements have much more noises than that of other measurements.

3. Analysis of snow depth

We choose to analyze the eigen-decomposition for snow-depth because the first 4 eigen-vectors explain 80% of the variance which indicates that the eigen-vectors for them make sense.

3.1 Eigenvector Interpretation

As the SNWD mean and eigen-vectors graph show below, we can observe that the snow season is from December to May, where the mid of February marks the peak of the snow-depth.



Then we interpret the eigen-functions.

According to the analysis above, the first eigen vector explains over 65% variance and the corresponding first eigen-function (eig1) has a shape very similar to the mean function just in the opposite direction. The interpretation of this shape is that eig1 represents the overall amount of snow below the mean.

eig2, **eig3** and **eig4** are similar in the following way. They all oscillate 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).

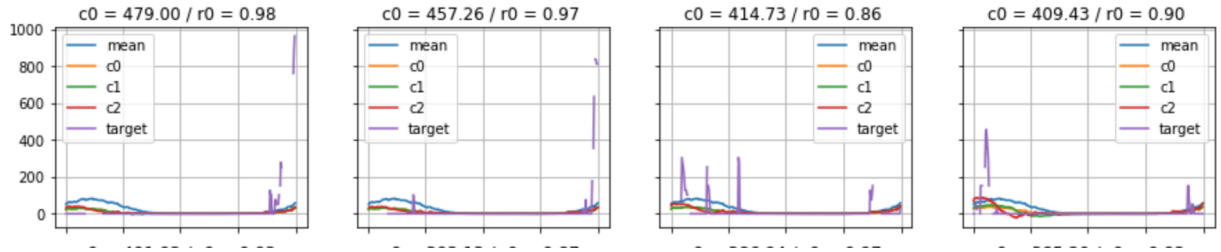
They can be interpreted as follows:

- **eig2:** more snow in Jan - mid Mar and less snow in mid Mar - Jun.
- **eig3:** more snow in Jan - Feb and Apr - May, less snow in Feb - Apr.
- **eig4:** less snow in Nov- Jan.

3.2 Reconstructions and Distribution of first 3 coefficients

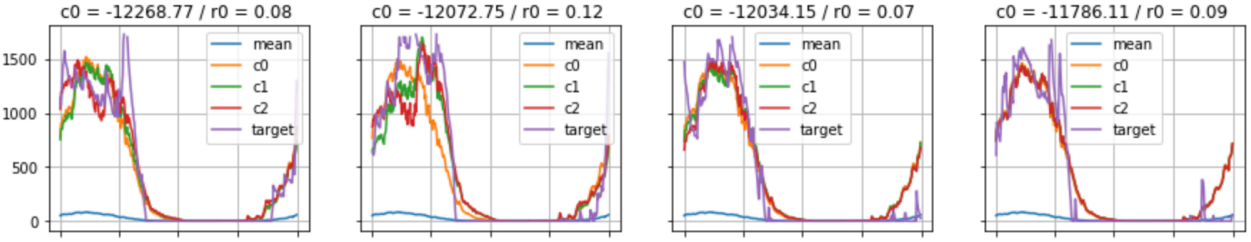
3.2.1 Coeff1

Coeff1: most positive



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

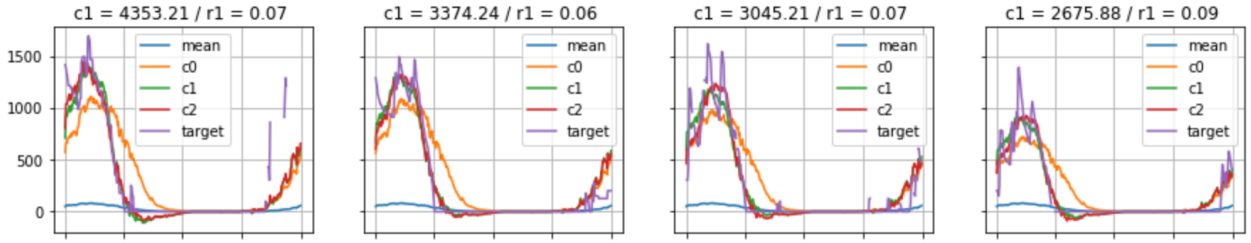
Coeff1: most negative



Low values correspond to more than average snow.

