W271 Group Lab 2

Analysis of Carbon Dioxide (1997 & Present)

Violet Davis, Yuri Kinakin, W. Sean McFetridge and Emanuel Mejia

1 1997 Report

1.1 Introduction

Charles D. Keeling discovered that the content of carbon dioxide in the atmosphere moves in a predictable and seasonal pattern with an upward trend. Carbon dioxide is a greenhouse gas that is transparent to visible light while being partially reflective to infrared-radiation, so as the levels of carbon dioxide increase the average global temperature will experience a corresponding increase. This rising temperature can have serious impacts on the environment and on human health. Therefore, the analysis of CO_2 data are critical to make informed decisions on how to address these changes to the climate. At the time of this paper (1997), the US government has not yet implemented any federal laws targeting CO_2 emissions.

We aim to develop a model to predict the growth of CO_2 in the atmosphere. We will analyze several models within this report, focusing on linear time trend and ARIMA time series models. These results should be reviewed at a later date to determine model performance, and hopefully to assess the impact of current and future legislation on reducing the emission of carbon dioxide into the atmosphere.

1.2 CO_2 Data

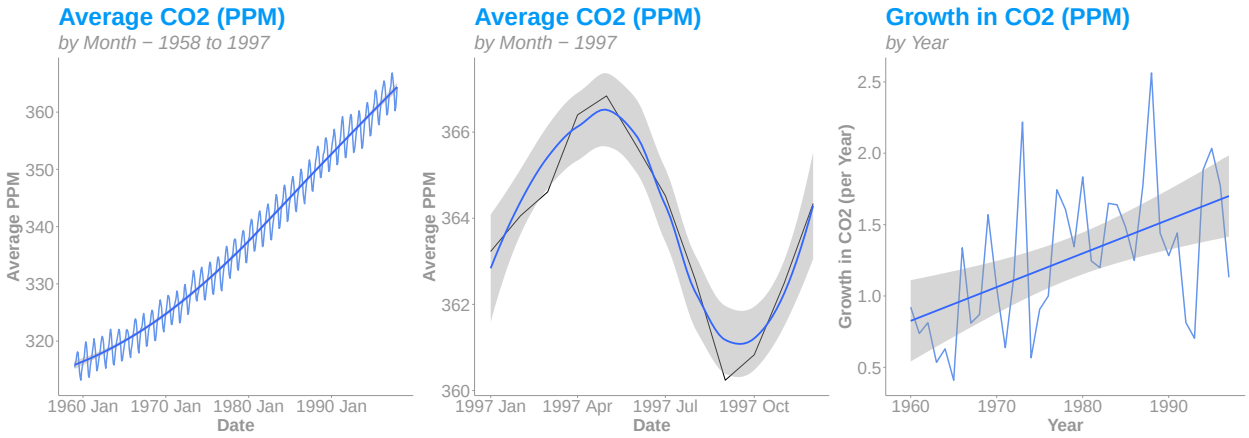
The data analyzed is from the *Global Monitoring Laboratory*, which was collected at Mauna Loa Observatory in Hawaii. As per the laboratory's website, they note that these CO_2 measurements represent a robust baseline of global atmospheric CO_2 concentration for three reasons:

1. The Observatory is near the summit of Mauna Loa at an altitude of 3400 m, making it well situated to measure air masses that are representative of very large areas.
2. All of the instruments are rigorously and frequently calibrated.
3. Ongoing comparisons of independent measurements at the same site allow measurement of variability and accuracy, which is generally better than 0.2 ppm.

Data are continuously collected by an infrared absorption spectroscopy instrument and reported as the "mole fraction" of CO_2 in dry air, defined as the number of carbon dioxide molecules in a given number of molecules of air after removal of water vapor; this is more commonly known as the "concentration" of CO_2 in the air. When we look at aggregated amounts (e.g. the monthly

concentration), we take that as the average amount measured during a set period of time versus a single point value during each time period.

