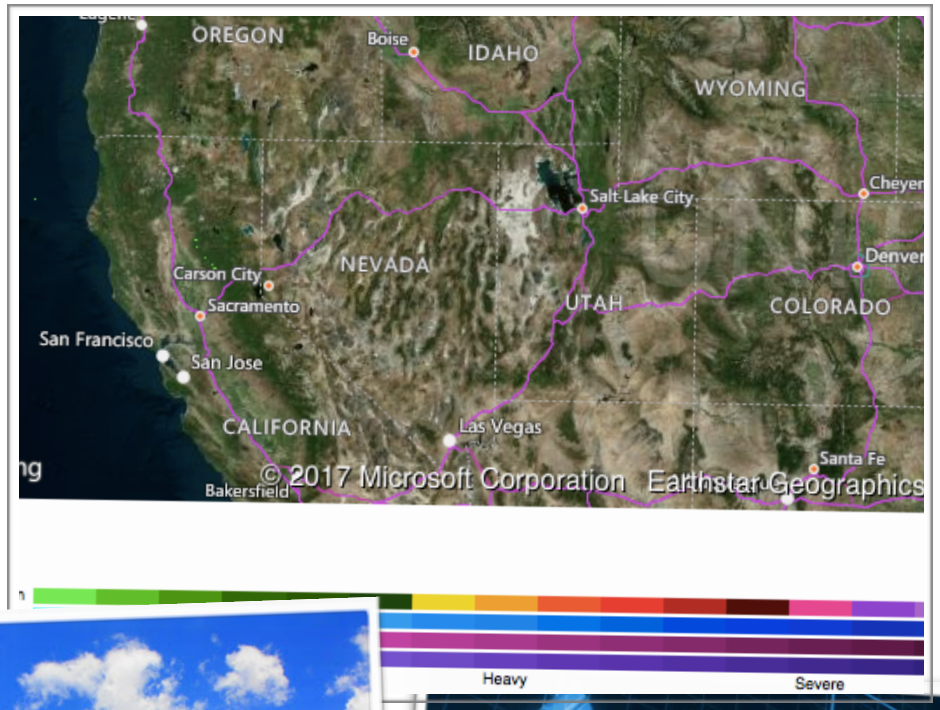
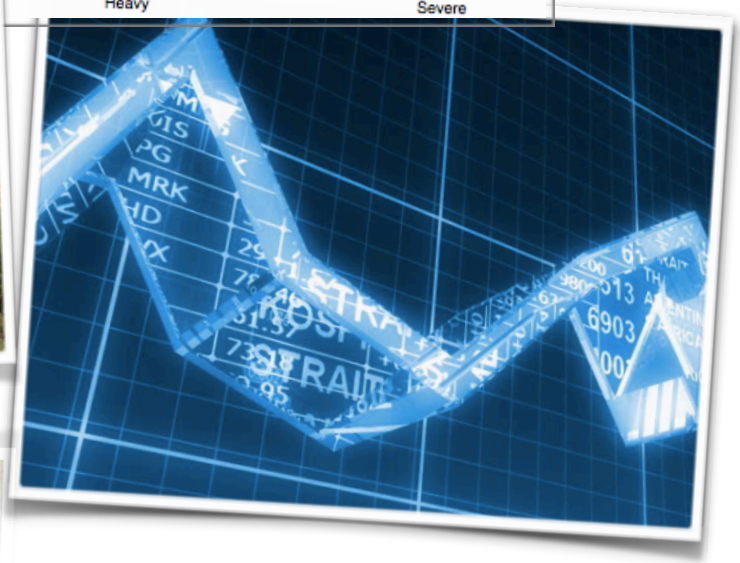
Student ID: A53227458

May 14, 2017

SSSBBSSB Region Weather Analysis





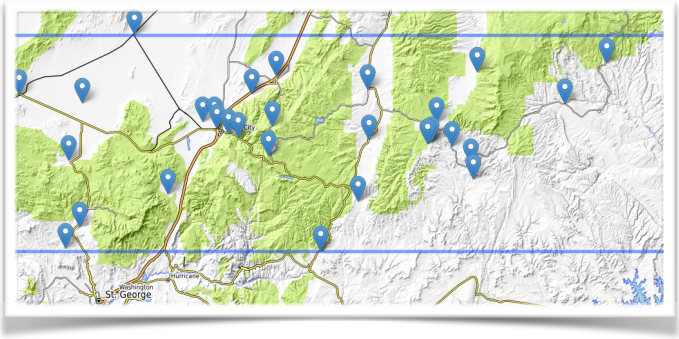
Summary Report Based on IPython Notebook Analysis

