Analysis and Conclusion
Global Temperature vs Precipitation Project

I. Introduction:

Climate change is really rampant these days with there being a constant increase in temperature, especially due to global warming. At the same time, water scarcity and reduced rainfall is definitely going to turn more dangerous in the near future. Hence, we will use this project to analyze if there is any relationship between global Temperature and Precipitation. We will also use these global temperatures to make predictions about precipitation, and hence utilize it to prevent ourselves from extreme water scarcity in the coming years.

The features of the model are each of the months of January, February, March, April, May, June, July, August, September, October, November, and December. The reason for choosing these features is because the trends in annual precipitation vary vastly with respect to monthly temperature depending on which month of the year it is.

A small practical example of this could be that these days we experience more fluctuations in precipitation in some months, when all of the months have almost the same changes in temperature due to global warming.

II. Data:

The data that we have consists of Monthly Air Temperatures (in oC), and Monthly Precipitation (in mm) respectively from January 1900 to December 2014 based on latitudes and longitudes of locations. Thus, our dataset contains Air Temperature (in oC) and Precipitation (in mm) for every month of these 115 years.

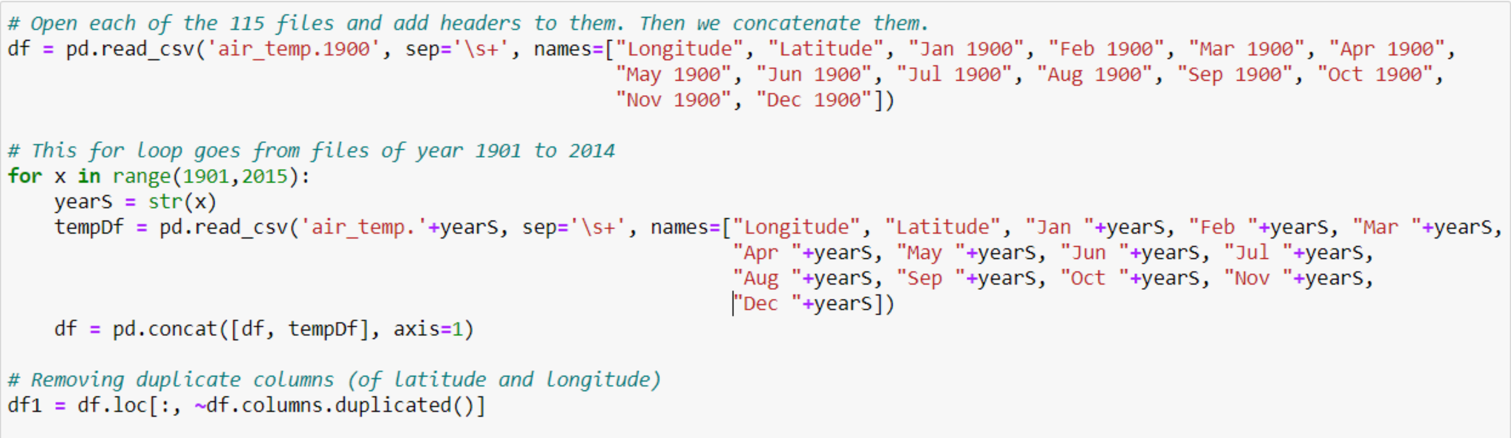
This dataset contains 2 files containing Monthly Air Temperatures and Precipitation, each from January 1900 to December 2014 in csv format.

Both of these datasets have a size of 85794 rows x 1382 columns. Lastly, this also means that we have data tracked for 85794 locations based on their latitudes and longitudes.

III. Methods:

We collected our data form after a really extensive research from the website https://climatedataguide.ucar.edu/climate-data/global-land-precipitation-and-temperature-willmott-matsuura-university-delaware which contains Terrestrial Temperature and Precipitation Data collected by the University of Delaware. Furthermore, the actual data file was in the form of .tar.gz files, which contained yearly files of Monthly Temperature and Precipitation from 1990 to 2014 based on latitude and longitude of the location.

