Team Control Number

12911

Problem Chosen

B

2022
HiMCM
Summary Sheet

Global warming is one of the most serious challenges the world faces today. Global warming causes drought, extreme weather, heat waves, and rising sea levels that disrupt ecosystems and harm people's livelihoods and communities. Scientists believe that the main cause behind global warming is the greenhouse effect, which occurs when greenhouse gases trap heat in the Earth's atmosphere. CO_2 is the greenhouse gas that scientists agree is the biggest contributor to global warming. A CO_2 concentration level above 500 ppm or a global temperature change of 1.5°C above pre-industrial levels indicates dangerous global warming, resulting in serious negative impacts on the environment and human health such as increased crop failures and spread of vector-borne diseases.

To better understand future global warming trends, we built several models to predict CO_2 concentration levels and global temperatures, as well as to investigate the relationship between them. We first analyzed CO_2 levels and found that March 2003, not March 2004 as claimed by National Oceanographic and Atmospheric Administration, resulted in the largest 10-year average increase up to that time. Then, to model CO_2 concentration levels and forecast future CO_2 level trends, we used Holt's linear trend, Autoregressive Integrated (ARI), Integrated Moving Average (IMA), and Autoregressive Integrated Moving Average (ARIMA) models. All of our models predict that CO_2 levels will not reach 685 ppm by 2050 but will reach at least 500 ppm by 2056. Based on model performance statistics, we found that ARI(8, 2) was the most accurate model, and it forecasts that CO2 levels will reach 500 ppm as early as 2048. Our sensitivity analysis verified that the ARI(8,2) model is robust because its model performance is not greatly affected by the amount of data or using data from different months.

To investigate the trend in global temperatures and forecast future global temperatures, we fit an ARIMA(3, 1, 3) model on global annual mean temperature changes. Our model predicts that the global temperature will change 1.5°C by 2052 and hence cross the threshold for dangerous global warming.

We then explored the relationship between CO_2 concentration levels and global temperatures. A Pearson's correlation coefficient of 0.96 suggests that a very strong positive relationship exists between them. In order to identify their temporal relationships, we analyzed multivariate time series data of CO_2 and temperature using a Vector Autoregressive (VAR) model and Granger causality tests. After performing Granger causality tests on our VAR model, we found that there is a strong Granger causal relationship from CO_2 levels to global temperatures and a weak Granger causal relationship from global temperatures to CO_2 levels. This shows that CO_2 and temperature are interconnected and that CO_2 levels have a statistically significant influence on global temperatures. The final fitted VAR(5) model uses past values of both CO_2 levels and global temperatures to forecast that in 2050, the CO2 level will be 512.85 ppm and the global temperature will change by 2.01°C, both indicating dangerous global warming. We determined that all predictions up to 2100 from VAR(5) should be reliable because of the narrow width of corresponding 95% confidence intervals. The concerns of our VAR(5) model include a small data size and that other factors besides temperature and CO_2 levels were not considered in the model. Our sensitivity analysis confirmed that the model is effective with smaller data sizes, and the predictions of the model are robust and not greatly affected by changing the sampling frequency to monthly or adding in other factors such as CH_4, N_2O, and SF_6.

Based on our research and model results, we conclude that global warming will reach a dangerous level soon, likely by 2050. If we wish to prevent dangerous global warming, CO_2 emissions will need to be greatly reduced, and the energy sector should be transformed to reduce emissions, such as replacing fossil fuels with renewable energy.

Keywords: CO_2, Global Warming, Holt's Linear Trend, ARIMA, VAR, Granger Causality

Global Warming and Carbon Dioxide Levels

Global warming has become increasingly serious over the past few decades, and with significantly increased temperatures, extreme weather, and droughts, climate change has greatly impacted many people's lives.

According to the United Nations, man-made greenhouse gas emissions are the primary cause of current climate trends. Greenhouse gases trap heat in the Earth's atmosphere, resulting in the greenhouse effect and causing the climate to warm abruptly with rapid increases in these gases. With rapidly expanding industrialization around the world, global production of these greenhouse gases has increased significantly. In particular, carbon dioxide emissions, which account for around three quarters of all greenhouse gases in the atmosphere, have recently seen patterns of serious exponential growth.

Using data from the National Oceanographic and Atmospheric Administration, our team created several models to forecast CO_2 levels that all verified that CO_2 levels will reach at least 500 ppm by 2056. A CO_2 concentration level of 500 ppm is well above 350 ppm, what scientists and government officials consider the safe level of carbon dioxide. CO_2 is agreed to be the biggest contributor to global warming and such a large quantity of carbon dioxide would result in dangerous global warming.

Another one of our models, created using temperature data from the National Aeronautics and Space Administration Goddard Institute for Space Studies, also verifies that global warming will reach dangerous levels soon. Our model predicts that the global temperature will change 1.5°C by 2052, crossing the threshold for dangerous global warming. Under the 2015 Paris Agreement, all countries agreed to try to limit global warming to 1.5°C compared to pre-industrial levels. However, if drastic actions are not taken soon, the current trend shows that this level will be reached by 2052.

We created another model using both CO_2 and temperature data in order to model the relationship between CO_2 levels and global temperatures. This model captures the temporal dynamics between these two factors and forecasts that in 2050, the CO_2 level will be 512.85 ppm and the global temperature will change by 2.01°C, both indicating dangerous global warming.

All of our models provide evidence that global warming will reach dangerous levels soon, likely around 2050, unless immediate measures to reduce greenhouse gas emissions are put into place. These rising temperatures would result in much more than just warmer weather. Sea levels will rise, and coastal cities will be at risk of flooding. Heat waves and extreme weather will occur more frequently. Increasing droughts with longer duration and greater intensity will threaten crops, wildlife, and freshwater supplies. From polar bears in the Arctic to marine turtles off the coast of Africa, biodiversity is at risk because of the changing climate and environment.

As CO_2 levels continue to rise, temperatures will only further increase, and steps must be made to reduce the already piling damages. With around 87 percent of human-produced CO_2 emissions resulting from the burning of fossil fuels, adopting sources of alternative energy would decrease emissions significantly.

Industries should switch their energy sources to using renewable energy, and solar, wind, hydroelectric, and nuclear power have all already seen safe and widespread usage. Public transportation and carpooling should be promoted to help reduce CO_2 emissions. Deforestation should be stopped, and we should work towards making agriculture more environmentally friendly and efficient. If we want to prevent dangerous global warming, effective measures like the ones we suggested need to be implemented immediately to reduce emissions of CO_2 and other greenhouse gases.

Contents

1 Introduction

1.1 Problem Background

Global warming is one of the most pressing issues in the world today [1]. Global warming poses a serious threat to ecosystems and livelihoods, and brings catastrophic weather conditions, among countless other consequences. For example, global warming causes sea levels to rise, resulting in saltwater intrusion that harms freshwater ecosystems. Heat waves caused by climate change can lead to water scarcity, drought, or wildfires, devastating both communities and ecosystems. Furthermore, these climate changes are expected to force over 200 million people to relocate by 2050 [2].

Global warming is mainly caused by greenhouse gases, and carbon dioxide (CO_2) is the primary greenhouse gas that contributes to global warming. CO_2 is released through natural processes like volcanic eruptions as well as through human activities, such as the burning of fossil fuels and deforestation. Although human sources of carbon dioxide emissions are much smaller than natural emissions, they have upset the natural balance, and most of the increase in CO_2 levels seen in the past decades can be attributed to human activities [3]. The primary human activity that contributes to CO_2 emissions is the burning of fossil fuels, which combines carbon with oxygen in the air to form CO_2. Once released into the atmosphere, CO_2 absorbs infrared energy from sunlight and emits it back as heat, greatly contributing to global warming if left in abundance.

Therefore, when studying global warming and evaluating measures to slow climate change, it is essential to predict both future CO_2 levels and temperature trends, as well as analyze the relationship between them.

1.2 Restatement of the Problem

Considering the background information and restricted conditions identified in the problem statement, we need to solve the following problems:

- Examine the patterns in CO_2 levels and forecast future CO_2 levels. Specifically, our sub-objectives are:
  a) Check if we agree with the claim that the March 2004 increase in CO_2 levels was the largest 10-year average increase observed over any previous 10-year period
  b) Use multiple models to examine CO_2 patterns and forecast future CO_2 levels until 2100, such as by which year the CO_2 level will reach 685 ppm
  c) Find the most accurate model for CO_2 level forecasting
- Examine global temperature patterns and forecast future temperature changes until 2100 compared to the average land-ocean temperature from 1951-1980
- Investigate the relationship between CO_2 concentration levels and temperature with sub-objectives listed below:
  a) Build a model to describe the relationship and predict future CO_2 levels and temperatures until 2100
  b) Review model reliability and concerns about forecasting accuracy

1.3 Our work

Our modeling process is shown as in Figure 1:

