Global Warming Analysis And Prediction

Aaryan Sharma
SRN:PES1UG20CS003
CSE Department
PES University
Bengaluru , Karnataka
eaaryansharma@gmail.com

Aditya N
SRN:PES1UG20CS021
CSE Department
PES University
Bengaluru, Karnataka
adinp2002@gmail.com

Abhay K Iyengar
SRN:PES1UG20CS004
CSE Department
PES University
Bengaluru , Karnataka
abzee2002@gmail.com

***Abstract*—In this report, we analyze global warming over a period of 185 years starting from 1825. By comparing the approaches taken over the course of the research, we can make predictions. In analyzing the problem statement, we have taken into account a variety of perspectives, including journal articles. We have performed all of the exploratory data analysis tasks that we set out to do. We have also included a select few of the visualizations that we performed, to gain a deeper understanding of certain features of the dataset better. To predict global warming patterns in the future, we have created a machine learning model using machine learning techniques.**

Keywords—forecasting, prediction, diagnostic analysis, autoregressive, seasonality, rigid-regression, long short-term memory (LSTM).

I. INTRODUCTION

The forecast of long-term global warming and weather conditions could be of huge significance in various fields, such as climate research, farming, electricity, medicine, and many more. The Global temperature reduction will benefit the entire globe because not only humans but also various animals suffer from global warming. In this regard, different techniques can be applied to evaluate global warming dynamics. Through the use of this type of analysis, one can make better predictions and gain a deeper understanding of the phenomenon.

There's a huge amount of data cleaning and preparation that goes into putting together a long-time study of climate trends. This analysis requires a long interval of historic data to capture various patterns in the data such as seasonality, trend, etc. We use a diagnostic analytics approach to analyse the environmental impact and a machine learning approach to forecast global temperatures.

The dataset which we are working on in this project is taken from Berkeley earth. Early data were collected by technicians using mercury thermometers, where any variation in the visit time impacted measurements. The Berkeley Earth Surface Temperature Study combines 1.6 billion temperature reports from 16 pre-existing archives. It is nicely packaged and allows for slicing into interesting subsets (for example by city). The goal is to analyse the trend and seasonality rise in global land and ocean temperatures and predict the global land and ocean temperatures for various cities around the world.

II. RELATED WORKS

In **[1]** The author deals with machine learning techniques for Global temperature prediction and analysis. Machine learning models tend to generalize better with shorter historic data compared to time series approaches. It is also claimed that regression approaches provide better results compared to time series methods as complicated patterns can be modelled. The author tries to best fit the data with a set of data points with a line and for this, the author uses linear ridge regression and polynomial ridge regression. for implementing this model, the equation used is:

1/n * || y - Xw ||^2 + alpha * || w ||^2

The author tries to model the data using different alpha values for better prediction. After this, he compares linear ridge regression with polynomial ridge regression and concludes that linear ridge regression gives better accuracy than polynomial ridge regression.

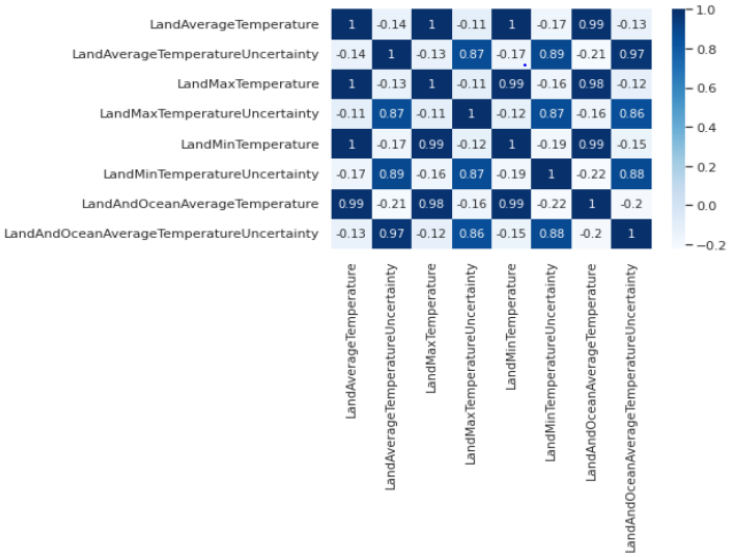
In **[2]**, The author predicts the Average Land Temperatures by predicting the values using the Random Forest model and c forest. Random forests (random decision forests) are an ensemble learning method for classification, regression, and other tasks that operate by constructing a multitude of decision trees.
The author does this by performing the 70:30 split on the dataset (70% for training and 30% for testing). To forecast errors is used the Root Mean Squared Errors (RMSE) as a measure of accuracy. The author has also used another model called C forest to predict the values of the Land's Average Temperatures. C forest is