Upon loading the results, we noted there were five months with no reported values (two in 1958 and three in 1964). These rows were removed from the data set and treated as "N/A". Three plots are shown below: a time series of average monthly values from 1958 to 1997, the average month rates during 1997, and the change in average annual co2 for each year (i.e. the growth).



From the graphs we can clearly see the data is non-stationary, with a strong upward trend overlain by a seasonal component. This is consistent with the findings from Charles D. Keeling in his original research study. The second graph is single year timeseries plot of CO_2 PPM by month, which shows there is the highest average concentration of CO2 in the atmosphere from March to July, and the lowest from August to November. Lastly, from the growth plot we can see a steady increase in growth over the years. This means that the rate of increase is not stationary, but rather is increasing over time.

1.3 Model Development

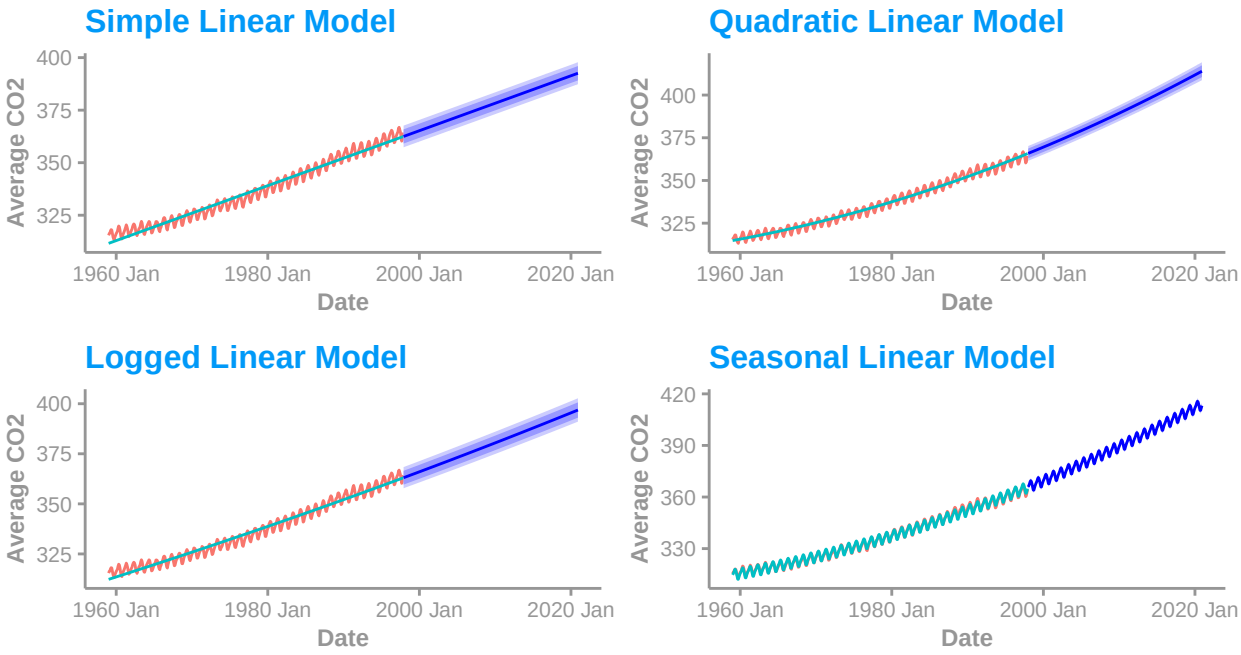
In order to predict the amount of carbon dioxide in the atmosphere we've developed two different classes of models. The first are a series of linear models and the second are a series of ARIMA models. Details of these models are summarized below.