SSSBBSSB Region Weather Analysis

A Report Based from IPython Notebook Weather Analysis on Historical Data

Executive Summary

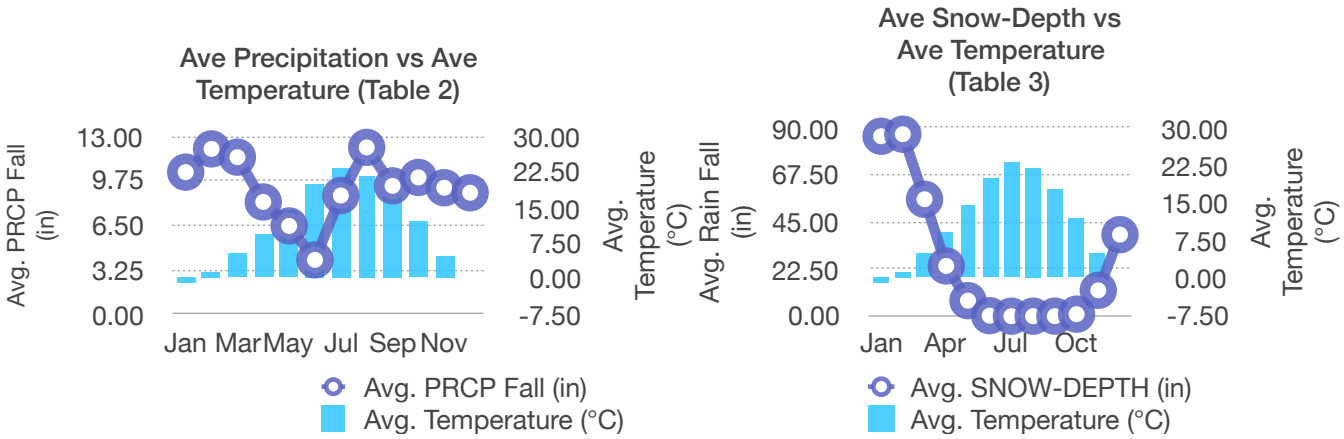
This is a summary report of the general weather pattern of a pre-defined geographic area (SSSBBSSB) that spans majority of south Utah region and over-extending west to the Nevada area. Like most northern areas, this region features snow-driven seasonal cycle. The area is made up of different terrain from forests (Dixie National Forest, Manti-La Sal National Forest), canyon, mountains, lakes, and deserts. In this region, the forest sections are typically famous for recreational activities.



On average, the over-all winter temperature ranges from 0 to 5 degrees Celcius. In spring, it climbs up to 20 degrees Celcius. The summer season is easy to spot because this is when everything heats up and average temperature rises up to

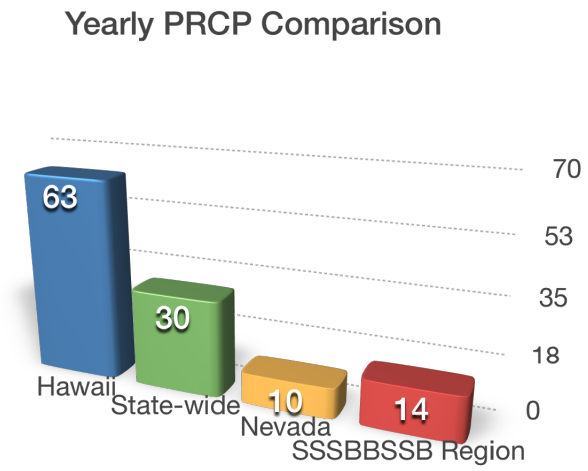
Table 1. **Monthly Average Statistics For Southern Utah Region**

	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
TMIN	-8.22	-6.27	-3.38	-0.10	4.22	8.55	12.69	12.23	8.05	2.53	-3.04	-7.40
TOBS	-0.88	1.39	5.17	9.33	14.53	19.96	23.03	21.74	17.98	11.99	4.97	-0.15
TMAX	5.17	7.33	11.46	15.91	21.29	26.99	30.53	29.28	25.51	19.39	11.88	6.14
SNOW	6.48	5.91	4.54	2.20	0.49	0.10	0.00	0.00	0.04	0.57	2.57	4.71
SNWD	85.99	86.97	55.89	24.00	7.42	0.25	0.00	0.00	0.02	0.97	12.36	38.98
PRCP	10.37	12.07	11.47	8.20	6.45	3.97	8.67	12.18	9.37	9.99	9.24	8.87



23 degrees Celsius and could go as high as 30 degrees Celsius. Not surprisingly, fall seasons transitions from both extreme temperatures and stay in the middle at around 11 degrees Celsius. (See Table 1-3)

State-wide Comparison



This region average precipitation is relatively lower compared to state-wide average. According to Current Report website, (https://www.currentresults.com/Weather/US/average-annual-state-precipitation.php) state-wide averages of annual rainfall including snowfall range is 63.7 - 9.5 inches and excluding Hawaii and Alaska, the average precipitation is 30.21 inches.