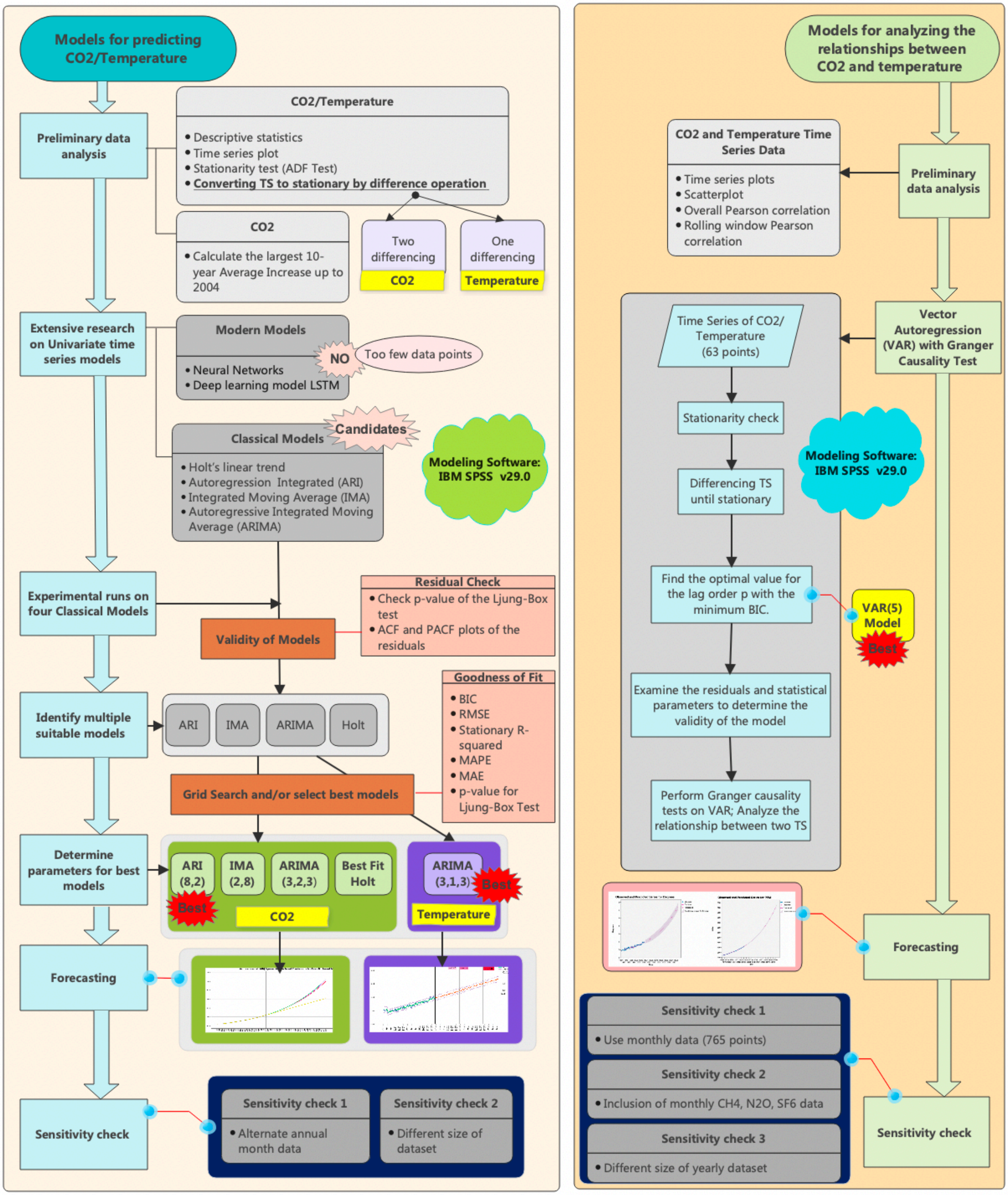


Figure 1: Our modeling process

2 Assumptions and Justifications

To simplify the problem of modeling CO_2 levels and temperature changes, we make the following assumptions.

Assumption 1: CO_2 and temperature levels are predictable.

Justification: Although unpredictable factors can be introduced in the future, we assume that changes in the environment will be gradual and will not drastically change the pattern of CO_2 or temperature levels. Extreme situations that drastically change CO_2 levels or global temperatures have a very small probability of occurring. Therefore, previous data would have autocorrelation with future CO_2 levels or temperatures, and models based on past CO_2 levels or past temperatures can be used to accurately forecast the future.

Assumption 2: The CO_2 levels data and temperature data used to fit the models are accurate.

Justification: The CO_2 levels data used come from the National Oceanographic and Atmospheric Administration (NOAA), which is a federal agency focused on the condition of the oceans and the atmosphere. The temperature data used comes from the National Aeronautics and Space Administration Goddard Institute for Space Studies (NASA GISS), which is an institute that conducts research in astrophysics, planetary atmospheres, and the climate. These institutions are very prestigious, and the data they collect are usually thoroughly inspected and very accurate.

Assumption 3: The variation in annual March averages of CO_2 levels can represent yearly variation in CO_2 levels.

Justification: The amount of CO_2 found in the atmosphere varies over the course of a year mainly because of the role of plants in the carbon cycle [4]. During spring and summer in the Northern Hemisphere, which has a greater land mass than the Southern Hemisphere, plants take up more carbon dioxide through photosynthesis than they release through respiration, leading to a decrease in CO_2 levels. However, yearly variation is typically the same from month to month because CO_2 level changes caused by plants are in an annual cycle. Therefore, variation in annual March averages of CO_2 levels can represent yearly variation, allowing predictions made by models fitted on annual March averages of CO_2 levels to be generalized to the entire year.

Assumption 4: The relationship between CO_2 levels and temperature does not change over time.

Justification: Adding more CO_2 to the atmosphere will always cause surface temperatures to rise [5]. Although the addition of extra CO_2 in the atmosphere gradually becomes less effective at trapping energy, temperature will still rise, and CO_2 levels must be very high for there to be a noticeable decrease in the rate of warming.

3 CO_2 Concentration Levels Data

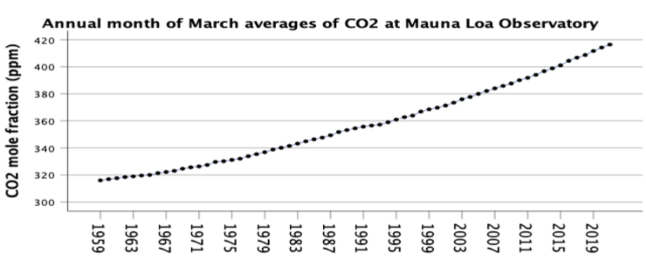


Figure 2: Time series plot of CO2 Concentration

We used CO_2 Data Set 1 to analyze changes in CO_2 levels over time and fit models for predicting future levels of CO_2 in the atmosphere. The CO_2 Data Set 1 is a time series dataset that recorded annual March averages of CO_2 levels from 1958 to 2021, ranging from 315.98 to 416.45 ppm with a mean value of 357. 34 ppm and standard deviation of 29.85 ppm. As shown in Figure 2, CO_2 levels clearly have an increasing pattern over time.

We calculated the 10-year average increase in CO_2 concentration levels using the Compound Annual Growth Rate (CAGR) Formula as follows [6]:

$$CAGR = \left(\frac{EV}{BV}\right)^{1/n} - 1 = \left(\frac{EV}{BV}\right)^{\frac{1}{10}} - 1,$$

where $\boldsymbol{EV}$ = the ending value, $\boldsymbol{BV}$ = the beginning value and $\boldsymbol{n}$ = the number of years.

After analyzing the 10-year average increases in CO_2 levels between 1969 and 2004, we concluded that March 2003, not March 2004, resulted in a larger increase than observed over any previous 10-year period. In 2003, the 10-year average increase was 0.513%, while in 2004, it was 0.510%, as shown in Figure 3. **Therefore, March 2003 resulted in the largest 10-year average increase in CO_2 levels compared to previous years, and we disagree with NOAA that the March 2004 increase of CO_2 was the largest increase up to that time.**

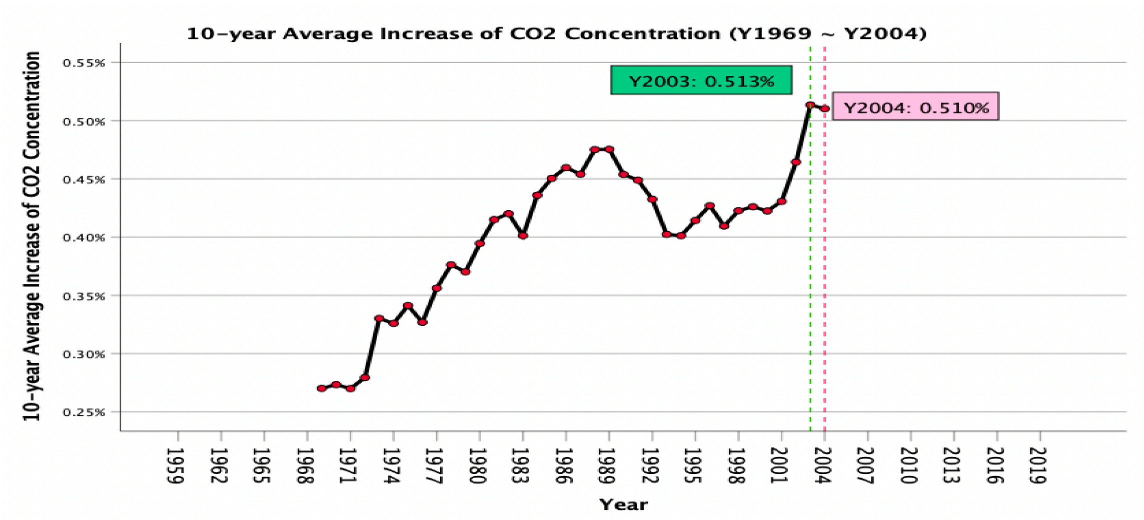


Figure 3: 10-year average increase in CO_2 Concentration (1969 to 2004)

4 Univariate Time Series Models

In this section, we introduce the univariate time series models that we use to model CO_2 concentration levels and global temperatures. These models describe past data and predict future values of a univariate time series, which is a set of observations of a variable recorded over time with equal time increments. When considering which models to use, we eliminated the possibility of using machine learning methods, such as neural networks and long short-term memory (LSTM) methods, because the small amount of data we had would result in poor accuracy.