1.3.1 Linear Time Trend Model

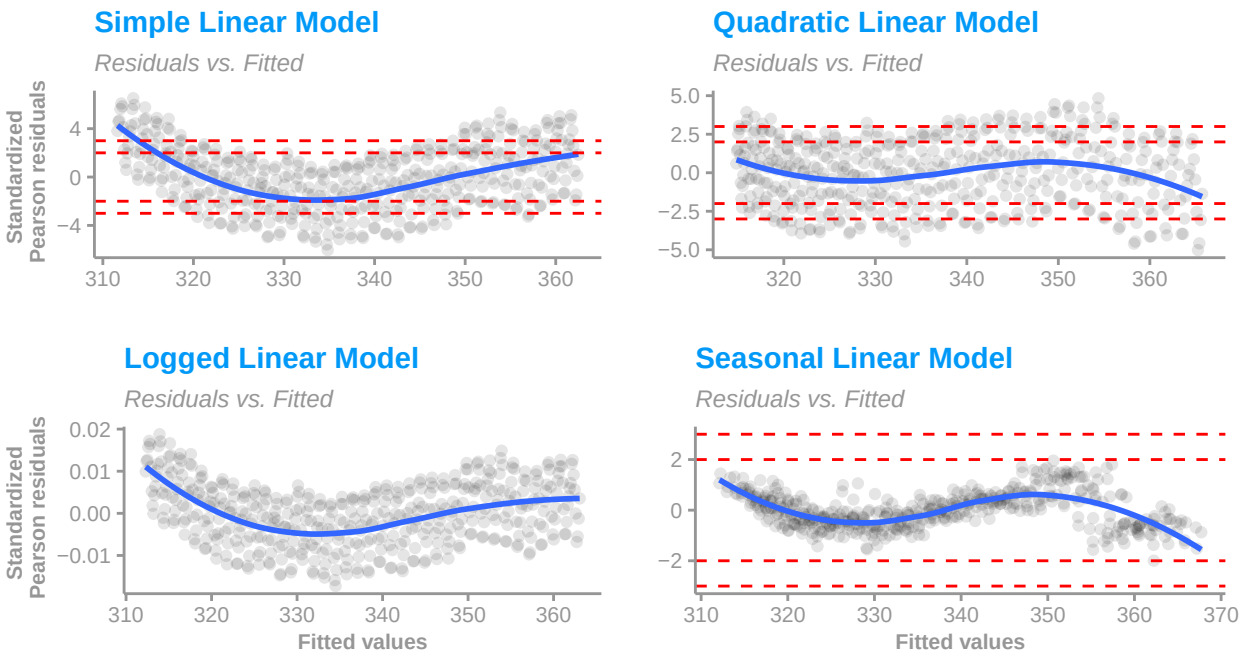
We ran four different linear models:

1. $Simple\ Linear : average = trend$
2. $Quadratic\ Linear : average = trend + trend^2$
3. $Logged\ Linear : log(average) = trend$
4. $Seasonal\ Linear : average = trend + trend^2 + season$

Each of these models are plotted below, including a projection out to 2020. The logged Linear model is only included for illustration, as when we look at the time series we don't notice a particular increase in stationary variance over time and would not suspect the logged model to be useful.