See below for other informative statistical figures:

Key Statistics

- Warmest month is July and temperature drops down to its
- Coolest month in January
- Pretty much rains all year round but weakens on June
- Snow are relatively absent during summer months

Historical Footnotes

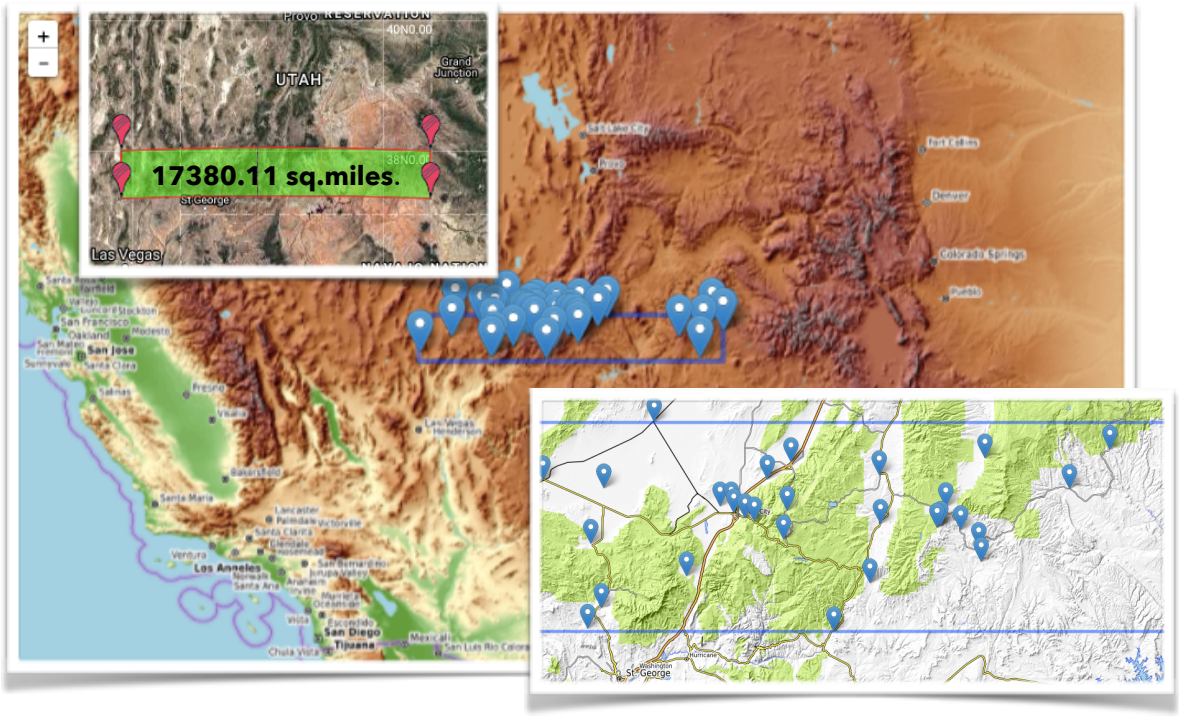
- Highest recorded temperature is 109.04 deg F in 1937.0
- Lowest recorded temperature is -13.0 deg F in 1942.0
- Strongest recorded snow activity is 1970 where depth reached 342.86 in
- Strongest recorded rainfall is in 2006.0 with staggering 166.82 inches

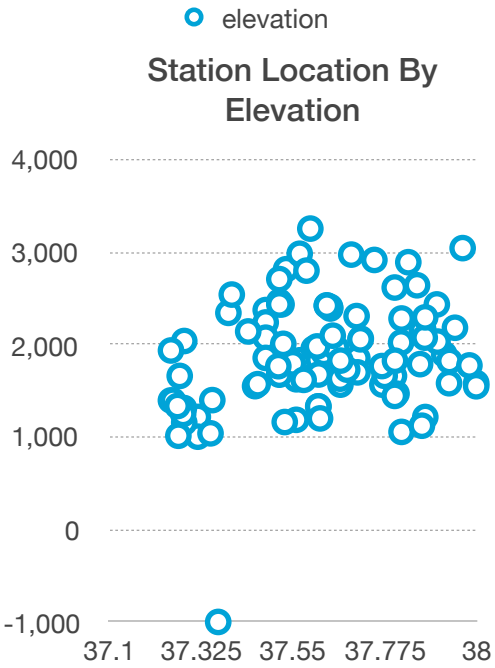
Reference: See modified version of 2.Small_PCA_computation_CLASS notebook

