

# E6893 Big Data Analytics:

## *The Impact of Global Warming from Big Geographical Data*

Team Members: Chandan Kanungo(ck2749)  
Wei Zhang(wz2363)  
Yizhou Shen(ys2840)



December 15, 2016

- Overview
- Technologies Used
- Dataset Overview
- Challenges / Struggles
- Architecture
- Dataset Visualization
- Sea Level Trend Visualization
- Demo
- Findings
- Conclusion
- Next Steps

The issue of Global Warming has been a controversial topic over the past decades. President Trump even claimed it is a made-in-China topic. The impact of global warming 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. 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.



**Donald J. Trump** ✓  
@realDonaldTrump

 Follow

The concept of global warming was created by and for the Chinese in order to make U.S. manufacturing non-competitive.

RETWEETS 104,260  
LIKES 65,675



- Spark/Scala, ML-Lib

We used Spark/Scala and ML-Lib to process data and produce models.

- Python/Matlab

We used Python/Matlab script to handle NetCDF and RLR file and pre-process data. Python code is used to also find the location of the ocean among the entire lat/long data.

- Linear Regression

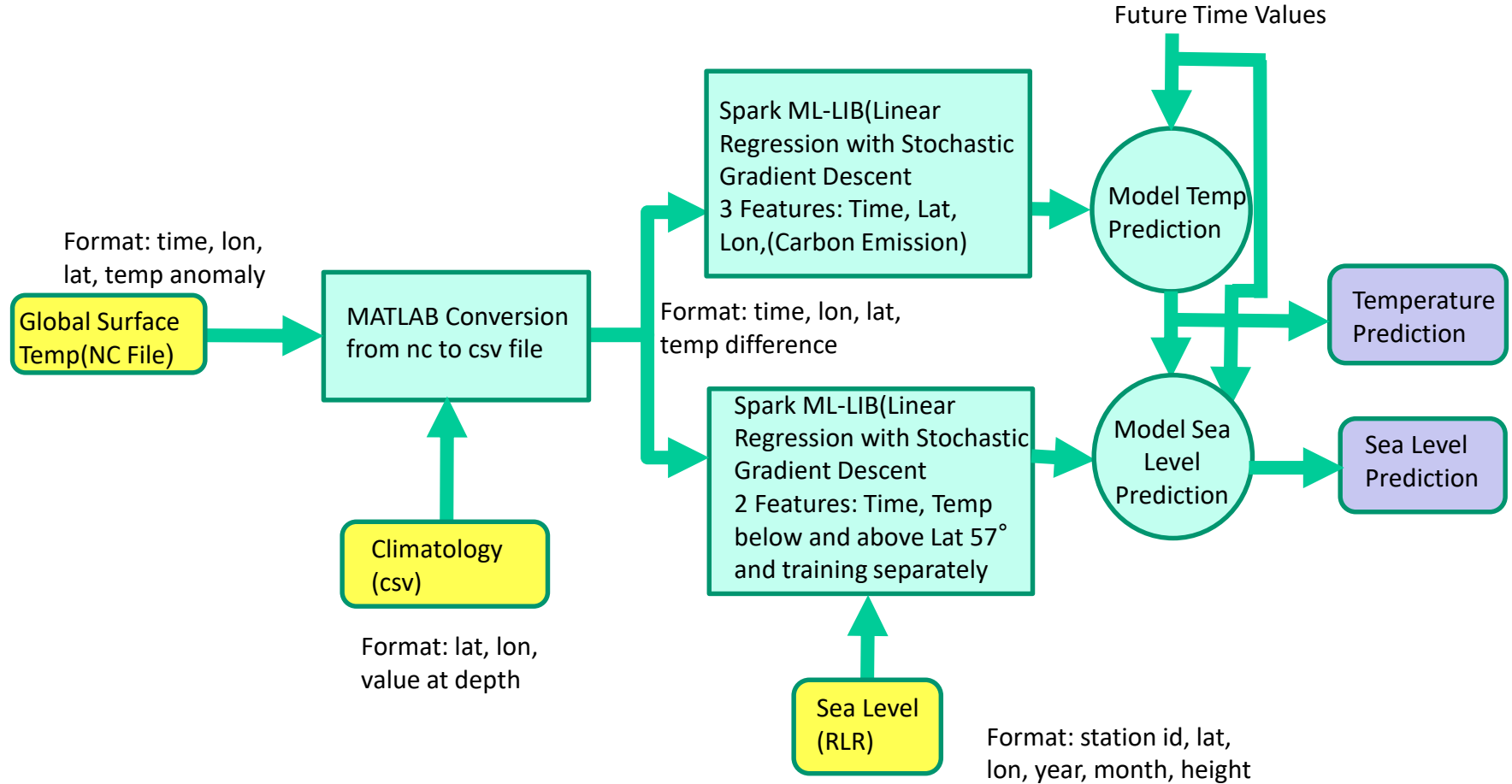
We take the knowledge of linear regression as the foundation of training.

