

# Present Global $CO_2$ Emissions

Meral Basit, Alex Hubbard, Mohamed Bakr

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Background . . . . .	2
1.2	Null Hypothesis . . . . .	2
<b>2</b>	<b>Measurement and Data</b>	<b>2</b>
2.1	Measuring Atmospheric Carbon . . . . .	2
2.2	Historical vs Present Trends in Atmospheric Carbon . . . . .	3
<b>3</b>	<b>Old Models Evaluation</b>	<b>3</b>
3.1	Linear Model Evaluation . . . . .	3
3.2	ARIMA Model Evaluation . . . . .	3
3.3	Linear vs ARIMA . . . . .	4
<b>4</b>	<b>Exploring New Models</b>	<b>4</b>
4.1	ARIMA Model . . . . .	4
4.2	Polynomial Model . . . . .	6
<b>5</b>	<b>Conclusion</b>	<b>7</b>

# 1 Introduction

## 1.1 Background

In the 1997 report, we explored the trend and variability in atmospheric  $CO_2$  levels using polynomial and ARIMA models. Our analysis indicated a significant upward trend in  $CO_2$  levels, with an accelerating rate of increase. In this report, we aim to re-evaluate those models and their predictions using the  $CO_2$  level data realized after 1997. The central question we are addressing is:

*Have the previous models accurately predicted the realized atmospheric  $CO_2$  levels?*

## 1.2 Null Hypothesis

The predicted atmospheric  $CO_2$  levels from the estimated models are equal to the realized atmospheric  $CO_2$  levels for the period from January 1998 to June 2024.  $H_0 : \hat{y}_t - y_t = 0$ . Where:  $\hat{y}_t$  is the forecasted atmospheric  $CO_2$  levels using the polynomial and ARIMA models and  $y_t$  is the realized atmospheric  $CO_2$  levels at time  $t$ .

# 2 Measurement and Data

## 2.1 Measuring Atmospheric Carbon

Up until April 2019, measurements were collected using infrared absorption. After April 2019, a new  $CO_2$  analyzer was installed which uses Cavity Ring-Down Spectroscopy (CRDS). This change in devices could impact the errors in our predictions. Additionally, since our last report, the volcano near the research center has erupted. Therefore, the measurements from December 2022 to July 4, 2023 are from the Maunakea Observatories, which are just over 20 miles north of the original observatory. There is also a note that the last several months worth of data is “preliminary” and therefore could be revised.

Table 1: Accounting Table

Cause	Number of Samples Available For Analysis (after removal for cause)	Number of Samples Removed
Start	2617	NA
Before 1998	1384	1233
July 2024	1383	1
Values = -1000 (*missing*)	1379	4

The data was obtained from the United States’ National Oceanic and Atmospheric Administration NOAA [accessible here]. The dataset contains 2,617 observations of weekly atmospheric concentrations of  $CO_2$  levels, span from the 3rd week of May 1974 to the 1st week of July 2024. As detailed in Table 1, we excluded 1233 observations before 1998, 1 observation for one week in July 2024 as the month data was incomplete, and 4 observations where average  $CO_2 = -1000$ , which indicates missing readings. We ended up with 1,379 observations after cleaning up the data. Then, we calculated the monthly averages from January 1998 to the Jun 2024 to address the 4 missing weekly values and to facilitate comparison with the forecast data from the 1997 report, which was also in by month.