Basic Set-up

Unfortunately, the general statistical figure above could not be completely checked against other specific external data available (for i.e. Dixie National Park weather data) because the systematic partitioning of weather data could have introduced different biases depending on dispersions of stations (clustering in one area vs isolated stations) , elevations, and variation of terrains. The more ideal scenario is to perform a secondary partition which could be achieved using K-means neighbor algorithm or any other unsupervised learning algorithm. However, if the result is compared nation-wide then the general-mean of the derived statistics could be meaningful.

Monitoring Station Locations





SSSBBSSB weather data were collected from monitoring stations located in a rectangular coordinates: [(37.2506, -115.2236) (38.0, -109.0828)]. The combined total square area of the region being monitored in this report is around: **17380.11 square miles**. See https://www.daftlogic.com/projects-google-maps-area-calculator-tool.htm The stations are dispersed throughout the area but mostly clustered on populated areas. And except for one station, all of the monitoring stations are located within 1,000 up to 3,000 feet of elevation. The analyzed dataset where downloaded NOAA ftp site: ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/.

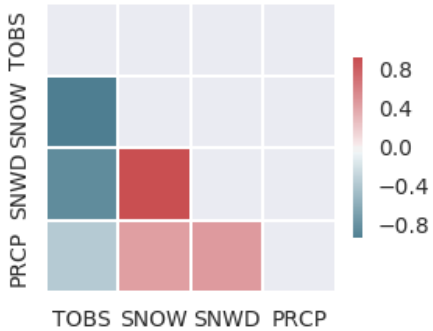
- data-source.txt: the source of the data
- ghcnd-readme.txt: A description of the content and format of the data
- ghcnd-stations.txt: A table describing the Meteorological stations.

Data Clean-Up

SSSBBSSB is a small part of a larger data-set that were pre-processed using the six measurements: ['TMAX', 'SNOW', 'SNWD', 'TMIN', 'PRCP', 'TOBS']. Only measurement-years that have at most 50 NaN entries were included in the master data-set.

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

The master data-set contains only measurements in the continential USA and partitioned into 256 geographical rectangles, indexed from BBBBBBBB to SSSSSSS where each section contains about 12,000 stations, year pairs.



Correlation Matrix - Figure 1

Weather Patterns

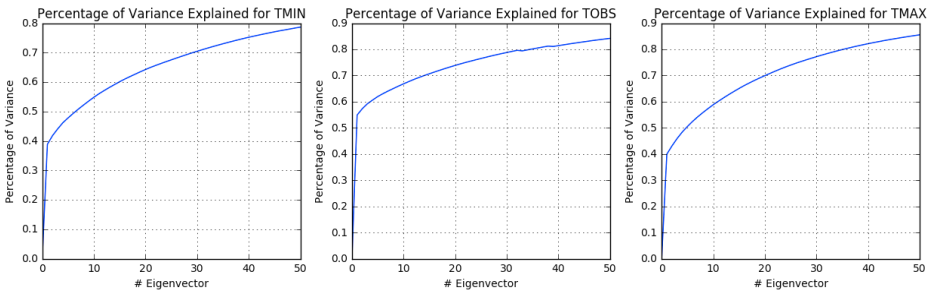
As expected, the general weather pattern revolves around the rise and fall of temperatures

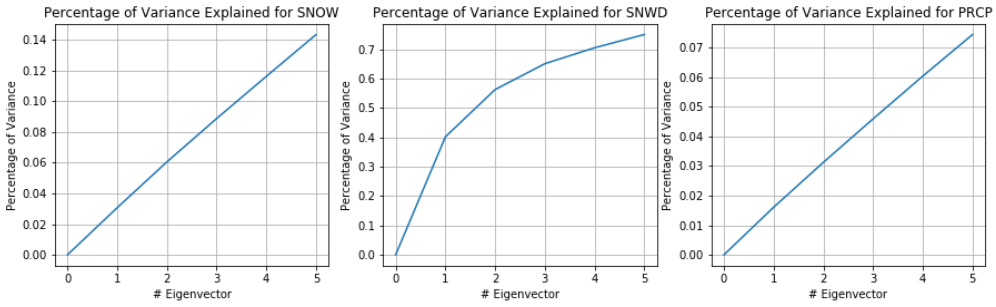
throughout the year. In Figure 1 (Correlation Matrix), the snow and snow-depth has negative correlation with temperature and precipitation has (slight) positive correlation with temperature since snow is also a product of precipitation which could happen in cold weather as well.

