Prediction of Carbon Dioxide Level Using Statistical Learning and Its Potential Correlation With Global Warming

Yuning Wang*, Yinlong Shen, Mingyi Hu

The High School Affiliated to Renmin University of China International Curriculum Center of RDFZ

Abstract. The Industrial Revolution caused a huge change in the climate of our planet. Since the 19th century, a high level of atmospheric carbon dioxide has contributed to global warming and other environmental problems. We first acknowledge the substantial correlations between the CO2 levels or temperatures and the years before creating our models. In this situation, we propose that the ARIMA model, which combines the auto-regression and moving average models, is essential for issue analysis. In order to estimate CO2 concentrations and land-ocean temperatures, we create polynomial models as well as an ARIMA model with seasonality. Following these hypotheses, we discover that the CO2 concentrations and temperatures have a significant direct link. In order to forecast the future relationships between CO2 concentrations and temperatures, we also attempt to employ polynomial function. We constantly reflect on and reexamine the issues as we construct these models in order to have a greater grasp of the circumstances. Each of our models is also evaluated, and the most precise one is used to make forecasts. Based on Matlab, we can quickly calculate the data, utilize iterations to determine the ideal model parameters, and then display our findings in diagrams.

1 Introduction

1.1 background

Our planet's climate has undergone a significant alteration after the Industrial Revolution. Then, since the 19th century, the atmosphere has had high carbon dioxide concentrations, contributing to global warming and other environmental issues.

1.2 Problem restatement

Cardon dioxide gas is a frequently occurring chemical in nature and accounts for 0.03% to 0.04% of the volume of the atmosphere. The following methods are employed in its creation: 1. Cardon dioxide can be released during the decomposition, fermentation, decay, and deterioration of organic materials. Both animals and plants are included in this. 2. Carbon dioxide is also produced by the burning of petroleum, paraffin, coal, and natural gas. 3. Coal and petroleum both emit carbon dioxide during the production of chemical compounds. Both manure and humid acid produce carbon dioxide as they age and ferment. 5. All animals breathe through the process of respiration, which involves taking in oxygen and exhaling carbon dioxide.

Prior to the industrial revolution Carbon dioxide has an atmosphere concentration of about 280 ppm. However, as a result of the industrial revolution's quick rise in productivity, the usage of internal combustion engines and other fuels rose greatly. The burning of these fuels resulted in massive amounts of CO2, which caused CO2 emissions to gradually rise. The 10-year average's highest level of 377.7 ppm of carbon dioxide was reached in the atmosphere by March 2004. The monthly average CO2 concentration peaked at 421 ppm in May 2022, per a study by the National Institute of Science, NOAA, and Scripps Institution of Oceanography (SIO). According to a research from the Organization for Economic Co-operation and Development, 685 ppm of CO2 will be present in the atmosphere by the year 2050.

Around 280 ppm of carbon dioxide were present in the atmosphere before the industrial revolution. However, the industrial revolution's rapid increase in productivity also led to a significant increase in the use of internal combustion engines and other fuels. These fuels produced enormous volumes of CO2 during combustion, which led to a steady increase in CO2 emissions. By March 2004, the atmospheric concentration of carbon dioxide had risen to its maximum value of 377.7 ppm for the 10-year average. According to research by the National Institute of Science, NOAA, and Scripps Institution of Oceanography, the monthly average CO2 concentration peaked at 421 ppm in May 2022. (SIO). The atmosphere will contain 685 ppm of CO2, according to research from the Organization for Economic

[*] Corresponding author: 15701571174@163.com

2 Historical CO2 Changes and Future Predictions

2.1 2004 CO2 increase

In problem 1.a, we calculated the increase in carbon dioxide for each decade given before 2004 by using linear regression. This is shown in the figures below.