The data cleaning was done by first providing appropriate headers to each of the yearly files. The python code used for data cleaning is depicted in Figure 1.



```
# Open each of the 115 files and add headers to them. Then we concatenate them.
df = pd.read_csv('air_temp.1900', sep='\s+', names=["Longitude", "Latitude", "Jan 1900", "Feb 1900", "Mar 1900", "Apr 1900",
                                                    "May 1900", "Jun 1900", "Jul 1900", "Aug 1900", "Sep 1900", "Oct 1900",
                                                    "Nov 1900", "Dec 1900"])

# This for loop goes from files of year 1901 to 2014
for x in range(1901,2015):
    yearS = str(x)
    tempDf = pd.read_csv('air_temp.'+yearS, sep='\s+', names=["Longitude", "Latitude", "Jan "+yearS, "Feb "+yearS, "Mar "+yearS,
                                                              "Apr "+yearS, "May "+yearS, "Jun "+yearS, "Jul "+yearS,
                                                              "Aug "+yearS, "Sep "+yearS, "Oct "+yearS, "Nov "+yearS,
                                                              "Dec "+yearS])
    df = pd.concat([df, tempDf], axis=1)

# Removing duplicate columns (of latitude and longitude)
df1 = df.loc[:, ~df.columns.duplicated()]
```

Figure 1: Code for data cleaning

We also checked if there are any null values, and while there were not any, we did spot some 0 values, which means the data was already cleaned before usage

This was done in order to clean Monthly Air temperature Data. The same process was followed for cleaning the Monthly Precipitation Data.

Finding the data was the first difficulty that we faced at the beginning of our project. The data we collected at the beginning was more focused on marine climate data, however that data did not have enough information on wave heights and even air temperatures. That is why we had to restart our search for datasets and we went through about 100 datasets after finalizing on the one we did right now. There was also a huge learning curve, as our expertise on finding good datasets grew exponentially, from when we had started, at the end of our search process.

IV. Analysis:

It was truly fascinating to explore this dataset as our findings were very astounding. During our Exploratory Data Analysis (EDA), we observed that there were huge changes in both the precipitation and temperature data for all of the locations together. This helped us in being a little more confident of our hypothesis, based on our exploration of data so far, although we do need to keep in mind that correlation does not mean causation.

We also saw a decrease in the monthly average precipitation. The line plots were more indicative of the trending decrease in average monthly precipitation for various locations based on latitude and longitude. We see in Figure 2 below, where we are plotting the mean annual precipitation of all years, along with mean annual precipitation in 1900, and mean annual precipitation (in mm) in 2014.

