IS 602 Final Project

EL NIÑO/LA NIÑA AND SNOWFALL/SNOWPACK

1 OVERVIEW

This project uses multiple data sets from the NOAA Global Historical Climatology Network (GCHN) to analyze historical snowfall and snowpack at weather stations in the United States and to look for correlations between these measures and historical El Niño/La Niña episodes between 1950 and 2010. The primary file used in this analysis is a tar file that is over 2.5 GB in size and contains over 93,000 weather station files. The Python programming language as well as the following Python packages were used for this project: Pandas, Tarfile, Os, NumPy, Itertools, Stats from Scipy, Curve_fit from Scipy.Optimize, Matplotlib.Pyplot, Matplotlib.cm, Pylab, and Shutil.

2 DATA

The source for all data used in this project is the National Oceanic and Atmospheric Administration (NOAA).

Station Data

The "ghcnd-stations.txt" file contains data for all weather stations and contains the following fields:
URL: ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/ghcnd-stations.txt

Variable	Columns	Type	Description
ID	1-11	Character	Station identification code
LATITUDE	13-20	Real	Station latitude
LONGITUDE	22-30	Real	Station longitude
ELEVATION	32-37	Real	Station elevation in meters
STATE	39-40	Character	U.S. postal code for the State
NAME	42-71	Character	Station name
GSN FLAG	73-75	Character	identifies if the Station is part of the GSN
HCN/CRN FLAG	77-79	Character	identifies if the Station is flag if part of the HCN
WMO ID	81-85	Character	World Meteorological Organization station number

Inventory Data

The inventory file is a list of the periods of record for each station and element type (e.g. snowfall, rain). This file is useful for determining which stations have data on snowfall and for how many years. This helped narrow down the relevant stations for this analysis.

Variable	Columns	Type	Description
ID	1-11	Character	Station identification code
LATITUDE	13-20	Real	Station latitude
LONGITUDE	22-30	Real	Station longitude
ELEMENT	32-35	Character	Element Type*
FIRSTYEAR	37-40	Integer	First year of unflagged data for the element.
LASTYEAR	42-45	Integer	Last year of unflagged data for the element.

***Element Type:** There are five core elements as well as a number of additional elements in the data set. However, for the purpose of this project the focus will be on two elements: SNOW = Snowfall (mm) and SNWD = Snow depth (mm). There is one record for each element type for which data was recorded. The element type is identified in the ELEMENT column. As an example there is one record that includes one month of SNOW data, and another record for the same month for SNWD.

Snowfall/Snowpack Data

The data source for snowfall and snowpack is the NOAA Global Historical Climatology Network (GHCN) Daily Data, specifically the -"ghcnd_all.tar.gz" file which contains daily climate summaries from land surface stations across the globe. ***This tar file is over 2.5 GB in size and contains over 93,000 weather station files.***

NOAA GHCN Daily Data URL: http://www.ncdc.noaa.gov/oa/climate/ghcn-daily/
NOAA GHCN Daily Readme File URL: http://www1.ncdc.noaa.gov/pub/data/ghcn/daily/readme.txt

This file is a TAR file of the GZIP-compressed "hcn" directory. The “hcn” directory includes all the ".dly" files. Each “.dly” file contains data for one station, with the name of the file corresponding to a station’s identification code. Each record contains one month of daily data. Each .dly file contains several fields; the fields that are relevant for this analysis are shown in the following table.

Variable	Columns	Type	Description
ID	1-11	Character	Station identificiation code
YEAR	12-15	Integer	
MONTH	16-17	Integer	
ELEMENT	18-21	Character	Element type
VALUE1	22-26	Integer	Value on the first day of the month
VALUE2	30-34	Integer	Value on the second day of the month
.	.	.	
.	.	.	
.	.	.	
VALUE31	262-266	Integer	value on the 31st day of the month*

*Value31 Note: If the month has less than 31 days, then the remaining variables are set to missing represented by -9999 (e.g., for April, VALUE31 = -9999, MFLAG31 = blank, QFLAG31 = blank, SFLAG31 = blank).

El Niño/La Niña Data

The final data set used in this project is the El Niño/La Niña record from the NOAA Climate Prediction Center which contains data from 1950-present.