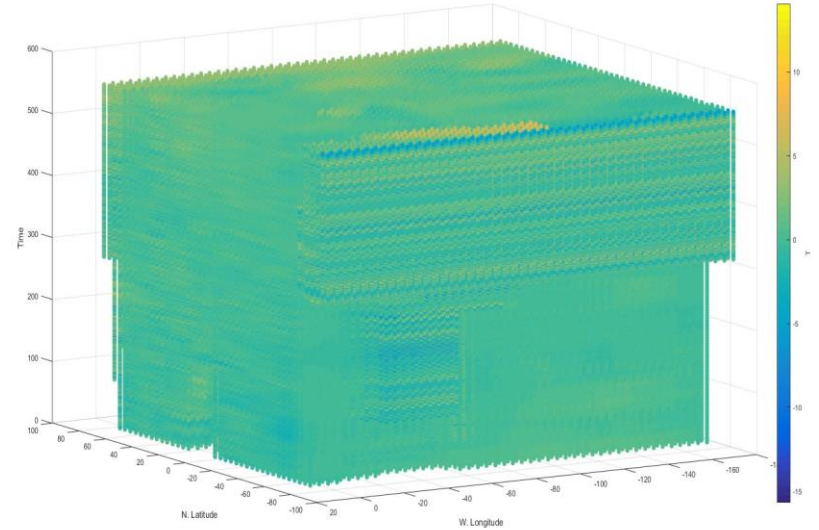
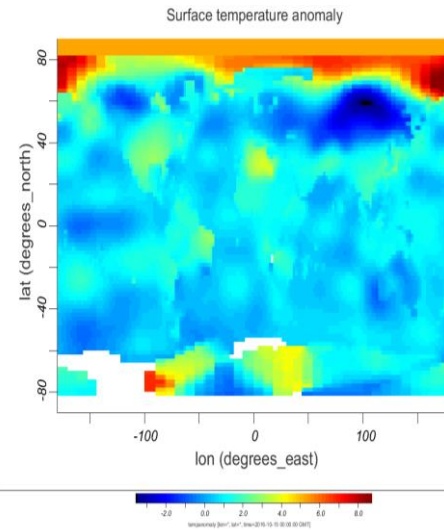
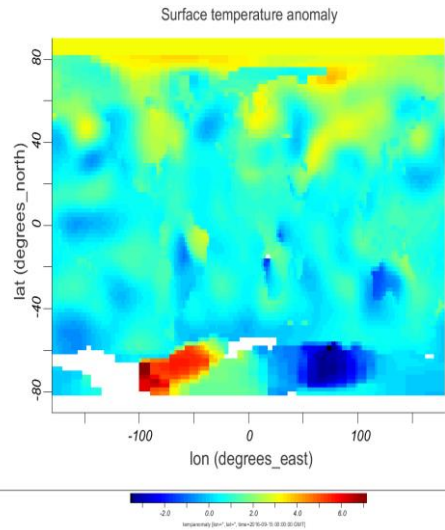
- Database Knowledge

We used the concepts of cross-join and self-join to deal with each table of data.