We decided to use the following four models to forecast CO_2 levels:

- **Holt's linear trend**
- **Autoregressive Integrated (ARI) model**
- **Integrated Moving Average (IMA) model**
- **Autoregressive Integrated Moving Average (ARIMA) model**

4.1 Holt's Linear Trend Model

Holt's linear trend method is a popular exponential smoothing model for forecasting data with a trend [7]. Holt's linear trend method continually revises a forecast based on recent data. Exponentially decreasing weights are assigned to data as they get older, resulting in recent data having more weight in forecasting than older data. Holt's linear trend method also results in the smoothing of random variability in the data. There are three equations used in the process: the first equation for smoothing the time series, the second equation for smoothing trend, and the third equation is a combination of the other two equations and used for forecasting. Two parameters are used: one for overall smoothing and the other for trend smoothing.

The equations to describe y_t, the value of the time series at time t are defined as:

Level equation $l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1})$

Trend equation $b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$

Forecast equation $\hat{y}_{t+h|t} = l_t + hb_t,$

where $\hat{y}_{t+h|t}$ is the forecast for the value of the time series at time $t+h$, l_t is the estimate of the level of the time series at time t, b_t is the estimate of the trend of the time series at time t, α is the smoothing parameter for the level $(0 \le \alpha \le 1)$, and β is the smoothing parameter for the trend $(0 \le \beta \le 1)$

4.2 ARI, IMA, and ARIMA Models

In this section, we introduce information on ARIMA models, as well as ARI and IMA models, which can be viewed as a subset of ARIMA models [8].

4.2.1 Stationarity

A stationary time series is a time series where the mean, standard deviation, and covariance do not vary with time. Stationarity means that the process generating the time series does not change over time. Stationarity is an assumption in many time series models because models predict stationary series more effectively.

Augmented Dickey Fuller Test

Augmented Dickey Fuller test (ADF test) is a statistical test used to test whether a time series is stationary. ADF test is a unit root test that tests the null hypothesis that a unit root, a characteristic that makes a time series non-stationary, is present in a time series.

Differencing

A non-stationary time series can be made stationary through differencing, which is computing the differences between consecutive time periods. For example, one order differencing for y_t is $y_t - y_{t-1}$. The time series is differenced until it is stationary, which can be determined using the ADF test.

4.2.2 Autoregressive Integrated Model

An Autoregressive Integrated model, or ARI, is a time series model that uses past data to predict future values of a time series. An ARI model can be denoted as $ARI(p, d)$ where p is the autoregressive order and d is the order of differencing.

$ARI\ (p, d)$ for y_t can be expressed as

$$y'_t = c + \phi_1 y'_{t-1} + \phi_2 y'_{t-2} + \cdots + \phi_p y'_{t-p} + \varepsilon_t,$$

where c is a constant, $\phi_1, \phi_2, \dots, \phi_p$ are autoregressive coefficients, ε_t is the error term at time t (generally assumed to be white noise), and y'_t is the value of the d-order differenced time series at time t.

4.2.3 Integrated Moving Average Model

An Integrated Moving Average model, or IMA, is a time series model that uses past forecast errors to predict future values of a time series. An IMA model can be denoted as $IMA\ (d, q)$ where d is the order of differencing and q is the moving average order.

$IMA\ (d, q)$ for y_t can be expressed as

$$y'_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q},$$

where c is a constant, $\theta_1, \theta_2, \dots, \theta_q$ are moving average coefficients, $\varepsilon_{t,} \dots, \varepsilon_{t-q,}$ are the error terms at time $t, \dots, t - q$ and y'_t is the value of the d-order differenced time series at time t.

4.2.4 Autoregressive Integrated Moving Average Model

An Autoregressive Integrated Moving Average model, or ARIMA, is a time series model that combines the ARI model and IMA model into one model. ARIMA uses both past data and forecast errors to predict future values of a time series. An ARIMA model can be denoted as $ARIMA\ (p, d, q)$ where p is the autoregressive order, d is the order of differencing, and q is the moving average order.

$ARIMA\ (p, d, q)$ for y_t can be expressed as

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t,$$

where c is a constant, $\phi_1, \phi_2, \dots, \phi_p$ are autoregressive coefficients, $\theta_1, \theta_2, \dots, \theta_q$ are moving average coefficients, $\varepsilon_{t,} \dots, \varepsilon_{t-q,}$ are the error terms at time $t, \dots, t - q$, and y'_t is the value of the d-order differenced time series at time t.

4.3 Evaluation of the model performance

Popularly used model performance measures are listed in Table 1 [8]. A model with high forecasting accuracy would be one with small RMSE, MAPE, MAE, and BIC and a high stationary R-squared and p-value for Ljung-Box Test. In this report, we used SPSS v29.0 for fitting these univariate time series models and calculating these measures. Statistical significance is set at p-value < 0.05.

Table 1: Statistics for Fitness of models

Statistic	Abbreviation	Optimal Range
Stationary R-squared		Higher values
Root Mean Square Error	RMSE	Lower values
Mean Absolute Percentage Error	MAPE	Lower values
Mean Absolute Error	MAE	Lower values
Bayesian Information Criterion	BIC	Lower values
p-value for Ljung-Box Test		Higher values

5 Models for Forecasting CO_2 Levels

To predict future levels of CO_2 in the atmosphere, we fit the univariate time series models described above on the CO_2 Data Set 1.

5.1 Holt's Linear Trend

We fit Holt's linear trend model on the CO_2 Data Set 1 and found α to be 0.968 and β to be 0.224.

Table 2: Parameter estimation results for Holt's linear trend model

Parameter	Estimate	Standard Error	T-value	p-value
α	0.968	0.128	7.543	< 0.001
β	0.224	0.099	2.252	0.028

Because the p-values for the model parameters are less than 0.05, each term in our model is statistically significant, indicating the validity of our model.

We also analyzed the autocorrelation and partial autocorrelation plots (Figure 4) of the residuals to check the validity of our model. Because the ACF and PACF plots indicate that most of the lags fall within the 95% confidence interval, the residuals appear to be mostly white noise, indicating that the model fit is appropriate.

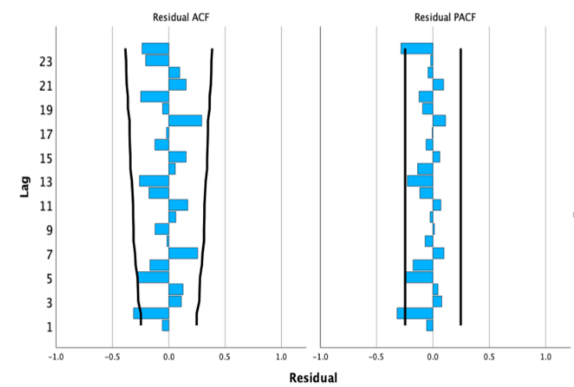


Figure 4: Residual ACF and PACF plots

Our model can then be expressed as:

$$l_t = 0.968y_t + 0.032l_{t-1} + 0.032b_{t-1},$$

$$b_t = 0.224l_t - 0.223l_{t-1} + 0.776b_{t-1},$$

$$\hat{y}_{t+h|t} = l_t + hb_t,$$

where $\hat{y}_{t+h|t}$ is the forecast for the value of the time series at time $t + h$, l_t is the estimate of the level of the time series at time t, b_t is the estimate of the trend of the time series at time t and y_t is the value of the time series at time t.

5.2 ARI, IMA, and ARIMA Models

We then fit ARI, IMA, and ARIMA models on the CO_2 Data Set 1. Because the data of the CO_2 levels was not stationary, we differenced the data twice until stationarity was reached, as tested by the ADF test.

Once the stationary assumption of the models was satisfied, we found the optimal parameters for each of the three models and then analyzed the ACF and PACF plots of the residuals to check model validity. For all the models, the lags are within the 95% confidence interval, indicating that there is no remaining pattern. Additionally, the p-values from the Ljung-Box test on the residuals of the three models are all much greater than 0.05, validating the model assumption that the residuals are not autocorrelated and therefore indicating that the model fit is appropriate. The final models found are ARI(8, 2), IMA(2, 8), and ARIMA(3, 2, 3).

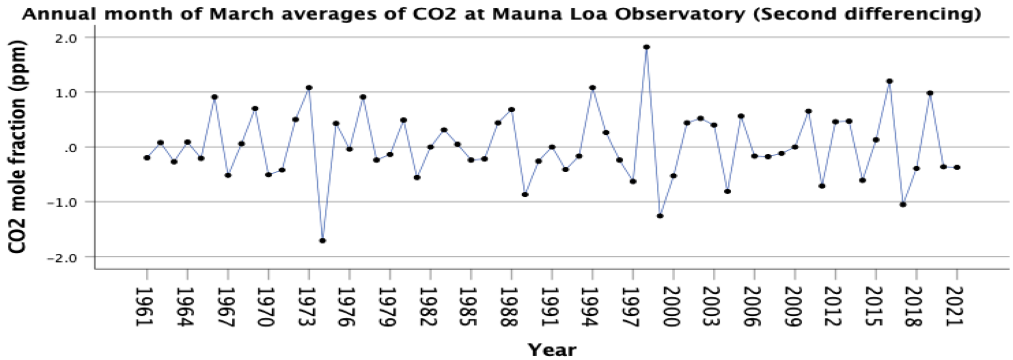
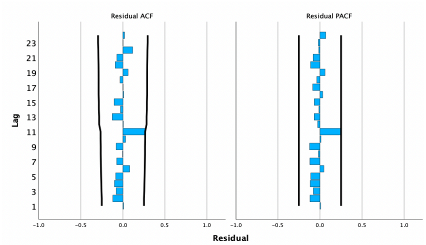
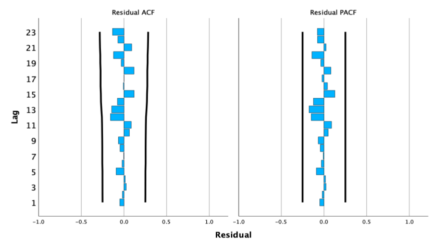