Technical Analysis

PCA Analysis

In this data-set, the eigen-vectors calculated from temperature measurements (TMAX,TMIN,TOBS) could not fully explain the percentage of variance in the data. However, from the eigenvectors of snow-depth, it takes only 5 vectors to explain 70% of the data variance. See middle chart below.





Since, there could be other statistical significance hidden in the layers of snow-depth data, a further analysis is performed using plotting of the first 3 eigen-vectors.

Analysis of snow depth¶

The chart above reveals that the snow depth starts to accumulate from mid-November to January then it retreats back to March, April, May and virtually gone in summer. It also shows that in the middle of February marks the peak of the snow-depth.

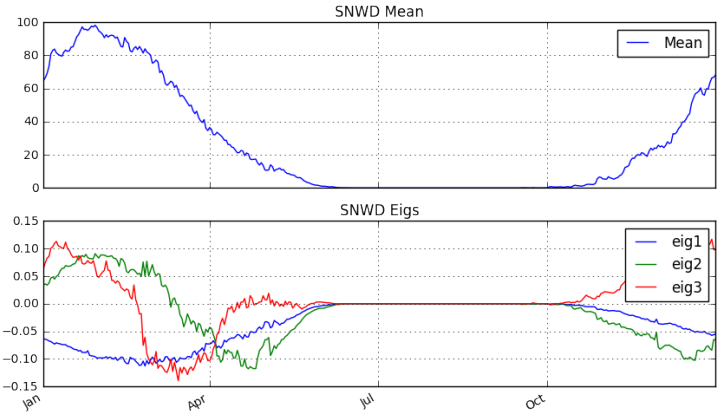


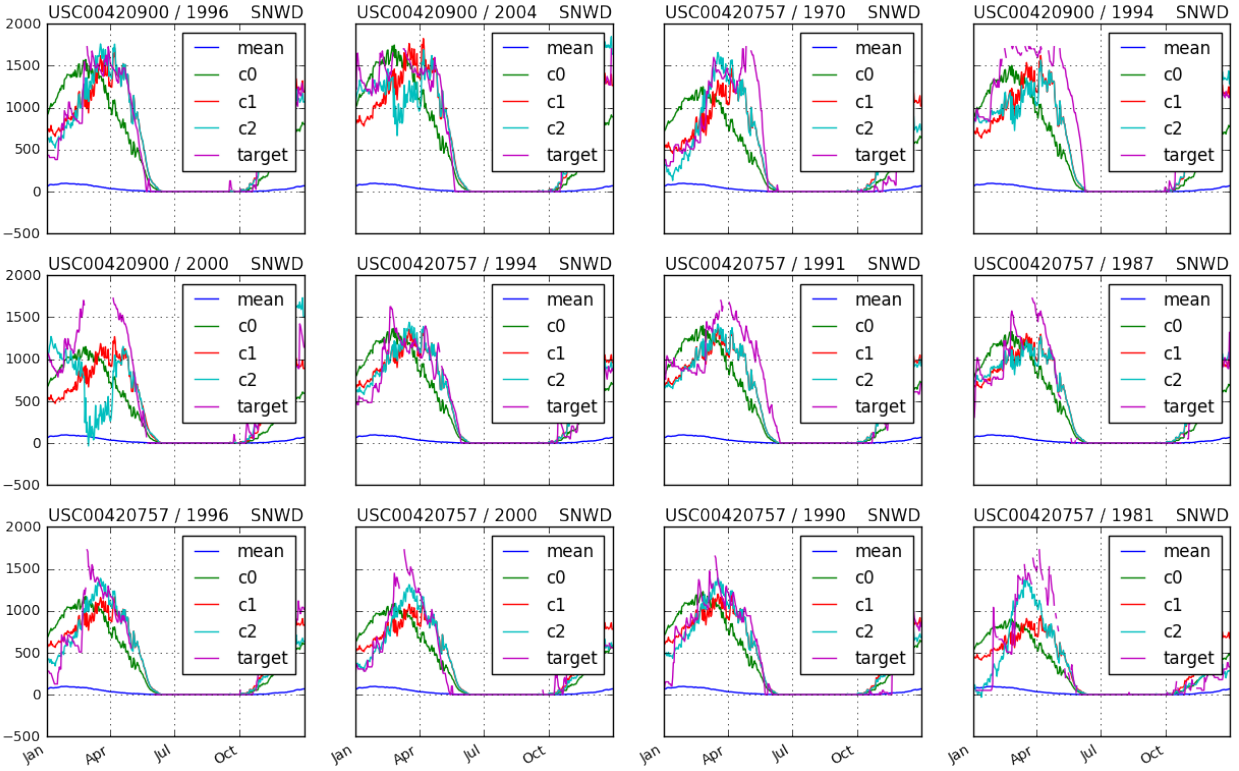
Figure 4. Snow-Depth Mean and Eigen-Vectors

