Team Control Number

12833

Problem Chosen

B

2022
HiMCM/MidMCM
Summary Sheet

Quantitatively Modelling CO_2 and Global Warming

Abstract

Keywords: Global Warming, Greenhouse Gases, CO_2, Forecast, Predictions, Environment, Temperature

Contents

1 Introduction **3**

1.1 Background . . . 3

1.2 Problem Analysis . . . 3

1.3 Keyword Definitions . . . 4

1.4 Assumptions and Justifications . . . 4

1.5 General Variables . . . 5

2 CO_2 - Modelling **5**

2.1 Pre-Analysis . . . 5

2.2 Linear Regression . . . 5

2.3 Exponential Regression . . . 7

2.4 Logistic Regression . . . 8

2.5 Prophet . . . 9

2.6 ARIMA . . . 10

3 CO_2 - Model Evaluation **12**

3.1 Procedure . . . 12

3.2 Accuracy Analysis . . . 13

3.2.1 MSE . . . 13

3.2.2 RMSE . . . 13

3.2.3 MAE . . . 13

3.2.4 PMCC, R^2 AND COVARIANCE . . . 13

3.2.5 Results . . . 14

3.3 Sensitivity Analysis . . . 14

3.4 Overall Evaluation . . . 14

3.5 Possible Improvements . . . 14

4 CO_2 - Results & Conclusions **15**

4.1 Forecast Results . . . 15

4.2 Comparison with External Claims . . . 15

5 Temperature - Modelling **15**

5.1 Time Series 15
5.2 Relationship with CO_2 15

6 Temperature - Model Evaluation **15**
6.1 Predictive Ability 15

7 Temperature - Results & Conclusions **15**
7.1 Forecast Results 15

8 References **15**
8.1 Program Code 15
8.2 Bibliography 15

1 Introduction

1.1 Background

The most significant greenhouse gas on Earth is carbon dioxide, which both absorbs and radiates heat. In contrast to oxygen and nitrogen, which together make up the majority of our atmosphere, greenhouse gases absorb heat emitted from the Earth's surface and re-emit it in all directions, including back toward the planet's surface. The natural greenhouse effect that keeps the Earth's atmosphere above freezing would be insufficient without carbon dioxide. People are accelerating the natural greenhouse effect and raising the earth's temperature by releasing more carbon dioxide into the atmosphere. The NOAA Global Monitoring Lab found that in 2021, carbon dioxide accounted for nearly two thirds of the total heating influence of all greenhouse gases created by humans.

Prior to the Industrial Revolution, carbon dioxide in the atmosphere was consistently around 280 parts per million (ppm). The concentration of CO_2 in the atmosphere reached 377.7 ppm in March 2004, resulting in the largest 10-year average increase up to that time. According to scientists from National Oceanographic and Atmospheric Administration (NOAA) and Scripps Institution of Oceanography (SIO) the monthly mean CO_2 concentration level peaked at 421 ppm in May 2022. An Organisation for Economic Co-Operations and Development (OECD) report predicts a CO_2 level of 685 ppm by 2050.

1.2 Problem Analysis

Problem 1: CO_2 level Forecasting

Modelling.

Prodice models that reflect existing CO_2 data and extrapolates to reasonable predictions.

Choose suitable mathematical models and fit each one to existing data.

Evaluate each model... Use different evaluation approaches including statistical accuracy, contextual reasoning, comparison with external predictions, etc.

Generate a conclusive model based on results obtained... Either pick the "best" model, or create new model based on multiple sub-parts.

Verify External Claims.

Whether CO_2 levels in 2004 had a "larger increase than observed over any previous 10-year period".

Determine how exactly the comparison is done with "any previous 10-year period"... Find supporting evidence from existing literature or make the best interpretation.

Whether CO_2 levels will reach 685ppm by 2050.

Testify this statement against all models.

Draw Conclusions.

Real-world implications based on predictions and results.

Problem 2: Temperature vs CO_2

Modelling.
Prodice models that reflect existing temperature data and extrapolates to reasonable predictions.