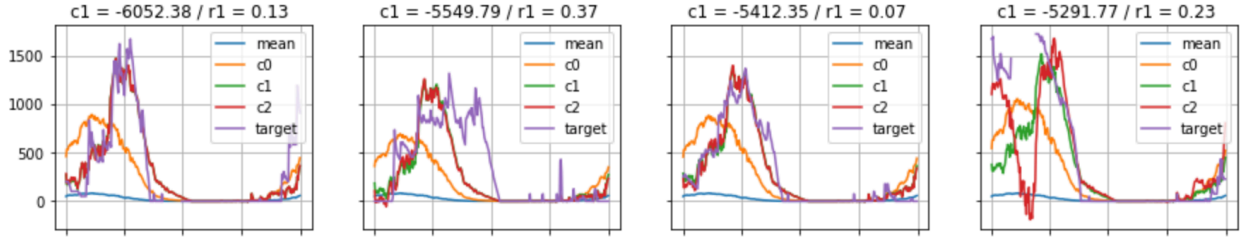
3.2.2 Coeff2

Coeff2: most positive



Large positive values of coeff2 correspond to a snow season with several peaks: one at the start of Jan, one at the middle of Feb and one at the start of Mar.

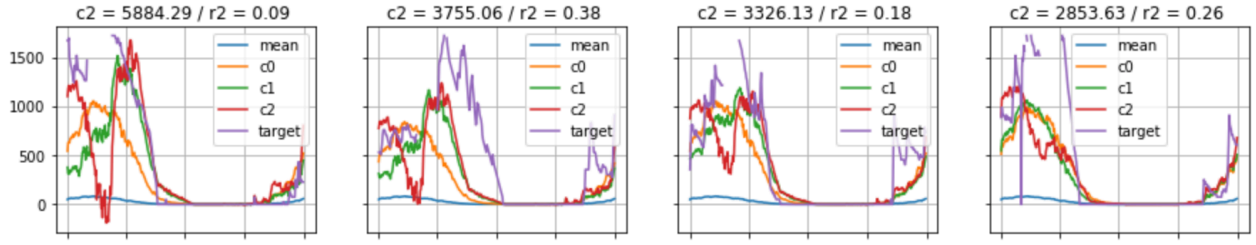
Coeff2: most negative



Negative values for coeff2 correspond to a late snow season (most of the snow is after mid of Mar).

