# Project 1 - Snowpack

Mohamad Ilham Mohd Rozie
Muhammad Zulfadli Che Zabri

February 28, 2020

# Contents

# 1 Introduction

The project 1 is part of the coursework of GPGN 268 and two persons have been assigned to complete the research and submission of the report. This project aims to predict the future snow depth at Berthoud Summit from current and historic data. In the society nowadays, people tend to confuse the snowfall and snow depth. Snowfall is the measurement of depth of snow since the previous snowfall observation, while snow depth is the total depth of snow on the ground, which taken at the scheduled time of observation with a measuring stick [1]. The area focused in this project is a mountainous terrain, or known as summit. Generally, in a mountainous terrain, there will be a situation where no snow is observed on south-facing slopes, while the snow is accumulated on north-facing areas. Hence, the recorder (person who in charge of measuring snow depth) has to use his/her good judgment to take the average of snow depth of several hundred yards from the weather station.

In this project, snow depth will be the subject to focus on and make a decent prediction of the future snow depth on March 28, 2020. Temperature data also will be included as an additional information to verify the factor contribution of snow depth. Snow depth in Colorado is very significant as 80% of Colorado water comes from snowpack [2]. The collected data from this project used a measurement called as SWE, or known as snow-water equivalent. This type of data allows to determine the volume of water obtained if all the snow was melted.

# 2 Methods and Analysis

Two variables were used and they were snow depth and temperature. These data were used simultaneously to each other in order to enhance our understanding of how the snow depth and temperature behave throughout the years.

## 2.1 Data Mining

Data mining is a process of analyzing a large set of data and converting them to a usable data. The initial approach to this is by plotting the data. For instance, snow depth in y axis is plotted again year in x axis. 10 years of historical data were obtained because trend can be vary within 10 years, and take the most similar trend expected to be in prediction. In addition, 10 years are the ideal time range to estimate the trend since they are still considered recent whereas time frame of above 10 years are quite not approaching to the prediction because of many environmental factors have changed that contribute to the future trend.

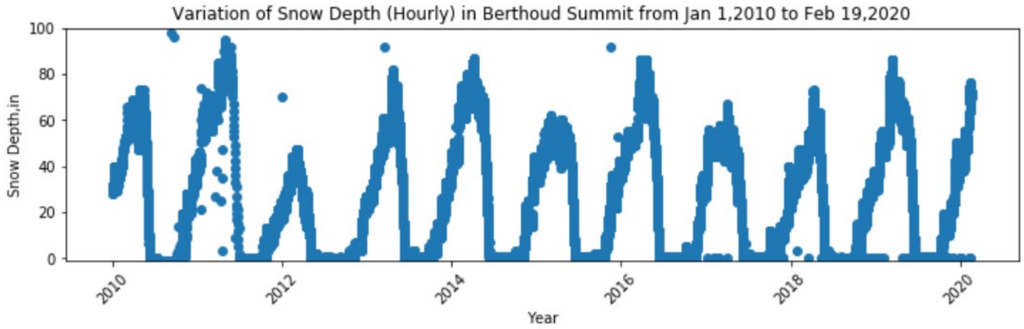


Figure 1: First Visualization of the Snow Depth Data in 10 Years.

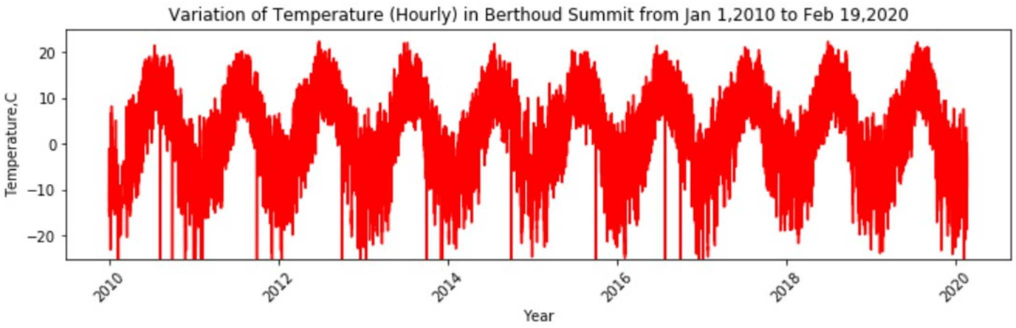


Figure 2: First Visualization of the Temperature Data in 10 Years.