another ensemble model giving conditional inference trees and comes with a 'party' package. Lastly, the author then compared the accuracy of both these models and concluded that random forest gives better accuracy than C forest.

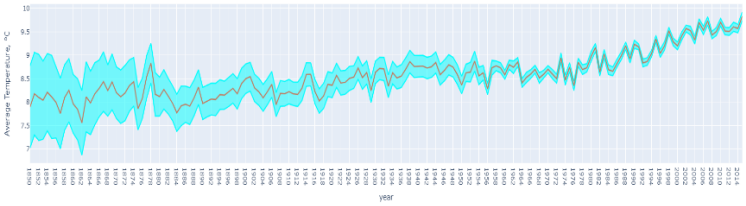
In **[3]**, The author has used the Long Short Term Memory model to predict the Land Average Temperatures. The LSTM model here learns a function that maps a sequence of past observations as input to an output observation. The author then transforms the sequence of observations into multiple examples from which the LSTM can learn. Then The accuracy here was calculated using RMSE.

III. Data cleaning, EDA and Vizualizations

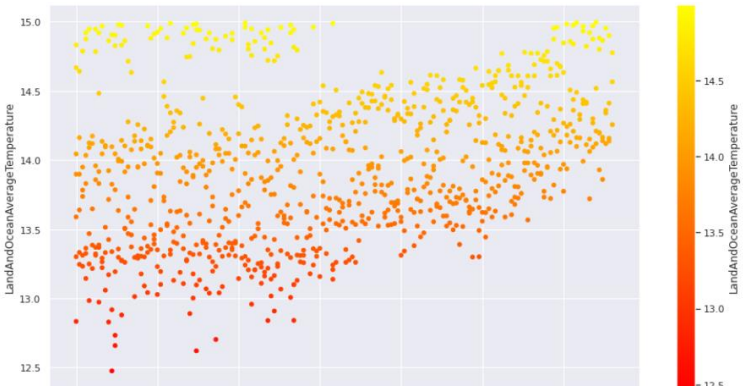
- We have performed the analysis and visualization of the global warming data for three datasets which were from the years 1825 to 2010.
- The data was initially available as three separate files, after which the two files were combined for EDA. The two files combined had data related to the temperatures of cities and data related to countries and their regions.
- There total of 3192 entries in the Global Temperatures dataset, for 9 different attributes.
- We have analysed the global warming trends in Rio de Janeiro from 1825 to 2000. There is a total of 2181 entries of data present for 7 different attributes.
- Then, we identified the missing values in each column of our dataset and dropped those rows that had missing values.
- We also found the number of outliers and checked for duplicate values. We noticed a few outliers in Land Ocean Average Temperature Uncertainty, Land Minimum Temperature uncertainty, Average Temperature, etc.
- To summarise the data, we calculated various summary statistics, such as correlation, heat maps, pair plots, and histograms.