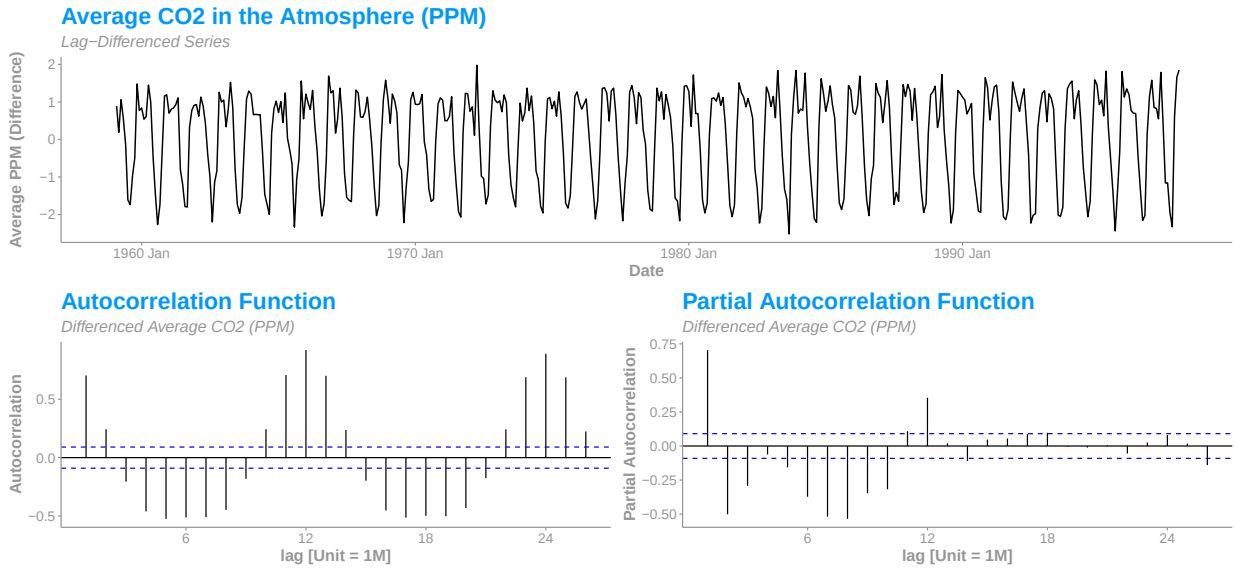
From the graphs above, we can see that the seasonal model appears to fit the data the best, and the simple linear model appears to fit the worst. As stated previously, there doesn't appear to be much gained by using a logged model as the results are similar to the quadratic model. However, we also want to review the standardized residuals vs. fitted values to ensure they're reasonable:



This conclusion is supported by looking at the residual plots, where the seasonal plot is more closely clustered together. However, there is an oscillation to the residuals which is concerning and will hopefully be corrected by an ARIMA model.

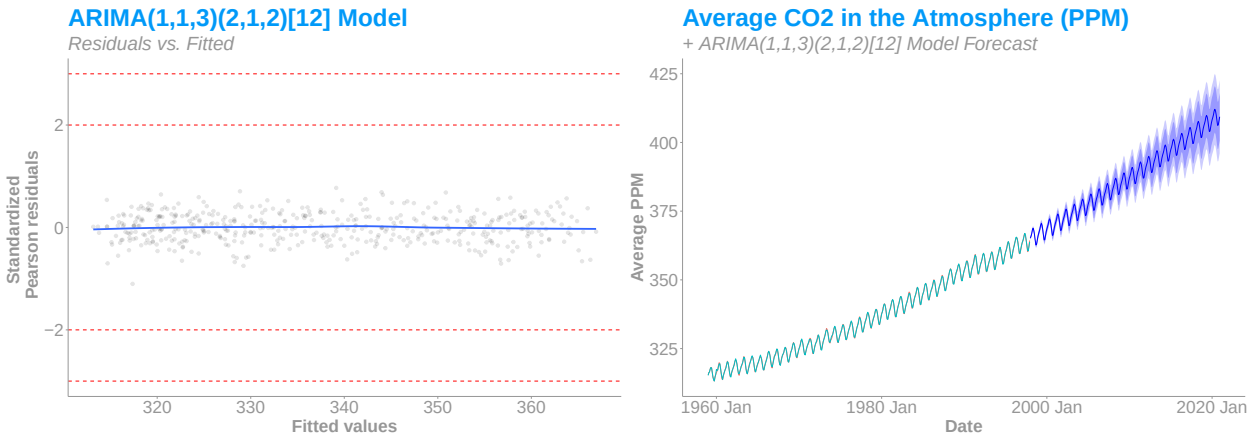
1.3.2 ARIMA Time Series Model

In order to determine an appropriate ARIMA model, we need to analyze the ACF and PACF graphs. However, since we have previously noted that the data are non-stationary, we first need to apply a differencing function. This is shown below:



After applying a differencing function, we can see that the data are now stationary. This is validated with the results of the kpss unit root test, where we fail to reject the null hypothesis ($p = 0.1$) after a first difference. The ACF shows an oscillation which is indicative of a seasonal ARIMA model. We therefore created a SARIMA model that included a differencing term for both the seasonal and non-seasonal elements, and iterated over various autoregressive and moving averages to best optimize the model.