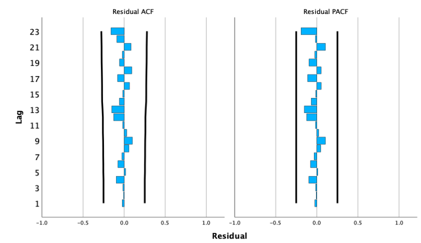
Figure 5: Second difference plot of annual March averages of CO_2 levels



ARI(8,2)



IMA(2, 8)



ARIMA(3, 2, 3)

Figure 6: Residual ACF and PACF plots for ARI(8,2), IMA(2,8) and ARIMA(3, 2, 3)

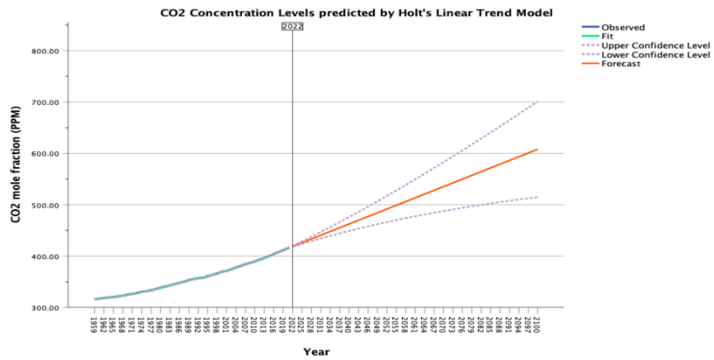
5.3 Forecasting

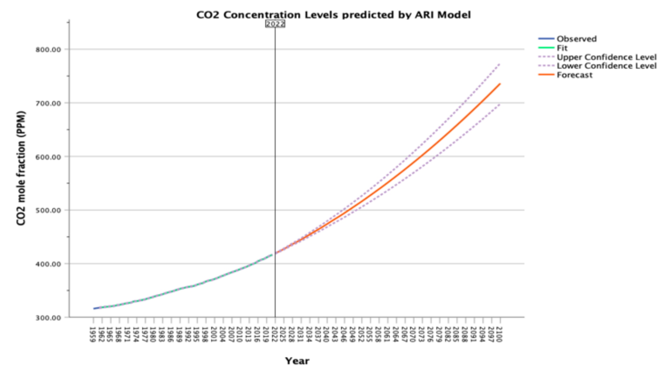
Once we fit Holt's linear trend, ARI, IMA, and ARIMA models on the CO_2 levels data, we then used these models to forecast future CO_2 levels as shown in Figure 7. Table 3 reports the forecasted CO_2 levels in 2050, 2075, and 2100, along with their corresponding 95% confidence intervals.

Table 3: Model predictions

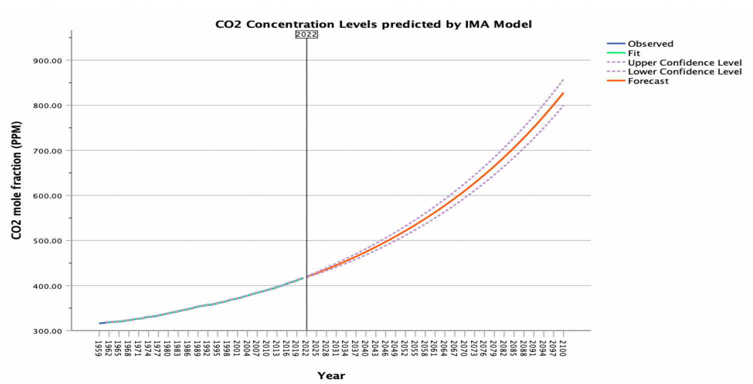
Model	Year	Predicted CO_2 level (ppm)	Lower 95% Confidence Limit	Upper 95% Confidence Limit
Holt's Linear Method	2050	486.89	463.46	510.13
	2075	547.43	492.99	601.87
	2100	608.06	514.84	701.28
ARI(8,2)	2050	508.95	499.51	518.38
	2075	611.44	589.32	633.56
	2100	736.21	698.25	774.17
IMA(2,8)	2050	513.16	503.63	522.86
	2075	639.74	622.14	657.7
	2100	828.15	799.38	857.7
ARIMA(3,2,3)	2050	511.76	502.95	520.69
	2075	635.78	618.75	653.17
	2100	819.09	789.44	849.56

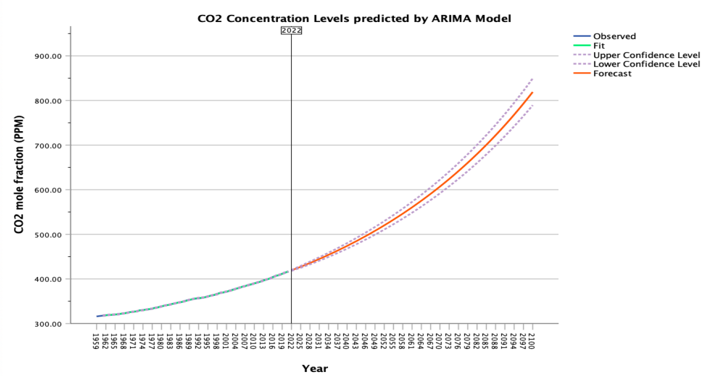
(1) (2)





(3) (4)





(5)

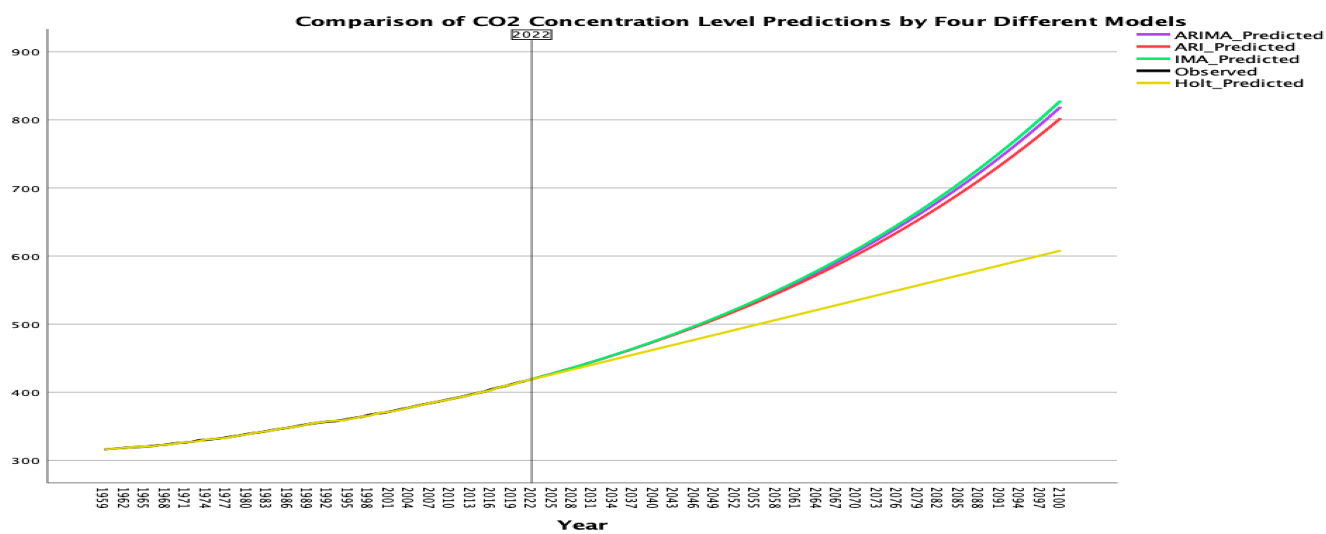


Figure 7: Forecasting plots - (1) Holt's linear trend (2) ARI (3) IMA, and (4) ARIMA models on the CO_2 levels data, and (5) Four models combined

We found that **none of our models agreed with the Organization for Economic Co-Operations and Development (OECD) that CO_2 levels will reach 685 ppm by 2050** [9]. Holt's linear trend predicted that CO_2 levels will reach 685 ppm by 2132, ARI(8, 2) predicted by 2091, IMA(2, 8) predicted by 2082, and ARIMA(3, 2, 3) by 2083.

The results from our models might differ from the findings of the OECD because we only considered data on CO_2 levels while the OECD also considered other factors, such as energy demand, that would affect CO_2 levels. As well, the statement made by the OECD that CO_2 levels will reach 685 ppm by 2050 was made based on data earlier than 2012. There are likely new patterns in the data we used, which included data up to 2021, resulting in our models predicting different patterns from the OECD.

5.4 Most Accurate Model

Table 4: Comparison of Model Statistics

Model	Stationary R-squared	RMSE	MAPE	MAE	Normalized BIC
Holt's linear trend	0.388	0.494	0.101	0.363	-1.28
ARI(8, 2)	0.8	0.318	0.066	0.232	-1.417
IMA(2, 8)	0.555	0.453	0.089	0.316	-0.977
ARIMA(3, 2, 3)	0.632	0.412	0.086	0.307	-1.233

*All the p-values from the Ljung-Box test were well above 0.05.

Comparing the time series models we used using the model performance measures in Table 4, we found that all the ARI(8, 2) model's statistics were better than the other models. We therefore concluded that **ARI(8, 2) is the best fit for the CO_2 Data Set 1** and is expected to have the best forecasting accuracy.