Table 1. 10-year period CO2 increase before 2004 in ppm

Years interval	PPM increase	1977-1986	17.2015505
1959-1968	16.26816851	1978-1987	17.27090524
1960-1969	16.30387948	1979-1988	17.34428359
1961-1970	16.34021873	1980-1989	17.41798698
1962-1971	16.3760585	1981-1990	17.48824123
1963-1972	16.41354142	1982-1991	17.55786445
1964-1973	16.45953083	1983-1992	17.62479033
1965-1974	16.50485152	1984-1993	17.68661278
1966-1975	16.55275474	1985-1994	17.74853216
1967-1976	16.598429	1986-1995	17.8130598
1968-1977	16.64913163	1987-1996	17.8800842
1969-1978	16.70333035	1988-1997	17.94422917
1970-1979	16.75677208	1989-1998	18.01124903
1971-1980	16.81451878	1990-1999	18.0791563
1972-1981	16.87584101	1991-2000	18.14658655
1973-1982	16.93821013	1992-2001	18.21574952
1974-1983	16.99775088	1993-2002	18.29129774
1975-1984	17.06333664	1994-2003	18.3760727
1976-1985	17.13157547	1995-2004	18.4606093

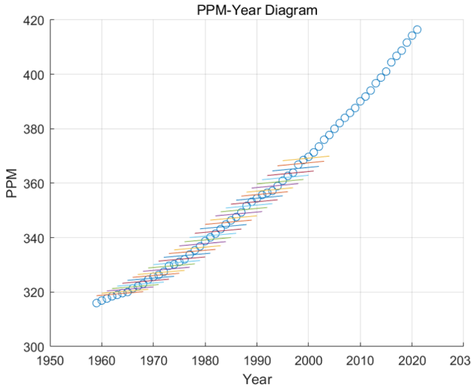

Figure 1 CO2 concentration in ppm and the increase rate of 10-year period before 2004

Based on the supplied data [3], linear regression was used every ten years from 1959 to 2022. Then, as illustrated in figure 1, construct a line for every ten years as a consequence of linear regressions. Each line's gradient serves as a gauge for the rise in CO2 concentration. According to table 1 and figure 1, the rise in carbon dioxide from 1995 to 2004 was 18.46, which was larger than the increase in each of the prior ten years. As a result, the CO2 spike in March 2004 caused a far larger increase than what was seen across any preceding 10-year intervals.

2.2 Models and Predictions

By incorporating historical data, our model seeks to analyze it while projecting future data. We decided to use data fitting to achieve this. Our models are shown here. [1]

Model 1: One-Degree Polynomial Fitting

The graph is initially drawn with all the discrete data points, and it is then clear that the points have a linear trend. Then we choose to fit the data using a one-degree polynomial.

The equation of one-degree fitting is f(x)=p1*x+p2. Then the algorithm fits the data and gives the results below (figure 2). In the one-degree fitting results, the r-square of the fitting function equals 0.9825.

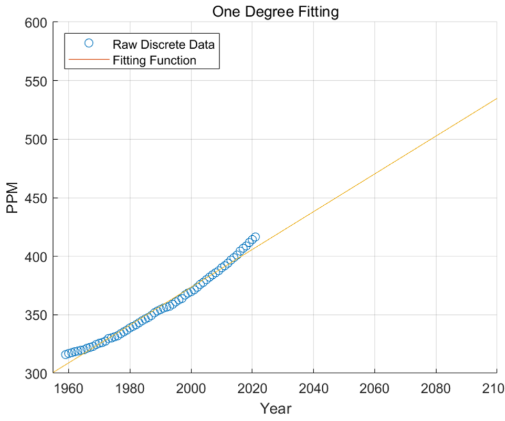

Figure 2 Diagram using one-degree polynomial fitting

Model 2:Two-degree Polynomial Fitting

Then, using the quadratic f(x)=p1*x^2+p2*x+p3, to fit the data. This produced a curve (Figure 3). The fitting function's r-square for the two-degree fitting results is 0.9994.

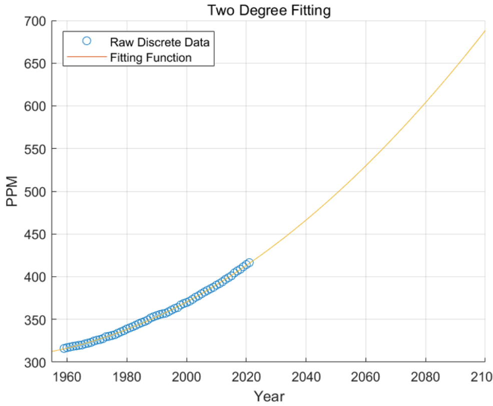

Figure 3 Diagram using two-degree polynomial fitting

By comparing the R-square, we can conclude that the two-degree polynomial fitting in this situation has higher quality.

Model3: ARIMA model

Finally, we used the Arima model where Arima(2,0,2). ARIMA model is a combination of AR(Auto-regressive) and MA(Moving Average).