All the data sets were attained from the U.S. department of Agriculture and this agency is the most reliable source of natural resource and climate data. Here are the step-by-step how the data were acquired:

1. Go to https://www.usda.gov
2. Search for SNOTEL (*SNOTEL is a stored system of snowpack data and related climate sensors regulated by Natural Resources Conservation Service [2]*)
3. Choose 'Snow Water Equivalent'
4. Select 'Stations', then go to 'Berthoud Summit'
5. Download the current and historic data from 'Site Page'

## 2.2 Data Cleaning

The data retrieved from the public source is usually not in proper order or in an order that following the coding script. Hence, the data needs to be cleaned first and put into a proper format for coding function to work. In this project, there were values written as -99.99, which was unrealistic. That values came up as a result of no presence of data, therefore the system replaced them with -99.99. In order to make the project viable, the -99.99 values then were replaced by 0. This was done to make sure all the data in a good shape; meaning when the equation curve will be applied eventually, the curve will not account for those negative values. At first, the -99.99 values were replaced by the average of the previous and forward value of that particular value, however this will take a ton of time. Due to that, converting those inapplicable values to 0 was the best option as it is time efficient. Below shows why the uncleaned data is not relevant to make a future prediction.

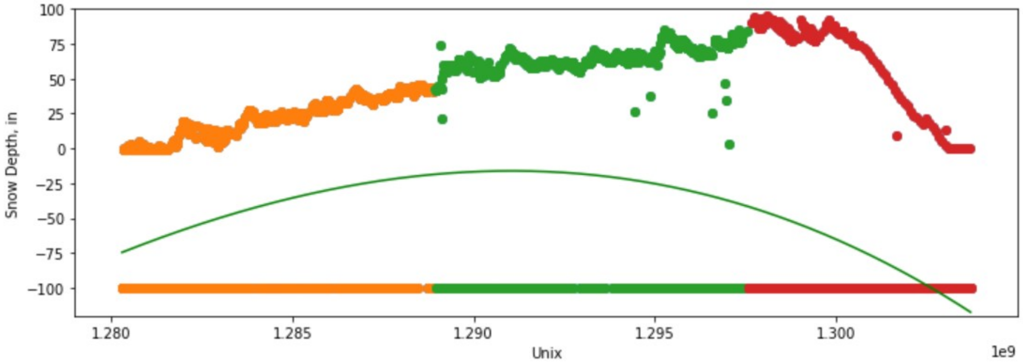


Figure 3: Applied Polynomial Regression Curve with Uncleaned Data.

The polynomial regression curve took an average of between two different curve of data sets. Not to mention, the average was even below 0 value which did not make sense since the snow depth should be higher during that period.

## 2.3 Enhance Visualization

From a quick glance to Figure 1, it was difficult to make a comparison of which curve should be referred to. The problem could be solved by enhancing the visualization, which changing the color for each of curve, thus it will be easier to choose which curve should be a reference to the future prediction.

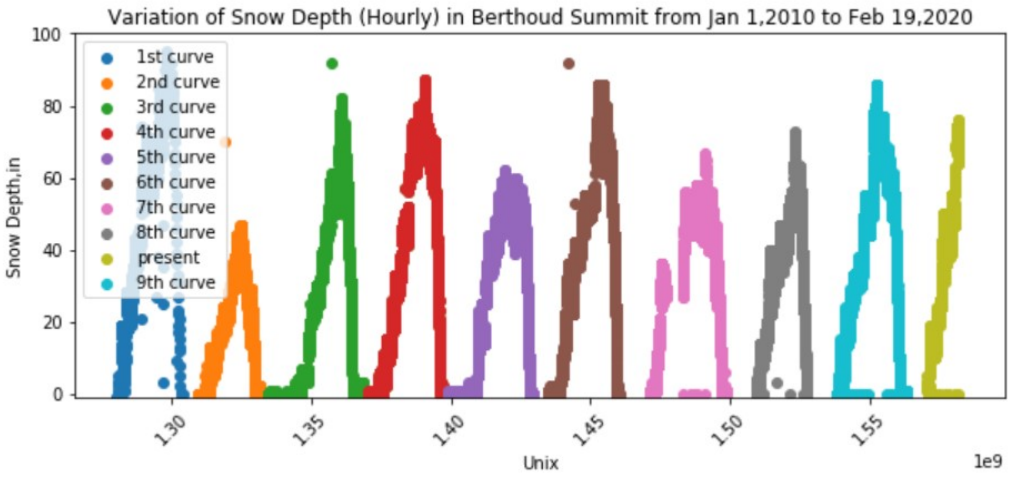


Figure 4: Variation of Snow Depth in Berthoud Summit from January 1, 2020 to February 19, 2020.