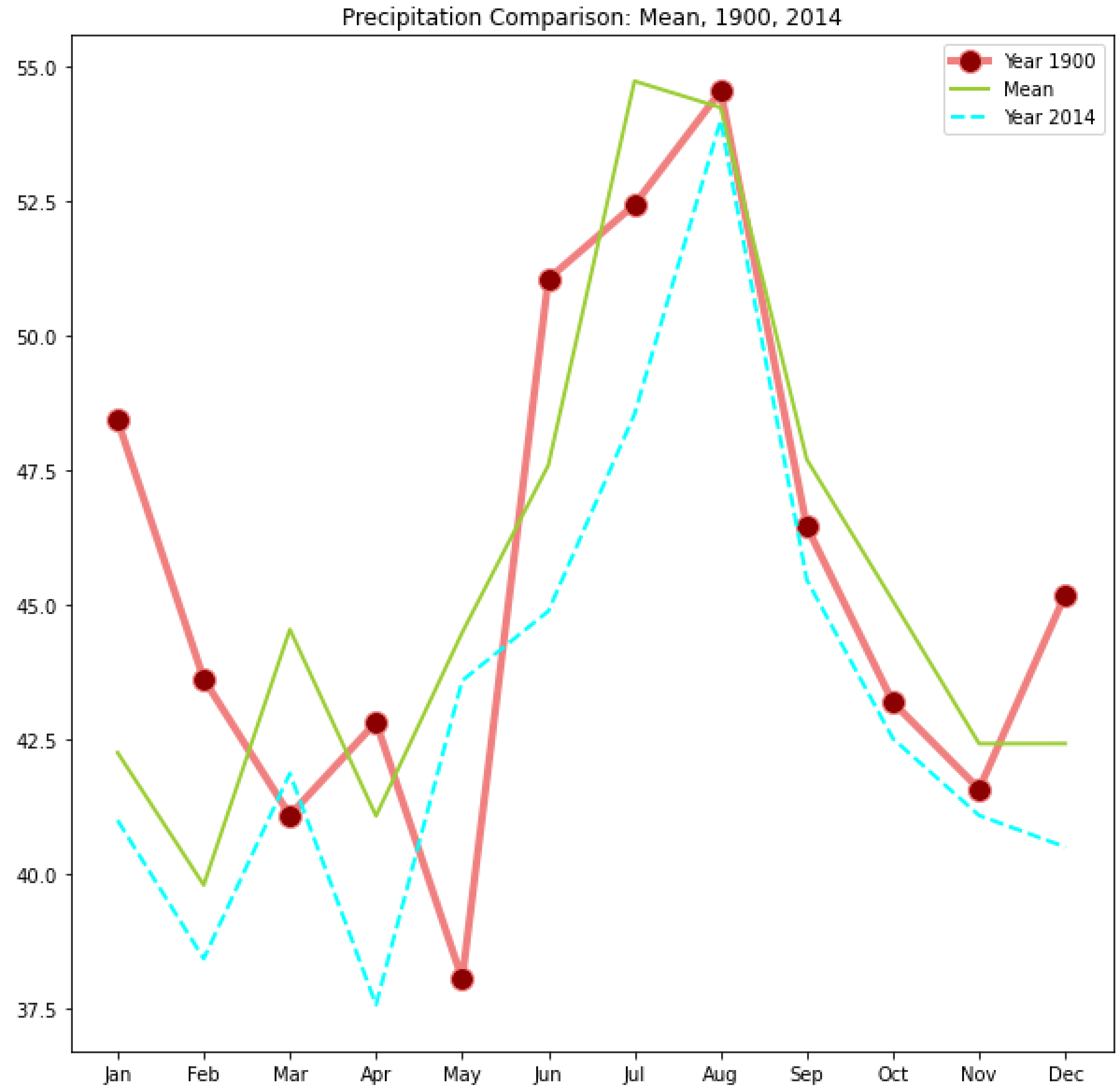


Figure 2: Comparison of Precipitation

We see that the annual precipitation in 2014 has greatly reduced in the first half of the year, which is generally hotter, compared to the second half of the year. Also, from the months of August to December in 2014, we again see that the rainfall has indeed decreased compared to mean rainfall and rainfall in 1900.

Also, in Figure 3 below, we are plotting the annual temperatures in 1900 and 2014 vs the mean annual temperature (in ℃) over all of these years.