5.5 Sensitivity Analysis

5.5.1 Sensitivity Analysis on Data Size

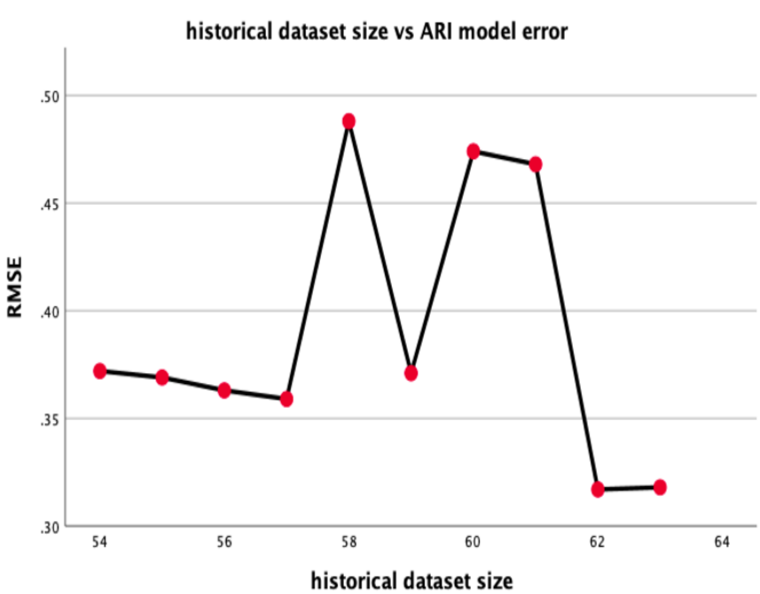


Figure 8: RMSE of ARI model using different dataset size

To judge the robustness of the ARI(8, 2) model, we explored the effect that data size has on the performance of the model in forecasting CO_2 prediction. The data we used to fit the model on only has 63 years of data, which is considered relatively small. Less data to fit on often decreases model performance, and we therefore run the ARI(8, 2) model on less data than before to analyze the impact on model performance.

As shown in Figure 8, as the amount of data is decreased from 63 time points to 54, there is a slight increase in model error. However, RMSE does not increase by a significant amount and we therefore determined that the ARI(8, 2) model still has high performance when the amount of data is decreased by up to 15%.

5.5.2 Sensitivity Analysis on Alternate Month Data

We fit the ARI(8, 2) model on annual March averages of CO_2 levels using Assumption 2 in section 2. To check if the ARI(8, 2) model fitted on annual March averages also describes the patterns in CO_2 levels from other months well, we fit this model using data obtained from NOAA on August averages of CO_2 levels, which represents a different season from March.

Table 5: Comparison of Model statistics for annual March vs August averages of CO_2 levels

Dataset	Stationary R-squared	RMSE	MAPE	MAE	Normalized BIC
Annual March averages	0.8	0.318	0.066	0.232	-1.417
Annual August averages	0.736	0.554	0.116	0.413	-0.458

When fitting ARI(8, 2) on annual August averages, the model performance is still satisfying. The performance statistics of this model is not much different from the model using annual March averages of CO_2 levels. Therefore, we conclude that using data from a different month has little impact on the ARI(8, 2) model performance.

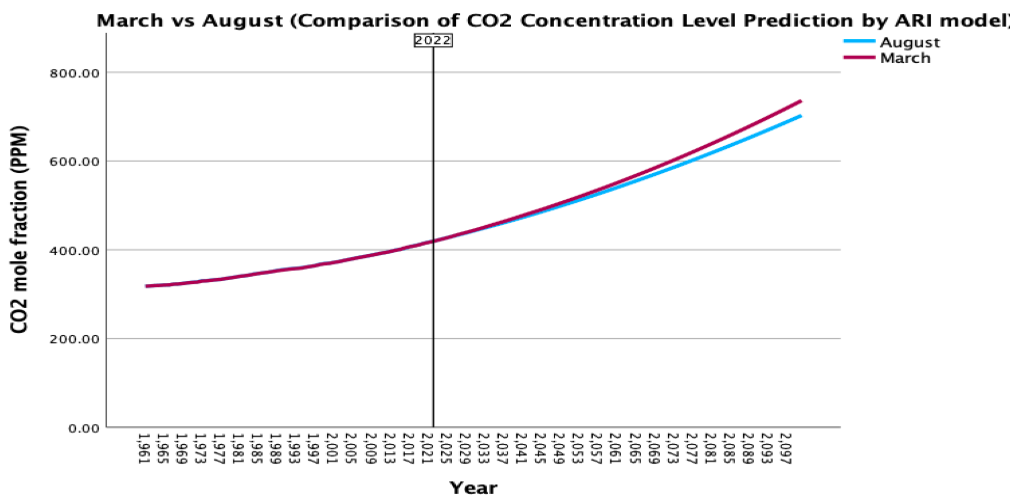


Figure 9: CO_2 Level Prediction by ARI(8, 2) using annual March vs August data

5.6 Model Strengths and Weaknesses

Model Strengths

1. The ARI(8, 2) model can account for non-stationary patterns.
2. The ARI(8, 2) model only requires univariate time series data to predict future values and can forecast without information on other factors.
3. The ARI(8, 2) model typically has powerful short-term prediction abilities.
4. The ARI(8, 2) model is computationally inexpensive, and future predictions and confidence levels can be easily obtained.
5. The sensitivity analysis of the model demonstrates its effectiveness with different amounts of data and alternate annual month data, proving the robustness of the model.

Model Weaknesses

1. The ARI(8, 2) model loses accuracy for long-term forecasting because the model is dependent on the accuracy of previous values.
2. The forecasting accuracy of the ARI(8, 2) model depends on the reliability of historical data and on future conditions being similar to the conditions at the time of the data used.
3. No other factors, such as energy demand, that affect CO_2 levels were considered in predicting CO_2 levels in our model.

6 Model for Forecasting Temperature Changes

To analyze temperature changes and predict future land-ocean temperatures, we fit a univariate time series model on global annual mean land-ocean temperatures.

We used Temps Data Set 2 to fit a model for predicting future temperature changes. Temps Data Set 2 is a time series dataset of global annual mean land-ocean temperature changes from 1958 to 2021 compared to the average temperature from 1951 to 1980. This dataset ranges from -0.2°C to 1.02°C and has a mean temperature change of 0.35°C and standard deviation of 0.32°C. The global mean land-ocean temperature is increasing overall with time as shown in Figure 10.

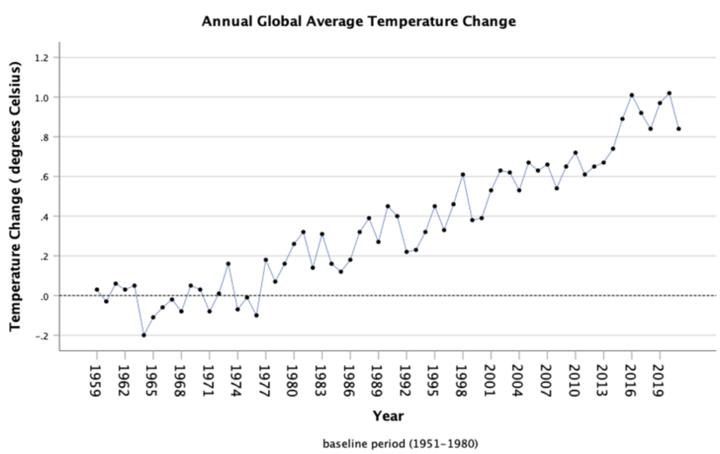


Figure 10: Time series plot of Temperature Change

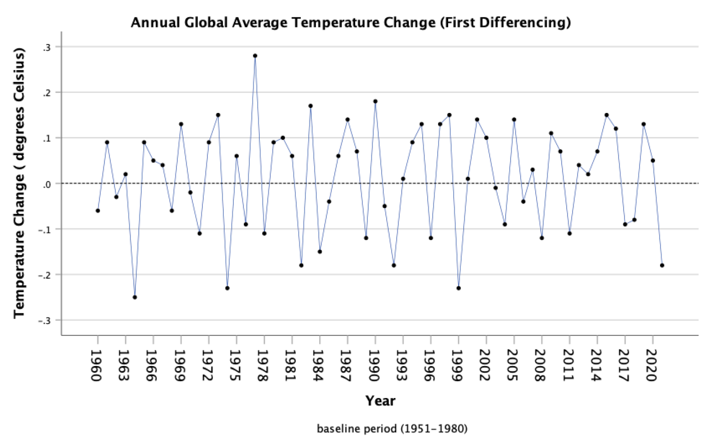


Figure 11: First difference plot of Temperature Change

We chose to use an ARIMA model for forecasting temperature changes after comparing the performance of Holt's linear trend, ARI, IMA, and ARIMA models in temperature forecasting. To build the model, we first differenced the temperature data once to reach stationarity (as shown in Figure 11), as tested by the ADF test.

Once the stationarity assumption of the models was satisfied, we found the optimal parameters for the ARIMA model and then analyzed the ACF and PACF plots of the residuals shown in Figure 12 to check model validity. All the lags are within the 95% confidence interval, as shown in the ACF and PACF plots, and the p-value for the Ljung-Box test on the residuals is much greater than 0.05, indicating that the assumption that the residuals are not autocorrelated is not violated. Other model performance measures shown in Table 5 suggest that the model describes this temperature time series well. The final model found is ARIMA(3, 1, 3).