## 2.4 Polynomial Regression Interpolation

There are two options in making a prediction of data; linear and polynomial equation. In this case, polynomial equation is used to produce a better result since it is more precise. Briefly, the difference between linear and polynomial curve is the linear curve measures a straight line curve that is plotted based on two different points in the data, while polynomial curve measures the whole data sets to better fit the non-linear data. There are two main reasons that contributed to the usage of polynomial curve method in this project:

1. **Scattered Data**
   The trend of data were scattered and non-linear; meaning the data were not having a constant distribution throughout the year. Logically, the snow will vary each year and it often takes place in winter (November to December) and spring (January to March) season.

2. **Precision**
   As mentioned earlier, polynomial curve accounts for most of the points in the data sets. The curve is non-linear and following the trend of the scattered data. Polynomial regression will improve the model's closeness to the data by increasing the relationship between the factors and the variable. Therefore, it will give more accurate result.

The polynomial equation requires us to use values, hence date time object are not accepted. The best option is to convert all the date time object to unix time. Unix time or also known as unix time stamp is a specific value to track time in total of seconds. This counts was started on January 1, 1970; known as Unix Epoch [3]. Therefore, the unix time is merely the number difference between the Unix Epoch and a particular date time. Here is the figure of polynomial regression curve that was applied on the data from July 2010 to July 2011.

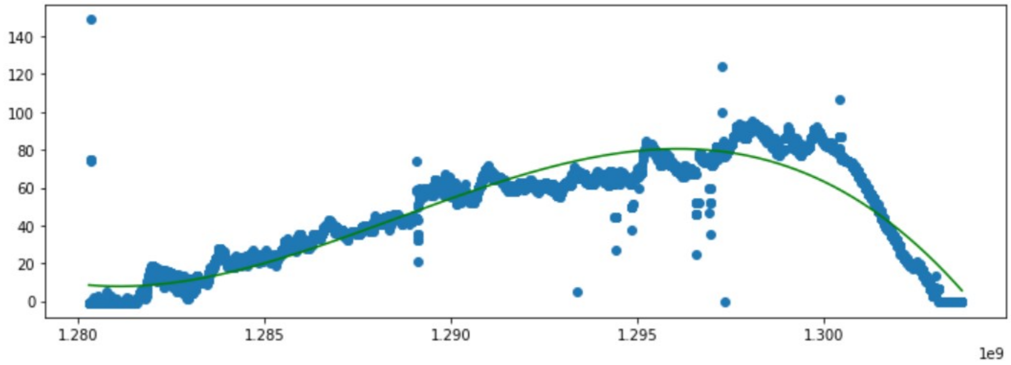


Figure 5: Polynomial Regression Curve from July 2010 to July 2011.

The curve went over most of the points on the plot, and it was non-linear since the curve bent or in a U-shape facing upward. If a linear curve was applied, there will be only one straight line that fits the whole data sets which will not give a satisfying result since there will be a large sets of data that were still left unaccounted for. Thus, polynomial regression curve is always the best option to proceed. However, there were certain points that the polynomial regression curve could not go through based on the Figure 5. In order to ensure the accuracy and reliability of the prediction process later, the curve should have been extrapolated in more efficient way to cover as much points as possible. Due to that, there were several alterations have been made to align to the objective of this project and these alterations would be discussed in the next section.

## 2.5 Prediction

Based on Figure 4, the first (1st) or dark blue curve was chosen to be applied to the future prediction because the curve is following the most of current trend in 2020. The first curve represented the second half of the 2010 and first half of 2011. In order to deliver the most relevant result that can be achieved, there are some alterations have been done and they were:

1. **Error Bar**
   The error bar shows a range of values that are possible to occur on that particular period. The range is calculated using standard deviation equation. Noted that any values in the range can be achieved at a particular time.