## Atmospheric CO<sub>2</sub> EDA plots 1998 – 2024

Upward trend with clear seasonal pattern

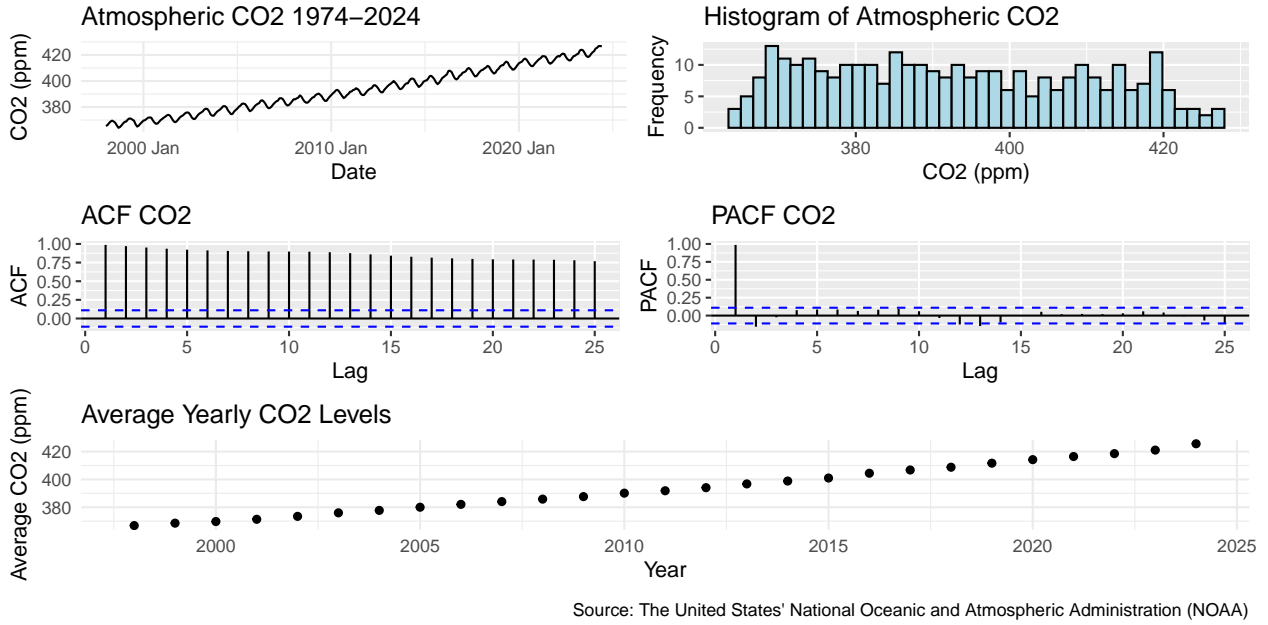


Figure 1: Atmospheric CO<sub>2</sub> EDA standard plots

## 2.2 Historical vs Present Trends in Atmospheric Carbon

Based on the EDA plots in Figure 1, we observed that the atmospheric  $CO_2$  levels continued to have a strong upward linear trend since 1997, as shown in the average yearly  $CO_2$  in the bottom plot in Figure 1. The histogram indicates a wide distribution of values with a slight right skew. Additionally, the ACF tails off very slowly while the PACF drops off after lag 1 but still has few lags above the significance level. This indicates that there may be some unit roots. As this timeseries is a continuation of our previous time series, we know that this data is also not stationary.

## 3 Old Models Evaluation

### 3.1 Linear Model Evaluation

Our polynomial model did a pretty good job at predicting  $CO_2$  up to 2024. It is missing the peaks and valleys, but it looks to capture the average yearly increase in  $CO_2$  levels as shown in Figure 2. Based on the results of the paired t-test, the mean difference = -0.0326 suggests that, on average, the predictions were slightly lower than the actual observed  $CO_2$  levels. However, this difference is not statistically significant with the p-value = 0.5 indicating that we fail to reject the null hypothesis  $H_0 : \hat{y}_t - y_t = 0$ , which means that there is no statistical significant difference between the values predicted by the polynomial model and actual  $CO_2$  levels.