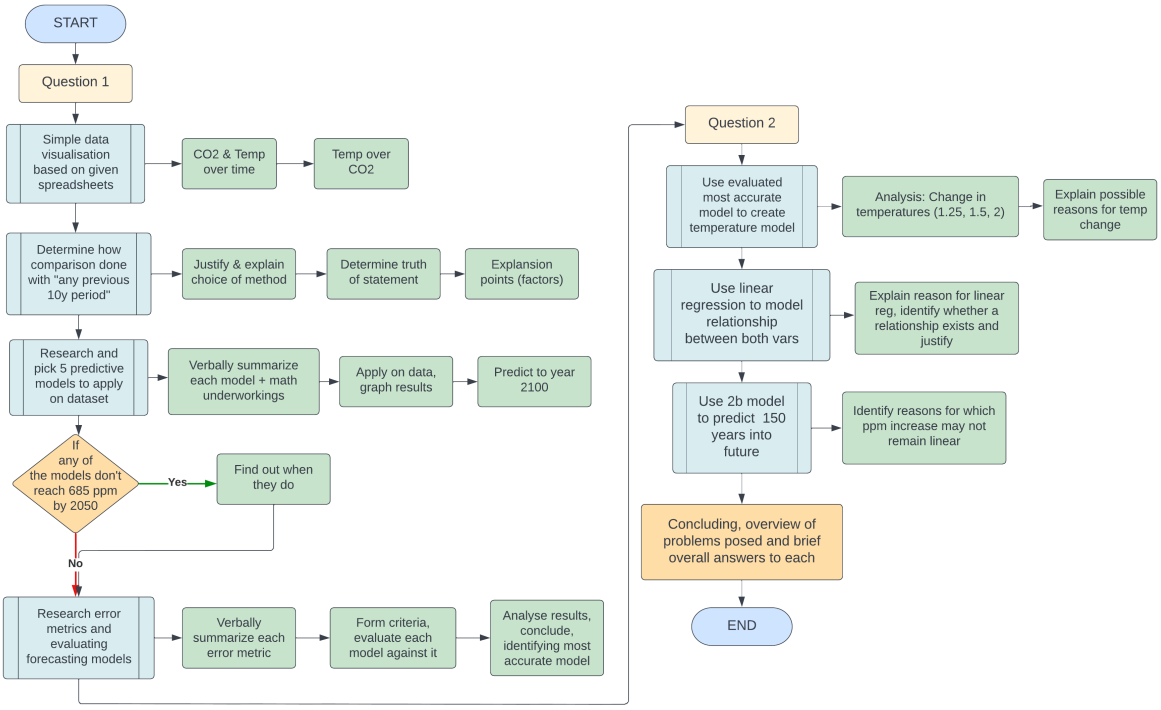
Direct relationship between CO_2 and temperature.

Temperature as a time series, and compare with CO_2 models from Problem 1.

Predictions.
Predict points in time where global temperature will have an average increase of 1.25°C, 1.5°C, and 2°C compared to the base period of 1951-1980.

Evaluation.
Longevity & Confidence; discuss the distance into the future that the model can still predict with reasonable accuracy.

Our thought process and plan is presented here below as a flowchart:



1.3 Keyword Definitions

CO_2 Level/Content: Global average concentration of CO_2 in the atmosphere, in parts-per-million (ppm).

Relative Temperature: Global average temperature, relative to the baseline average temperature from 1951 to 1980, in degrees Celsius.

1.4 Assumptions and Justifications

Assumption 1: CO_2, as a greenhouse gas, has a true case-and-effect relationship with temperature.
Justification: Most existing literature and evidence suggest that CO_2 is a greenhouse gas, and

that greenhouse gases traps heat and leads to increases in temperature on a global scale. This fact is taken for granted to make more confident predictions that relate temperature to CO_2 levels.

Assumption 2: Both datasets provided for CO_2 emissions and temperatures are reliable and accurate. **Justification**: Accurate predictions are always based on accurate data. Since official data is given along the problem, they will be taken and treated assuming the absence of mistakes and errors.