2. **Three Polynomial Regression Curves**
   Since the initial plot with single polynomial curve showed less precision because of the fluctuation of the curve through the year, we came to a decision to use multiple polynomial curve to get as close as the curve could to the large sets of data. Based on the Figure 5, there were three distinct slopes observed, hence creating a polynomial curve for each slope would suffice.

3. **Starting Point of Polynomial Equation**
   Mother nature hits at different period every year, but it can be predicted based on either a solstice (winter and summer) and an equinox (spring or autumn). Briefly, a solstice is a state when the sun reaches the most southerly or northerly in the sky, while an equinox is a state when the sun passes by the earth's equator [4]. In terms of the project, the snow depth in 2011 was made as reference to the snow depth in year 2020. In this case, the snow depth in April 2011 was assumed to show the same trend in February 2020. Back in 2011, 70 in of snow depth was measured to be around 1297584000, which was quite same with the snow depth in the mid-February. Therefore, the polynomial regression equation used were shifted to the point where the snow depth of April in 2010 to be applied to the starting point of prediction in mid-February in 2020.

Below is the plot of snow depth from July 2010 to July 2011. The red (or third) curve was used to generate the polynomial equation, then applied to extrapolate a prediction curve in 2020. The red curve started from April 24, 2011 with a unix time at 1297468800.

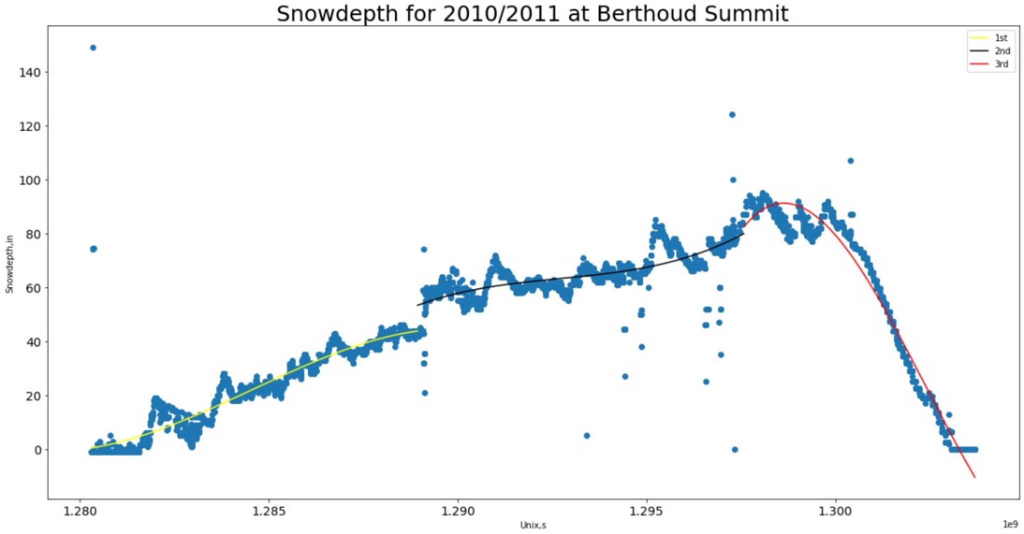


Figure 6: Separation of Three Slopes from July 2010 to July 2011.

The equation that has been generated from the red (or third) curve from Figure 5 was

$$y = -4.777e^{-18}x^3 + 2.269e^{-8}x^2 - 35.94x + 1.897e^{10} \tag{1}$$

In making a comparison, below is the recent data running from October 12, 2019 to February 19, 2020. There were three different colors that corresponded to different colors in Figure 6. Figure 6 represents data from 2010 to 2011, while Figure 7 represents data from 2019 to today (February 19, 2020).