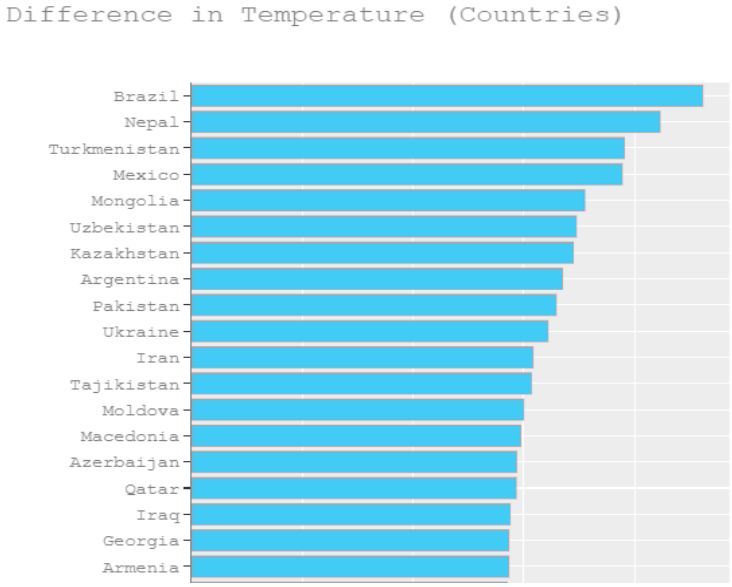
- We then visualize the Land's Average Temperature and its error through the years.



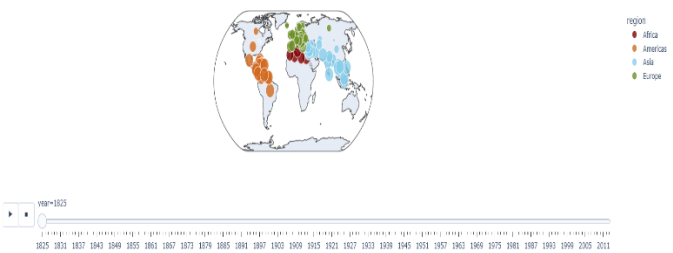
- Scatterplots have been used to analyse the average Land and Ocean Average Temperatures against the years for a certain range of degrees (above or below a threshold).



- We grouped the data based on temperature increase and ranked them based on it.

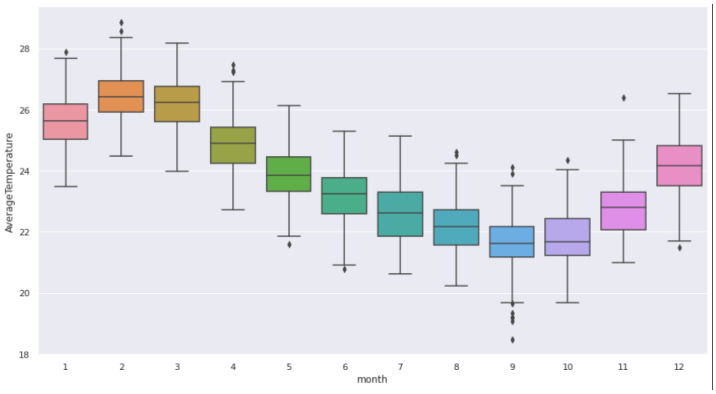


- An interactive globe map has been created using Plotly to show the rise in temperatures over the years across the globe.

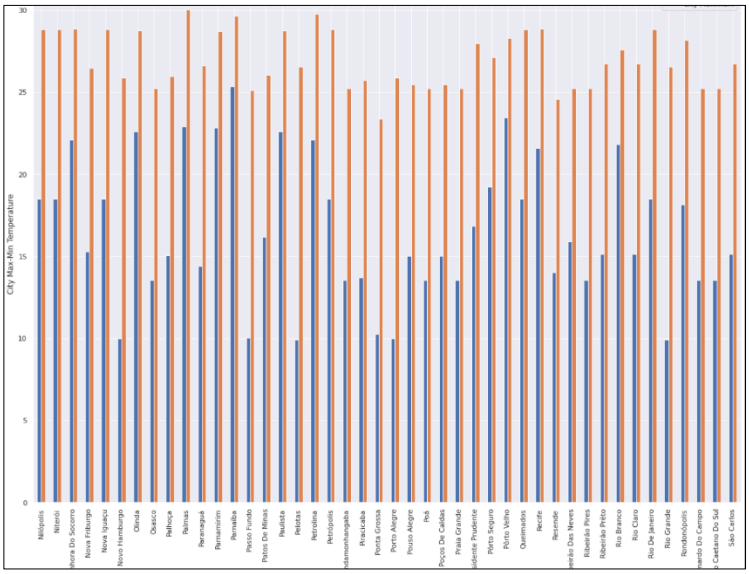


- We have also plotted an interactive globe map using Plotly to show the difference between the mean and the maximum temperatures across the globe.

