**2022 HiMCM**

Team Control Number

Problem Chosen

**B**

**Summary Sheet**

**Keywords:**

**Contents**

1. Introduction
   1. Background
   2. Question Restatement
2. Assumptions and Justifications
3. Modeling
   1. Linear Regression
      1. Introduction to the Model
      2. Variables of the Model
      3. Why This Model Work
      4. Building the Model
      5. Results and Prediction
      6. Reflection
   2. Stochastic Model
      1. Introduction to the Model
      2. Variables of the Model
      3. Why This Model Work
      4. Building the Model
      5. Stop and Reflect
   3. Normal Stochastic Model
      1. Building the Model
      2. Results and Predictions
      3. Linking to Other Models
4. Comparison of the Models
5. Errors
6. Conclusion
7. Appendix
   1. Code
   2. Reference
8. **Introduction**

**1.1 Background**

As the world’s economy continues its expeditious expansion, the concomitant emission of greenhouse gases has been the source of tribulation for many countries as the exacerbation of extreme weather, prolongation of droughts, and flooding of coastal cities due to rising sea levels have both catastrophic and insidious effects on agriculture, infrastructure, and the welfare of people. In particular, reducing carbon emissions has been the primary focus for most governments because it is by far main byproduct of energy consumption via fossil fuels such as oil, natural gas, and coal. Therefore, the attenuation of CO2 concentrations within the atmosphere specifically, has elicited an indispensable exigence. However, the sanction of reduction policies rests heavily upon the veracity of carbon predictions as the government has to mitigate GHG emissions without abnegating economic output. Whence, an accurate divulgence of future CO2 concentrations would be highly conducive to the construction of reduction plans as it would predicate the amount of GHG the government must reduce, and the minimum economic output that must be sacrificed.

**1.2 Restatement of problem**

The purpose of this project is model future atmospheric CO2 concentrations and predict land-ocean temperature changes through justified mathematical models by employing data from previous years. Upon these models, a recommendation is further constructed for Scientific Daily to explain our results and mitigate the results predicted through our findings.

**Question 1a**: Using the given data, we need to ascertain whether the March 2004 increase of CO2 was greater than previous 10-year periods.

**Question 1b - 1d**: By using the conclusions gathered in 1a, we are required to administer various models to derive predictions of future CO2 concentrations in the atmosphere and determine when the concentrations will reach 685 ppm. Then, an evaluation must be given of the models' accuracy.

**Question 2a**: Assuming there is a correlation between atmospheric CO2 concentrations and land-ocean temperature, we must derive a model to predict changes in average land-ocean temperature with respect to the base period of 1951-1980

**Question 2b-2c**: Based on the results of 2a, we are to build an additional model to analyze the type of relationship between CO2 concentrations and land ocean temperatures and dictate its reliability as well as any flaws undisclosed within the model.

**Question 3**: In a single page, we must explain our findings and offer propositions to Scientific Today established from our results.

It is obvious that the emission of CO2 is caused by many aspects of our society. The population of human beings, for example, could directly contribute to the emission of CO2 into the atmosphere since the action of exhaling produces CO2. Meanwhile, human actions such as powering transports, producing goods, and destructing forests all cause changes in the environment, releasing large amounts of CO2 while removing the plants that are responsible for the uptake of CO2.

1. **Assumptions and Justifications**

**Assumption 1**: The trend of carbon dioxide emission is not completely random, but rather predictable and possible to attach models on owing to human effects.

**Justification 1**: As a result of humans triggering a notable halt in the natural CO2 concentration cycle, concentration records since the 1960s indicate an evident spike in concentration. This could also be explained by the existence of global warming and climate change, where scientists are recognizing an irregularity in the CO2 trend that cannot be explained by normal carbon dioxide cycles. As a result, it is critical to consider the carbon dioxide trend as something that can be studied and predicted with models in order to determine the actual condition mankind is in and offer adequate guidance to the globe.

**Assumption 2**: Influences that are relatively less significant will not be taken into consideration.

**Justification 2**: With the whole world being the most complex system, the change in carbon dioxide levels is influenced by an endless number of variables. As a consequence, including every single variable becomes simply impractical; as a result, we would only evaluate the most relevant variables that might considerably impact the trend of carbon dioxide.