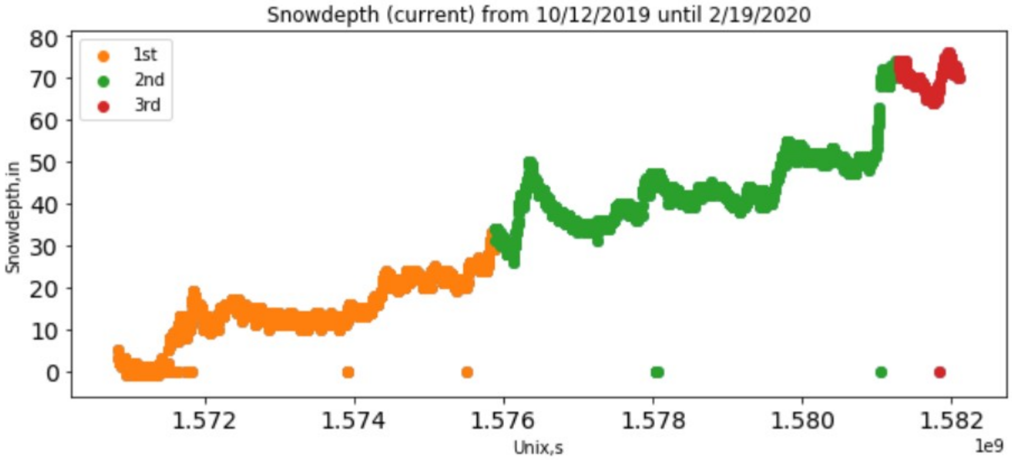


Figure 7: Current Snow Depth Data from December 10, 2019 to February 19, 2020 (Today).

And here is the plot for the prediction of snow depth in March 28, 2020. The prediction by extrapolation is just inserting the input into the polynomial equation from February 19, 2020 (1582000400 seconds) to March 28, 2020 (1585400400 seconds). The snow depth obtained at 1 PM on March 28, 2020 is **77.06 in**. The error bar is included to show the possibility of ranges of snow depth on March 28,2020 at 1pm. The error calculated is **19.31 in**.

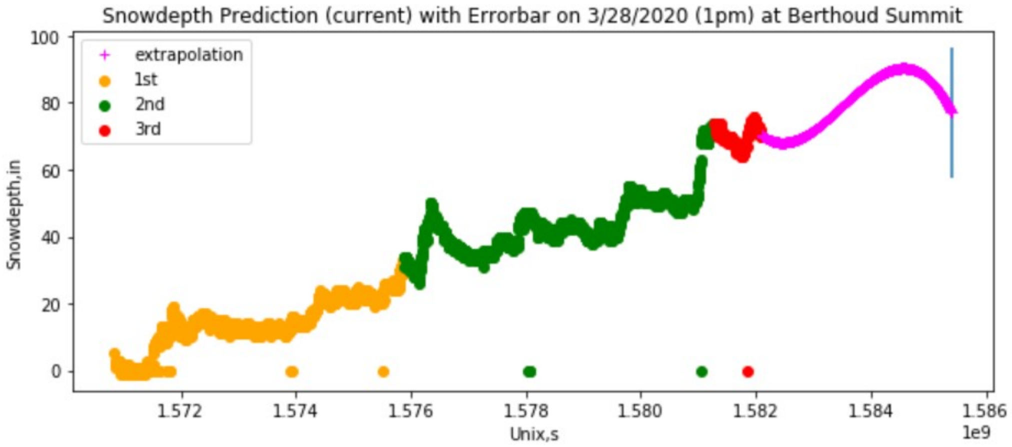


Figure 8: Prediction of Snow Depth at 1 PM on March 28, 2020.

## 2.6 Analysis

Firstly, there is no correlation between snow depth and temperature because temperature plot shows a constant trend each year and they have not changed much in 10 years. Besides, snow depth shows a pattern for each year but they were distinct in values between each other. Unlike temperature, snow depth varies from year to year and this non-constant trend made the prediction process became more difficult to interpret. The snow depth trend somehow showed a relationship with wind-temperature-weather related pattern, or commonly known as El Nino and La Nina occurrences. Based on historical data from National Oceanic and Atmospheric Administration (NOAA), El-Nino had occurred in 2002–03, 2004–05, 2006–07, 2009–10, 2014–16 and 2018–19 [5]. In addition to that, International Research Institute for Climate and Society (IRI) had predicted that an EL-Nino Southern Oscillation (ENSO) neutral event will continue throughout 2020 based on Figure 10. ENSO neutral event simply means that there will be no event of El-Nino or La-Nina for that period of time [6]. Therefore, we make an assumption that the trend of snow depth for this year will follow specifically to the year that had an ENSO neutral event. In our case, we determine that the snow depth trend in 2019/2020 will follow the snow depth trend in 2010/2011. That is why we use the polynomial regression data for 2010/2011 and apply it to predict the snow depth in 2020 specifically at 1pm on March 28,2020.