Table 5: ARIMA(3, 1, 3) Statistics

Stationary R-squared	RMSE	MAPE	MAE	Normalized BIC
0.535	0.085	52.095	0.067	-4.398

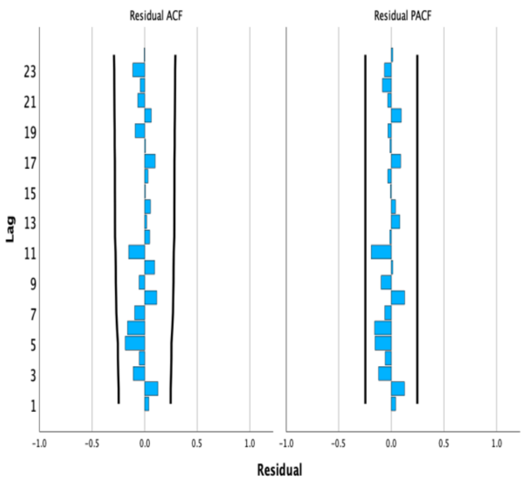


Figure 12: Residual ACF and PACF plots for ARIMA

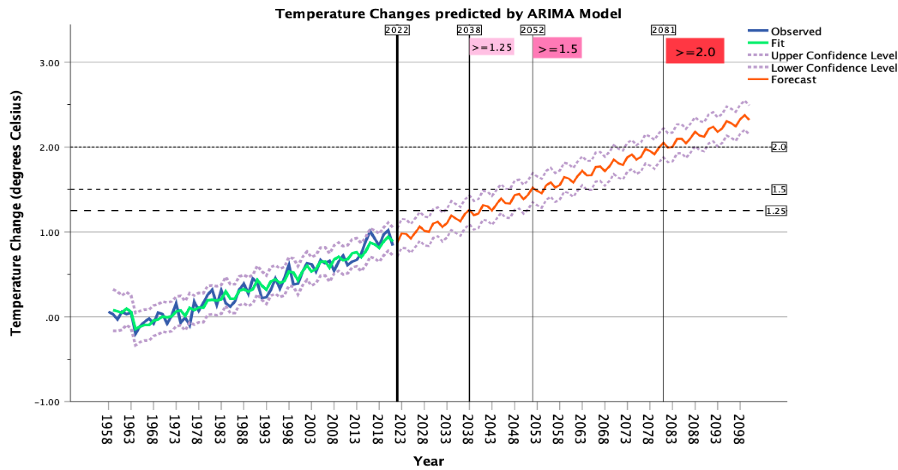


Figure 13: Forecasting of ARIMA models on the Temperature data

We then used our ARIMA(3, 1, 3) model to forecast future land-ocean temperature changes as shown in Figure 13. According to our model's predictions, the average land-ocean temperature will change by 1.25°C in 2038, 1.5°C in 2052, and 2°C in 2081 when compared to the average temperature from 1951 to 1980. Our predictions show that the threshold for dangerous global warming will likely be crossed by 2052, resulting in serious negative impacts on the environment and human health.

7 Relationship between CO_2 and Temperature

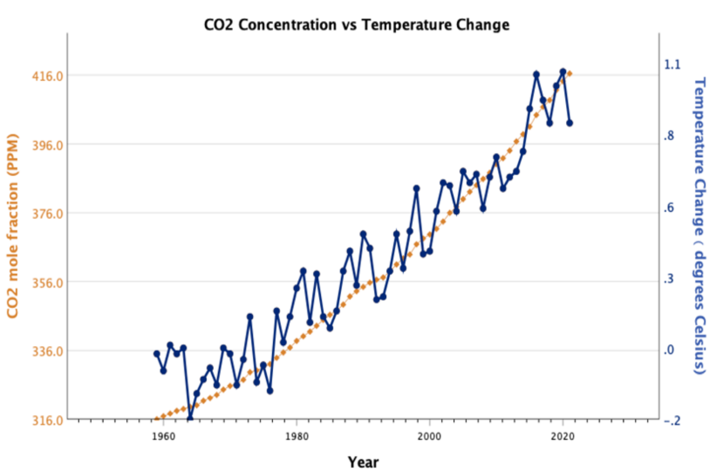


Figure 14: Time series plots (CO_2 vs Temperature)

Many scientists think that there is a relationship between warming global temperatures and the concentration of CO_2 in the atmosphere. CO_2 is known to be a major contributor to global warming and the primary greenhouse gas emitted by human activities. According to observations by the NOAA Global Monitoring Lab, carbon dioxide was responsible for about two-thirds of the total heating influence of all human-produced greenhouse gases in 2021. On the other hand, scientists say that temperature also affects CO_2 levels, but human activities contribute to most of the increase in CO_2 levels. Rising temperatures result in warming oceans that release CO_2 into the atmosphere because CO_2 becomes less soluble in warmer water. Warmer temperatures also result in the release of carbon stored in permafrost and increase the frequency of wildfires, which also release CO_2. Figure 14 clearly implies that CO_2 levels and temperature co-vary over time and appear to have a positive relationship.

To analyze the relationship between CO_2 levels and temperature, we first examine the overall relationship through Pearson's correlation. Then, to forecast the future and explore the possible intercausal relationship between CO_2 levels and temperature, we build a Vector Autoregression (VAR) model, which can discover key temporal relationships in multivariate time series data.

7.1 Pearson Correlation

The Pearson correlation measures the strength and direction of the linear relationship between two variables. The Pearson correlation coefficient (r) has a value between -1 and 1, with a value of -1 meaning a perfect negative linear correlation, 0 being no linear correlation, and 1 meaning a perfect positive correlation.

We normalized the CO_2 levels data and temperature changes data to meet the normality assumption and then calculated r to be 0.96, indicating that CO_2 levels and temperature have a very strong positive correlation.

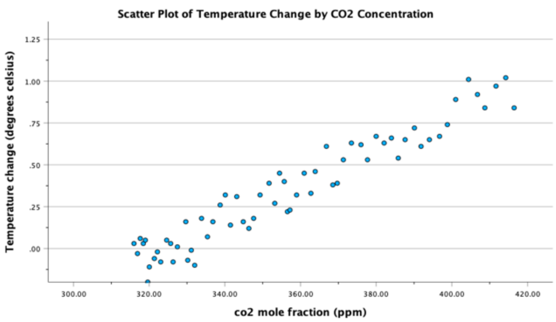


Figure 15: Scatterplot of Temperature vs CO_2

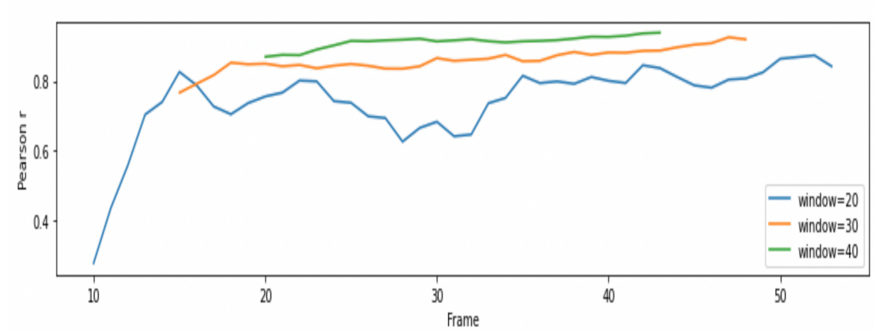


Figure 16: Rolling Window Correlation

We also examined local synchrony using Pearson correlation by measuring the Pearson correlation for a small portion of the data and repeating this process along a rolling window until all the data had been examined. As shown in Figure 16, when the amount of data used is 30 or 40 time points, r already exceeds 0.9, showing that there's a strong positive correlation between CO_2 levels and temperature even over small timespans.

7.2 Vector Autoregression Model

A Vector Autoregression model, or VAR, is a multivariate time series model that is used when two or more time series influence each other [8]. The VAR model can explain relationships among multiple variables over time and predict future observations. The model predicts future values of a time series using past values of the time series along with other related time series. A VAR model can be denoted as $VAR\ (p)$ where p is the lag order.

A 2-dimensional $VAR\ (5)$, which we use to analyze the relationship between CO_2 (y_1) and temperature (y_2), can be expressed as

$$y_{1,t} = c_1 + (\phi_{11,1} y_{1,t-1} + \cdots + \phi_{11,5} y_{1,t-5}) + (\phi_{12,1} y_{2,t-1} + \ldots + \phi_{12,5} y_{2,t-5}) + \varepsilon_{1,t}$$

$$y_{2,t} = c_2 + (\phi_{21,1} y_{1,t-1} + \cdots + \phi_{21,5} y_{1,t-5}) + (\phi_{22,1} y_{2,t-1} + \cdots + \phi_{22,5} y_{2,t-5}) + \varepsilon_{2,t}$$

where c_1 and c_2 are constants, $\phi_{ij,l}$ represents a coefficient, $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$ are error terms, and $y_{1,t}$ and $y_{2,t}$ are the values of the time series y_1 and y_2 at time t.

7.2.1 Granger Causality Test

The Granger causality test is a statistical hypothesis test that is performed within the framework of the VAR model [8]. The Granger causality test helps analyze the relationship between time series in the data. This test determines whether one time series is useful in forecasting another by testing whether values of one variable X provide statistically significant information about the future values of another variable Y. Granger causality is based on the notion that causes precede and help predict their effects. X is said to Granger-cause Y if the VAR model for Y using past values of X is significantly more accurate than the VAR model for Y without X. The Granger causality test uses F-statistic tests to test if the coefficients of all lags of X are equal to zero in the VAR model for Y. If the null hypothesis that all the coefficients of X equal zero is rejected, the Granger causality test concludes that X Granger-causes Y, and past values of X can help explain Y.

7.2.2 Modeling Steps

The following steps can be applied to fit a time series to a VAR model:

1. If any of the time series in the data is non-stationary, the data needs to be differenced until stationary before fitting a VAR model to the data. Stationarity is tested using the ADF test.
2. Find the optimal value for the lag order p by fitting $VAR\ (p)$ models on various values of p and then selecting the value of p associated with the minimum BIC.
3. Examine the residuals and statistical parameters of the VAR model to determine the validity of the model.
4. Perform Granger's causality test to further confirm that all time series in the data significantly contribute to better model fitting. Granger's causality test is also used to analyze the relationship between time series in the data.
5. Invert the transformation of the data in the VAR model to return the data to its original scale. Each difference done to the original data needs to be inverted.
6. Forecast future time series using the model.

