The Impact of Global Warming from Big Geographical Data

Chandan Kanungo Wei Zhang Yizhou Shen
Department of Electrical Engineering
Columbia University
New York, NY
ck2749@columbia.edu wz2363@columbia.edu ys2804@columbia.edu

Abstract—**There are several existing global that have been developed in the past decade, including the one from United States Environmental Protection Agency. Some of them are trying to vision the issue of global warming, whereas the others are trying to relieve the impact of global warming. Thus, in this project, we develop an unbiased yet innovative model to predict the impact of global warming. The datasets used in this project are all from official government resources with no bias on the formation of data. We are trying different models on the data so as to get a better fit to predict the future. As a result, we claim that the impact of global warming will be severe to our planet in the future, specially to the coastline cities. At the bottom line, we should at least start paying attention on the issue from now on so as to prevent our planet from devastating damage.**

Keywords-global warming; big data; prediction; sea level trend

I. INTRODUCTION

The issue of global warming has been a controversial topic over the past decades. Some of researchers provide with some probabilistic model and claim that the impact of global warming is severely devastating for human beings and is coming sooner than most researchers have concluded. On the other hand, some of expert have claimed that the impact of global warming is neglectable and should not be paid too much attention on. Some of them even cast political ink on this topic. For example, President Trump even claimed it is a made-in-China topic.

Other than those influential judgements, the impact of global warming is still a meaningful topic not only for governmental purpose but also for people in general. It is known that no one can predict future. However, we can take advantage of probabilistic model to predict future as realistic as possible. The impact of global warming is the one of interest since it can truly be devastating to our planet. For example, 40% of the population in Netherlands are exposed to the risk of drowning. Growing sea level resulted from global warming can lead to submerging city's land like Manhattan. Those things can possibly have impact on everyone's life.

As a result, we plan to conduct a comprehensive study on the issue of global warming. We assume that the global temperature can be the reflector of global warming effect and thus result in the increase of sea level, which demonstrate our project flow in a logical way. In order to conduct an unbiased study, we try our best to obtain the most impartial data from governmental sources. After this stage, we use the method of cross-validation to fine-tune our model parameter so as to achieve the best performance. We used evidence from big geographical data and evaluated the impact of global warming. We predicted the global temperature and the resulting sea level trend around the whole world.

In addition, there are two factors that a probabilistic model usually can't take into account, which is the effect of economic growth and natural resources. To take those two into account as well, we make separated analytic models to represent the impact from economic growth and natural source factor. It should be noticed that economic growth does not come as a linear way in the past hundreds of years, which is more like an exponential model. The factor of natural source behaves pretty much the same way since the exploitation of natural sources are getting dramatically denser in the past ten years than the ten years prior to the past ten years. Eventually, we believe that with those factors taken into account, our model should be sophisticated enough to make a truthful prediction in genera.

II. RELATED WORKS

Let's first begin exploration of the background work involved which led to the development of the current project work. First we decided upon considering a unique dataset that is not commonly employed by people considering the Big Data Analytics project because this would help us learn the Big Data Analytics in a way we explored it more out of a unique way of our innate understanding rather than getting influenced by the huge amount of information available over the internet. We figure out climate change as our topic. because recent news President elect and other great leaders in United States have claimed that climate change is a made up issue, so it grew our interest in exploring this particular topic that whether it is really a made up issue or a real pressing issue. We looked up at some National Oceanography websites and seen that they provide data regarding that. We did find really valuable resources that helped us understand the data that we used in the project. Especially there are lot of tools available to explore this fascinating area and we learn more about it. This was just the tip of the iceberg as we got the data but no way how to

use this immense data to produce some useful conclusions. We took quite some time in understanding the data and the respective websites along with some models descriptions from Spark official tutorials helped us understanding how to produce useful models out of the immense data, but the way we used the data to produce the model and the analysis was quite original and innovative, so we were able to finally start using the data to predict the temperature but now the problem came once we start understanding the sea level data because earlier when we tried to plug in the data to the model it shows incorrect results, so we further analyzed sea level results and as the sea level data clearly indicates the past results that there is a decrease in sea level above latitude 57 degrees north and increase below latitude 57 degrees north therefore we considered this into account our model and it produces useful results.

Also, the documentation of spark was proved to be helpful in understanding the regression model. So we started looking into various model and for the cause of simplicity due to time constrained looked into linear regression as analysis from part results clearly showed somewhat linear relation. This background analysis helped us understand the data related to the climate and sea level and regression in more detail and has played a pivotal role in materializing the interesting project work that we have produced.