By using a loop in MATLAB, we tried different combinations of the parameters in Arima model. Then, we choose the combination with smallest variance, which is Arima(2,0,2). By using ARIMA model, we conclude that PPM will reach 650.42 by 2100. PPM will reach 685 by 2110 instead of 2050 (figure 4).

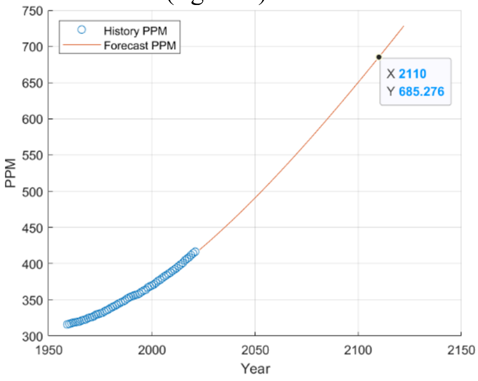

Figure 4 Prediction of CO2 concentration in 2110

2.3 Results Analysis

We tested these three models to determine their correctness after having them all. The first x number of data points are fitted using one of the three techniques described above, and the average absolute value of the difference between the fitting function's value and the original value of the last (64-x) data points is then calculated. These values are known as "diff." The "diff" values for the one-degree polynomial fitting techniques, the two-degree polynomial fitting methods, and the ARIMA model are 4.0936, 0.6453, and 0.5745 respectively. As a result, we may conclude that the ARIMA model is the most precise model that can produce the most precise forecast in our trials.

3 Relationship Between Temperature and CO2

3.1 Future Temperature Predictions

The picture below depicts the scatter plot we created using the global temperatures for each year and the data from Temps Data Set 2. [1]

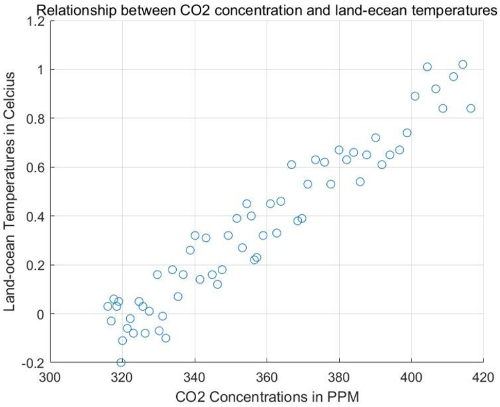

Figure 5 Scatter plot of past land-ocean temperature

We are motivated to create an ARIMA model since this data set is non- stationary, which makes it clear that a linear or polynomial model cannot adequately represent it. (Figure 5)

The statistical method known as Auto Regressive Integrated Moving Average, or ARIMA, is a common one for forecasting time data. Before extending the signal into the future to create forecasts, the ARIMA model functions as a "filter" that tries to separate the signal from the noise.

In general, the use of ARIMA always contains several steps (Figure 6):

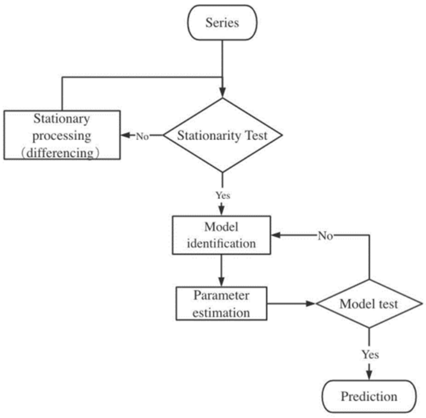

Figure 6 General process for using ARIMA model

Statistics defines stationarity of data as the persistence of the distribution of the data across time. Non-stationary data must be transformed before being analysed because of the changes brought on by trends. We evaluate the stability of this data set using the Augmented Dickey-Fuller Test (ADF test). The Dickey- Fuller test equation is enlarged by the ADF test to include higher-order regression processes. It evolved out of the unit root test. In general, if the ADF test's p-value is higher than a predetermined critical percentile (in this example, we'll pick 1%, 5%, or 10%), the result should be considered significant. [4]

The first-time stationarity test yields a p-value greater than 0.05, necessitating the stationary processing. The most effective approach to do this is via differentiating. Differentiation in statistics reduces seasonality, trends, and fluctuations in the level of the time series, stabilizing the mean. To remove seasonal components, for instance, seasonal time series are subjected to seasonal differences.

In this issue, we use first-order differencing to stabilize data in time series.