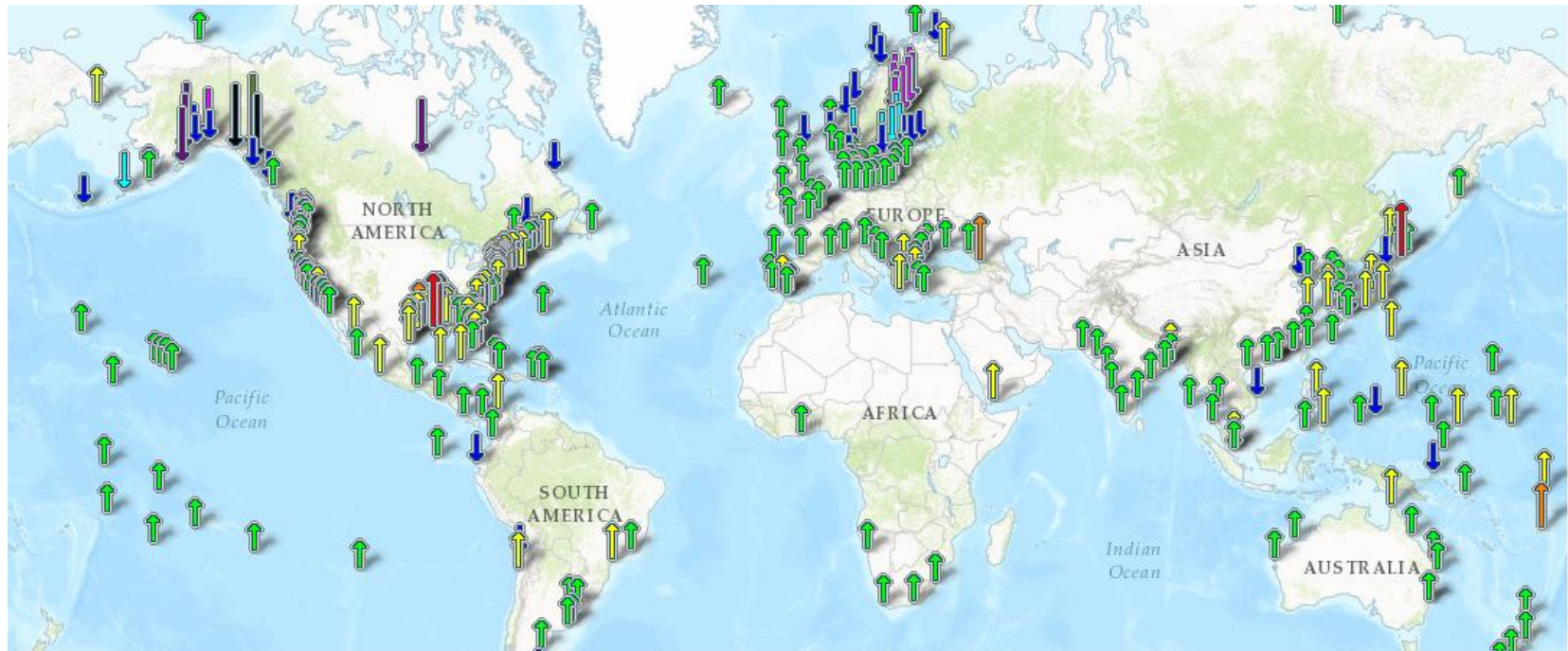
- Climatology Data from NOAA  
<https://www.nodc.noaa.gov/access/index.html>
- Global Surface Temperature Data from NASA  
<http://data.giss.nasa.gov/gistemp/>
- Sea Level Trends Data from NOAA  
<http://tidesandcurrents.noaa.gov/sltrends/sltrends.html>
- Sea Level Trends Data From PSMSL  
<http://www.psmsl.org/data/obtaining/>