III. Dataset and System Overview

A. Dataset Overview

There are three main datasets we are using: The climatology dataset and the sea level trends dataset from National Oceanic and Atmospheric Administration, and the global surface temperature dataset.

In the climatology dataset, we have temperature of oceans in different depth and different regions around the world including: The Northern North Pacific (NNP) Region, The Northwest Atlantic (NWA) Region, The Greenland Region, The Iceland and Norwegian Seas (GINS) Region, The Arctic Region, The East Asian Seas Region, and The Gulf of Mexico (GoM) Region. The example visualizations are shown below as provided by NOAA.

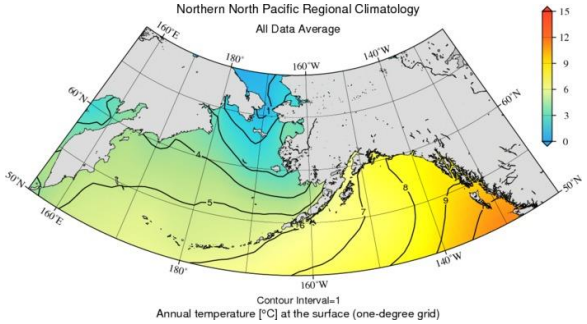

Figure. 1. NNP Region Temperature

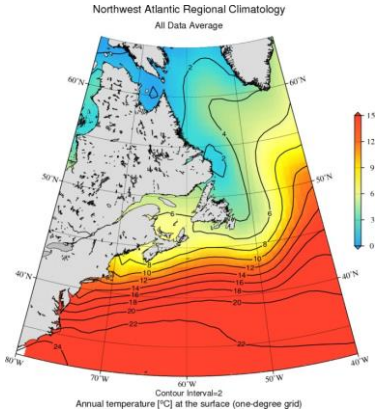

Figure. 2. NWA Region Temperature

The global surface temperature dataset contains global temperature anomaly from 1880 to 2016 in every regular 2x2 degree grid on the world map. We test the raw dataset via ncBrowser as shown below. Red shows a positive temperature anomaly and blue shows negative anomaly. We used ncBrowser to plot the temperature anomaly during 2016.9 and 2016.10 as shown below.

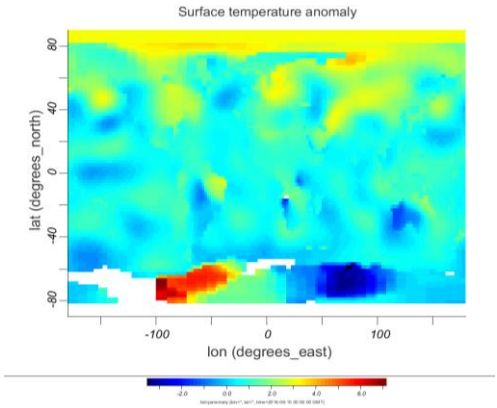

Figure. 3. Global Temperature Anomaly in 2016.9

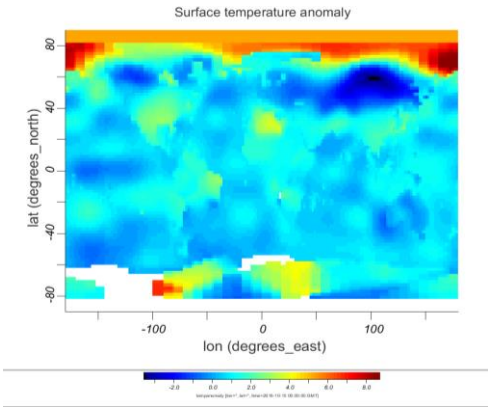

Figure. 4. Global Temperature Anomaly in 2016.10

We also use MATLAB to preview part of our dataset. We can see that in the below figure, yellow stands for positive temperature anomaly while blue stands for negative temperature anomaly. The top layer is the most recent year and the bottom layer is temperature of 1880. Each layer represents temperature around the world. There seems to be a very slow trend that temperature around the world is going up.

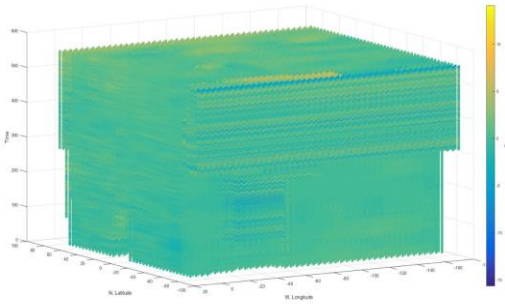

Figure. 5. Global Temperature Anomaly Trend

The sea level trends dataset includes 1485 stations' sea level trends around the world from 1800s to 2000s. By just examining the annual sea level trend around the world, we found that the sea level above latitude 57 degrees north goes down annually, while the sea level below latitude 57 degrees north goes up annually as shown in Figure 6 provided by NOAA.