7.3 Model for Relationship between CO_2 and Temperature

We built a VAR model to model the relationship between CO_2 levels and temperature. To build our model, we first differenced the data until it was stationary, as tested by the ADF test, and then found the lag order p by varying p and choosing the model with the best normalized BIC. The model we selected to model the relationship between CO_2 levels and temperature was VAR(5). Specifically, VAR(5) shows the following relationship between CO_2 (y_1) and temperature (y_2):

$$\begin{aligned} y_{1,t} = &-18.18 + 0.81y_{1,t-1} + 0.16y_{1,t-2} + 0.16y_{1,t-3} + 0.16y_{1,t-4} - 0.24y_{1,t-5} \\ &+0.86y_{2,t-1} - 2.26y_{2,t-2} - 0.70y_{2,t-3} - 1.13y_{2,t-4} - 0.91y_{2,t-5} \\ y_{2,t} = &-6.04 - 0.03y_{1,t-1} + 0.06y_{1,t-2} + 0.02y_{1,t-3} + 0.02y_{1,t-4} - 0.05y_{1,t-5} \\ &+0.25y_{2,t-1} - 0.26y_{2,t-2} - 0.26y_{2,t-3} - 0.08y_{2,t-4} - 0.36y_{2,t-5}, \end{aligned}$$

The fit statistics for VAR(5) model suggest that it models both temperature and CO_2 levels well, with both BIC values being very low.

Table 6: Fit Statistics

Predicted time series	RMSE	RMSPE	BIC	R-Squared
CO_2 level	0.41	0.00	-70.36	1.00
Temperature	0.08	0.97	-256.79	0.95

We then performed Granger causality tests using our model to further examine the temporal causal relationship between CO_2 levels and temperature.

Table 7: Granger causality tests results

Output Series	Input Series	p-value for Granger causality
CO_2 levels	CO_2 levels	0
	Global temperatures	0.011
Global temperatures	CO_2 levels	1.033E-5
	Global temperatures	0.021

The Granger causality tests show that when predicting CO_2 levels, past CO_2 levels are the most significant predictor (p-value = 0) while past temperatures contribute less to the prediction (p-value = 0.011). However, when predicting global temperatures, past CO_2 levels are more important in prediction (p-value = 1.033E-5) than temperature itself (p-value = 0.021). Therefore, there is strong Granger causality from CO_2 levels to global temperatures, but weak Granger causality in the opposite direction. These results indicate that CO_2 levels and global temperatures both influence the other, but CO_2 levels have a greater effect on temperature than temperature does on CO_2 levels.

7.3.1 Forecasting using VAR(5)

Using our VAR(5) model, we forecast future CO_2 levels and global temperatures until 2100.

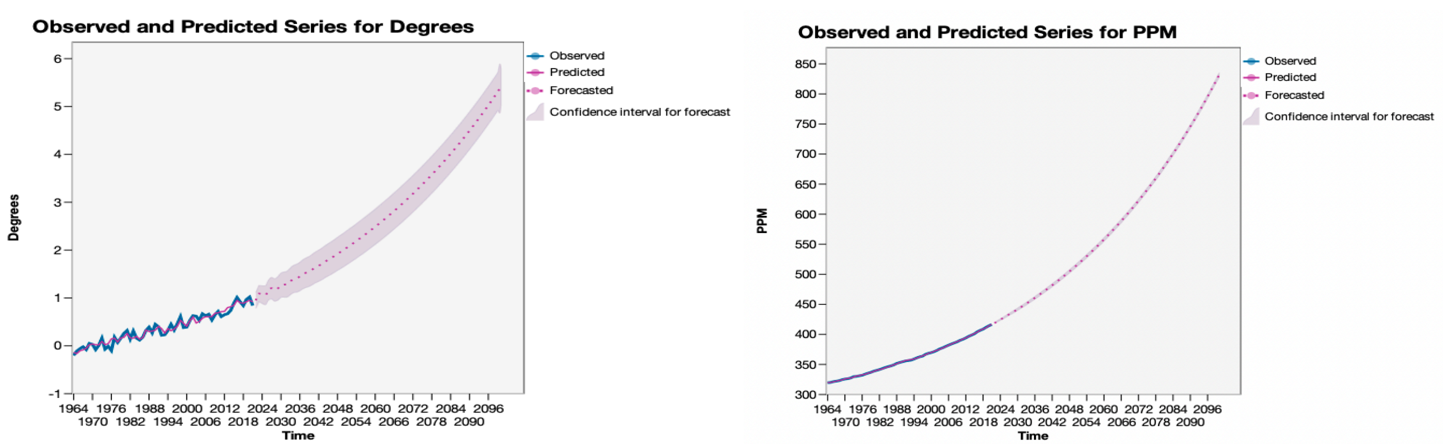


Figure 17: Future CO_2 levels and global temperatures forecasting till 2100 using VAR(5)

The predictions from VAR(5) should be reliable because the model has good fit statistics and small confidence intervals for the forecast. As well, the model considers the mutual influence between CO2 and temperature. After analyzing our predictions up to 2100, we believe that our model is reliable up to 2100 because we have a small margin of error for all the predictions. The 95% confidence intervals shown in Table 8 suggested that we are 95% confident that the predicted temperature should be within 0.395°C of the actual temperature up to 2100, and the predicted CO_2 level should be within 3.055 ppm of the actual CO_2 level up to 2100.

Table 8: Predicted CO_2 level and global temperature change in years 2050, 2075 and 2100 using VAR(5)

Time Series	Year	Prediction	Lower 95% confidence limit	Upper 95% confidence limit
CO_2 level	2050	512.85	510.14	515.55
	2075	640.56	637.58	643.54
	2100	831.36	828.31	834.42
Global temperature	2050	2.01	1.66	2.36
	2075	3.37	2.99	3.76
	2100	5.41	5.01	5.8

7.4 Sensitivity Analysis

The amount of data used to fit VAR(5), 63 time points, is relatively small. More historical data results in better model performance because the model coefficients are better optimized to model the relationship between CO2 levels and temperature. In addition, our VAR model does not consider other factors besides CO2 levels and global temperatures. Other variables that influence CO2 levels and temperature would provide more information and increase forecasting accuracy. To evaluate the seriousness of these concerns that we have with our model, we performed several sensitivity analyses.

7.4.1 Sensitivity Analysis on Data Size

To verify that our model performance is not greatly affected by data size, we explored the effect that data size has on the RMSE and BIC of VAR(5).

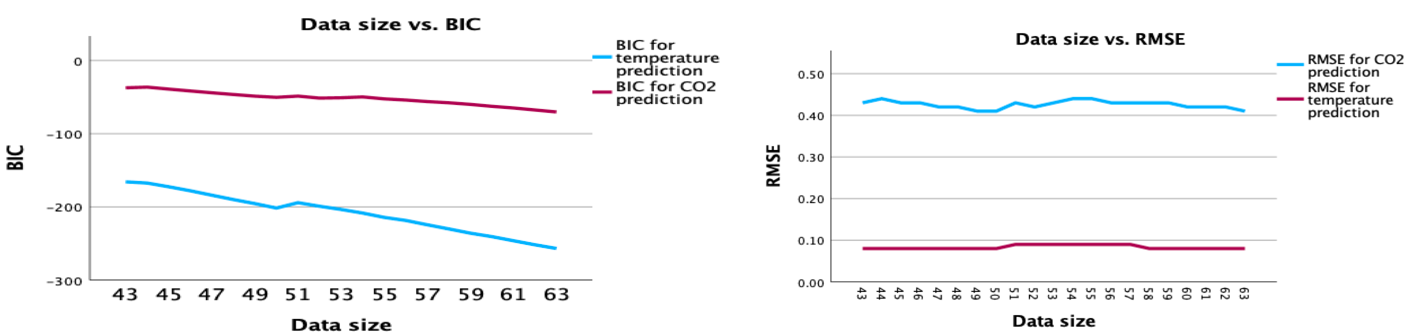


Figure 18: BIC and RMSE of VAR(5) for different data size

As shown in Figure 18, as the amount of data is decreased from 63 time points to 43, RMSE and BIC worsens slightly. However, RMSE and BIC both do not worsen significantly, and model performance is still high. Therefore, we determined that it has little impact on our model performance when the data size decreases within a range, and concerns over the amount of data available are not major.

We also tested how Granger causal relationships in the model were affected by data size. To test the effect of data size on Granger causation, we fit the VAR(5) model on different amounts of data and examined the p-values from the Granger causation tests. We found that the p-value of the Granger causation test for CO_2 levels predicting temperature remained zero when the data size was from 63 to 56 time points and was 0.001 when the data size was from 55 to 43 time points. This shows that even with a smaller timespan, there is always strong Granger causality from CO_2 levels to temperature.

However, the p-value of the Granger causation test for temperature predicting CO_2 levels is greater than an alpha level of 0.05 when the data size is decreased to 59 time points. Therefore, there was not enough evidence to support the weak Granger causation from temperature to CO_2 levels when the data size is decreased by 4.

Table 9: p-value of the Granger causation test for temperature predicting CO_2

Data size	63	62	61	60	59
p-value	0.011	0.017	0.048	0.048	0.066

Therefore, data size has little influence on the strong Granger causation from CO_2 levels to temperature, and a data with size greater than 59 time points still preserves the weak Granger causation from temperature to CO_2 levels.