**Assumption 3**: Within the predictions, the future will not have any massive events that could change the entire state of carbon dioxide emission, such as nuclear war or emergent government regulations.

**Justification 3**: As one of the uncontrollable variables during prediction, if government regulations and sudden events that could change the whole trend of carbon dioxide were taken into consideration, the prediction would simply become meaningless. Trying to find a logical prediction under every possible circumstance not only will be incredibly challenging, but it will also fail the goal to deliver the necessary counsel. As a result, the prediction made in this paper are all coming from the assumption or premise that the situation remains relatively similar to the overall development that is occurring under the status quo.

**Assumption 4**: Fossil energy consumption and carbon dioxide emission is directly correlated.

**Justification 4**: Since the use of fossil fuels is a major contributor to carbon dioxide emissions into the atmosphere, it is reasonable to assume that a change in fossil energy use would directly result in a change in carbon dioxide emissions. Due to its unique carbon dioxide emissions and impact to global warming, only energy consumption resulting from the use of fossil fuel is considered. All of the other factors that may lead to a shift in fossil energy use, such as population increase, economic growth, or political constraints, can culminate in this one component, causing contact with the carbon dioxide concentration in the atmosphere.

1. **Modeling**
   1. **Linear Regression**
      1. **Introduction to the Model**

This model is used to predict the relationship between two variables by applying a linear equation to the given data. It is commonly used for predictive analysis that includes an explanatory (independent) variable and a dependent variable. But before using this model, it is essential to prove that there is a relationship between the two variables.

* + 1. **Variables of the Model:**

x: the time (year number)

y: the level of CO2 in the atmosphere (in ppm)

* + 1. **Why This Model Works:**

To justify the use of linear regression, all we have to do is to justify that there is a close relationship between the year number and the level of CO2 in the atmosphere. As a visualized graph, a scattered plot may be used to indicate such correlation:

图表, 折线图

描述已自动生成[Graph 1]

As shown above, the trend of the line is sloping upwards, indicating that there is a positive relationship between the year number and the level of CO2: The level of CO2 increases as the year increases. And since the line is relatively flat, it seems reasonable for us to use linear regression as a model to predict the future level of CO2 concerning the year.

Now we have the idea of a correlation, we should also use a correlation coefficient to prove numerically that our conclusion is not only a conjecture. Correlation coefficient formulas are used to measure the relationship between two variables. It will return a value between -1 and 1. The magnitude of the value means how strong the relationship is (with 1 indicating the strongest relationship). The sign of the value means how the two values are related. If it results is negative, then the two variables have a negative relationship; if the result is positive, then the two variables have a positive relationship. If the result is 0, then there is no relationship at all.

The most commonly used type of formula is Pearson’s correlation coefficient formula, which is the one we will use here:

is the sample covariance.

is the standard deviation of values of x.

is the standard deviation of values of y.

Expanding the formula will give us:

Where n is the sample size.

The result is 0.9912, which means that there is an extremely strong and positive correlation between the two variables. Thus, our use of this model is valid.

* + 1. **Building the Model**

To insert our regression line into the XY plot, we use the method of least squares. This process determines the best-fitting line for the given data by reducing the sum of the squares of the vertical deviations from each data point to the line. It starts with a modeled linear equation:

where and are the coefficients of this equation, and thus the variables to be found during this process.

With that being the case, the calculation for sum of squares error would be as following:

where is the real value of concentration at *i*, and n is the degree of the equation, which should be 1 in current case of linear regression.

After coming up with this formula, the only step last is to find the get to a solution where the sum of squares error becomes its minimum value for the input data.

Continuing with some more calculation to solve for the and value that provides the minimum Q, the equation becomes:

Now, with this equation, we can simply substitute in the statistics that requires regression for the linear model and get the result of the equation needed.

After getting the equation needed, it is also necessary to measure the coefficient of determination to see how accurate the line is to the given data. We used the method of r2\_score which processes the data through the following formula:

where represents the mean of all y values.

The line is the more accurate to the given data when R2 gets closer to 1, which means that we would want R2 or the coefficient of determination to be as close to 1 as it could possibly be.

* + 1. **Results and Prediction**

图表, 折线图

描述已自动生成[Graph 2]

Fitted equation: ;

;

;

Coefficient of determination = 0.9824637498719381.