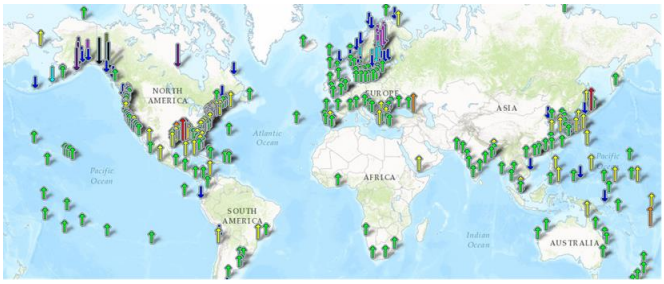

Figure. 6. Global Sea Level Trend

So we separated the world map into two regions: the "upper region" which is above latitude 57 degrees north and the "lower region" which is below latitude 57 degrees north. We claimed and later proved that while the temperature goes up globally, the upper region's sea level will go down and the lower region's sea level will go up.

B. System Block Diagram

The block diagram below displays the flow of our project. So we input Global Surface Temp file to MATLAB and convert the data to csv and then the output of the MATLAB code is input to the model that we developed in Scala using Linear Regression with Stochastic Gradient Descent. The first model that we used comprises of 3 features: time, latitude, longitude. It generated model for predicting temperature over the future. The second model comprises of 2 features Time and Temperature below and above latitude 57 degrees north and along with uses sea level data in rlr format and thus generates the sea level prediction model which predicts the sea level values over the future. By inputting future time to both the models the values of future temperature and sea level can be determined.

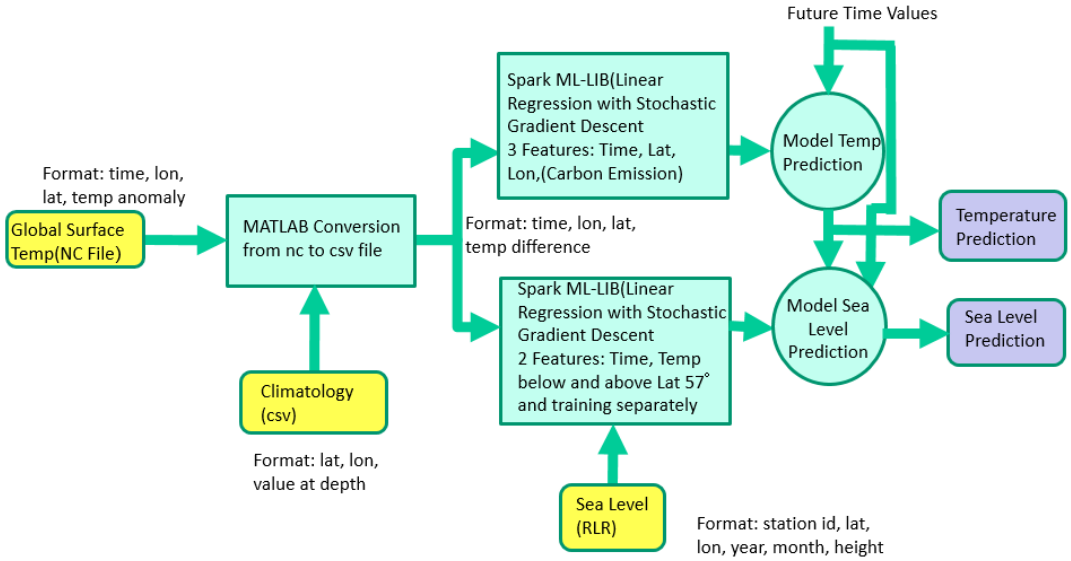

Figure. 7. Block Diagram of our System

IV. ALGORITHM

A. *Data Pre-processing*

We used Python/Matlab script to handle NetCDF and RLR file and pre-process data. Python code is used to create such a special .csv file that does not have a matrix form. That is, the number of columns varies among each rows. In this case, Matlab does not work as efficient as expected while Python plays the role.

We have totally four data sources, which can be categorized into two divisions. First, Climatology Temperature and Global Temperature dataset coming from NOAA and NASA, respectively, provide us with all the information of temperature anomaly. The two datasets from NOAA and PSMSL about sea level not only enable us to have the information of sea level but also provide us with the linking between sea level and temperature anomaly data, which is the critical relationships that we plan to build a model on.