### 3.2 ARIMA Model Evaluation

In the 1997 report, we used an ARIMA(3,1,0)(2,1,0)[12] model to forecast atmospheric  $CO_2$  levels. Since 1997, we visually observed that linear trend of the Keeling Curve very slightly increased in slope around the year 2000. Both the ARIMA model forecast and the the realized  $CO_2$  show an

upward trend with seasonal pattern. However, the model predicted lower values than the realized values in the long term as shown in Figure 2. Based on the results of the paired t-test, the mean difference = -3.66 suggests that, on average, the predictions were lower than the actual observed  $CO_2$  levels by 3.66 units. The difference is statistically significant with the p-value < 0.05 indicating that we reject the null hypothesis  $H_0 : \hat{y}_t - y_t = 0$ , which means that the difference between the  $CO_2$  levels predicted by the ARIMA model and the actual are statistically significant.

#### Reliaized vs Forecasted Atmospheric CO2 plots 1998 – 2024

The polynomial model performs better than the ARIMA model, when compared to actual data

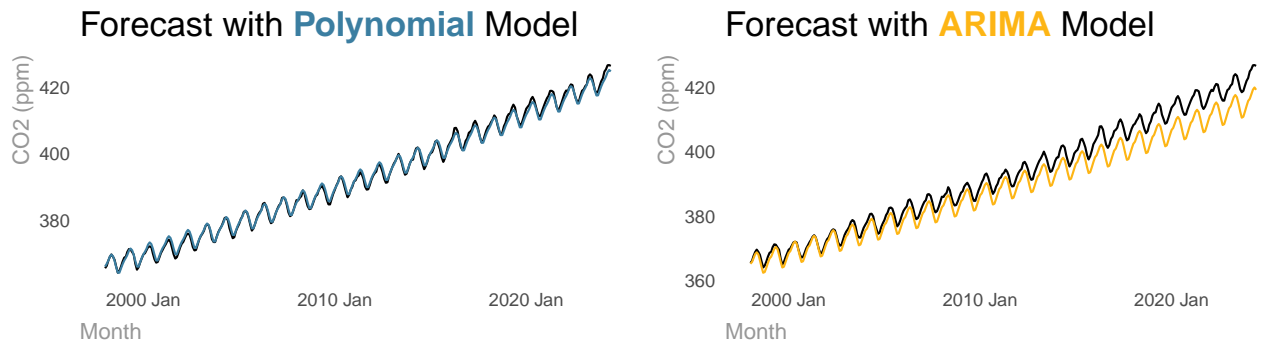


Figure 2: Comparing Realized Atmospheric CO2 Levels with Polynomial and ARIMA Models

### 3.3 Linear vs ARIMA

In 1997, we predicted that  $CO_2$  levels would reach 420 PPM by 2025 March using our ARIMA model, and 2022 May using the Polynomial model. However, actual  $CO_2$  levels reached 420 PPM by 2022 Apr. Our polynomial model was much closer to the true outcome in this case.

Our ARIMA model with log transformation produces  $RMSE = 4.267$ , while the Polynomial model produces an  $RMSE = 0.831$ . Considering the descriptive analysis, the t-test results, threshold-prediction results, and the RMSE comparisons, we can conclude that despite the fact that the ARIMA model performed better with the 1997 data, the Polynomial model outperformed the ARIMA model in the long-term forecast.

## 4 Exploring New Models

In this section we will explore new models and train them on the present data then evaluate their performance to see if we can estimate a better model than what we estimated in the 1997 report.

### 4.1 ARIMA Model

We made two copies of the aggregated NOAA data and seasonally adjusted one of them using STL decomposition. After that, we split both datasets (the seasonally adjusted (*SA*) and the non-seasonally adjusted (*NSA*)) into training and test sets, using the last two years as test sets.

We used the KPSS test to determine whether the *SA* and the *NSA* data are stationary. For both series, the first test yields a p-value of 0.01, leading us to reject the null hypothesis, indicating that the data is not stationary. After taking one difference, the p-value for both series rose to 0.1, and we failed to reject the null hypothesis, suggesting that both datasets are stationary after one difference.