CPD data URL: http://www.cpc.noaa.gov/products/analysis_monitoring/ensostuff/ensoyears.shtml

The NOAA description of El Niño and La Niña is:

> El Niño is characterized by unusually warm ocean temperatures in the Equatorial Pacific, as opposed to La Niña, which is characterized by unusually cold ocean temperatures in the Equatorial Pacific. El Niño is an oscillation of the ocean-atmosphere system in the tropical Pacific having important consequences for weather around the globe. Among these consequences are increased rainfall across the southern tier of the US and in Peru, which has caused destructive flooding, and drought in the West Pacific, sometimes associated with devastating brush fires in Australia.
>
> To provide necessary data, NOAA operates a network of buoys which measure temperature, currents and winds in the equatorial band. These buoys transmit data daily which are available to researchers and forecasters around the world in real time.
>
> (Source: http://www.pmel.noaa.gov/tao/elnino/el-nino-story.html)

The 1997-1998 El Nino remains the strongest on record with the second strongest having occurred in 1982-1983.

The Oceanic Niño Index (ONI) is a record of anomalies in sea surface temperatures: the values in the ONI represent the difference between a three month average for a particular period (e.g. January-February-March) and a 30 year average for the same three month period.

Surface temperatures are based on the Extended Reconstructed Sea Surface Temperature (ERRSST) version 3b. The ERSST is a global monthly sea surface temperature analysis derived from the International Comprehensive Ocean–Atmosphere Dataset with missing data filled in by statistical methods. Source: http://www.ncdc.noaa.gov/data-access/marineocean-data/extended-reconstructed-sea-surface-temperature-ersst-v3b.

Rather than using the most recent 30 year period (1981-2010) multiple centered 30 year periods are used as the base comparison for the ONI. As an example, ONI values for 1950-1955 are based on the 1936-1965 base period and likewise ONI values for 1956-1960 are based on the 1941-1970 base period. Source: http://www.cpc.noaa.gov/products/analysis_monitoring/ensostuff/ONI_change.shtml.

As illustration, the following table contains two years of the NOAA CPD Oceanic Niño Index (ONI) data set:

Year	DJF	JFM	FMA	MAM	AMJ	MJJ	JJA	JAS	ASO	SON	OND	NDJ
1950	-1.4	-1.3	-1.2	-1.2	-1.1	-0.9	-0.6	-0.5	-0.4	-0.5	-0.6	-0.7
1951	-0.8	-0.6	-0.4	-0.2	0.0	0.4	0.6	1.0	1.1	1.2	1.1	0.9

The red values indicate warm periods (El Niño) while the blue values indicate cold periods (La Niña). As described above the ONI is a record of anomalies, for historical purposes the threshold required to indicate a warm/cold period is +/- 0.5°C anomaly for a minimum of 5 consecutive overlapping seasons.

3 Methodology and Coding

Analysis for one station: Each ".dly" file has an identical structure. Therefore, a core part of the coding work was to first design a program that would work for one station which could then be applied to other stations.

1. **Monthly Average:** Compute a total snowfall and a snowpack average for the "snow months" defined in this project as October – April for all years (1950 to the present).

2. **30 Year Monthly Average:** Compute an average monthly snowfall (an average of the monthly snow totals) and snowpack for the snow months (October – April) for the 30 year period 1981 – 2010. This is the most recently completed 30 year period and is commonly used to calculate climatological "normals" in meteorology.

   It should be noted that the ONI data is based on multiple centered thirty year periods; however, that is not the standard practice in meteorology and it would significantly complicate the analysis for this project to follow the ONI method. Therefore only the 1981-2010 thirty year period was used to compute snowfall and snowpack averages.

3. **Snowfall/Snowpack Delta (Monthly Snowfall vs 30 Year Average):** Calculate the snowfall/snowpack deltas between the values in step 1 (monthly averages) and the values in step 2 (30 year averages). This step will create an array of monthly delta values (October-April) for the weather station, for each year in the ONI record (1950 –present).