We merge into two datasets from four data sources so as to prepare for the future data pre-processing stage. We do not expect we have clean data by this stage but do expect this helps data pre-processing in the next stage runs smoothly.
We used the concepts of cross-join and self-join to deal with each table of data.

We primarily utilized Matlab to pre-process data in order to feed into our model. This process can be divided into two parts. First of all, in order to retain the global temperature data, Matlab is used to read .nc file. The .nc file is such a file that one can easily visualize but hardly be used for modeling. Four tables contained in the .nc file are useful for our modeling, timestamps, longitude, latitude and temperature anomaly. The timestamps table are timed from year 1880, lasting 135 years with one month as a step. And longitude and latitude tables are just across our planet with one-degree increment. Last, the temperature anomaly table are a 4-dimension table with longitude, latitude, timestamp and temperature anomaly. To be clear on the table dimensions, a table containing dimension specification is created below to demonstrate our data scale.

	TIme	Longitude	Latitude	Temperature Anomaly
Dimensions	1640×1	180×1	90×1	1640×180×90

Table. 1. Input csv data file format for temperature

After extracting each table from .nc file, our next step is to join each table from the above into one table. That is, for each location, varying by time, we have a temperature anomaly, indicating how different this year's temperature behaves compared with a baseline. In this task, there are two challenging tasks. The first one is apparently the size of data, which is quite large. Since the unzipped data size comes around 2 GB, the processing time of making this table is approximately 5 hours. In this period, memory overflow issues may sometimes happen. We solve them by controlling the usage of our heaps in order to avoid memory overflow issue. The second challenge is about the data in the timestamps table. It is originally recorded as the span of days from year 1800. For example, each data point in timestamps table represents how many days have elapsed since the year of 1800. We need to carefully handle this since there are leap years in our case. Correctly converting them to the format of yyyy-mm-dd is a fundamental basis for our model to work properly.

The above steps conclude the pre-processing stage for the training data to our temperature prediction model. Our next step is to build the relationships between temperature anomaly and sea level over the time.

Our sea level data is also different from temperature anomaly data. The temperature anomaly data is coming out globally. However, this is not the case for sea level, which is reasonable, indicating the number of stations capturing sea level data is far less than the number of locations capturing temperature anomaly data. The comparison between temperature anomaly data and sea level data can be found in the table below.

	Sea Level	Temperature Anomaly
# of location	1024	180×90
Time stamps	135 years since 1880 with one month as a step	135 years since 1880 with one month as a step

Table. 2. Input csv data file format for sea level

As we can see from the table above, we have much less locations of sea level than those of temperature anomaly. It should be known that prediction of sea level should be based on the temperature around that location instead of taking far-away lands into account. This results in extra work on pre-processing our data. That is, we need to match each location in sea level dataset to temperature anomaly dataset based on their corresponding longitude and latitude. In this step, we use the formula in Matlab code:

```
Target_long = round(long_sea_level/2)*2+1
Target_lat = round(lat_sea_level/2)*2+1
```

where target_long is the longitude that we can find in temperature anomaly dataset and long_sea_level is what we have in the longitude in the sea level dataset, so do target_lat and lat_sea_level.

As a result, we have been able to form a new table where each longitude and latitude pair should have its corresponding temperature anomaly over time and its matching sea level over time as well. At this point, we have completed the data pre-processing step, which is an essential step for building our models in next stage.

B. Linear Regression Algorithm

We then used Spark/Scala and ML-Lib to process data and produce models. There are three models that we built. The first model is to predict future temperature anomaly based on three input features: the timestamp, the longitude, and the latitude. Since we have found that the sea level trend above latitude 57 degrees north is slowly going down annually and the sea level trend below latitude 57 degrees north is slowly going up annually, so we separated the planet into two regions: the upper region and the lower region as mentioned in the previous part. Therefore, we treated these two regions separately, and thus, we built two different models to predict sea level for these two regions. So the second model is to predict sea level above latitude 57 degrees north based on two input features: the timestamp and the temperature anomaly, and the third model is similar to the second model. It is used to predict the sea level below latitude 57 degrees north also based on the timestamp and the temperature anomaly.