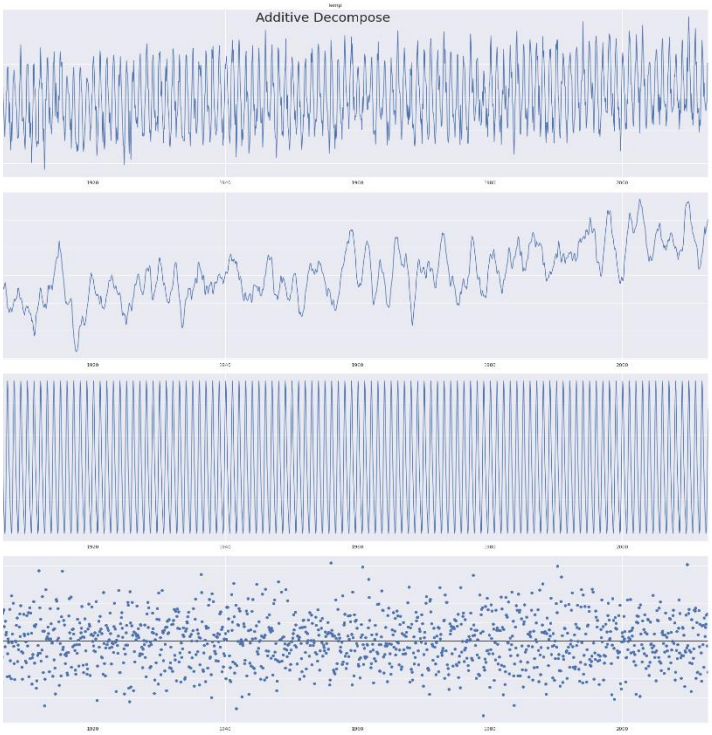
- box plots have been plotted for the average temperature for every month in a year.



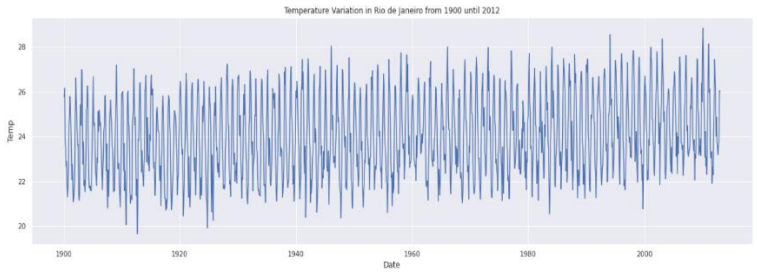
- Bar Plots have been used to depict the minimum and maximum temperature of each city for a particular country (in our case Brazil)



- We tried to decompose the data using both multiplicative and additive models for the temperature across the years.



- As per the results seen in the above figure, there seems to be a seasonality in the temperature over the years.



IV. METHODOLOGY

Once pre-processing of the data, exploratory data analysis, and visualization were done, a good understanding of the data was obtained. Our testing involved using machine learning models and techniques.

SARIMA

ARIMA is a popularly used model in time series data.
Autoregression (AR) refers to a model that shows a changing variable that regresses on its own lagged, or prior, values.
Integrated (I) represents the differencing of raw observations to allow for the time series to become stationary.
Moving average (MA) incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations.