4. **Correlation Analysis:** Use scipy.stats.pearsonr to compute the correlation between each array (snowfall and snowpack) in step 3 and the ONI data set. Each data point in the ONI table is a three month rolling average so a January-February-March 2010 value would be compared to a February 2010 value from the snowfall/snowpack arrays.

   NOTE: According to the Scipy notes on "pearsonr", Pearson's correlation requires that each dataset be normally distributed. To simplify this analysis it was assumed that each dataset is normally distributed although it is possible that they are not.

Step 4, Regression analysis: Create a scatter plot and calculate a linear regression for snowfall delta and snowpack delta vs. the ONI. This step is only executed if a strong correlation exists.

Expanding to more than one station: Once the code was completed and successfully tested for one station, the next step was to look at additional stations.

Data Extraction and Processing: NOAA provides information in the readme file for the column structure of each data set; this was very helpful in analyzing and extracting data.

1. The inventory and stations files were used to check which weather stations had data for snowfall and snowpack for the years 1950 and 2010. For the purpose of narrowing the scope of this project only North American weather stations were considered. These were identified by non-blank values in the "State" field which only exist for North American stations. There are 4466 North American stations in the tar file.

2. The inventory file does not provide information about data for the interim years between 1950 and 2010 (see challenges section below). To determine if a station had data for all interim years the files had to be extracted and the data checked file by file. Therefore, this process is the most time consuming of the processes used in this analysis. The computer used for this project required 45 minutes to run this part of the code. Running this process with multiple cores may be one way to speed this up. Only 83 of the 4466 North American files contain a complete data record as defined for this project: all years, all months between 1950 and 2010. This result was a little surprising, but made the remaining analysis easier as it limited the number of files that needed to be analyzed.

3. The next step in the data extraction and processing was to identify only those files where either the snowfall delta, snowpack delta, or both contained a significant Pearson correlation with the Oceanic Niño Index. A significant Pearson correlation is indicated by a p-value of less than 0.05, but for this analysis in order to allow for more stations to be analyzed up to 0.10 was allowed. There were 37 files that contained a p-value less than or equal to 0.10, 27 of these contained a p-value less than or equal to 0.05.

4. The final step in the analysis was to convert the directory with these 37 files to a tar file, which could then be used to analyze correlations and create scatter plots, without requiring the user to run through the entire file extraction and processing.

4 VISUALIZATION

Visualizations include scatter plots for each station analyzed and a graph of the regression equation for snowfall and snowpack vs the Oceanic Niño Index.

5 CODING AND ANALYSIS CHALLENGES

Several challenges had to be overcome while working on this project:

1. **Missing Data**
   The weather station used as the test case to create the code to analyze a particular weather station was USW00094705, BARRE MONTPELIER AP, VT. While it was possible to create working code, it

soon became clear that this particular station was missing data for at least ten of the years between 1981 and 2010, making the computed average a 20 year average instead of a 30 year. This prompted further analysis of the data to determine if interim years, months, or days were missing from other weather stations as well. This analysis found that only a small number of stations contained all years and all months between 1950 and 2010, significantly reducing the number of weather stations that could be used for this analysis.

The weather stations used in the final analysis were those which contained a complete record, all years, all months for the years 1950 through 2010 (the years with Niño index data). The station and inventory files only provide data on the first and last year with recorded data for each element type, but there is no information in these files about interim years. The only way to determine if a file has data for interim years is to extract the data from the file and run through an analysis.

To simplify this project daily missing data was ignored. A complete scientific analysis would have to address and resolve this issue. Missing data is marked as "-9999" in the NOAA files. This was converted to "np.nan" for Python analysis purposes.

2. **File Extraction**
   Since it was possible to narrow the relevant files from over 93,000 files to just 4,466 using the inventory file, it seemed like the fastest method for extracting files was to extract only the 4,466 files from the tar.gz file. However, because this required a search through over 93,000 files to find each of the 4,466 this process was so slow that it would have taken days to extract the 4,466 files. It was actually much faster to extract all of the 93,000 files from the tar.gz archive, and then copy the 4,466 files to another directory. The computer used for this analysis took about 10 minutes to extract all 93,000 files, and another 10 minutes to copy the relevant 4,466 files to a new folder for further analysis.