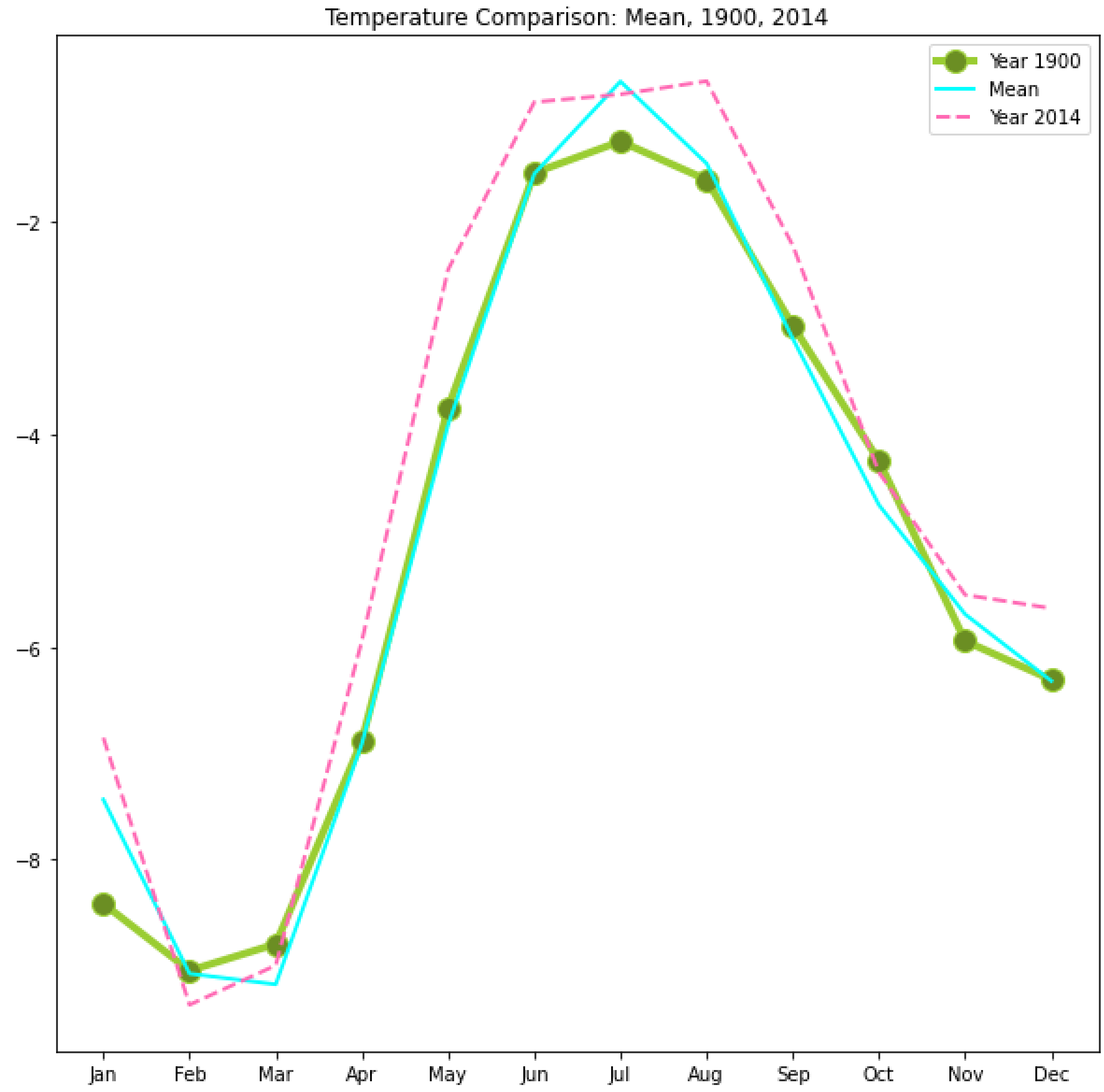


Figure 3: Comparison of Temperature

We again see a similar case as earlier, wherein the temperature in 2014 is really high as compared to the mean temperature and the annual temperature in 1900. Also, we see that, unlike other months, the months of February and March have actually experienced a decrease in their annual temperature in 2014 as compared to previous years.