Next, the error bar is obtained by calculating the standard deviation of snow depth from December 10, 2019 to February 19, 2020. We believe that to calculate the possible error in our snow depth prediction, it is best to assume that the value will deviate from the mean snow depth based on the current snow depth trend in 2019/2020. We also believe that if we take into account of the snow depth trend in different period such as the 1st curve or the 2nd curve in Figure 4, the deviation will not perfectly describe the actual digression for the current 2019/2020 season.

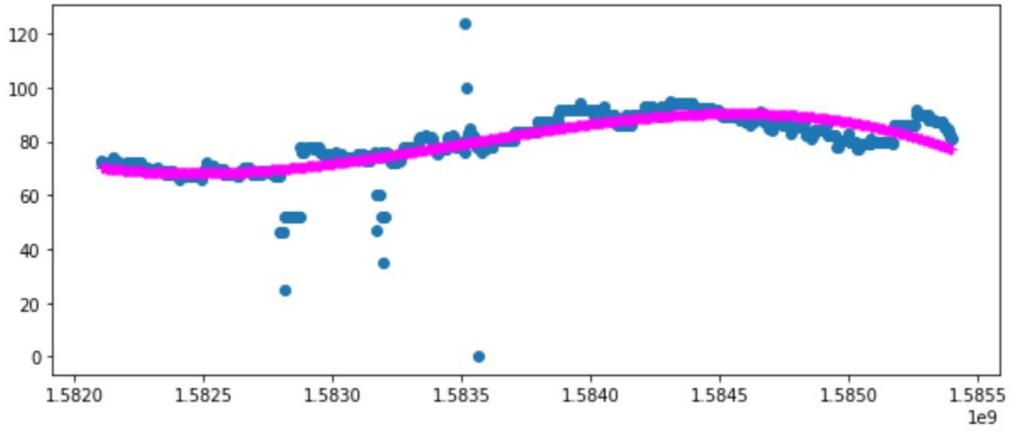


Figure 9: Polynomial Regression Curve of Prediction Data from February 20, 2020 to March 28, 2020.

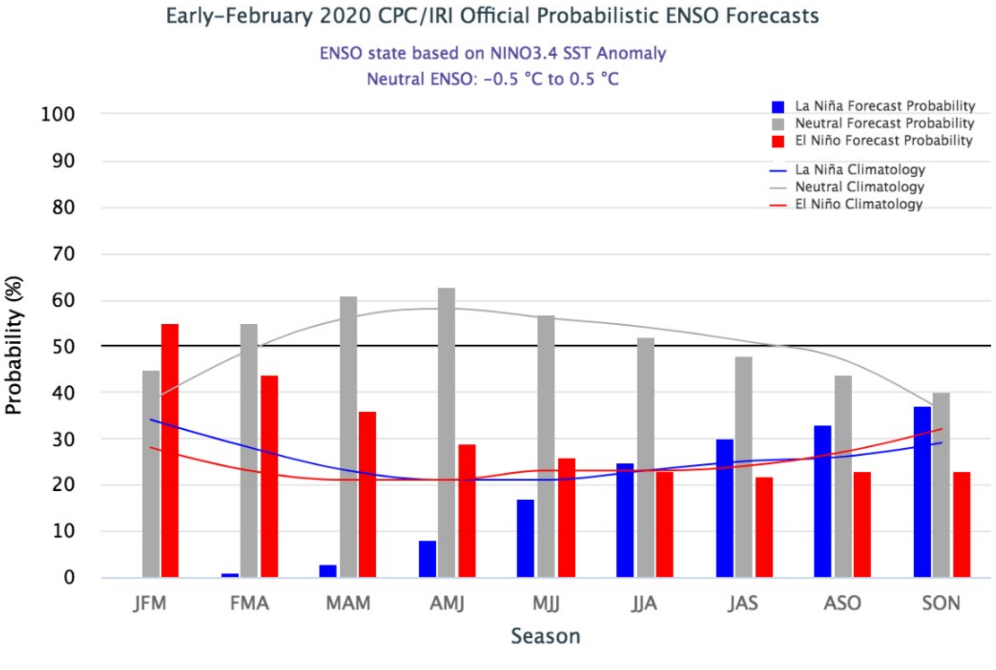


Figure 10: El Nino and La Nina Forecasts in 2020.

# 3 Obstacles

Upon the completion of the project, there were some obstacles that we encountered. These obstacles were not affecting the result of the prediction because there was another alternative taken to counter these obstacles. The first obstacle is using date time object to plot the data into a graph. The polynomial curve requires a single value to measure the trend of the data and creates an equation based on the data. Therefore, date time object could not be used as variables in x-axis. That is the reason why we use unix time instead.