Table 2: ARIMA Models Comparison - Seasonally Adjusted CO2 values

.model	sigma2	log_lik	AIC	AICc	BIC
arima_fit_log.search	0.00	1625	-3238	-3238	-3216
arima_fit.search	0.11	-88	188	188	210

Based on the EDA performed earlier, it was challenging to estimate the  $p$  and  $q$  terms for the ARIMA model just by looking at the ACF and PACF plots. We estimated two non-stepwise ARIMA models for each dataset, one with log transformation and one without.

Using the information criteria AICc in Table 2, we see that the best *SA* model was an ARIMA(1,1,1)(2,0,0)[12] with drift and a log transformation. This model has one AR term, one MA term, and a first-order differencing term. Additionally, it has two seasonal AR terms with a period of 12 months. This model is invertible (MA coef = -0.60). The best *NSA* model was an ARIMA(1,1,2)(0,1,1)[12] with a log transformation. This model has one AR term, two MA terms, and a first-order differencing term. It also has one seasonal MA term and one seasonal differencing term with a period of 12 months. This model is not invertible (MA coef = -1.41, 0.44, -0.88). We selected both of these models because they had the lowest AICc.

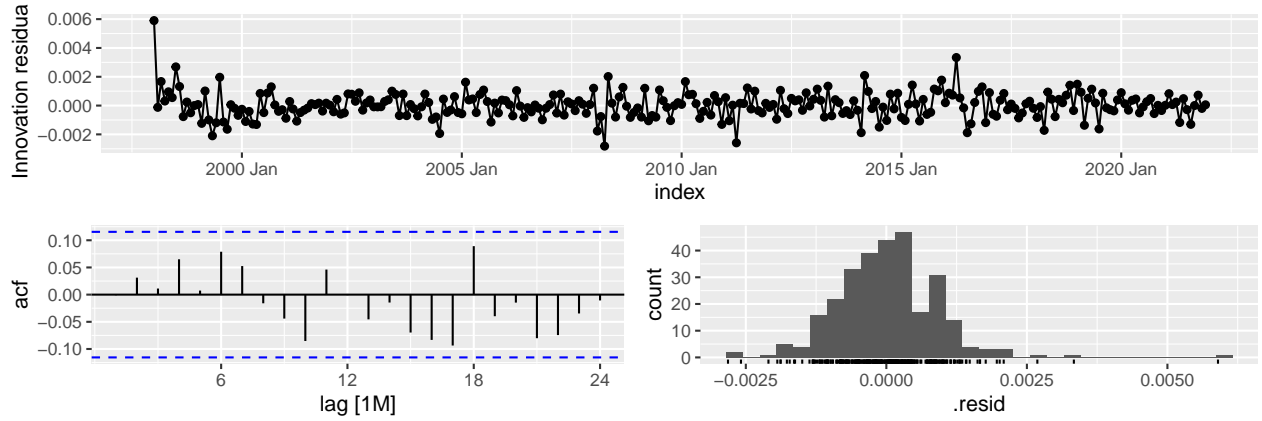


Figure 3: SA Trained ARIMA(1,1,1)(2,0,0)[12] with drift Model Residuals

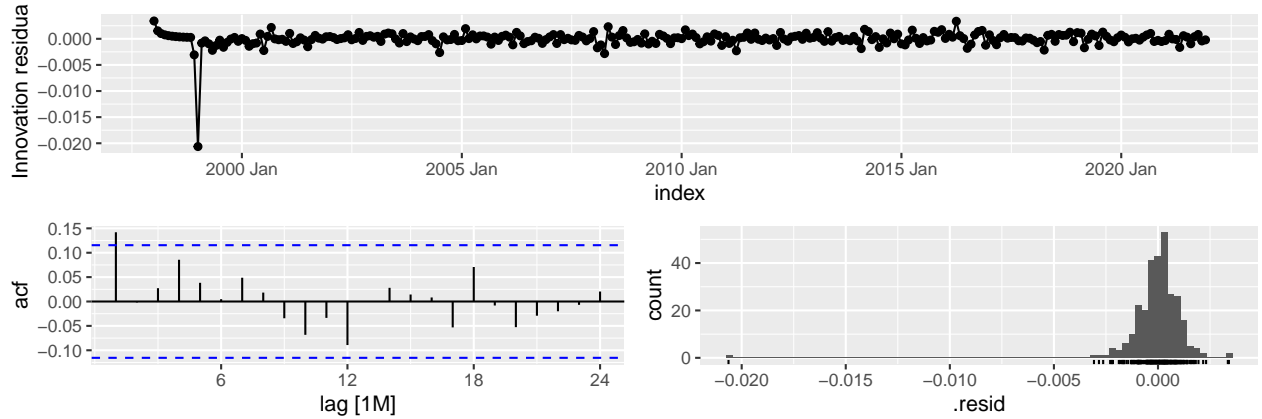


Figure 4: NSA Trained ARIMA(1,1,2)(0,1,1)[12] Model Residuals

The residuals line and ACF plots of the *SA* ARIMA(1,1,1)(2,0,0)[12] model Figure 3 show that all autocorrelations are within the threshold limits, indicating that the residuals are behaving like white noise which is confirmed by Ljung Box test with a p-value = 0.717 we fail to reject the null hypothesis  $H_0$  : data is independently distributed. Similarly, the residuals line and ACF plots