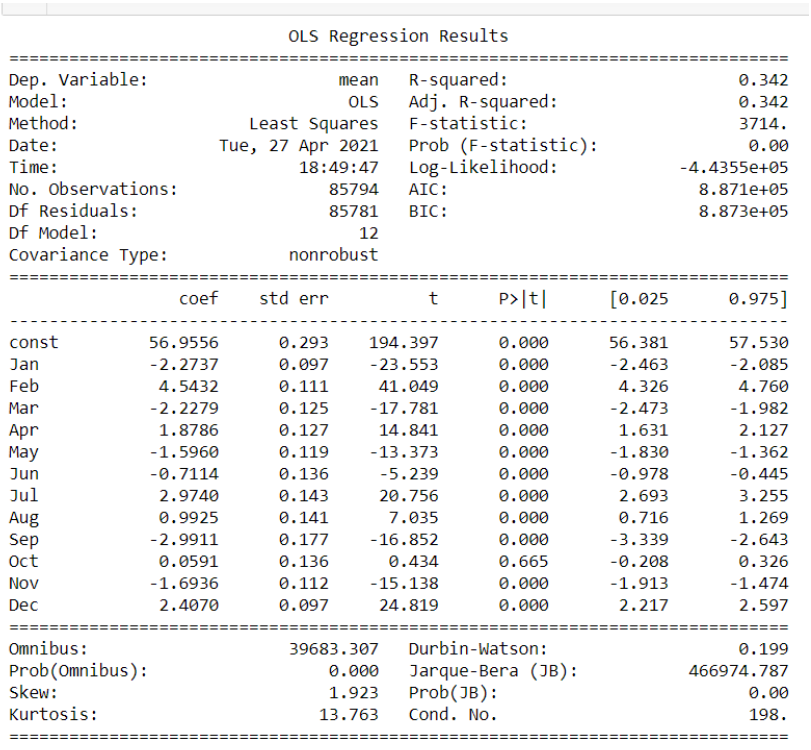
We used a Linear Regression model as our predictions are actual Precipitation values (in mm). We find the model using 80% of the dataset for training and 20% for testing giving us the best predictions. We specifically use the multivariate version of linear regression in order to accurately make our predictions. Also, we tested the results on various other train-test splits such as 90%-10%, 85%-15%, 75%-25% and 70%-30%, but found that the 80%-20% train-test split gave us the best results.

We found the mean monthly temperature for every location, and use it as our input (X). Similarly, we found the mean yearly precipitation for every location and used it as our y in order to train our model.

V. Results:

Using statsmodels.api, we find that the Adjusted R squared is 0.342 thus explaining the model can explain 34.2% of variations seen in training data. However, as we are attempting to predict the relationship between temperature and precipitation, which is more randomized, this model explaining more than 25% is optimum.