3. **Run time**
   Some of the modules created for this project take a long time to run. To address this issue a tar file of just the 37 files was created for the correlation and scatter plot/regression line analysis without having to run through all of the extraction and processing modules. The user is presented with the option to either run through all of the extraction and processing modules or to use the pre-processed files.

6 RESULTS AND CONCLUSIONS

This analysis found 83 North American weather stations with a complete data record (all years, all months between October and April) for the years 1950 through 2010. Of these 83 stations, 37 showed significant correlations (p-value <= 0.10, 27 for p-value <=0.05) for snowfall, snowpack or both vs. the Oceanic Niño Index. The results for p-value <= 0.05 appear in Table 1. Scatter plots with regression lines for two stations appear in Figure 1.

Table 1 shows that 26 of the 37 weather stations show a negative correlation between snowfall and/or snowpack and the ONI while 11 of these stations show a positive correlation. The inconsistency in the direction of the correlation may imply that this is a case where correlation does not imply causation.

On the other hand, with only 83 stations having a complete data record, the available data may not be sufficient to draw a conclusion. It is conceivable that if there were hundreds or thousands of stations with a complete data record either the negative or positive correlations would be shown to be the anomaly and one of these directions could be established as the norm.

This project defined snow months as October through April which may have limited the number of weather stations with a complete record. A revision of this analysis could limit the snow months to just December through February which might result in more weather stations with a complete record and a different correlation result.

Weather is a complex phenomenon controlled by many factors. Even if this analysis had shown a definitive correlation between snowfall/snowpack and El Niño/La Niña, that would not definitively establish that this correlation was caused by El Niño/La Niña. Correlations could be caused by some other weather or climatic feature, or a combination of features, not studied in this project.

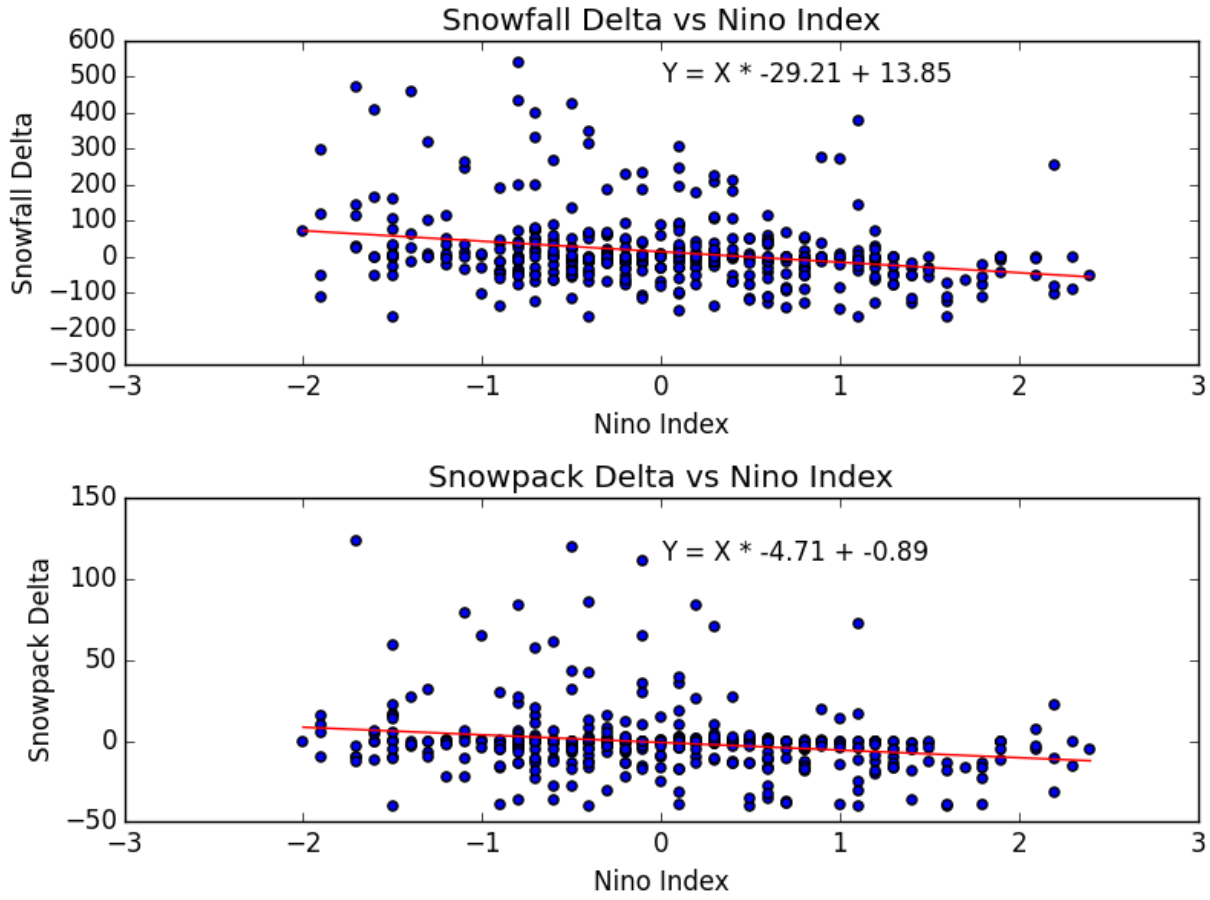
Table 1: North American Weather Stations with Complete Data Record and Significant Correlation