Assumption 3: Statement **Justification**: blah blah

1.5 General Variables

See table 1.1:

Variable	Definition
C_i	CO_2 level at year i
T_i	Temperature at year i
C_i^{-1}	Year where (predicted) CO_2 is at level i
T_i^{-1}	Year where (predicted) Temperature is at i

Table 1.1: General Variables

2 CO_2 - Modelling

This section includes a walk through of all mathematical models used to model the CO_2 levels as a time series and how they specifically apply to the given data.

2.1 Pre-Analysis

Firstly, this problem takes the form of a single-variable time series; CO_2 levels evolve as time passes, and we are trying to model the relationship between time and CO_2. Based on simple logic and scientific reasoning, there is no cause-and-effect relationship between time and carbon dioxide levels, so the models will assume a relationship that is purely a statistical correlation.

The given CO_2 data is graphed to visualize the rough correlation and trend present in the data - 2.1. It is very apparent that there is a strong correlation with minimal variance between time and carbon dioxide levels. The shape of the curve seems exponential. These ideas will be help guide future mathematical modelling. Auto-correlation is calculated to identify any seasonality within the data - 2.2. However, it is apparent that there is no clear seasonality within the data, as shown by the lack of an oscillating correlation. This is expected since the data comes in annual resolutions; yearly seasonality could be expected due to seasonal effects, but would only be observed with monthly data.

2.2 Linear Regression

As the simples model, a linear regression is performed first.

Linear regression is an approach used to model the linear relationship between an independent variable (x) and dependent variable (y), by fitting a linear equation. It is used to present the past values and predict others, in this case future CO_2 emissions. One of the most common and

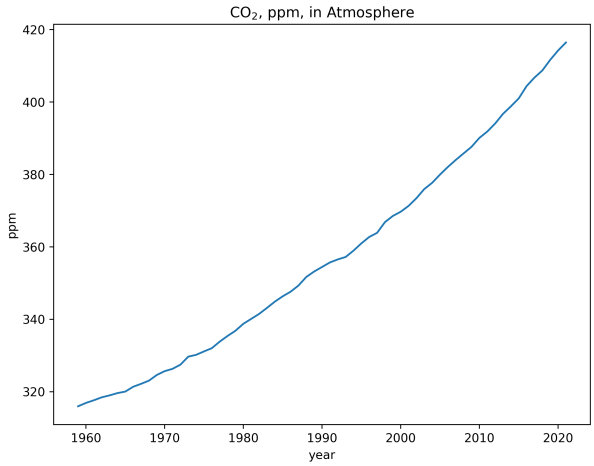


Figure 2.1: Graph of given CO_2 data

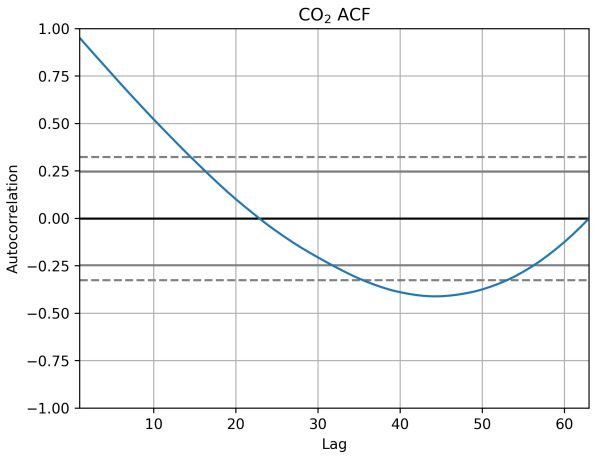


Figure 2.2: CO_2 Auto-correlation

simplest methods used to obtain a regression line is the Ordinary Least Square (OLS) method. In short, OLS minimizes the Square-Error for each point against a given linear function by adjusting the function's parameters, which in the end produces an equation for the linear line of best fit.