The data predicted through this line tells us that the carbon dioxide concentration would be 434.7312 ppm by 2050 and 534.8825 ppm by 2100. It could be clearly seen that the concentration by 2050 did not achieve the value of 685 ppm, but was achieved in the year 2193

**3.1.6 Reflection**

The linear regression model outputs the line of best fit for the data given. With the equation of , it provides us with a straight line and a constant slope. But in reality, multiple aspects of life will affect the data. The outcome could not be summarized as a linear equation. In fact, many things happen without a specific pattern. Therefore, we provide the stochastic model.

* 1. **Stochastic Model**

**3.2.1 Introduction to the Model**

Stochastic models are used when the variables have certain levels of unpredictability or randomness. The word “stochastic” comes from the Greek word “stokhazesthai,” meaning to aim or guess. Within our discussion, there seems to be random factors in the changes of the level of CO2 in our atmosphere. Thus, we have decided to include this model as well.

**3.2.2 Variables of this Model**

x: the elapsed year starting from 1959.

y: the change in the level of CO2 in the atmosphere (in PPM)

**3.2.3 Why This Model Works**

It is easy for us to justify this model with an example. Suppose a man is lying on his bed and not doing anything. Suddenly he feels an urge for the consumption of chips. He figures no chips are left in his house and decides to travel by car to the nearest market in order to buy some. On the road, the car will emit CO2 into the atmosphere. Such actions are barely able to be predicted, if not say it is completely random. However, the effect, or the CO2 emitted in this case, is relatively small in comparison to the global emission of CO2. By using the stochastic model, random variables are included to make the data more realistic.

Then, we have to justify that our data are suited to using such a model. Since the level of CO2 in the atmosphere is positively correlated to the time, we have to process the data to gain the randomized factors. Thus, we subtracted one year’s level of CO2 in the atmosphere from the previous year’s level of CO2 in the atmosphere.

We then get this data:

表格

低可信度描述已自动生成[Table 1]

The graph of the change in the level of CO2 in the atmosphere in respect to elapsed year looks like this:

[Graph 3]

Since the result seems to be an oscillating line, it is obvious that there are random components within the graph.

**3.2.4 Building the model**

To create our stochastic model, we need to set the range of our sample space. However, our data seems to be progressing with an upward trend that makes us unable to set a specific range. Thus, we decomposed this data into the equation shown below:

is the change in the level of CO2 in the atmosphere.

is the elapsed year; is the line of best fit of the data, indicating the total trend.

is the unpredictable variable that is necessary for indicating how far the data is away from the line of best fit.

[Graph 4]

As indicated above, the equation for could be denoted as:

After that, we can get the values of by subtracting from . The values obtained is shown below:

[Graph 5]

We are now able to find the maximum and minimum values of the variable K. With help of computer programing, we obtain a set of randomized values for K within the range between the maximum value and the minimum value of K.

We can then transform it into the corresponding .

This leads us to the predicted data:

[Graph 6]

**3.2.5 Stop and Reflect**

Even though the predicted data of the change in the level of CO2 in the atmosphere has an increasing trend, there is still something queer about the data. Starting from our prediction, the average amplitude of the oscillation seems to be a lot larger than the real data. After discussing the matter, we concluded that the error exists because the probability of gaining any number within the range of sample space is the same in our model. In reality, the data is normally distributed, meaning that the probability of getting a number is weighted.

图表, 直方图

描述已自动生成[Graph 7]

The x axis stands for the difference between the true data and the modeled data, while the y axis stands for the number of these data that satisfies the corresponding,

differences.

For example, the graph would show that there are 20 points had a difference from 0 to 0.5 between the modeled data and the true data.

With analysis done to the given data, we are able to determine that the data are normally distributed to our equation modeled with the change in concentration. The above figure shows the existence of normal distribution, and thus provide us with the logical proof to apply normal distribution for the improvement on our stochastic model.

Thus, we decided to optimize our model.

**Model 3: Normal Stochastic Model**

**2.3.1 Building the improved model**

This model is mostly the same as the Stochastic Model, only that the generated values of the uncertain variable will undergo a normal stochastic process, making it a gaussian variable.

For the process to occur, we calculated the average and the standard deviation of our data:

After that, we generated the gaussian random variable .