Eigen functions are good tools in re-constructing or approximating any original function that could be derived in its space. Simple plot of the first eigen-function (eig1) compared against the mean function requires subtracting the latter to the former if the plot shows a high-degree of similarity. In the chart above, the obvious main difference is that the eigen-function is close to zero during October-December timeframe while the mean is not. The interpretation of this shape is that eig1 represents the overall amount of snow above/below the mean without changing the distribution over time.

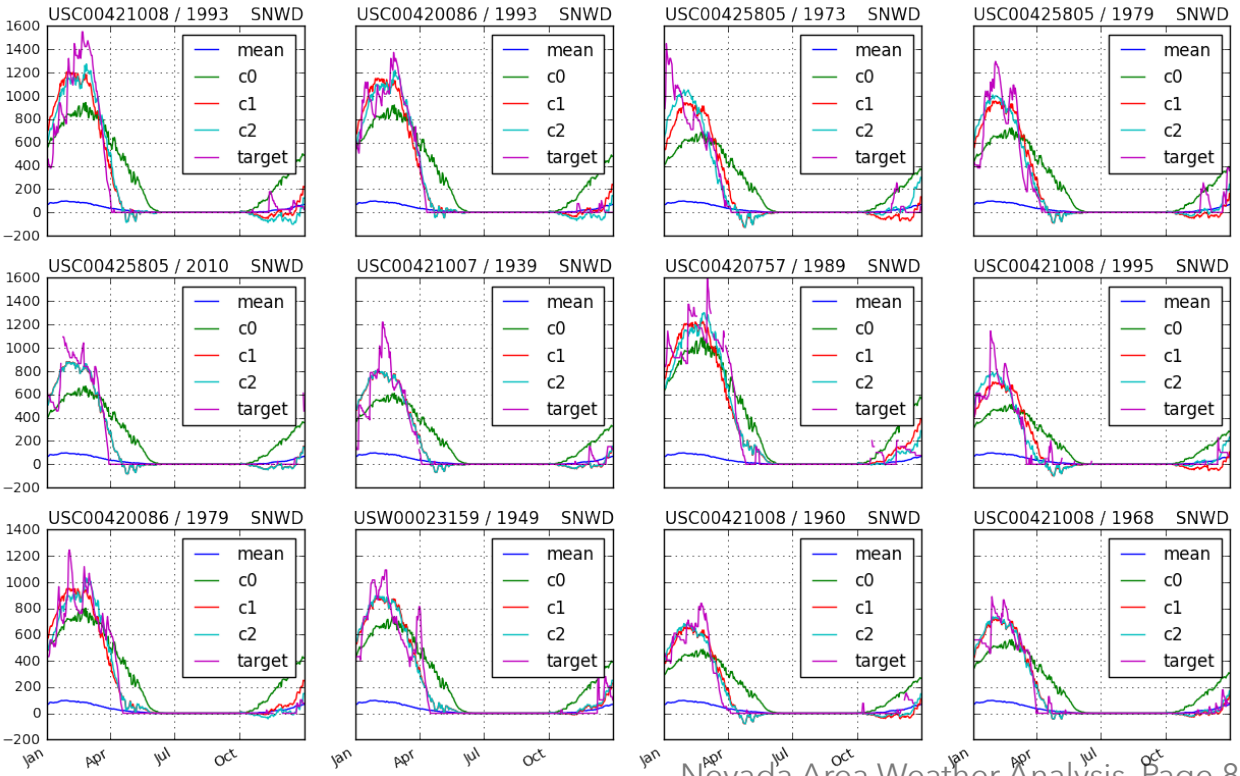
The remaining eig2,eig3 vectors behaves in similar ways. Both oscillate between positive and negative values corresponding to changing the distribution of the snow depth over the winter months without affecting the total distribution.