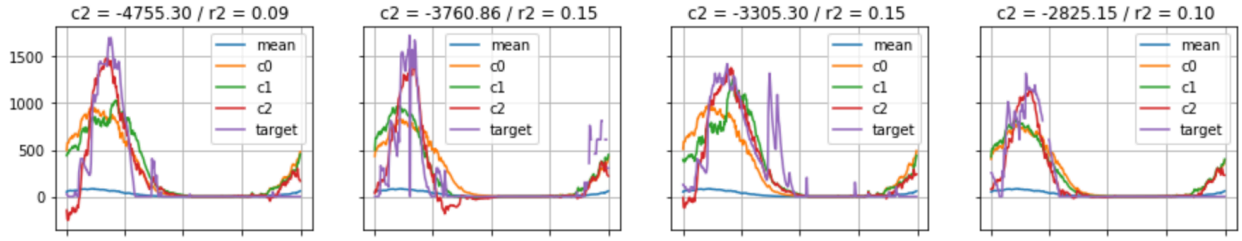
3.2.3 Coeff3

Coeff3: most positive

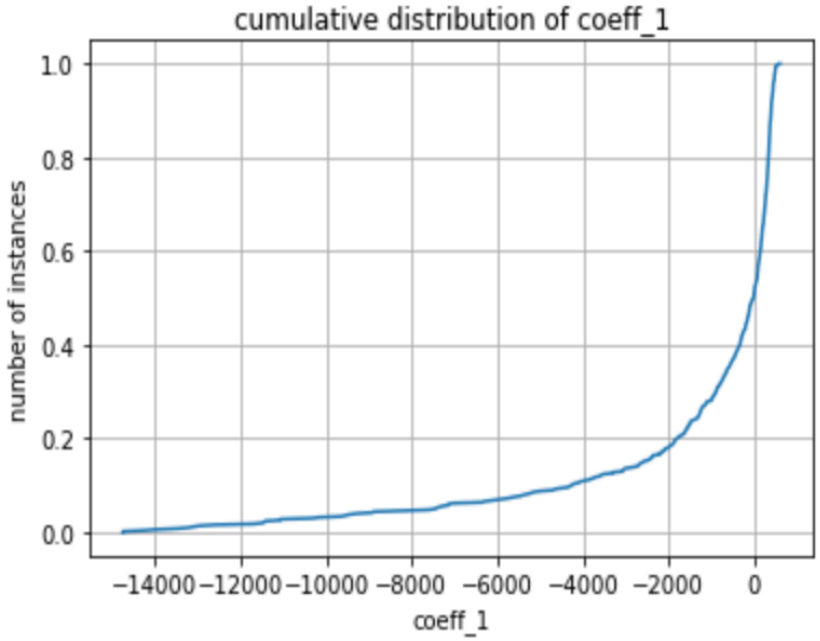


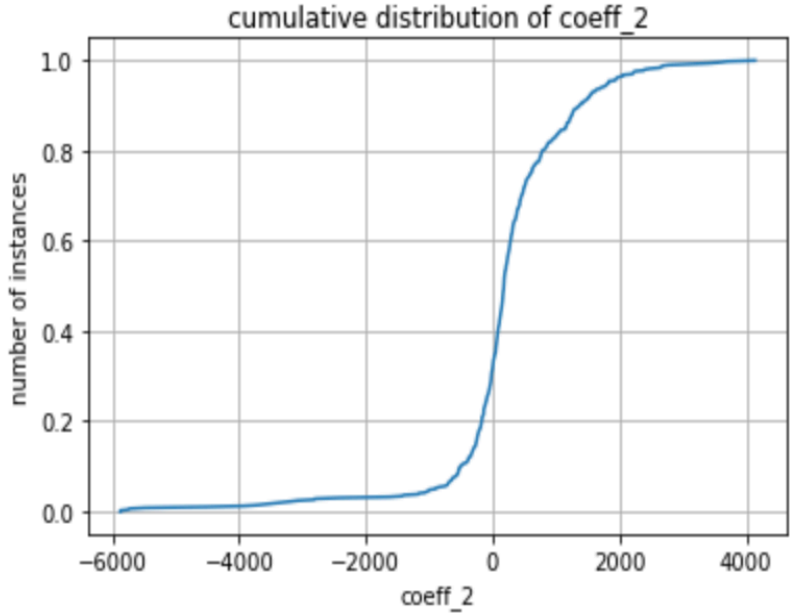
Large positive values of coeff3 correspond to several small peaks which indicate a long duration of heavy snow.

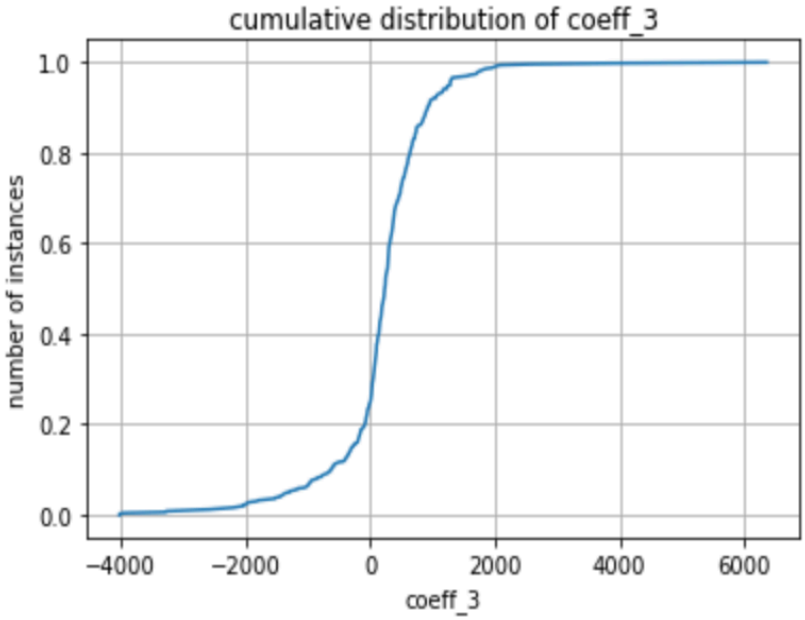
Coeff3: most negative



Large negative values of coeff3 correspond to a slightly late snow season (most of the snow is after Mar).