For the three models, we all used the regression algorithm to predict future values. Since we found that the temperature anomaly and the sea level changes in a linear trend. We applied linear regression using Spark/Scala ML-Lib. Linear least squares is the most common formulation to solve regression problems while there are also other ways to do the regression such as Lasso and ridge regression. We chose linear least squares since we do not have large number of input variables. The mathematical formulation of linear regression is shown below:

$$f(\mathbf{w}) := \lambda\, R(\mathbf{w}) + \frac{1}{n} \sum_{i=1}^{n} L(\mathbf{w}; \mathbf{x}_i, y_i)\ .$$

With loss function:

$$L(\mathbf{w}; \mathbf{x}, y) := \frac{1}{2}(\mathbf{w}^T \mathbf{x} - y)^2.$$

Here, the first function is the objective function since we need to find the minimizer of a convex function f that depends on variable w, which is weight that has d entries. x is the training data examples in d dimension for 1<=i<=n, and y is the corresponding label to predict. The loss function L can be expressed into a function of y and $(w^T)x$ as shown above, so the method is linear. The mean square error shows the average loss or training error and it can be calculated by:

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^T x_i - y_i)^2,$$

The coefficient of determination, R squared value, can be calculated by:

$$R^2 = 1 - \frac{MSE}{\mathrm{VAR}(\mathbf{y})\cdot(N-1)}$$

The coefficient of determination is a key output of linear regression analysis. It is the square of the correlation between predicted y scores and actual y scores. Ranging from 0 to 1, it shows the percentage of the variance in y that is predictable from x. While building our models, we also calculated the MSE and the R squared value to evaluate the performance of our model.

C. Models Overview

For the first model, we input our pre-processed dataset in csv format into our linear regression algorithm in spark-Scala. The first three columns in the csv dataset file are the three features that we input are the timestamp, the latitude location, and the longitude location. The fourth column in the csv file is the temperature anomaly corresponding to a certain location at a certain time. We used spark ML-Lib and regression method to predict temperature. We used Linear Regression with stochastic gradient descent: LinearRegressionWithSGD to build a linear model to predict label value temperature anomaly based on timestamp, latitude, and longitude location. The following is the output of the model file, which shows three coefficients that corresponding to timestamp, latitude and longitude.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<PMML xmlns="http://www.dmg.org/PMML-4_2" version="4.2">
   <Header description="linear regression">
      <Application name="Apache Spark MLlib" version="2.0.1"/>
      <Timestamp>2016-12-14T16:13:17</Timestamp>
   </Header>
   <DataDictionary numberOfFields="4">
      <DataField name="field_0" optype="continuous" dataType="double"/>
      <DataField name="field_1" optype="continuous" dataType="double"/>
      <DataField name="field_2" optype="continuous" dataType="double"/>
      <DataField name="target" optype="continuous" dataType="double"/>
   </DataDictionary>
   <RegressionModel modelName="linear regression" functionName="regression">
      <MiningSchema>
         <MiningField name="field_0" usageType="active"/>
         <MiningField name="field_1" usageType="active"/>
         <MiningField name="field_2" usageType="active"/>
         <MiningField name="target" usageType="target"/>
      </MiningSchema>
      <RegressionTable intercept="0.0">
         <NumericPredictor name="field_0" coefficient="0.006913942000254928"/>
         <NumericPredictor name="field_1" coefficient="6.525776868104444E-5"/>
         <NumericPredictor name="field_2" coefficient="7.489349294553797E-4"/>
      </RegressionTable>
   </RegressionModel>
</PMML>
```

We also tested this model to predict the temperature anomaly at year 1980 and location longitude 177 and latitude 77 by the following code (note that 100 means year 1880+100 = year 1980, 400 means 1880+400 = year 2280):

```
// Test model on training data
predictedValue = model.predict(Vectors.dense(100,177,77))
```

```
// Predict temperature anomaly at random time and location
predictedValue = model.predict(Vectors.dense(400,144,44))
```