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 and more snow in march.



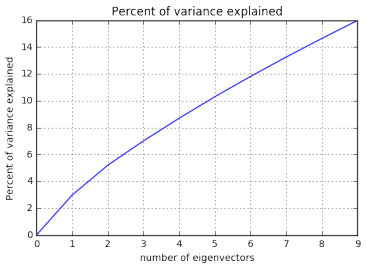
It could be extrapolated that large positive values of coeff2 correspond to a late snow season (most of the snowfall is after mid feb.) Negative values for coeff2

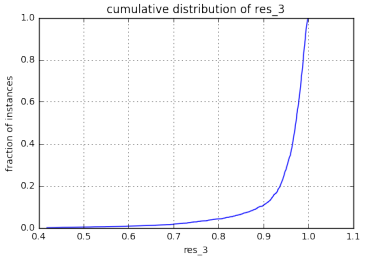


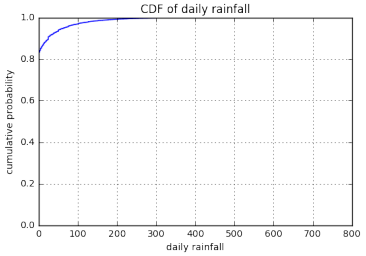
correspond to an early snow season (most of the snow is before mid-Feb)

Analyzing PRCP through Residuals

Plot the percent of residual variance on average



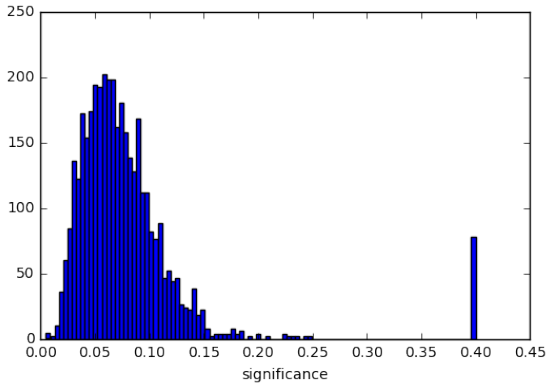




The chart shows precipitations/rainfall typically slows down at the end of July and fluctuates in a volatile fashion towards the end of the year. However, the plot of the third 3 eigen function which is the subtracted residual variance after the mean and the first two eigen-vectors shows a much stabler oscillation between positive and negative levels. It could very well represent multiple occasions of less precipitation frequency than the average suggests. (PRCP Eigs and Mean chart)

Looking at the chart explaining the percent of variance, it would take more than 9 eigen-vectors to explain majority of the variance.

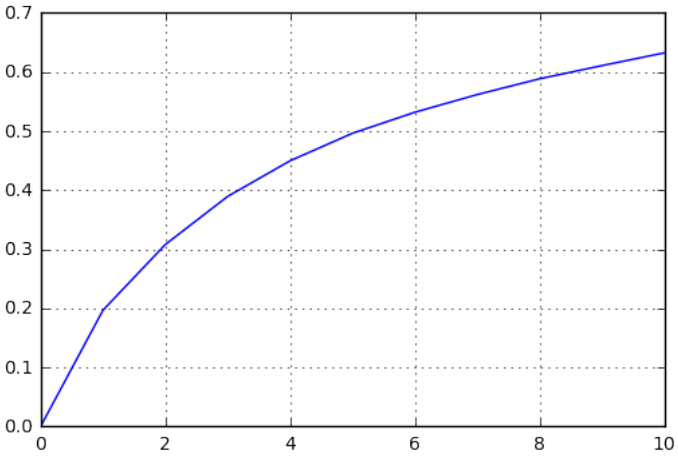
Because amounts of rain vary a lot between even close locations, it is unlikely to find correlations between the amount of rain on the same day in different stations. It is more reasonable to try to compare whether or not it rained on the same day in different stations. As we see from the graph above, in this particular region it rains in about a third of the time. (CDF of daily rainfall chart above)