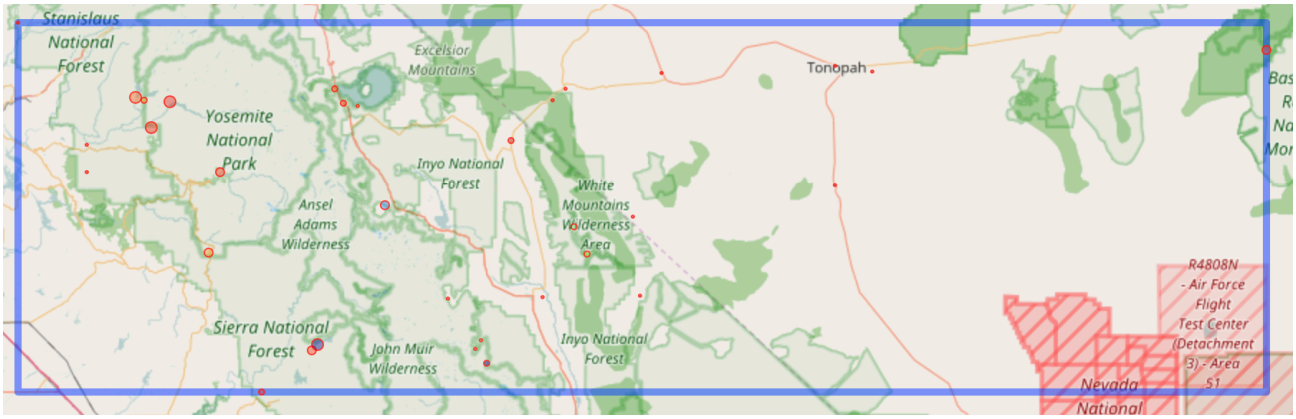


To sum up, according to the three graphs above, we can conclude that distributions of first 3 coefficients are normal distributed.

3.3 Estimating the effect of the year vs the effect of the station

Then, we mainly focus on the effects of time factor and place factor towards the snowfall.

3.3.1 Visualized Distribution of stations



The above graph shows the station places distributed in the map, we could find that the topography of each station is relatively different.

Since the difference exists in each station's topography, this may contribute some impact towards the snowfall and that's the reason why we want to analyze this factor.

3.3.2 Statistical Analysis

Since the first three eigen-vectors of SNWD contains about 80% information, here we choose these 3 coeffs to represent SNWD.

To estimate the relative importance of location-to-location variation relative to year-by-year variation, we build a matrix with year as row label, station as column label and different coeffs as item in the matrix.

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 = 11289133.92

MS removing mean-by-station= 3536527.61, fraction explained=68.7%

MS removing mean-by-year = 6277212.17, fraction explained=44.3%

coeff_2

total MS = 1511743.17

MS removing mean-by-station= 1188617.63, fraction explained=21.37%

MS removing mean-by-year = 738384.66, fraction explained=51.16%

coeff_3

total MS = 1106973.62

MS removing mean-by-station= 1038254.48, fraction explained= 6.2%

MS removing mean-by-year = 482709.30, fraction explained=56.4%

We see that, for coeff1, the variation by station explains more than the variation by year. However, this effect is opposite for coeff2 and 3. For coeff2 and 3 variation by station explains less than the variation by year, which indicates that these coeffs have more effects on timing factors of snowfall.

3.4 Analyzing Residuals to find out relations between statations

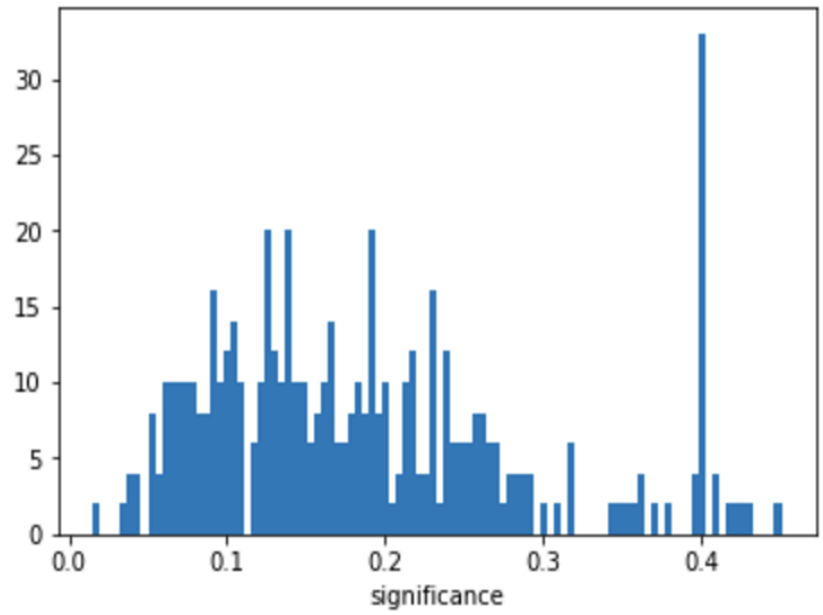
Finally, we begin to analyze the relationships between stations using residual analysis.

3.4.1 Testing Hypothesis

Null Hypothesis: the snowfall in two stations is independent.

By calculating probability of every two stations, we build a probability matrix as the following graph shows.

In this graph, the dark color means that two stations are strongly correlated, in contrast, the light color means that two stations are relatively uncorrelated.



The figure above displays a histogram of the resulting p-values, and from the histogram we can conclude that the majority of stations are correlated, because most of the p-values are larger that a significance of 0.05.