As depicted in Figure 4, we see that only the mean temperatures in February, April, July, August, and December have an impact on the mean precipitation. This is because the t-values of these variables are considerably bigger than 0. The mean squared error on this model is also the least among all the other models.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   mean   R-squared:                       0.342
Model:                            OLS   Adj. R-squared:                  0.342
Method:                 Least Squares   F-statistic:                     3714.
Date:                Tue, 27 Apr 2021   Prob (F-statistic):               0.00
Time:                        18:49:47   Log-Likelihood:            -4.4355e+05
No. Observations:               85794   AIC:                         8.871e+05
Df Residuals:                   85781   BIC:                         8.873e+05
Df Model:                          12
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         56.9556      0.293    194.397      0.000      56.381      57.530
Jan           -2.2737      0.097    -23.553      0.000      -2.463      -2.085
Feb            4.5432      0.111     41.049      0.000       4.326       4.760
Mar           -2.2279      0.125    -17.781      0.000      -2.473      -1.982
Apr            1.8786      0.127     14.841      0.000       1.631       2.127
May           -1.5960      0.119    -13.373      0.000      -1.830      -1.362
Jun           -0.7114      0.136     -5.239      0.000      -0.978      -0.445
Jul            2.9740      0.143     20.756      0.000       2.693       3.255
Aug            0.9925      0.141      7.035      0.000       0.716       1.269
Sep           -2.9911      0.177    -16.852      0.000      -3.339      -2.643
Oct            0.0591      0.136      0.434      0.665      -0.208       0.326
Nov           -1.6936      0.112    -15.138      0.000      -1.913      -1.474
Dec            2.4070      0.097     24.819      0.000       2.217       2.597
==============================================================================
Omnibus:                    39683.307   Durbin-Watson:                   0.199
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           466974.787
Skew:                           1.923   Prob(JB):                         0.00