7.4.2 Sensitivity Analysis on Sampling Frequency

When our model was built, we fit the model on yearly data from 1959 to 2021. However, using monthly data provides more data as well as information on seasonal patterns. We fit VAR(5) on monthly data of CO_2 levels and global land-ocean temperatures from 1959 to 2021 obtained from NOAA[10] and NASA GISS[11] to verify that changing the sampling frequency does not affect our model prediction. Using the model fit on monthly data, we found that in 2029, the CO_2 level was predicted to be 431.3 ppm, which is very close to the 438.8 ppm predicted by the model fit on yearly data, and temperature was predicted to be 1.12°C, which is very close to the 1.2°C predicted by the model fit on yearly data. Therefore, because the predictions by the model fit on monthly data and the model fit on yearly data are similar, we determined that our model fit on yearly data is robust and captured the important patterns in monthly data.

Table 10: Granger causality test results using VAR(5) fit on monthly data

Target Series	Input Series	p-value for Granger causality
CO_2 levels	CO_2 levels	0
	Global temperatures	9.441E-7
Global temperatures	CO_2 levels	3.049E-9
	Global temperatures	0

We also checked that changing the sampling frequency does not greatly change the Granger casual relationships we find in the model.

The results of our Granger causality tests show that CO_2 levels still Granger-cause global temperature and temperature still Granger-causes CO_2 levels. As well, the Granger casual strength from CO_2 levels to temperature is still significantly stronger compared to the reverse direction. Therefore, we verified that the Granger causal relationships found are robust and are preserved with a finer sampling frequency.

7.4.3 Sensitivity Analysis with Other Factors

When our model was built, we only used data on CO_2 levels and global land-ocean temperatures. However, other factors are also useful in forecasting CO_2 levels and global temperatures and would provide more information. We therefore added in monthly data from 2001 to 2022 obtained from NOAA[10] on three important greenhouse gases, CH_4, N_2O, and SF_6, and fit a VAR(5) on the data to verify that the inclusion of new variables does not greatly change our model prediction. Using the model fit on the data with the new variables, we found that in 2029, the CO_2 level was predicted to be 436.92 ppm, which is very close to the 438.8 ppm predicted by the model fit on the original data, and temperature was predicted to be 1.19°C, which is very close to the 1.2°C predicted by the model fit on original yearly data. Therefore, because the predictions by the model fit on the data with the new variables and the model fit on the original data are similar, we determined that the inclusion of new variables does not greatly change the prediction, and the predictions made by the model using only CO_2 levels and global temperatures are accurate and capture most of the information provided by the new factors.

We also checked how the inclusion of new factors changed the Granger casual relationships we find in the model.

Table 11: Granger causality test results using VAR(5) fit on data with new variables

Output Series	Input Series	p-value for Granger causality
CO_2 levels	CO_2 levels	0
	Global temperatures	0
Global temperatures	CO_2 levels	0
	Global temperatures	0

The results of our Granger causality tests show that CO_2 levels still Granger-cause global temperature and temperature still Granger-causes CO_2 levels. However, the Granger casual strength from CO_2 levels to temperature is now the same as the casual strength in the reverse direction (all p-values are extremely small shown as 0). We therefore determined that the bi-directional Granger causal relationship between CO_2 levels and temperature remains after the inclusion of new factors.

7.5 Model Strengths and Weaknesses

Model Strengths

1. The VAR(5) model can account for non-stationary patterns.
2. The VAR(5) model can predict multiple time series variables using a single model.
3. The VAR(5) model can capture temporal relationships between time series.
4. The VAR(5) model typically has powerful short-term prediction abilities.
5. The VAR(5) model is easily implemented, and future predictions and confidence levels can be easily obtained.
6. The sensitivity analysis of the model demonstrates the effectiveness of the model under different data sizes, proving the robustness of the model. As well, the predictions of the model are robust and are not greatly changed by inclusion of monthly data or new factors.

7. The sensitivity analysis of the model demonstrates that Granger causal relationships found in the model are robust and are not affected by the sampling frequency. As well, strong Granger causal relationships are not affected by data size and the inclusion of new factors does not affect the bi-directional Granger causal relationship between CO_2 levels and temperature.

Model Weaknesses

1. The VAR(5) model loses accuracy for long term forecasting because the model is dependent on the accuracy of previous values.
2. The forecasting accuracy of the VAR(5) model depends on the reliability of historical data and on future conditions being similar to the conditions at the time of the data used. When there are drastic changes in the pattern of CO_2 levels or land-ocean temperatures, the model's prediction will likely not be reliable.
3. Granger causality does not mean true causality and does not address hidden variables.
4. The amount of data used to fit the model, 63 time points, is relatively small.
5. The model does not consider other factors besides CO_2 levels and global temperatures. Other variables like volcanoes and solar radiation that influence CO_2 levels and temperature would provide more information and increase forecasting accuracy.
6. Our current model assumes that the relationship between CO_2 levels and global temperatures stays the same over time. However, this relationship could be subject to change. For example, the relationship between CO_2 levels and global temperatures during the pre-industrial period was different than the relationship during the post-industrial period. Our current model can be extended to model such a relationship if given more data.

8 Conclusion

To better understand future global warming trends, we created several models to predict CO_2 concentration levels and global land-ocean temperatures.

We first analyzed the CO_2 levels data and discovered that the March 2003 increase of CO_2 resulted in a larger increase than observed over any previous 10-year period instead of the March 2004 increase that the NOAA claimed. We then fit four different models, Holt's linear trend, ARI(8, 2), IMA(2, 8), and ARIMA(3, 2, 3), on annual March averages of CO_2 levels and used these models to forecast the future CO_2 levels up to 2100. Holt's linear trend predicted that CO_2 levels will reach 685 ppm by 2132, ARI(8, 2) predicted by 2091, IMA(2, 8) predicted by 2082, and ARIMA(3, 2, 3) by 2083. Our results disagreed with the OECD's claim that the CO_2 concentration level will reach 685 ppm by 2050. After comparing the model performance statistics of all four models, the best model for describing patterns in the CO_2 levels was the ARI(8,2) model and hence we think this model is the most accurate one. Our sensitivity analysis verified that the ARI model is robust and model performance is not greatly affected by the amount of data or using data from different months.

In order to forecast future global temperatures, we then fit an ARIMA(3, 1, 3) model on global annual mean temperature changes. ARIMA(3, 1, 3) forecasted that the global average temperature will change by 1.25°C in 2038, 1.5°C in 2052, and 2°C in 2081 when compared to the average temperature from 1951 to 1980.

Finally, to investigate the relationship between CO_2 and temperature, we first examined Pearson's correlation and determined that there's a strong positive relationship between CO_2 levels and temperature. We then used VAR(5) to model the temporal causal relationship between CO_2 concentration levels and global temperatures. After performing Granger causality tests with our VAR model, we found that there is a strong Granger causal relationship from CO_2 levels to global temperatures and a weak Granger causal relationship from global temperatures to CO_2 levels. VAR(5) forecasts that in 2050, the CO_2 level will be 512.85 ppm and the global temperature will be 2.01°C. We determined that all predictions up to 2100 from VAR(5) should be reliable because of the narrow width of corresponding 95% confidence intervals. The concerns of our VAR(5) model include a small training data size and that other factors besides temperature and CO_2 levels were not considered in the model. To test the seriousness of these concerns and the robustness of the VAR model, we performed several sensitivity analyses. We verified that the model is effective with smaller data sizes in terms of forecasting and identifying Granger causal relationships. In addition, the predictions of the model and strong Granger causal relationships are robust and not greatly affected by changing the sampling frequency to monthly or the inclusion of new factors such as CH_4, N_2O, and SF_6.

All our analysis and model results confirm that CO_2 levels and global temperatures are steadily increasing and verify that CO_2 levels greatly influence global warming. To protect the environment and lessen global warming, CO_2 emissions should be greatly reduced and new policies to reduce emissions should be enacted, such as restricting the use of fossil fuels.

References

[1] United Nations. (n.d.-b). *Climate Change*. https://www.un.org/en/global-issues/climate-change

[2] *Population Pressure and the Climate Crisis.* (n.d.). https://www.biologicaldiversity.org/programs/population_and_sustainability/climate/

[3] *Sources of Greenhouse Gas Emissions | Greenhouse Gas (GHG) Emissions | US EPA.* (n.d.). https://climatechange.chicago.gov/ghgemissions/sources-greenhouse-gas-emissions

[4] U.S. *EPA* Student's Guide to Global *Climate Change*. U.S. Environmental Protection Agency. https://archive.epa.gov/climatechange/kids/documents/carbon-through-the-seasons.pdf

[5] National Academies of Sciences, Engineering, and Medicine. 2020. Climate Change: Evidence and Causes: Update 2020. Washington, DC: The National Academies Press. https://doi.org/10.17226/25733

[6] Mark J. P. Anson; Frank J. Fabozzi; Frank J. Jones (3 December 2010). The Handbook of Traditional and Alternative Investment Vehicles: Investment Characteristics and Strategies. John Wiley & Sons. pp. 489–. ISBN 978-1-118-00869-0.

[7] Holt, C. E. (1957). Forecasting seasonals and trends by exponentially weighted averages (O.N.R. Memorandum No. 52). Carnegie Institute of Technology, Pittsburgh USA.

[8] Hamilton, J. (1994). Time Series Analysis. Princeton University Press. ISBN 9780691042893.

[9] Organisation for Economic Co-Operations and Development. (2012). The OECD environmental outlook to 2050 [Internet]. https://www.oecd.org/env/cc/Outlook%20to%202050_Climate%20Change%20Chapter_HIGLIGHTS-FINA-8pager-UPDATED%20NOV2012.pdf.

[10] The Global Monitoring Laboratory (GML) of the National Oceanic and Atmospheric Administration at https://gml.noaa.gov/aftp/products/trends/

[11] National Aeronautics and Space Administration Goddard Institute for Space Studies at https://data.giss.nasa.gov/gistemp/