Next, we also encountered obstacle in trying to extrapolate the data up until March 28,2020. We realized that we cannot extrapolate the data if we do not have the snow depth in any time after 19 February 2020. Based on this limitation, we came to conclude that we need to make assumption that the snow depth in 2019/2020 trend will follow exactly the trend in 2010/2011 based on past historical data and current ENSO prediction by IRI.

# 4 Conclusion

In conclusion, the predicted snow depth that is obtained for this project is 77.06 in with a standard deviation of 19.31 in. Moving forward with this project, there are few improvements that we believe that we can do to better enhance this project. First, we believe that we should take into account of more data beyond the past 10 years in order to understand the snow depth trend better in ENSO neutral events, El-Nino events and La-Nina events. Second, we believe that we should also take into account of the other nearby recording stations so that we can better visualize the snow depth trend and see if there are any correlation between two or more different recording station. Third, machine learning is one of the precise ways to estimate the trend of snow depth because the system within the machine learning has a capability of interpreting a range of data and then forecasting them. This method could be done by using an open source Python library named as SciPy, which can be used for both scientific computing and technical computing. If we have more time or has been exposed with SciPy in the class, than we might be able to increase the confidence level of our snow depth prediction for this project.

# References

[1] N. W. Service, “Snow measurement information.”

[2] C. Reppenhagen, “This is how colorado snowpack is measured.”

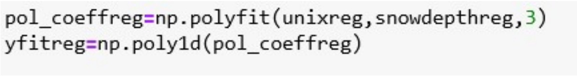
[3] T. Converter, “Epoch unix time stamp converter.”

[4] C. Boeckmann, “When do the seasons start in 2020?”

[5] N. W. Service.

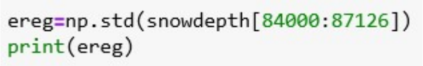
[6] NOAA.

# Appendices



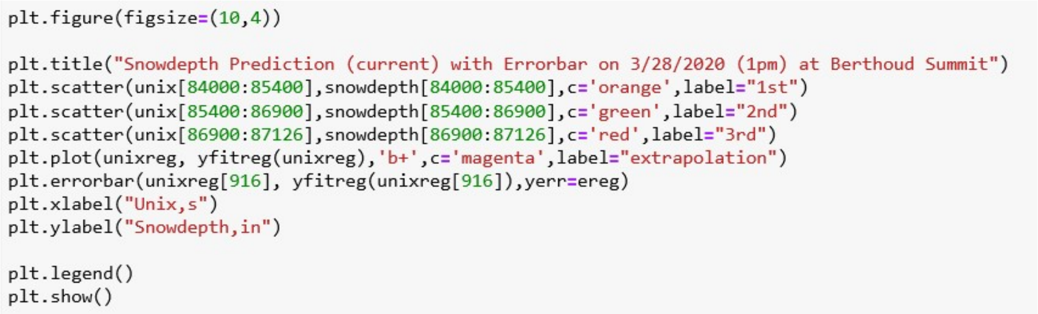
```
pol_coeffreg=np.polyfit(unixreg,snowdepthreg,3)
yfitreg=np.poly1d(pol_coeffreg)
```

Figure 11: Code for Polynomial Regression Curve.



```
ereg=np.std(snowdepth[84000:87126])
print(ereg)
```

Figure 12: Code for Standard Deviation or Error Bars.



```
plt.figure(figsize=(10,4))

plt.title("Snowdepth Prediction (current) with Errorbar on 3/28/2020 (1pm) at Berthoud Summit")
plt.scatter(unix[84000:85400],snowdepth[84000:85400],c='orange',label="1st")
plt.scatter(unix[85400:86900],snowdepth[85400:86900],c='green',label="2nd")
plt.scatter(unix[86900:87126],snowdepth[86900:87126],c='red',label="3rd")
plt.plot(unixreg, yfitreg(unixreg),'b+',c='magenta',label="extrapolation")
plt.errorbar(unixreg[916], yfitreg(unixreg[916]),yerr=ereg)
plt.xlabel("Unix,s")
plt.ylabel("Snowdepth,in")

plt.legend()
plt.show()
```

Figure 13: Code for Prediction Plot.