Some time series clearly exhibit a cyclical oscillation because to seasonal variations (quarter, month, etc.). Since the temperature change contains a seasonal component, we decide to use a SARIMA variation of the ARIMA model (Seasonal ARIMA).

Two models, AR and MA, make up SARIMA. Three parameters are used to define the model, which is expressed as SARIMA(p,d,q) where:

1) d = degree of first differencing involved
2) p = order of the AR part
3) q = order of the moving average part

Seasonality is a unique characteristic in SARIMA that distinguishes it from ARIMA (written as s).

The value of p may be calculated more precisely using the ACF plot, which shows the autocorrelations that quantify the link between an observation and its predecessor. The value of d, the order of integration, may be calculated by counting how many transformations are necessary to make the time series stationary. The value of q may be determined using the PACF plot. The econometric modeler toolbox in MATLAB helps us determine p=7 and q=4 with seasonality=1 in this case.

After we automatically determine the values of p and q, we can next synthesis the ARIMA and SARIMA models with d=1,2 and accordingly.

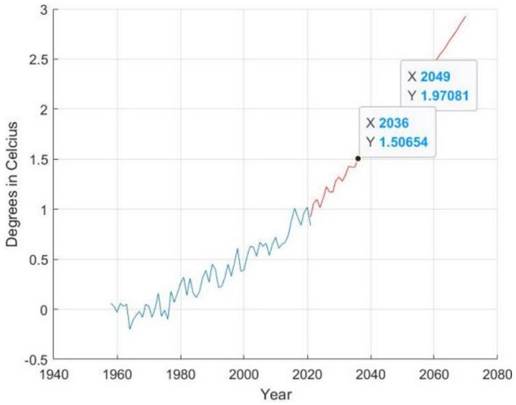

Figure 7 SARIMA model when d=2

When d=2 (Figure 7),

Two charts produced and two forecasts made, however, do not always imply that they are acceptable or correct. The goodness of fit has to be investigated. The average absolute value of the difference between the function value of the last 23 data points and the original value of the last 23 data points reflects how well the model fits the data set. In specifically, two models were employed to fit the first 40 data points in turn. These test results are displayed.

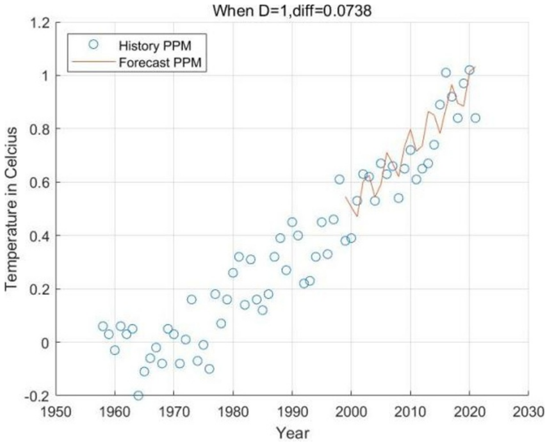

Figure 8 Test for SARIMA model with d=1

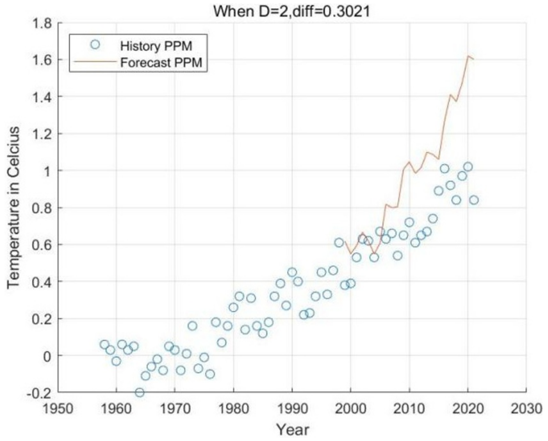

Figure 9 Test for SARIMA model with d=2

However, the existence of two charts and two forecasts does not automatically suggest that they are good or accurate. It is necessary to look at the goodness of fit. How well the model fits the data set is shown by the average absolute value of the difference between the function value of the last 23 data points and the original value of the last 23 data points. Two models were used especially to fit each of the first 40 data points. These test outcomes are shown. (Figure 8 and Figure 9)