of the *NSA* ARIMA(1,1,2)(0,1,1)[12] model Figure 4 show that all autocorrelations are within the threshold limits, indicating that the residuals are behaving like white noise which is confirmed by Ljung Box test with a p-value = 0.339 we fail to reject the null hypothesis  $H_0$  : data is independently distributed.

#### SA vs NSA Atmospheric CO2 ARIMA Models

The NSA trained model outperforms the SA trained. However SA successfully captures the trend.

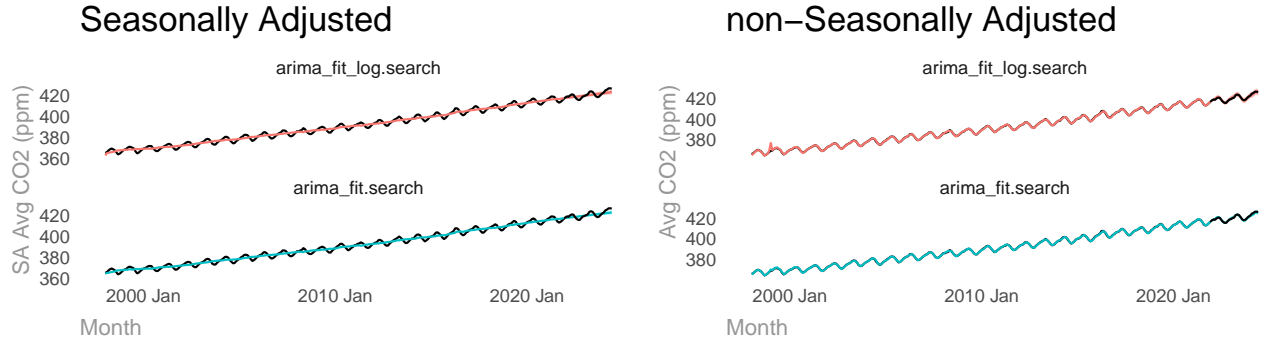


Figure 5: Comparing SA vs NSA data sets ARIMA models performance

The selected *NSA* trained model produced RMSE = 0.59 which outperforms the selected *SA* trained model that produced RMSE = 0.63

## 4.2 Polynomial Model

We estimated a Polynomial model of the form:  $CO2_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \epsilon_t$  and trained it on the *SA* data. Based on the model fit results, the estimated coefficient  $\beta_1 = 0.14$  indicates that the  $CO_2$  levels increase by  $\approx 0.14$  units per month which is lower than the rate of  $\approx 0.0674$  we estimated in the 1997 report. The p-value of the time index is  $< 0.05$  which suggests that the coefficient is statistically significant. We reject the null hypothesis that the coefficient  $\beta_1 = 0$ . This also provides evidence that the  $CO_2$  levels continue to have an upward linear trend. The estimated quadratic term coefficient  $\beta_2 = 0.000141$ . The positive coefficient suggests that the rate of increase in  $CO_2$  levels is accelerating at a higher rate than the model estimated in 1997 report. The p-value of the quadratic time index is  $< 0.05$  which suggests that the coefficient is statistically significant. We reject the null hypothesis that the coefficient  $\beta_2 = 0$ .