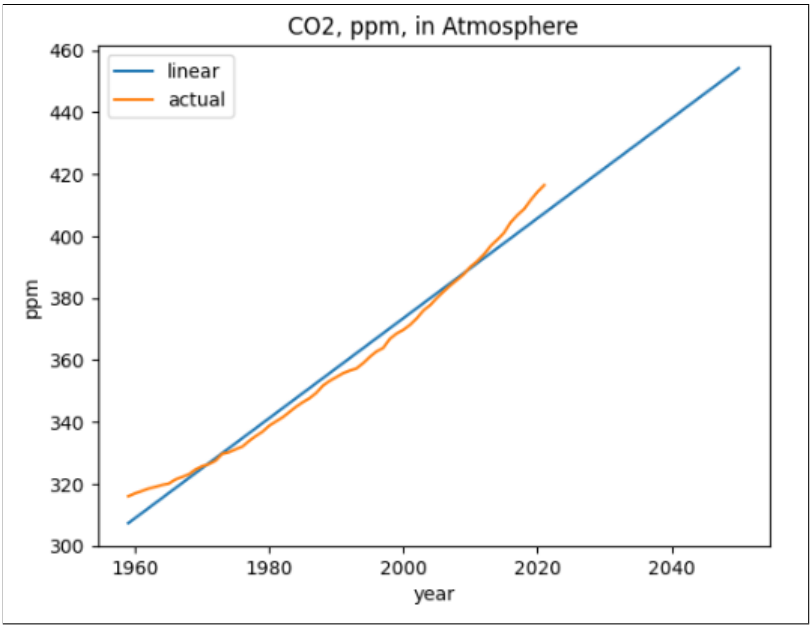
To perform the method, a difference is first set between the dependent variable (CO_2) and its estimation: $(y - \hat{y})$. The difference is then squared, and a summation taken for the entire data set: $S = \sum(y - \hat{y}^2)$. To obtain the parameters that make the sum of square difference become minimum, a partial derivative is taken for each parameter, then equated to 0: $\frac{dS}{d\alpha} = 0$. The final formula for OLS that is obtained is denoted as:

$$S = \sum (y - \hat{y})^2 = \sum (y - ax - b)^2 \tag{2.1}$$

The formula was applied to the given data set to obtain the following line equation:

$$y = 1.6140361x - 2854.59326421 \tag{2.2}$$

y was then set to 685 to find the year at which CO_2 emissions would reach 685ppm; it was found that it would reach this level at the year 2193. The regression line was then graphed and used to predict the next 30 years of CO_2 emissions, the actual data was also plotted for comparison:



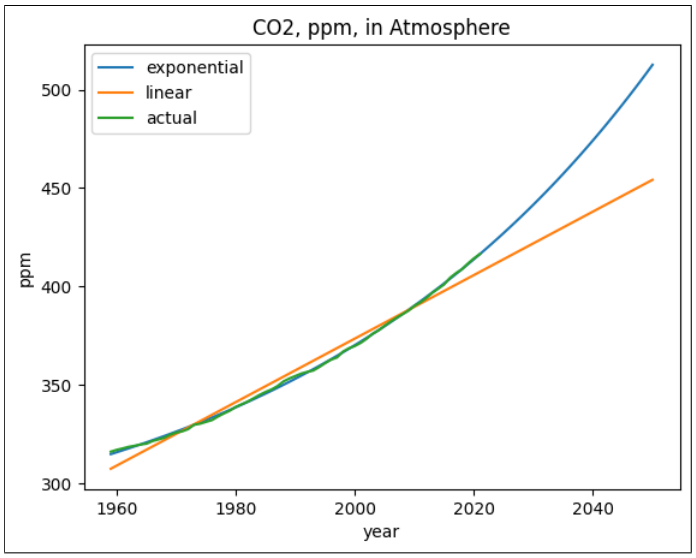
Linear regression is useful in relation to the problem as it is simple to interpret and portray, allowing the prediction of data to be accurate during interpolation. However, if the data to be predicted is outside of the range, such as predicting future CO_2 levels, extrapolation may be inaccurate due to a false assumption of the trend. Furthermore, if the variables plotted provide a non-linear relationship, a linear regression line may inaccurately represent and predict values, which is the case in the data provided.