```
MSE = 0.42943047935729134
RMSE = 0.6553094531267585
predictedValue: Double = 1.2606128146500999
predictedValue: Double = 3.307927055688089
```

Figure. 8. First Model's Performance and Prediction

As we can see in the above picture, using our model, we first test our model on the training data. We found that the temperature anomaly is 1.26 in year 1980 at longitude 177 degree east, and latitude 77 degrees north. The actual value of the year average temperature anomaly is 0.92, which is only 0.34 difference with our model output. We also predicted that the temperature anomaly is 3.31 in year 2280 at longitude 144 degree east, and latitude 44 degrees north as shown in the picture.

For the second and third models to predict sea level above and below latitude 57 degrees north, we applied the similar linear regression method. However, here, we used average temperature in each region instead of different temperatures in different locations since the difference of the sea level at difference places is very small in upper or lower region, and we considered the sea level trend globally. To predict the sea level above latitude 57 degrees north, we first input the average annual temperature anomaly in the upper region and the corresponding timestamp into spark. Then, in a similar linear regression method as used to predict temperature anomaly, a model file will be generated as shown below:

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<PMML xmlns="http://www.dmg.org/PMML-4_2" version="4.2">
   <Header description="linear regression">
      <Application name="Apache Spark MLlib" version="2.0.1"/>
      <Timestamp>2016-12-14T18:03:02</Timestamp>
   </Header>
   <DataDictionary numberOfFields="3">
      <DataField name="field_0" optype="continuous" dataType="double"/>
      <DataField name="field_1" optype="continuous" dataType="double"/>
      <DataField name="target" optype="continuous" dataType="double"/>
   </DataDictionary>
   <RegressionModel modelName="linear regression" functionName="regression">
      <MiningSchema>
         <MiningField name="field_0" usageType="active"/>
         <MiningField name="field_1" usageType="active"/>
         <MiningField name="target" usageType="target"/>
      </MiningSchema>
      <RegressionTable intercept="0.0">
         <NumericPredictor name="field_0" coefficient="-2.300719060446825"/>
         <NumericPredictor name="field_1" coefficient="-0.006793432845823683"/>
      </RegressionTable>
   </RegressionModel>
</PMML>
```

We can see the two coefficients of features are -2.3 and -0.0068 in the above model file output. We also tested our model by the following code (note that 100 means year 1880+100 = year 1980, 400 means 1880+400 = year 2280. 1.7 and 5 are corresponding temperature anomaly.):

```
// Test model on training data
predictedValue = model.predict(Vectors.dense(100, 1.7))
// Predict sea level by random time and temp
predictedValue = model.predict(Vectors.dense(400, 5))
```

```
MSE = 1911.910044371454
RMSE = 43.72539358738185
R-squared = 0.7979850872027365
predictedValue: Double = -230.08873944030617
predictedValue: Double = -920.3427295821022
```

Figure. 9. Second Model's Performance and Prediction

We can see the results as shown above. We tested our model and got the result that the average sea level above latitude 57 degrees north of year 1980 went down 230mm as compared to that in year 1880. Also, the average sea level above latitude 57 degrees north of year 2280 went down 920mm as compared to year 1880. Here, we can also see the coefficient of determination R squared value is calculated and equals to 0.79 which means a relatively high 80% of variance in the sea level that is predictable from the input temperature and timestamp.

In a similar way, we also performed the same algorithm to predict sea level below latitude 57 degrees north to get the third model. The model file generated is shown below.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<PMML xmlns="http://www.dmg.org/PMML-4_2" version="4.2">
   <Header description="linear regression">
      <Application name="Apache Spark MLlib" version="2.0.1"/>
      <Timestamp>2016-12-14T18:10:30</Timestamp>
   </Header>
   <DataDictionary numberOfFields="3">
      <DataField name="field_0" optype="continuous" dataType="double"/>
      <DataField name="field_1" optype="continuous" dataType="double"/>
      <DataField name="target" optype="continuous" dataType="double"/>
   </DataDictionary>
   <RegressionModel modelName="linear regression" functionName="regression">
      <MiningSchema>
         <MiningField name="field_0" usageType="active"/>
         <MiningField name="field_1" usageType="active"/>
         <MiningField name="target" usageType="target"/>
      </MiningSchema>
      <RegressionTable intercept="0.0">
         <NumericPredictor name="field_0" coefficient="1.680470687144002"/>
         <NumericPredictor name="field_1" coefficient="0.0016035267191445305"/>
      </RegressionTable>
   </RegressionModel>
</PMML>
```

We can see the two coefficients of features are 1.68 and 0.0016 in the model file output. We also tested our model for sea level below latitude 57 degrees north by the following code (note that 100 means year 1880+100 = year 1980, 400 means 1880+400 = year 2280. 0.1 and 5 are corresponding temperature anomaly.):

```
// Test model on training data
predictedValue = model.predict(Vectors.dense(100,0.1))
// Predict sea level by random time and temp
predictedValue = model.predict(Vectors.dense(400,14))
```