Our polynomial model performs well in sample, but appears to underpredict out-of-sample. The polynomial model produced RMSE = 52.67 which is much higher than the ARIMA model out of sample RMSE = 0.63.

As stated previously, in 1997, we predicted that  $CO_2$  levels would reach 420 PPM by 2025 March using our ARIMA model, and 2022 May using the Polynomial model. However, actual  $CO_2$  levels reached 420 PPM by 2022 Apr. We also previously predicted that  $CO_2$  levels would reach 500 ppm by the 2050 May. Our updated prediction is that we will reach 500 ppm by 2052 Apr. This indicates that the updated data may support a slowing of the growth in atmospheric  $CO_2$ . By 2122, our model predicts that  $CO_2$  levels will reach 751.99 ppm. As with the 1997 report, we have low confidence in these predictions, as we are well beyond the range of our data, and have no way of

Polynomial Model In-Sample vs Out-of-Sample Atmospheric CO<sub>2</sub> Forecasted based  
 In-sample forecast had a better results than the out-of-sample forecast.

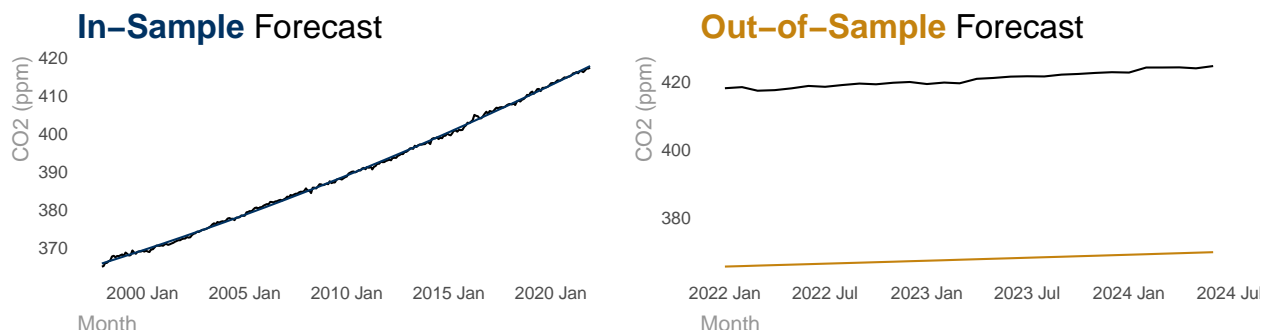


Figure 6: SA trained Polynomial Model - In-Sample vs Out-of-Sample

accounting for improvements in efficiency, grid electrification, etc.

## 5 Conclusion

The Polynomial model estimated in our 1997 report accurately predicted the realized atmospheric CO<sub>2</sub> levels, and we failed to reject our null hypothesis. However, the ARIMA model had a significant difference between the predicted and actual atmospheric CO<sub>2</sub> levels and we had to reject the null hypothesis.

The new models trained using seasonally adjusted data was outperformed by those trained using non-seasonally adjusted data. However, our best model with non-seasonally adjusted data was not invertible.

The updated atmospheric CO<sub>2</sub> data shows that the increasing trend in CO<sub>2</sub> levels continued roughly as expected. We still see significant coefficients in our estimated predictive models using present data, and forecast that CO<sub>2</sub> levels will continue to rise into the future, barring any significant intervention. As follows, we again reject our null hypothesis.