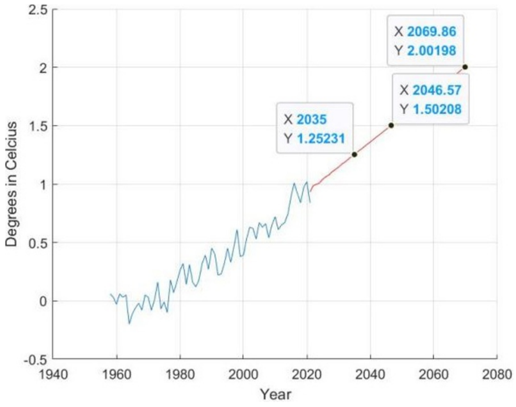

Figure 10 Using SARIMA(7,1,4) to predict

This model clearly shows that the average temperature will increase by 1.25 degrees Celsius by 2035, 1.50 degrees Celsius by roughly June 2046, and 2 degrees Celsius by October 2069. (Figure 10)

3.2 CO2 and Temperature Correlations

We need to know if there is a model that completely matches the position and relationship of each point in the CO2 concentration-land ocean temperatures scatter diagram to assess whether CO2 concentrations and land-ocean temperatures have changed since 1959. The scatter plot we originally created using the data on CO2 levels and land- ocean temperatures is depicted here.

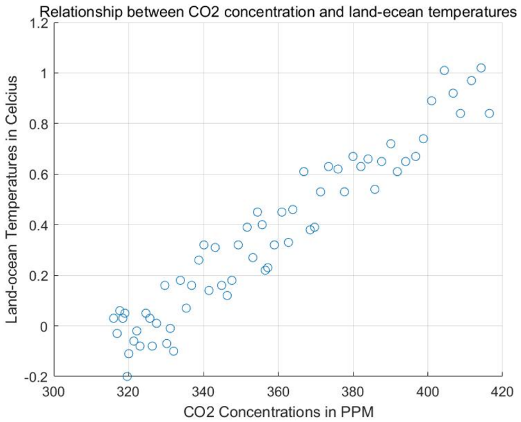

Figure 11 Scatter plot of Co2 concentration and land-ocean temperatures

We use the MATLAB corrcoef model to analyze the link between land- ocean temperatures and CO2 concentrations. (Figure 11)

Corrcoef function:

The corrcoef model (written as R=corrcoef(A,B)) may be used to assess the relationship between two sequences of the same length, sequence(A) and sequence(B). [2]

The correlation coefficient, often known as corrcoef, is a measure of the linear dependency between two random variables. The Pearson correlation coefficient is described as, where N is the number of scalar observations for each variable,

$$\rho(A,B) = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{A_i-\mu_A}{\sigma_A}\right)\left(\frac{B_i-\mu_B}{\sigma_B}\right) \tag{1}$$

μ_A : the mean of sequence A

μ_B : the mean of sequence B

σ_A : standard deviation of A

σ_B : standard deviation of B

In that case, we can also define the correlation coefficient in terms of the covariance of A and B:

$$\rho(A,B) = \frac{cov(A,B)}{\sigma_A\sigma_B} \tag{2}$$

The correlation coefficient matrix of two random variables is the matrix of correlation coefficients for each pairwise variable combination:

$$R = \begin{pmatrix} \rho(A,A) & \rho(A,B) \\ \rho(B,A) & \rho(B,B) \end{pmatrix} \tag{3}$$

As ρ(A,A) and ρ(B,B) = 0 since a sequence is always correlated to itself, so:

$$R = \begin{pmatrix} 1 & \rho(A,B) \\ \rho(B,A) & 1 \end{pmatrix} \tag{4}$$

The signal effect of R in determining the relationship between sequence A and B:

ρ(B,A) or ρ(A,B) (ρ(A,B)= ρ(B,A))is between 1 and -1, and A, B are positively correlated if ρ(A,B) is smaller than 0 while A, B are negatively correlated if ρ(A,B) is greater than 0. As the ρ(A,B) get close to 1, the correlation between A and B becomes stronger.

In this problem, ρ(A,B)=0.9613. This shows a strong positive correlation between land-ocean temperatures and CO2 concentration.

3.3 Reliability of Predictions

When the data is scaled up and the oscillations are ignored as they get more subtle, first order linear equations are easy to use for fitting. Because there is a considerable positive correlation between land-ocean temperature and CO2 concentration, as shown by figure 12, we utilize a linear function. As a result, it is possible to employ a linear model that has been simplified.