```
MSE = 239.2669128290895
RMSE = 15.468255002717324
R-squared = 0.9387854541109344
predictedValue: Double = 168.04722906707212
predictedValue: Double = 672.2107242316689
```

Figure. 10. Third Model's Performance and Prediction

We can see the results as shown in the above picture. As in the second model, we tested the model and got the result that the average sea level below latitude 57 degrees north of year 1980 went up 168mm as compared to that in year 1880. Also, the average sea level below latitude 57 degrees north of year 2280 went up 672mm as compared to year 1880. Here, we can see the coefficient of determination, R squared value, is calculated. It equals to 0.939 which means a very high 93.9% of variance in the sea level below latitude 57 degrees north that is predictable from the input temperature and timestamp.

V. SOFTWARE PACKAGE DESCRIPTION

Based on the model we previously generated, we then using matlab to build a more complicated function that is used to generate future temperature and sea level trendline. In the output of our Matlab Function, x-axis represents the years starting from 1880 as zero to 500 years later, and y-axis represents the temperature anomaly in Celsius. We implement our prediction model under four different considerations. The baseline is an ideal case where we thought the future temperature trend will follow the similar overall trend that happened in the past 135 years. However, it is easy to recognize that this might not be the case since the economic growth that happened in the later 50 years of 135 years had contributed much more than the time before as the factors leading to global warming. This encourages us to develop our second consideration. Nevertheless, this might still be a naive consideration since the factor of natural resources consumption are yet not taken into account. Thus, we use a cross-validation strategy to come up with a parameter that can represent such a factor. Accordingly, we have two more scenarios corresponding to the first two cases. All in all, this resulted in four different scenarios for both below 57 degrees and above 57 degrees.

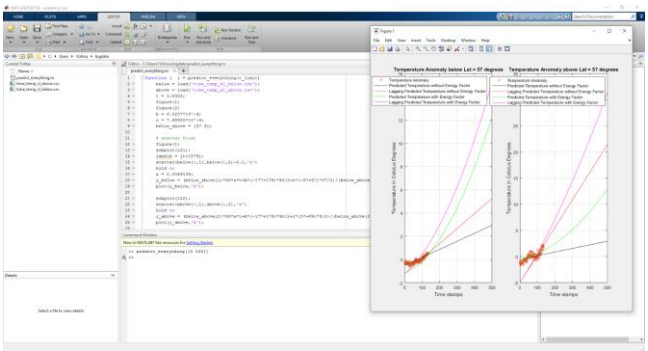

Figure. 11. Our Matlab Funtion and Prediction Output

Similarly, since we predict our sea level variations based on our temperature prediction. Thus, we also have four different considerations in this model. The below and above 57 degrees also matter for sea level prediction as we can see from the plots.