Then, we can obtain the change in the level of CO2 in the atmosphere with the same procedure as the stochastic model. The result would be:

[Graph 8]

The graph seems much more natural, indicating that our prediction becomes more realistic.

We can then predict the level of CO2 in the atmosphere with the following equation:

2.5 Results and predictions

Our data and predicted result would be the following:

表格

描述已自动生成

[graph 9]

According to our prediction using this model, the level of CO2 in the atmosphere will be 700.04 ppm by the year 2100. This model also denies the prediction of reaching 685 ppm by the year 2050. It predicts that the world’s level of CO2 in atmosphere will reach 685 ppm by the year 2097.

2.6 Similarities with other models

Our Stochastic Model is also similar to the Quadratic Model. Within a quadratic model, the rate of change of the values on the y axis in respect to the rate of change of the values on the x axis could be defined as the derivative of the equation. The derivative is denoted as:

With the regular formula of a quadratic equation is , the derivative of the formula would be . This is a linear equation, which means that within a quadratic equation, the change in the values of the y axis is directly proportional to the change in the values of the x axis.

Our Stochastic Model has similar ideas. In our model, the change in the values of the y axis is the change in level of CO2 in the atmosphere. The change in the values of the x axis is the time elapse. And in our model, the basics of the derivative is also a linear equation (indicated in graph 4). Thus, we can note a significant similarity between these two models.

图表, 折线图

描述已自动生成Now we add the prediction and some variety with normal random function to the line:

**3.1 Time Series Forecasting with ARIMA**

Auto Regressive Integrated Moving Average (ARIMA) is a time series forecasting model to predict future values. We attempt to use ARIMA to predict the future CO2 level as well as land-ocean temperatures.

**3.1.1 Stationary Time Series**

Firstly, because only a stationary data can be used with ARIMA model to forecast the future data, we need to confirm whether our history data is a stationary time series. A stationary time series is the one whose properties do not depend on the time at which the series is observed. Thus, time series with trends or seasonality, are not stationary because the trend and seasonality will affect the value of the time series at different times.

There are couple of methods to test whether a data series is stationary, for instance, ADF (Augmented Dickey–Fuller) testing or ACF (AutoCorrelation Function) plotting.

However, the easiest way to is just to plot the data and visualize the trend judged by eyes. If it is not stationary, we need to make the time series stationary - compute the differences between consecutive observations. This is known as **differencing**. Because differencing can help stabilize the mean of a time series by removing changes in the level of a time series, and therefore eliminating trend and seasonality.

In our initial test, we observed that the ppm/year (history CO2 level) is not stationary, and even its n-order (n=1,2,3…) differenced data is not stationary either. However, we observed that although the raw temperature (degree/year) data is not a stationary time series (see the 1st figure below) their 2-order differenced temperature data is a stationary series (the 2nd figure below).

图表, 折线图

描述已自动生成

图表, 折线图

描述已自动生成

**3.1.2 p, d, q Value**

An ARIMA model is characterized by 3 terms: p, d, q, where p is the order of the Auto Regressive (AR) term, d is the number of differencing required to make the time series stationary, q is the order of the Moving Average (MA) term.

**3.1.2.1 d Value and ARIMA(p,d,q) Model**

Obviously, in the previous step, we got to know that d=2. Now let’s decide the other two model parameters p (AR) and q (MA). This will lead to the concepts of AR model and MR model.

A pure Auto Regressive (AR only) model is one where Yt depends only on its own lags. Which means Yt is a function of the ‘lags of Yt’. This AR(p) model can be depicted as below

Yt = c + ϕ1Yt−1+ ϕ2Yt−2 +⋯+ ϕpYt−p + εt ,

where, Yt is the lag 1 of series, ϕ1 is the coefficient of lag 1 that the model estimates, and c is the intercept, εt is the white noise (error).

While a pure Moving Average (MA only) model is the one where Yt depends only on the lagged forecast errors ε. The MA(q) model can be depicted as below

Yt = c + εt + θ1εt−1 + θ2εt−2 +⋯+ θqεt−q ,

where, where εt is white noise, each value of Yt can be thought of as a weighted moving average of the past few forecast errors.