2.3 Exponential Regression

The second approach decided upon was exponential regression, which models the non-linear relationship between an independent variable (x) and dependent variable (y). This choice was made based on visual indicators of the data given's trend, exhibiting a possible exponential curve. The regression curve portrays data rapidly increasing; at an exponential rate. In exponential regression, the correlation can be denoted as:

$$\log(y) = \log(A) + Bt$$

An exponential function contains a base and an exponent where: $y = ab^x$. This function was then applied to the given data set to form the equation: $y = 1.005717^(2.8427x - 4854.4) + 256.024$. y was then set to 685 to find the year at which CO_2 emissions would reach 685ppm; it was found that it would reach this level at the year 2079. The equation was graphed alongside the actual values and linear regression line for comparison:



Visually, the exponential function better aligns with the actual values; the predicted values also seem to fit with the general trend. An advantage of exponential regression as a predictive model is that it provides high quality forecasts, which increase the accuracy of predicted values during interpolation. However, a key drawback is that a large data set is necessary to carry this method out, as a reasonable amount of continuity is needed to accurately predict future values, especially during extrapolation.

2.4 Logistic Regression

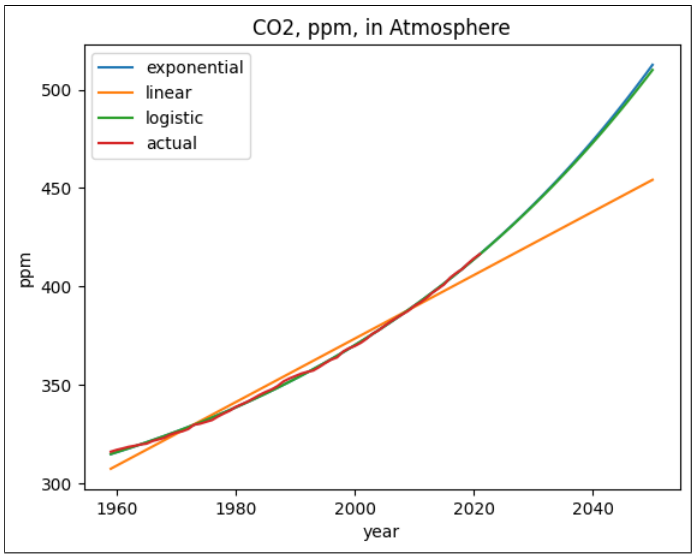
The third predictive model selected shares similarities with the exponential model. One key difference however is that an exponential curve is J-shaped, whereas a logistic curve is sigmoid, the growth rate of the y-variable (CO_2) increases during its lag phase, and eventually reaches a stationary phase. The logistic function has the equation:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

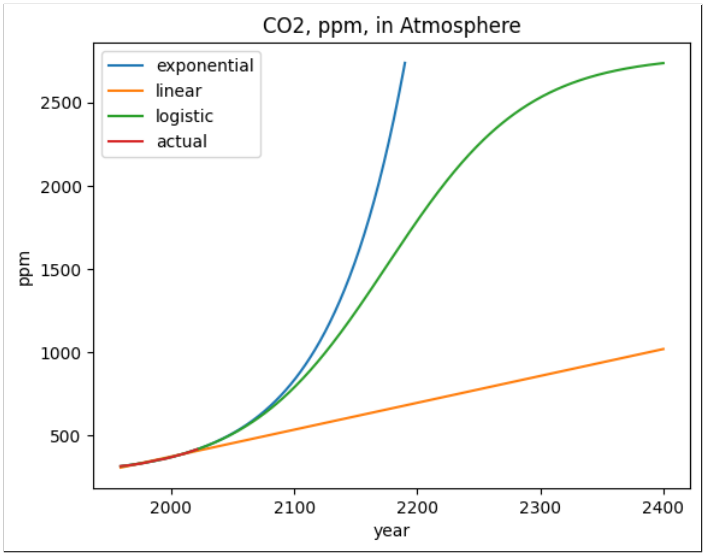
Where:

- L = Curve's maximum value
- k = Logistic growth rate
- x_0 = X value of the sigmoid point
- x = Real number

The function's parameters were then manually adjusted to best suit the current data set, giving the following equation: $y = 2523.6/(1 + e^{(-0.017587*(x-2175.5))}) + 260.0180641636529$. y was then set to 685 to find the year at which CO_2 emissions would reach 685ppm, it was found that it would reach this level at the year 2193. The equation was then graphed alongside with the regression lines of the two previous models to obtain the following graph:



As can be seen, the logistic curve follows a very similar path to that of the exponential function. However, the 1959-2050 year range is too small to depict the sigmoid point, so the functions were graphed to reach the year 2400:



2.5 Prophet

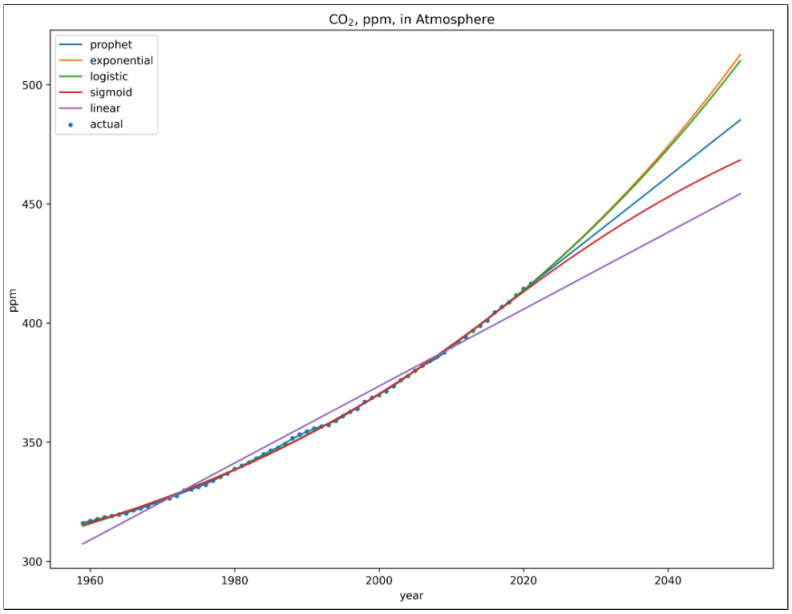
Prophet is a non-linear regression model released by Facebook, it is a procedure for forecasting time series data, working best with series that have strong seasonal effects of historical data. The graph for CO_2 levels over time can be plotted. The regression model is in the form:

$$y_t = g_t + s_t + h_t + \epsilon_t$$

Where:

- g_t = Piecewise-linear trend
- s_t = Seasonal patterns

- h_t = Holiday effects
- ϵ_t = White noise error



Prophet is fast and generally maintains a high accuracy when predicting values. It can be used in a range of different contexts and it is robust to outliers, shifts in the overall trend and missing data. A key disadvantage is that a larger set of data is needed to accurately depict a trend-line.

2.6 ARIMA

An additional time-series prediction model that was investigated was ARIMA, which is based on ARMA (AutoRegressive Moving Average). These models are widely used in datasets that demonstrate non-stationarity, where the series'statistical properties such as mean, variance and autocorrelation change over time. ARIMA assumes the input data to be stationary, so any non-stationary data has to be made stationary through a reversible process. Usually, the transformation involves finding the general trend with methods such as regression and then using differencing to remove the trend from the dataset. With the trend eliminated, an ARIMA model can then be constructed and its optimal parameters found.

Another appropriate model to use in regards to the data set is ARIMA. ARIMA models are generally denoted as ARIMA (p, d, q) where:

- p = Number of Auto-Regressive (AR) terms
- d = Number of differencing
- q = Number of Moving Average (MA) terms

The functions AR(p) and MA(q) are defined below as:

————THERES SUPPOSED TO BE A TABLE HERE————-

Before tuning the parameters p and q, the number of differencing required to make the data stationary must be found out. To evaluate whether the current dataset is stationary, an Augmented Dickey-Fuller (ADF) test was performed.