Note: If no value is shown for the correlation it is because the p-value is greater than 0.05

StationCode	StationName	ST/ Prov	SnowCorr	SnowPval	SnwdCorr	SnwdPval
USC00358746	UNION EXP STN	OR	-0.26	0.00	-0.21	0.00
USW00025339	YAKUTAT STATE AP	AK	-0.20	0.00	-0.20	0.00
USW00024131	BOISE AIR TERMINAL	ID	-0.17	0.00		
USW00014845	SAGINAW MBS INTL AP	MI	-0.15	0.00	-0.12	0.01
USW00014913	DULUTH	MN	-0.15	0.00	-0.13	0.01
USW00094008	GLASGOW INTL AP	MT	-0.15	0.00	-0.11	0.02
USW00094014	WILLISTON SLOULIN INTL AP	ND	-0.14	0.00	-0.11	0.02
USW00014898	GREEN BAY	WI	-0.14	0.00	-0.16	0.00
USW00024143	GREAT FALLS INTL AP	MT	-0.13	0.01		
USW00025501	KODIAK AP	AK	-0.13	0.01		
USW00014827	FT WAYNE INTL AP	IN	-0.12	0.01		
USW00026510	MCGRATH AP	AK	-0.12	0.01	-0.17	0.00
USW00014839	MILWAUKEE MITCHELL AP	WI	-0.12	0.02		
USW00024033	BILLINGS LOGAN INTL AP	MT	-0.11	0.02		
USW00014826	FLINT BISHOP INTL AP	MI	-0.11	0.02		
USW00023169	LAS VEGAS MCCARRAN AP	NV	-0.10	0.04	-0.10	0.04
USW00023155	BAKERSFIELD AP	CA	-0.10	0.04		
USW00013985	DODGE CITY	KS	0.11	0.03		
USW00093721	BALTIMORE WASH INTL AP	MD	0.11	0.03	0.11	0.02
USW00013741	ROANOKE RGNL AP	VA	0.11	0.02	0.13	0.01
USW00023061	ALAMOSA SAN LUIS AP	CO	0.12	0.02	0.11	0.03
USW00013966	WICHITA FALLS MUNI AP	TX	0.13	0.01		
USW00023047	AMARILLO	TX	0.16	0.00	0.16	0.00
USW00023042	LUBBOCK	TX	0.16	0.00	0.15	0.00
USW00013733	LYNCHBURG RGNL AP	VA			0.14	0.00
USW00013894	MOBILE	AL			0.11	0.02
USW00013967	OKLAHOMA CITY WILL ROGERS AP	OK			0.11	0.02
USW00023062	DENVER-STAPLETON	CO			0.11	0.03

Figure 1: Weather Station Plots

1a: USC00358746 UNION EXP STN in OR (negative correlation)



1b: USW00013985 DODGE CITY in KS (positive correlation)