If we combine differencing with autoregression and a moving average model, we obtain a non-seasonal ARIMA model. The full model can be written as

Y′t = c + ϕ1Y′t−1 + ⋯ + ϕpY′t−p + θ1εt−1 +⋯+ θqεt−q + εt ,

where Y′t is the differenced series (it may have been differenced more than once, e.g. 2-order). The “predictors” on the right-hand side include both lagged values of Yt and lagged errors εt. We call this an ARIMA( p, d, q) model.

2.1.2.2 p value

By inspecting the Partial Autocorrelation (PACF) plot., we can get the value of p order.

PACF can be imagined as the correlation between the series and its lag after excluding the contributions from the intermediate lags. So, PACF sort of conveys the pure correlation between a lag and the series.

We plot the PACF of d-order (d = 2 as got above) differenced temperature time series with Python tool, as the figure blow.

图片包含 图表

描述已自动生成

We can select the order p based on significant spikes from the PACF plot above. so it can concluded that this PACF plot has significant spikes at lags 2, hence we can try **p = 2**.

**2.1.2.3** q Value

In contrast to the AR model, we can select the order q for model MA(q) from ACF (Autocorrelation Function) if this plot has a sharp cut-off after lag q.

So, we plot ACF for the d-order (d = 2) differenced temperature time series with Python tool, see figure below:

图表, 箱线图

中度可信度描述已自动生成

Similar to determining p for the AR model, in order to select the appropriate q order for the MA model, we analyze all spikes higher than the blue area (95% confidence by default). So, we can try to use q = 2.

**2.1.2.4** Modeling

Now, we have selected p, d and q order values, but typically, in order to find the best combination of p, d and q, we need to measure model performance on a validation set. Usually, we can use AIC (Akaike information criterion)[[1]](#footnote-1) and BIC (Bayesian information criterion)[[2]](#footnote-2) for that purpose. The lower the value of these criteria, the better the model is.

Based on above observation, we can try multiple models with combinations, for example, p = 0,1,2, and q = 0,1,2. We can simply get AIC/BIC values for each of them and select the best one model, which is (***2,2,1***) because the absolute values of AIC/BIC are the lowest among those models. See figure below after executing Python code:  
# by Default, use generalized least squares(GLS) to fit the model  
modelfit = sm.tsa.ARIMA(degree.dropna().to\_numpy(), order=(2,2,1)).fit()

print(modelfit.summary())

表格

描述已自动生成

Furthermore, we review the residual plots as shown below.

图表, 直方图

描述已自动生成

where, we can interpret this plot diagnostics:  
**Top Left:** The residual errors seem to fluctuate around a mean of 0 and have a uniform variance.

**Top Right**: The density plot indicates a normal distribution with mean near-zero.

**Bottom left**: All the dots looks fall in line with the red line. No significant deviations.

**Bottom Right**: The Correlogram(ACF) plot shows the residual errors are not autocorrelated (all lags seat inside of blue area)

Overall, it seems to be a very good fit with ARIMA(6,1,1) model. And also we tried to plot the Actual vs Fitted, see below, as a comparison.

图表, 折线图

描述已自动生成

For cross-validation of our forecast model, we also tried to split the data with training and test data set (from 1959 to 2021), e.g., 80% training data and 20% test data. And plot the predict results as below. you can see that the green line indicates forecast data which is pretty close to actual data set with trend and variations.

图表, 折线图

描述已自动生成

Now, we can forecast for future years starting from 2022, as it is noticeable that in a long run, the prediction trend seems to be a straight line. This is not a surprise, because ARIMA mode is able to forecast very short time series, and more actually, most time series model do not work well for very long time series.

图形用户界面, 图表, 折线图

描述已自动生成

So with this ARIMA (2,2,1) model, we get the results: the temperature will increase by 1.25 degrees in 2032, by 1.5 degrees in 2041, and by 2 degrees in 2060.

**3.1.2.5 Reflection**

From the graph above, we can observe that after about 2030, the predicted data line is almost a linear straight line. Thus, it is obvious that ARIMA model has a serious flaw: it can only precisely and accurately predict a short time into the future. The concern is that, this model is not reliable at all after the year of 2030. We should use other models if we need further data.

* 1. Normal Stochastic Model

1. [↑](#footnote-ref-1)
2. [↑](#footnote-ref-2)