ADF tests expand on the original Dickey-Fuller test by including higher-order autoregressive processes to form the equation given below:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t$$

Where:

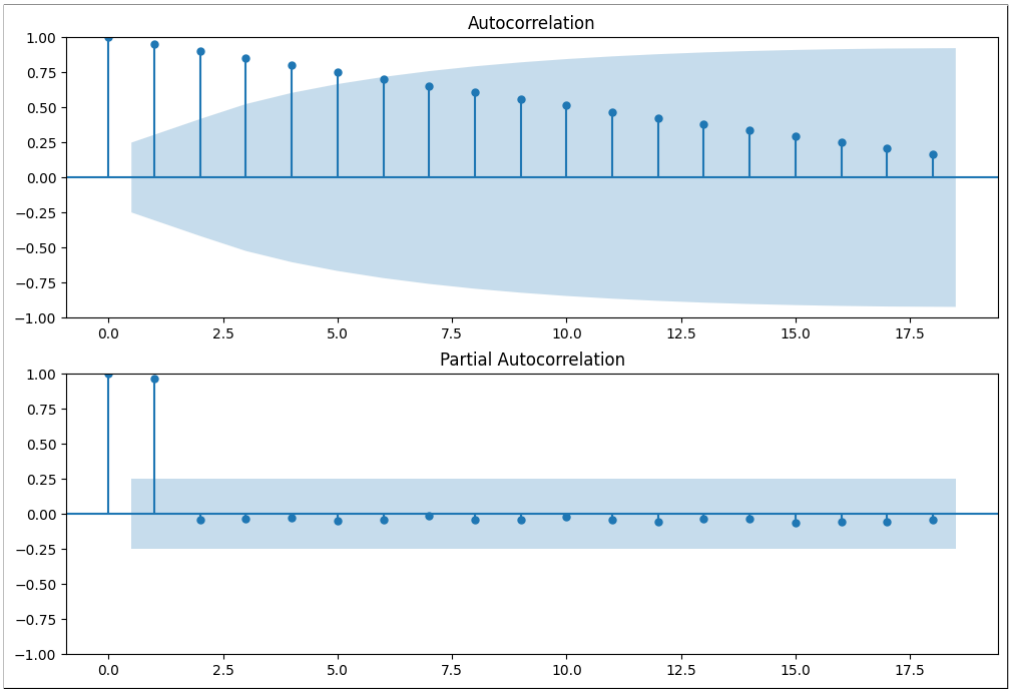
- y_t = Value of the time series at time t
- α = constant
- β = Coefficient on a time trend
- p = Lag order of autoregressive process

The results of the ADF test applied on the given dataset is presented below:

———-THERES SUPPOSED TO BE ANOTHER TABLE HERE———

The p-value obtained is greater than the significance level 0.05 and the ADF statistic is higher than any of the critical values, hence it can be concluded that the time series has a unit root and is non-stationary. The high p-value signifies that a high order of differencing will need to be used.

To further confirm the data's stationarity, autocorrelation and partial autocorrelation graphs were also used. AFC and PAFC functions are measures of correlation between past and present data, and indicate which past data values are most useful in predicting future ones. The results of these functions are then used to select the most optimal parameters for p and q. Both functions were applied on the given dataset and the graphed results displayed below:



For both graphs the x-axis represents lag, whereas the y-axis indicates the correlation strength. ACF graphs represent the correlation between data values that are n intervals apart. PACF graphs are similar in that they represent the same information, however they also account for the values of the intervals in between.

The correlation that can be seen in the AFC graph is negative, indicating that large current values correspond with small values at the specified lag. The absolute correlation values represent the strength of the relationship; to construct an ARIMA model, the expected trend of these values should be random. In this case, the initial relationship between past and present values is strong, but gradually decreases over lags, indicating a clear trend of decreasing correlation strength.

If the autocorrelation follows a random non-linear trend, then AR(p) and MA(q) can be applied to the graphed functions to obtain the optimal parameter values for p and q. However, this is only assuming the data does not have a trend or seasonality component, which does not apply to the given dataset.

The autocorrelation trend as well as the high order of differencing required to transform the data to stationary (as seen in the ADF test) both demonstrate that the given dataset is unsuitable for constructing an ARIMA model, and therefore this prediction model has been rejected and will not be used as part of the predicted CO_2 and temperature values.