- Alignment of all dataset from different source. (Based on GPS location)
- Dealing with NetCDF file. Easy to visualize but hard to process.
- Prediction of temperature and regional relation between temperature and sea level.
- The size of data introduces some memory issues.









ubuntu [Running] - Oracle VM VirtualBox

File Machine View Input Devices Help

LibreOffice File Edit View Insert Format Sheet Data Tools Window Help

Home spark-project src main resources exported\_pmmml\_models

test linearregression.xml linearregression\_above.xml linearregression\_below.xml

```
sheny@sheny-VirtualBox: ~/spark-project/src/main/resources/spark_shell_exporter
853a4c
MSE = 239.2669128290895
RMSE = 15.468255002717324
R-squared = 0.9387854541109344
predictedValue: Double = 168.04722906707212
predictedValue: Double = 672.2107242316689

scala> :load /home/sheny@sheny-VirtualBox:~/spark-project/src/main/resources/spark_shell_exporter/linearregression_sl_above.scala
Loading /home/sheny@sheny-VirtualBox:~/spark-project/src/main/resources/spark_shell_exporter/linearregression_sl_above.scala...
import org.apache.spark.mllib.regression.LinearRegressionWithSGD
import org.apache.spark.mllib.regression.LabeledPoint
import org.apache.spark.mllib.linalg.Vectors
import org.apache.spark.mllib.evaluation.RegresstionMetrics
import org.apache.spark.mllib.regression.{LabeledPoint, LinearRegressionWithSGD}
import org.apache.spark.rdd._
data: org.apache.spark.rdd.RDD[String] = ../datasets/time_temp_sl_above.csv MapPartitionsRDD[140] at textFile at <console>:112
parsedData: org.apache.spark.rdd.RDD[org.apache.spark.mllib.regression.LabeledPoint] = MapPartitionsRDD[141] at map at <console>:114
warning: there was one deprecation warning; re-run with -deprecation for details
regression: org.apache.spark.mllib.regression.LinearRegressionWithSGD = org.apache.spark.mllib.regression.LinearRegressionWithSGD@128ce1ad
res41: org.apache.spark.mllib.optimization.GradientDescent = org.apache.spark.mllib.optimization.GradientDescent@7d779fc2
16/12/14 21:10:01 WARN LinearRegressionWithSGD: The input data is not directly cached, which may hurt performance if its parent RDDs are also uncached.
16/12/14 21:10:02 WARN LinearRegressionWithSGD: The input data was not directly cached, which may hurt performance if its parent RDDs are also uncached.
model: org.apache.spark.mllib.regression.LinearRegressionModel = org.apache.spark.mllib.regression.LinearRegressionModel@128ce1ad
del: intercept = 0.0, numFeatures = 2
valuesAndPreds: org.apache.spark.rdd.RDD[(Double, Double)] = MapPartitionsRDD[180] at map at <console>:120
metrics: org.apache.spark.mllib.evaluation.RegresstionMetrics = org.apache.spark.mllib.evaluation.RegresstionMetrics@282dc952
MSE = 1911.910044371454
RMSE = 43.72539358738185
R-squared = 0.7979850872027365
predictedValue: Double = -460.1659300447744
predictedValue: Double = -690.2655389776339

scala>
```