SARIMA stands for Seasonal-ARIMA and it includes seasonality contribution to the forecast. The importance of seasonality is quite evident and ARIMA fails to encapsulate that information implicitly. The AR, I, and MA parts of the model remain as that of

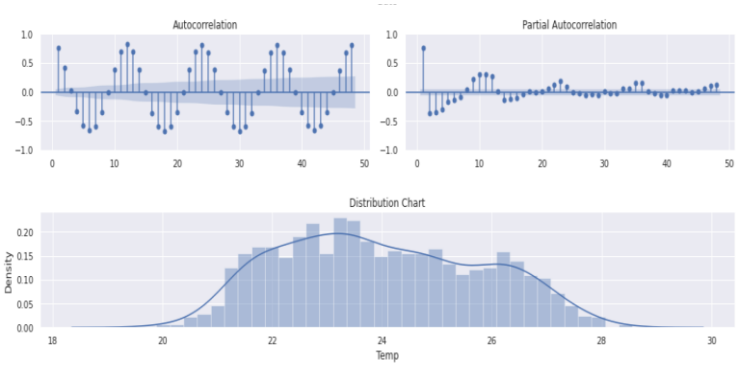
ARIMA. The addition of Seasonality adds robustness to the SARIMA model.
It's represented as SARIMA (p, d, q) (P, D, Q, S)

Similar to ARIMA, the P, D, and Q values for seasonal parts of the model can be deduced from the ACF and PACF plots of the data.

p: it is the lag observations number and it can also be called lag order.
d: count of raw observations is differenced; also known as the degree of difference.
q: the size of the moving average window; also known as the order of the moving average.

P: seasonal p
D: seasonal d
Q: seasonal q
S: seasonal length in our data

We perform additive and multiplicative decomposition on our dataset and we check for seasonality. We observed that there is seasonality in Global Temperatures which is why we choose to use the SARIMA model in our dataset. Later, we checked for stationarity using Augmented Dickey-Fuller Test and observed that the test statistics are lower than the 5% critical value. Therefore, it was inferred that the series is stationary.



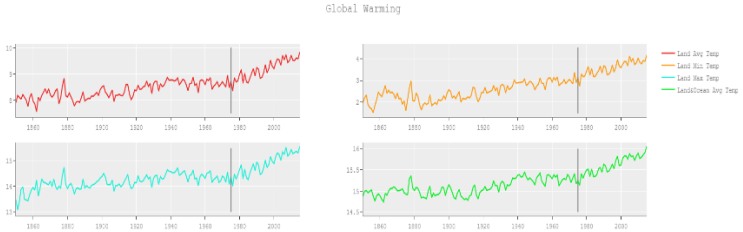

```
Results of Dickey-Fuller Test:
Test Statistic                  -3.7599
p-value                          0.0033
Lags Used                       23.0000
Number of Observations Used    1272.0000
Critical Value (1%)             -3.4355
Critical Value (5%)             -2.8638
Critical Value (10%)            -2.5680
dtype: float64

The Test Statistics is lower than the Critical Value of 5%.
The serie seems to be stationary
```

The other parameters are determined using ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) plots. The parameters p, d, q has been taken as (3,0,0) and seasonal parameters P, D, Q, S (0,1,1,12) after using ACF and PACF.
The RMSE of the SARIMA (3,0,0), (0,1,1,12),'c' model was 0.7874 Celsius degrees.

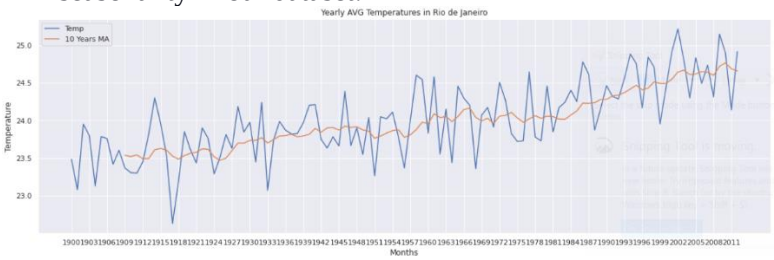
We observe a decrease of -40.71% in the RMSE compared to the baseline model.