Kurtosis:                      13.763   Cond. No.                         198.
==============================================================================
```

Figure 4: Regression Results

Thus, our Equation for Linear regression model looks without random error like this:

y = -2.2737 * X1 + 4.5432 * X2 - 2.2279 * X3 + 1.8786 * X4 - 1.5960 * X5 - 0.7114 * X6 + 2.9740 * X7 + 0.9925 * X8 - 2.9911 * X9 + 0.0591 * X10 - 1.6936 * X11 + 2.4070 * X12 + 56.9556, where Xi stands for the mean monthly temperature in i^{th} month of the year.

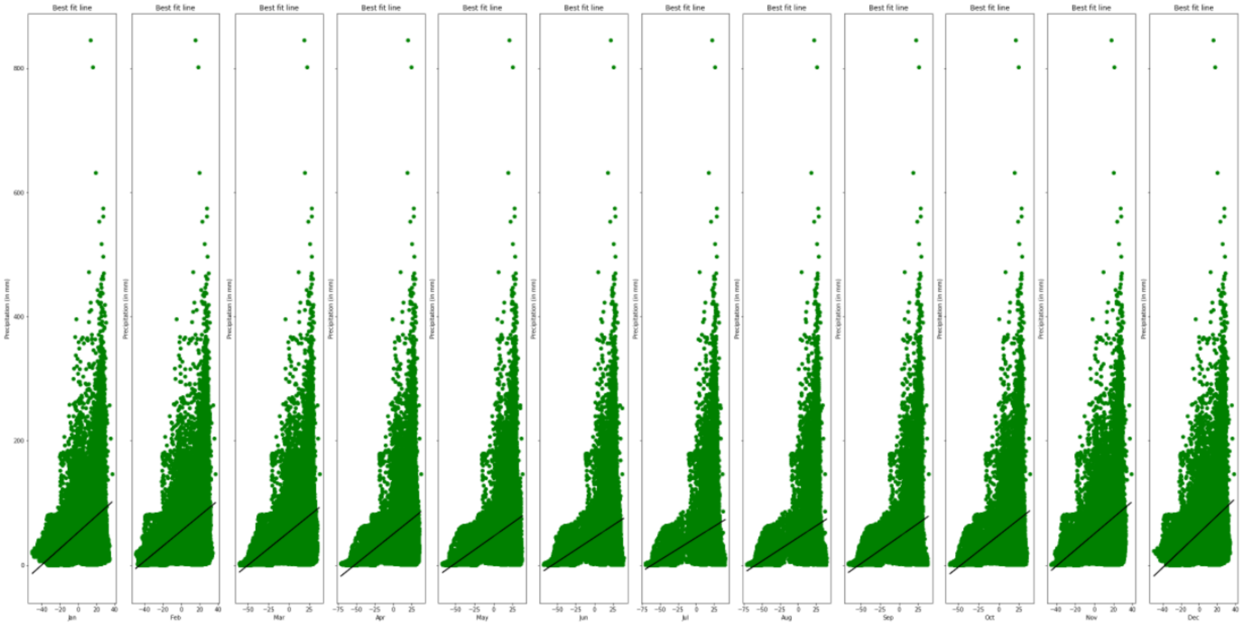


Figure 5: Glimpse of the Regression Plot for every feature

In Figure 5 above, we are plotting linear regression for each feature, that is, every month of the year. We see that as the Temperature increases, our linear regression model faces more outliers, which is what accounts for more errors for the model.

At the very end, let us plot our predicted y and actual test y to see how they coincide for the first 100 locations of our test data in Figure 6.

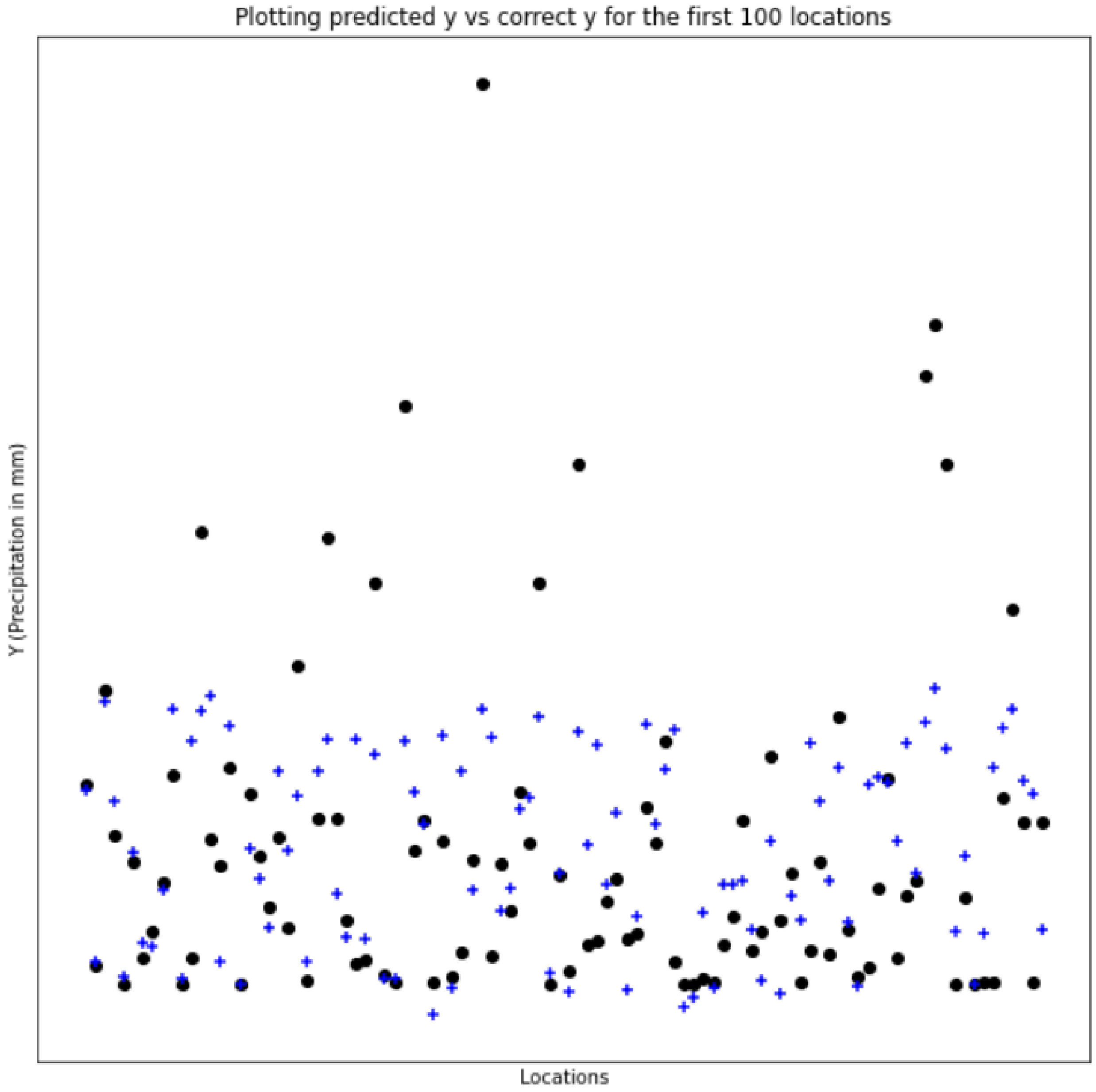


Figure 6: Plot of predicted y vs real y

We thus see that the model predicts the values correctly for those locations who have a lower precipitation value. As the value increases and as there are more outliers, the model somewhat fails to predict those values correctly.

VI. Conclusion:

Mean Annual Precipitations is influenced by the mean monthly temperatures in the months of February, April, July, August, and December. Therefore, it might be more probable to experience sudden, unexpected or excessive rainfall in these months over a year. Lastly, there is a relationship between the global mean monthly temperatures and annual precipitation based on the results of our model. In conclusion, our findings would be extremely helpful in deciding the plan of action taken by humans in order to protect from the dangers of climate change.

VII. Appropriate Links for further information:

1. Link to the full data in downloadable form: https://drive.google.com/drive/folders/1MUz0PMaH1Hh_UWIHAVNBXKylccKzPrCs?usp=sharing
2. Link for Temperature Data Readme: http://climate.geog.udel.edu/~climate/html_pages/Global2014/README.GlobalTsT2014.html
3. Link for Rainfall Data Readme: http://climate.geog.udel.edu/~climate/html_pages/Global2014/README.GlobalTsP2014.html