3 CO_2 - Model Evaluation

To rank the 4 accepted predictive models on their mathematical accuracy, the following error metrics are to be calculated for each model and then compared with each other:

————ANOTHER TABLE————————

3.1 Procedure

Separation of Known Values into Testing and Training data

75% of known data values are allocated to the testing data group, remaining 25% are allocated to the training group. The predictive model will take in the training group data values as its sole input, and will predict the remaining 25%. The model's predicted values are then to be compared with the training group's values and the forecast errors obtained. A forecast error is defined as the difference between an observed and its forecasted value; the formula for a single forecast error can be modified to suit multiple data values, and it is denoted by the following equation: q1 $e_{T+h} = y_{T+h} - \hat{y}_{T+h|T}$

Where:

- e_{T+h} = Forecast error
- y_{T+h} = Actual value of the h-step observation
- $\hat{y}_{T+h|T}$ = Actual value of the h-step forecast

[ADD GRAPHED RESULTS HERE + CALCULATED FORECAST ERRORS]

It is important to note that although a model may fit the training data well, it does not necessarily mean the model will forecast well, therefore it is important to take the other error metrics into consideration.

3.2 Accuracy Analysis

3.2.1 MSE

Mean Squared Error (MSE) is a measure of the quality of a predictor or of an assessor, its definition differing accordingly. It is the average squared distance between the actual and predicted values, measuring the variance of the residuals. In this case, only the quality of the predictor is to be assessed. It involves taking the average squared distance between the actual and predicted values, measuring the variance of the residuals. The within-sample MSE of a predictor can then be denoted as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y_i}\right)^2$$

Where:

- $\frac{1}{n}\sum_{i=1}^{n}$ = Mean
- $\left(Y_i - \hat{Y_i}\right)^2$ = Squares of the errors

MSE also has a differentiable graph so it makes it easier to perform mathematical operations in comparison to MAE. MSE is more sensitive to outliers compared to MAE.

3.2.2 RMSE

Root Mean Squared Error (RMSE) is the square root of the MSE, measuring the standard deviation of the residuals. The higher the RMSE value, the larger the deviation between actual and predicted values. By proxy, the lower the RMSE value, the lower the deviation, hence the model is more accurate. Building on the last error metric formula, RMSE can be denoted as:

$$\text{RMSE} = \sqrt{MSE}$$

When outliers are exponentially rare, such as this situation, RMSE is generally preferred over MSE as it provides a better evaluation of model performance, as it uses the same units as the Y axis (CO_2 emissions).

3.2.3 MAE

Mean Absolute Error (MAE) is the sum of the absolute difference between the actual and predicted values. A perfect prediction model would yield a 0 as its MAE value. The further away from 0 the MAE value is, the more errors the model makes and hence the less accurate the model is. The formula to calculate MAE is denoted as:

$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n}$$

Where:

- $\sum_{i=1}^{n}|y_i - x_i|$ = Sum of absolute error
- n = Number of errors

3.2.4 PMCC, R^2 AND COVARIANCE

The last 3 error metrics are explained in further detail when exploring the relationship between temperature and CO_2 emissions

3.2.5 Results

The MSE, RMSE, MAE, PMCC, R^2 and covariance were calculated for each accepted model, rounded to 5 s.f, and compiled into a results table below. For each error metric, the model with the value closest to and furthest from a perfect value were highlighted:

——————A TABLE, AGAIN.————————- ——————LO AND BEHOLD, ANOTHER TABLE——————

3.3 Sensitivity Analysis

blah blah

3.4 Overall Evaluation

Strength 1: asdf

Strength 2: asdf

Weakness 1: asdf

3.5 Possible Improvements

blah blah

4 CO_2 - Results & Conclusions

4.1 Forecast Results

we're doomed

4.2 Comparison with External Claims

they're all wrong

5 Temperature - Modelling

5.1 Time Series

Temperature over time.

5.2 Relationship with CO_2

100% causation

6 Temperature - Model Evaluation

6.1 Predictive Ability

our model definitely retains its accuracy 1000 years into the future.

7 Temperature - Results & Conclusions

7.1 Forecast Results

time to get baked.

8 References

8.1 Program Code

Result data generated:

```
text data stuff
```

Python program code:

```
    # pass
```

8.2 Bibliography