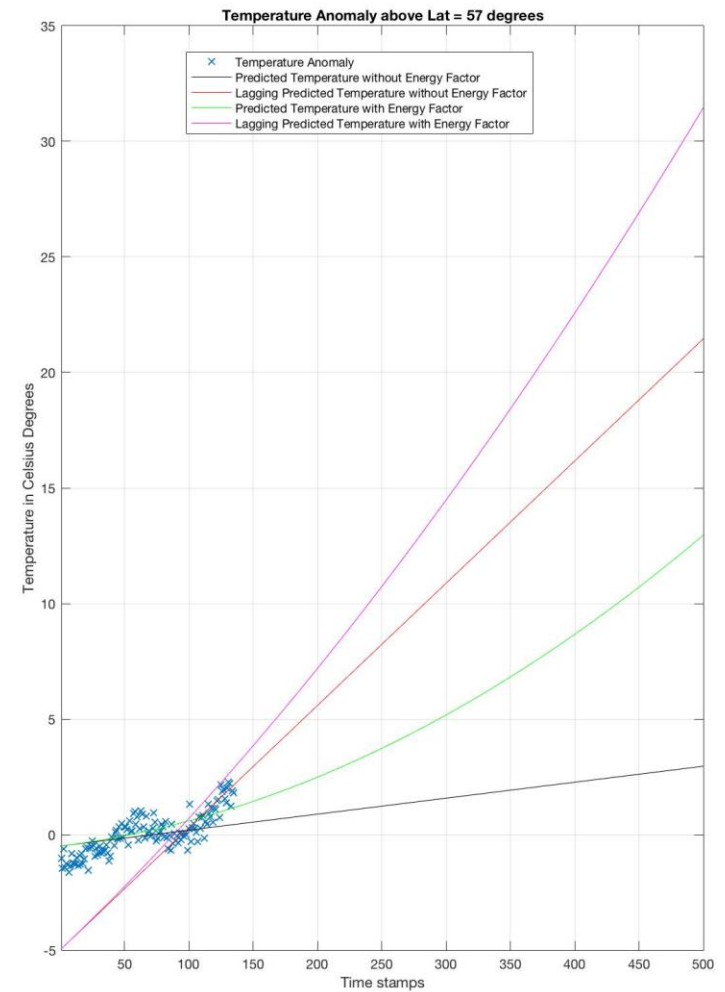
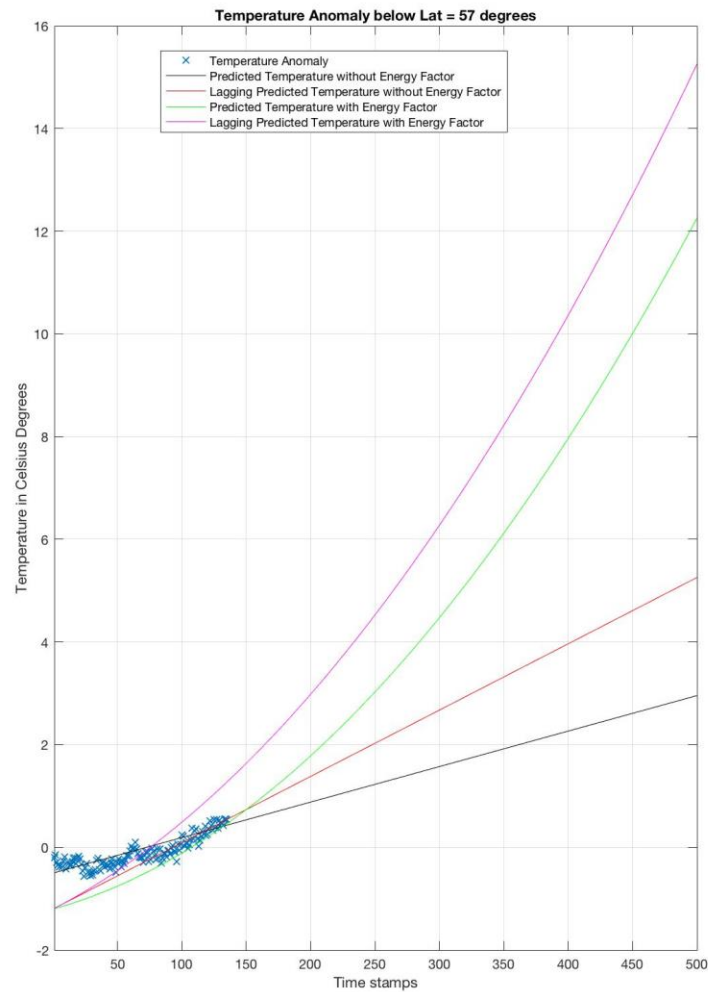
"test" selected (containing 10 items)