```
>> predict_everything([0, 500])
>> predict_everything([400, 401])
>> predict_everything([0, 1000])
```

By simply typing the year range, as shown above, into our function, both the temperature anomaly and the sea level prediction will be generated. The detailed results will be explained in the next part.

VI. EXPERIMENT RESULTS

In this part, we will explain our experiment results generated by our Matlab function.

A. Temperature Prediction

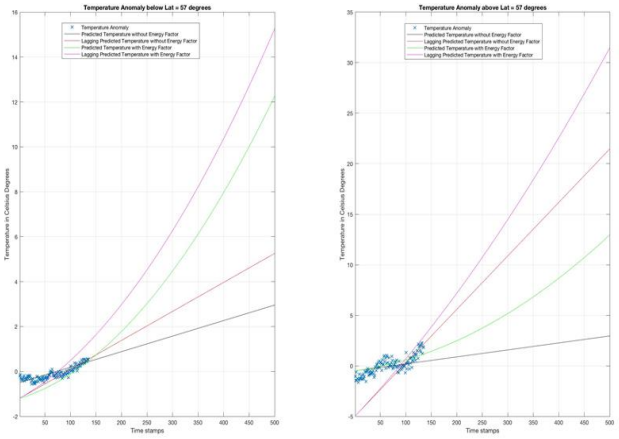

Figure. 12. Our Matlab Output for Temperature Prediction
Left: temperature prediction for latitude less than or equal to 57 degrees under four different scenarios.
Right: temperature prediction for latitude larger than 57 degrees under four different scenarios

The graphs above show the temperature prediction. The left hand side graph shows less than or equal to latitude 57 degrees north whereas the right curve shows the temperature prediction above latitude 57 degrees north and comprises of 4 graphs. The cross marked blue points show the temperature data from 1880-2015.The graph in the black shows the output of the model predicted by the model generated considering all the past years whereas the red curve shows the prediction made by model considering the last 50 years only. Also the green curve is generated by black curve model considering the other energy factors for instance the carbon emission whereas the purple curve is generated by red curve model considering the other energy factors for instance the carbon emission.

In both the cases the temperature is expected to increase over the future years. Over the next 500 years the temperature will rise up to 3-15°C for the latitude below 57 degrees and up to 30°C for latitude above 57 degrees.

B. Sea Level Prediction

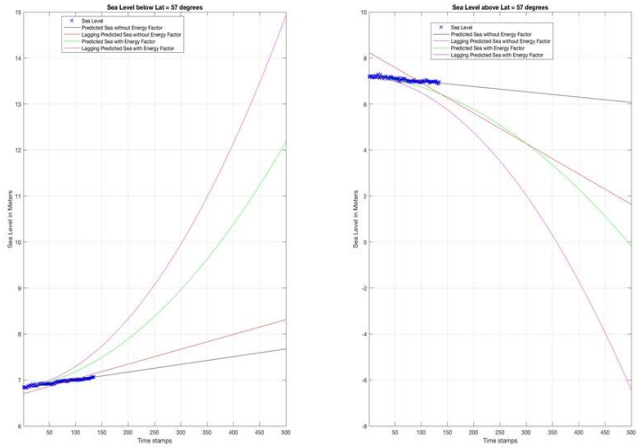

Figure. 13. Our Matlab Output for Sea Level Prediction
Left: sea level prediction for latitude less than or equal to 57 degrees under four different scenarios.
Right: sea level prediction for latitude larger than 57 degrees under four different scenarios

The graphs above show the sea level prediction. The left hand side graph shows less than or equal to latitude 57 degrees north whereas the right curve shows the temperature prediction above latitude 57 degrees north and comprises of 4 graphs similar to the one in the temperature graph. Here we find an interesting result: the sea level below latitude 57 degrees north will be increasing in the future but an interesting observation is that the sea level is going to decrease for latitude greater than latitude 57 degrees north., but as most of the earth and for that matter most of the oceans are below latitude 57 degrees north the sea level is going to increase over the future and this result is of high concern. Over the next 500 years, the temperature and other factors will lead to a 6-13 meters decrease of the sea level where Latitude is greater than 57 degrees and up to 8 meters increase of the sea level in rest of the planet.

VII. CONCLUSION

By analyzing the data and using regression to predict the future values of temperature and sea level, it can be concluded that Mr Trump should reconsider his stance as his statement do not aligns with the scientific data that has been generated by the National Oceanography agencies.

The temperature will show an increase of an average by 3-15°C in the next 500 years for the latitude below 57 degrees north and for latitude above 57, the temperature will rise up to 30°C. Also, the temperature and other factors have led to decrease of the sea level where Latitude is greater than 57 degrees and increase of the sea level in rest of the planet. Over the next 500 years, the temperature and other factors will lead to a 6-13 meters decrease of the sea level where Latitude is greater than 57 degrees north and up to 8 meters increase of the sea level in rest of the planet.

Also, our project is of high commercial as well as moral values. We as humans have great responsibility towards the wellbeing of our planet as some call it “Mother Earth”, and without it we cannot have any future as a species. According to Mr. Stephen Hawking we have just 1000 more years on the planet. Huge commercial values must be associated on this area so that many people start valuing this fragile ecosystem that is still the only known life supporting cocoon in the entire universe.

All the authors contributed their best towards the project and helped each other as it was a team work. Wei was focusing on preprocessing data with MATLAB and nc data, and built a Matlab function that predicts both temperature and sea level. Yizhou was in charge of the algorithm part and also helped a lot for building the Matlab function for prediction. Chandan focused on the data and result analysis, problem statement, and some part of preprocessing with python.

ACKNOWLEDGMENT

The Authors would like to Thank Professor Ching-Yung Lin for this opportunity and offering constant motivation to pursue the project as when we were discussing with him initially he appreciated us for choosing a not so common topic. Also, Eric Johnson always motivated us and provided support in learning the tools.

REFERENCES

[1] "Evaluation metrics - RDD-based API,". [Online]. Available: https://spark.apache.org/docs/latest/mllib-evaluation-metrics.html. Accessed: Dec. 23, 2016.

[2] "Linear methods - RDD-based API,". [Online]. Available: http://spark.apache.org/docs/latest/mllib-linear-methods.html. Accessed: Dec. 23, 2016.

[3] N. N. Centers and E. Information, "Access data," 2015. [Online]. Available: https://www.nodc.noaa.gov/access/index.html. Accessed: Dec. 23, 2016..

[4] "Data at PSMSL,". [Online]. Available: http://www.psmsl.org/data/. Accessed: Dec. 23, 2016.