3.4.2 Correlated Stations Grouping Analysis

The following map shows 4 coeffs' scale for every station, and we are going to use these 4 coeffs to group correlated stations.

We apply PCA to obtain the first 4 eigen-vectors, and using these vectors to build the probability matrix like the one shown above.

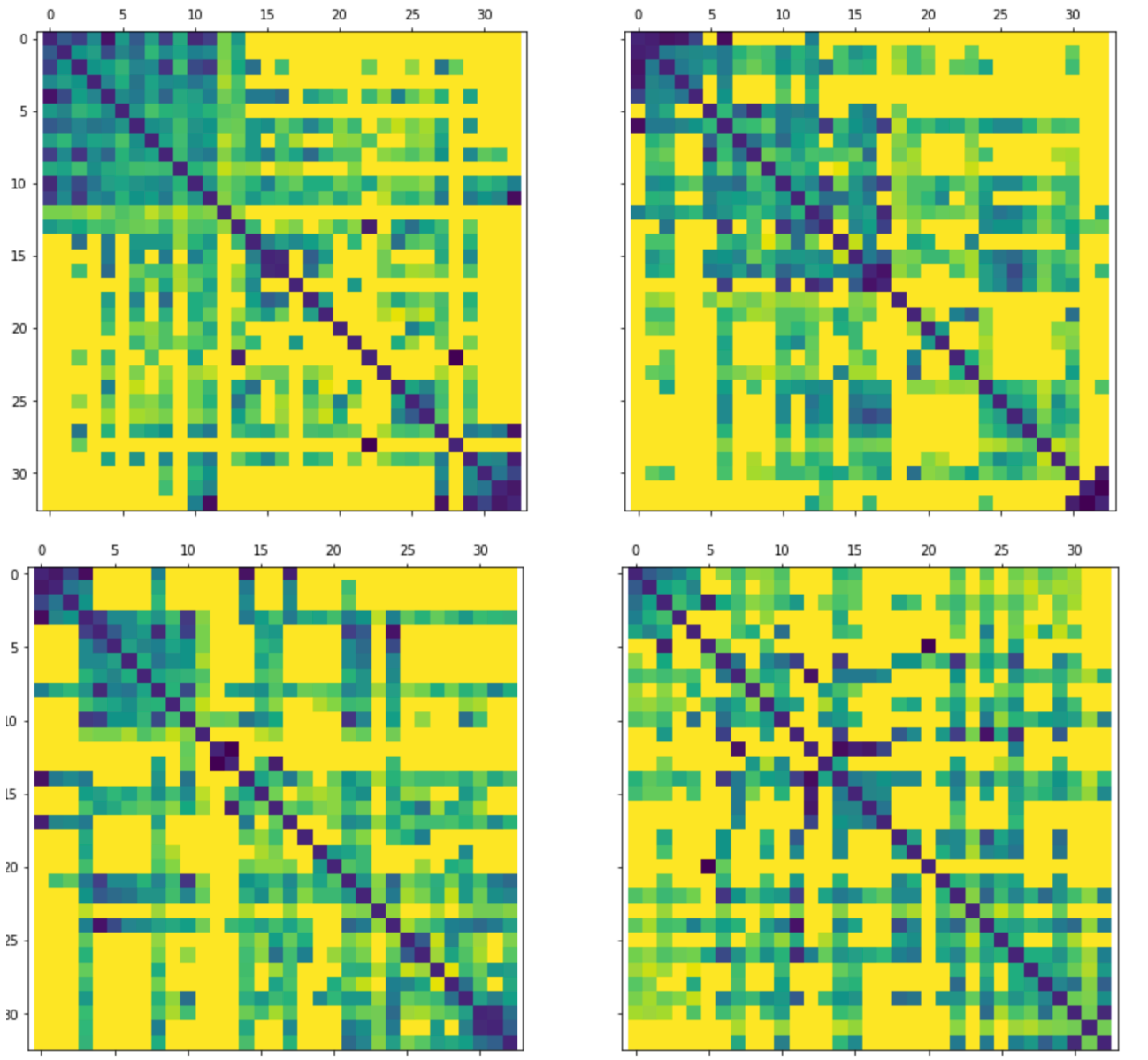


Figure above shows 4 probability matrix applying grouping.

After we recorded the rows and columns using one of the eigenvectors, we can find some station groups. For example, the bottom right corner of the first matrix. Stations at position 0-12, 15-19 and 28-32 are clearly strongly correlated with each other. And in the second matrix, stations at position 0-4, 5-17 are strongly correlated. For other two matrices there also exists such grouping patterns.