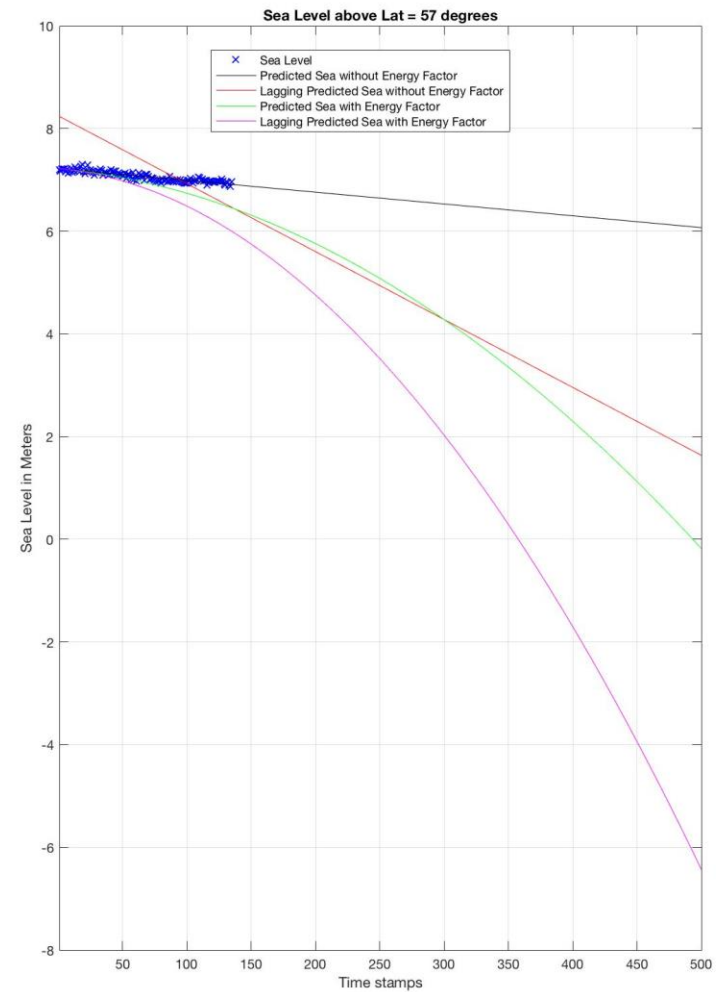
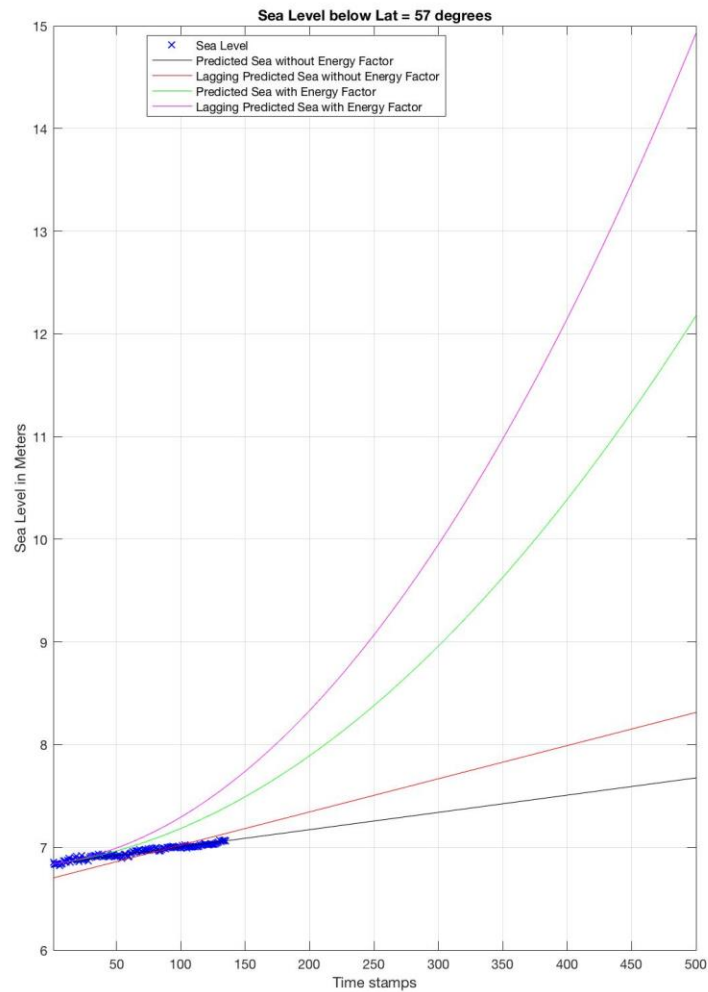
time\_lat\_lon\_tempavg2.csv - LibreOffice Calc