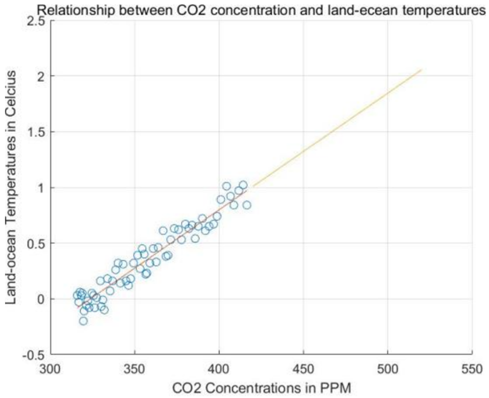

Figure 12 Using linear function to fit

Meanwhile, such a model is not particularly reliable. First off, a linear model cannot account for the seasonal variations that characterize the temperature shift. It is expected that the temperature would fluctuate in the future since we study the Seasonal ARIMA model in 2(a). As a result, when utilized for prediction, such a move could only provide an approximate value.

Second, even if land-ocean temperature and CO2 concentration have a strong positive connection, we cannot draw the conclusion that the two variables can be fitted by a simple linear function since the correlation coefficient is 0.9613 rather than 1. Because of this, using a single function for a lengthy period of time is limited and inaccurate (a specific time period cannot be given yet).

Additionally, the effects of outside factors have been disregarded because the model used in the previous question only considers one possibility. For instance, the effects of future changes in vegetation and the fact that many countries seek to achieve carbon neutrality in this century have not been taken into account. Both the rate of change in CO2 concentration and the rate of change in temperature are anticipated to slow down in the future. The model can thus be improved further and is only for reference. [5]

4 Conclusion

The Industrial Revolution caused a huge change in the climate of our planet. The atmosphere has also had high carbon dioxide concentrations since the 19th century, which has contributed to global warming and other environmental problems.

We focused on CO2 levels and land-ocean temperatures as the two main markers of the Earth's environmental status in our study. We use the PPM (parts per million) shorthand for carbon dioxide content and degrees Celsius for land-ocean temperature. PPM stands for mole fraction in dry air.

We began by researching the levels of carbon dioxide. Analyzing past data and predicting future data were topics that both of us were interested in. When compared to any other ten-year period in the historical data, we found that the CO2 spike in 2004 caused a sizable increase. To show this, we used mathematical methods like linear regression to calculate the ppm growth over each ten-year period. Next, we demonstrated our discovery. We utilized three alternative mathematical models to suit the scatter point on the PPM-Year diagram for the most crucial future forecast. Our objective was to forecast CO2 levels for the year 2100 and the year that the ppm reaches 685.

We began by researching the quantities of carbon dioxide. Analyzing past data and predicting future data were something that both of us were interested in. When compared to any other ten-year period in the historical data, we found that the CO2 rise in 2004 produced a significant increase. To show this, we used mathematical methods like linear regression to determine the ppm rise for each ten-year period. We then showed out our discovery. For the most important future projection, we fitted the scatter point on the PPM- Year diagram using three different mathematical models. Forecasting CO2 levels for the year 2100 and the year the ppm hits 685 was our goal.

Finally, we combined the conclusions from the preceding section's CO2 concentration and temperature data. We found a large and unmistakable correlation between CO2 concentrations and land-ocean temperatures. We also used a simple polynomial function to fit the data and provide a forecast for the future. Making forecasts based on past data is a beneficial method for resolving environmental issues, even though it may seem difficult. Future phases of our study will continue.

References

1. Lenssen, N.J.L., Schmidt, G.A., Hansen, J.E., Menne, M.J., Persin, A., Ruedy, R. and Zyss, D. (2019). Improvements in the GISTEMP Uncertainty Model. Journal of Geophysical Research: Atmospheres, 124(12), pp.6307–6326. doi:https://doi.org/10.1029/ 2018jd029522.
2. Mathwork.com. (2022). Available at: http://mathwork.com [Accessed 15 Nov. 2022].
3. NASA (2010). Data.GISS: GISS Surface Temperature Analysis, GISTEMP/v3. [online] Nasa.gov. Available at: https://data.giss.nasa.gov/gistemp/.
4. www.csdn.net. (2022). CSDN. [online] Available at: http://csdn.com [Accessed 15 Nov. 2022].
5. Zheng, Harvey. "Analysis of global warming using machine learning." Computational Water, Energy, and Environmental Engineering 7.03 (2018): 127.