V. RESULTS

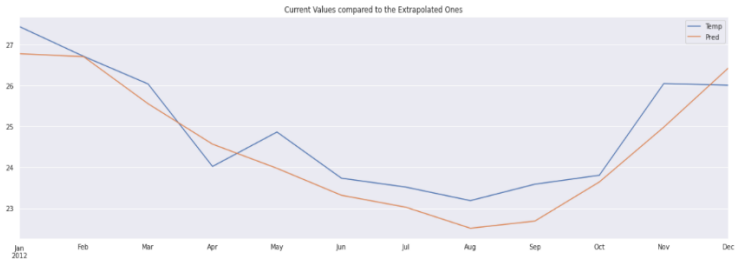
- From the analysis performed on the dataset, we learn that there has been a constant increase in average temperatures from 1850 onwards, but from 1975 there was a drastic increase in the level of global warming across the globe.

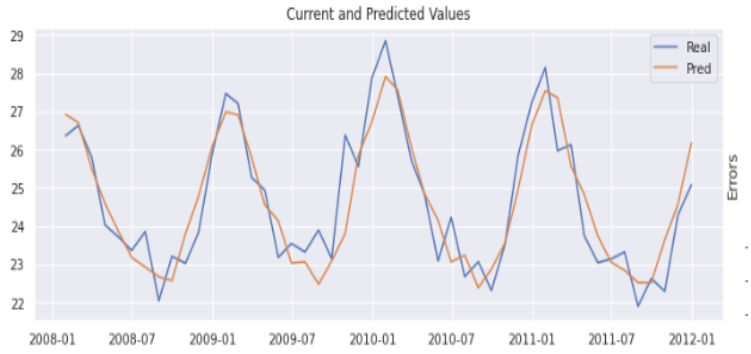


- We have observed the presence of trends and seasonality in our dataset.



- Also, we can observe that the most affected countries due to global warming are countries like Brazil, Nepal, Turkmenistan, and Mexico.
- From the various visualizations plotted we can also conclude that the continent most affected by global warming is Africa, followed by Asia and Europe.
- From the interactive map showing the increase in temperature over the years, we can conclude that Europe, South America, and West Asia have had greater temperature increases compared to other places.
- From the interactive map plotted for the difference between mean and maximum temperatures we can conclude that the countries in the northern hemisphere had a greater increase in temperature compared to the countries in the southern hemisphere.
- From the SARIMA model created for prediction, we split the dataset into training, validation, and test set, and the RMSE score is calculated. The RMSE of the baseline that we try to diminish is 1.32 Celsius degrees.
- The RMSE of the SARIMA(3,0,0),(0,1,1,12) model was 0.7874 Celsius degrees. It is a decrease of -40.71% in RMSE from baseline RMSE.
- The predicted values are plotted and the graph below is observed against the actual values.





VI. CONCLUSION

After analysing the global land and ocean temperatures of the last 185 years it can clearly be concluded that there has been a global increasing trend in temperature, particularly since 1975. The rate of global warming has increased rapidly since 1975 which has been the result of the increase in industrialisation, deforestation and due to the release of greenhouse gases into the atmosphere.

After predicting the average temperatures of the future by using a machine learning model we can clearly see that mankind must reflect and take the phenomenon of global warming seriously and take all necessary actions to reduce global warming.

ACKNOWLEDGMENT

We would like to acknowledge our Data Analytics Course Professor Anand M S for providing constant guidance during each phase of our project. We would also like to acknowledge the teaching assistants for preparing the course content and for constantly providing resources to practise the learnt concepts.

REFERENCES

[1] https://www.kaggle.com/code/costantinomarco/ridge-regression-on-climate-data

[2] https://www.kaggle.com/code/nithya22/predict-land-average-temperature-random-forest

[3] https://www.kaggle.com/code/lordxerxes/uk-cities-climate-change-prediction

[4] https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html

[5] https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/

[6] https://plotly.com/python/

[7] https://www.geeksforgeeks.org/getting-started-with-plotly-python/