	A	B	C	D	E	F	G
1	0	-177	-67	0.58	S		
2	0	-177	-63	0.564444			
3	0	-177	-59	0.63			
4	0	-177	-55	0.62			
5	0	-177	-51	0.65667			
6	0	-177	-47	0.77333			
7	0	-177	-43	0.89583			
8	0	-177	-39	0.87917			
9	0	-177	-35	0.85917			
10	0	-177	-31	0.87083			
11	0	-177	-27	0.88			
12	0	-177	-23	0.85333			
13	0	-177	-19	0.7625			
14	0	-177	-15	0.72			
15	0	-177	-11	0.74667			
16	0	-177	-7	0.82583			
17	0	-177	-3	0.83833			
18	0	-177	1	0.825			
19	0	-177	5	0.515833			
20	0	-177	9	0.20333			
21	0	-177	13	-0.07167			
22	0	-177	17	-0.11667			
23	0	-177	21	-0.05167			
24	0	-177	25	0.0825			
25	0	-177	29	0.17333			
26	0	-177	33	0.25833			
27	0	-177	37	0.34417			
28	0	-177	41	0.38917			
29	0	-177	45	0.30417			
30	0	-177	49	0.21667			
31	0	-177	53	0.035			
32	0	-177	57	-0.13583			
33	0	-177	61	-0.7429			
34	0	-177	65	-1.134			
35	0	-173	-67	0.58			
36	0	-173	-63	0.66375			
37	0	-173	-59	0.66917			
38	0	-173	-55	0.62			
39	0	-173	-51	0.6125			
40	0	-173	-47	0.7			
41	0	-173	-43	0.79			

Sheet 1 of 1

# Findings—Temperature Anomaly





- By analyzing the data and using regression to predict the future values of temperature and sea level, it can be concluded that Mr President Elect 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 and for latitude above 57, the temperature will rise up to 30° C. Also, the temperature and other factors have lead to decrease of the sea level where Latitude is greater than 57° 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 2-8 meters decrease of the sea level where Latitude is greater than 57° C and up to 15 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.

- The impact from iceberg melt is not completely taken into account but should be.
- The consumption of natural resource should also be taken into account.
- Thus, the model should be more complex than what it is now.
  - It is possible a more complicated non-linear regression
  - Inference technique can be more sophisticated such as using variational inference
  - Model parameter can be fine-tuned better such as using cross-validation
- Datasets should be more comprehensive since currently we still miss some data due to accessibility issue.

# Thank You