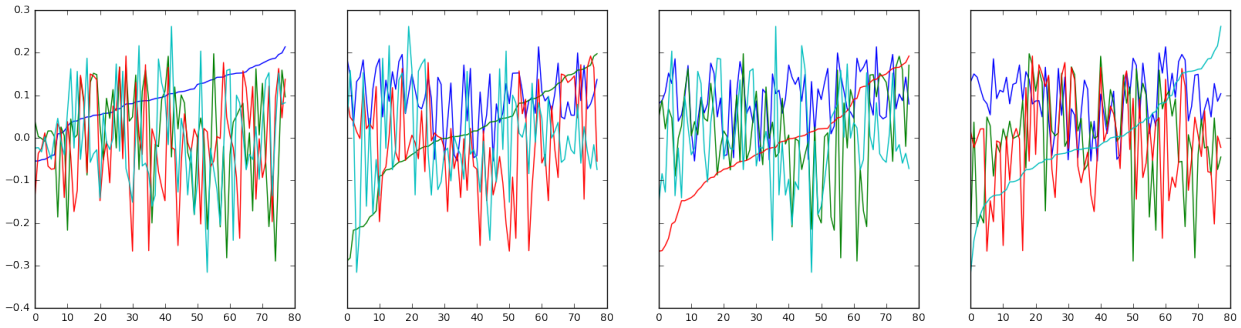


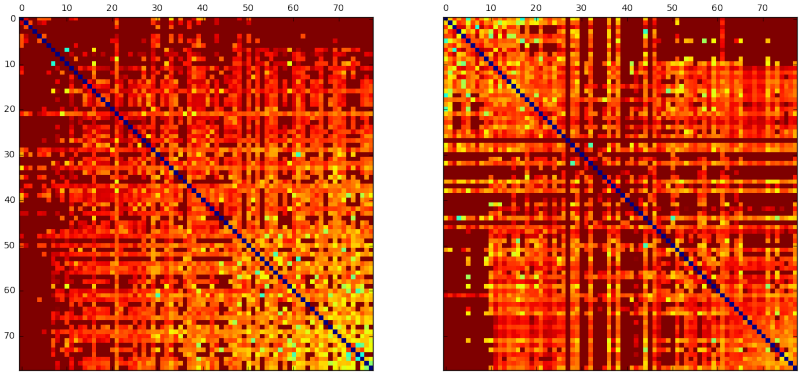
The graph above calculates the normalized log probability for each pair of stations.

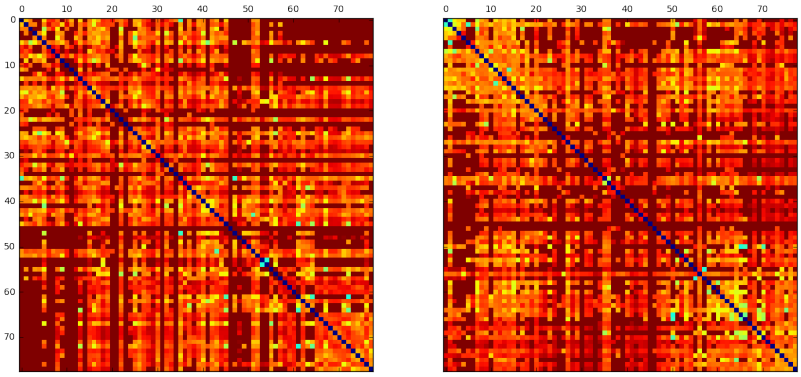
The matrix above shows, for each pair of stations, the normalized log probability that the overlap in rain days is random. The chart also shows that there are few stations that are highly correlated with each other. Further correlations of stations could be achieved by using SVD (the term PCA is reserved for decomposition of the covariance matrix). As we shall see that

the top 10 eigenvectors explain about 65% of the square magnitude of the matrix. (See chart on the right)









Reordering the rows and columns of the matrix using one of the eigenvectors, the grouping of the stations becomes highlighted. For example, consider the upper left corner of the second matrix (the upper left one). The stations at positions 0-30 are brightly colored and a clear indication of correlation.

Key Take-Aways / Conclusion

Although precipitation in the form of (rain or snow) is always present (random), it could be argued that there is a general stability in the seasonal climate changes in the SSSBBSSB region. Since there are pockets in the area where recreational activities could be performed, winter months bring a good amount of snow which are perfect for skiing and winter activities. On the other hand, there could be plenty of window of opportunities for warm months in between which are suitable for summer activities.

On the technical side, one lesson learned is how simple it is to generate more historical records by employing map/reduce parallelization processing to compute for the max or min value of a specific feature on a large number of records. It would be interesting to apply the same procedure to a much bigger data set to extract outliers and other extreme values.

This exercise shows that Principal Component Analysis (PCA) is not only an excellent tool for reducing dimension or reconstruction by approximation but could also be used as unsupervised learning tool by applying Singular Value Decomposition (SVD) to group together un-labeled dataset into clusters of common attributes. It would also be interesting to apply the same procedure to a much bigger data set using Hadoop technology.

All procedures/algorithms developed in this exercise is a good base template for further research.