Our optimized model was an ARIMA(1,1,3)(2,1,2)[12] model. This is a seasonal model, with a frequency of 12 months. Reviewing the residuals, we can see that between all of the models analyzed so far, it has the best presentation of residuals which are clearly centered around zero with no outputs being over two standard deviations away, and the projection continues to capture the seasonal movements and upward trend.

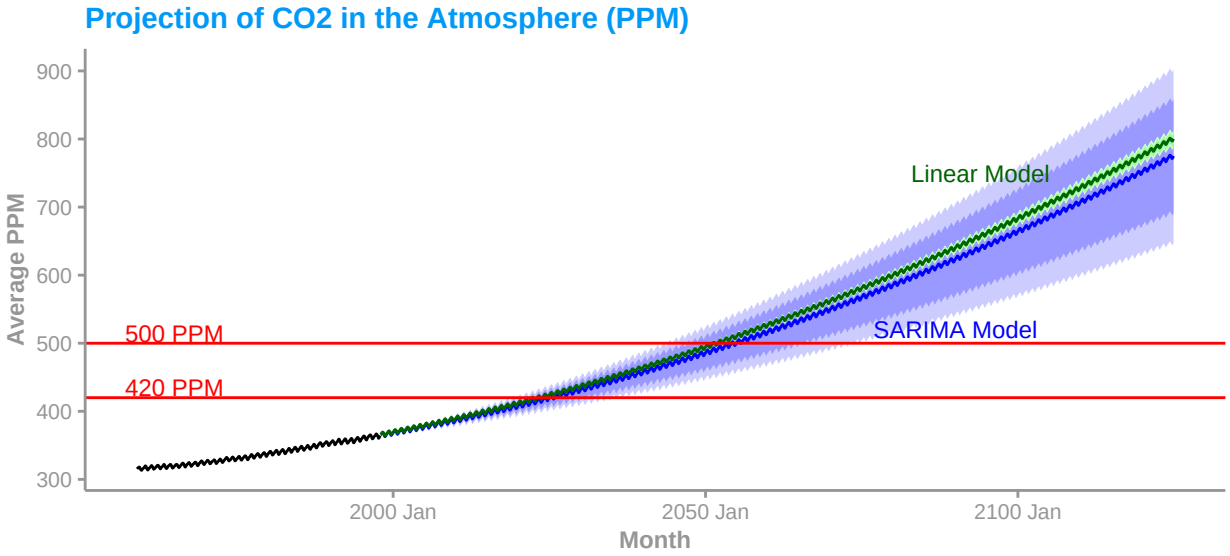


When we run the Box-Ljung test, we get a favorable p-value of 0.8086 at lag 1 and 0.6967 at lag

10. Since the p-value is above 0.05 in both cases, we fail to reject the null hypothesis that the data is independently distributed.

1.4 Forecast Atmospheric CO_2 Growth

The graph below shows the projection of both the linear model (incl. seasonal and quadratic components) and the SARIMA model projected several decades into the future:



The models produce somewhat different predictions. As a result the models are expected to hit 420 and leave 500 in approximately the following years:

Model	420 PPM (Point)	420 PPM (First)	420 PPM (Last)	500 PPM (Point)	500 PPM (First)	500 PPM (Last)
Linear	2022 May	2022 Mar	2023 Apr	2051 Apr	2033 May	2052 Apr
SARIMA	2025 Apr	2019 Apr	2035 Apr	2056 Mar	2044 May	2076 Apr

The "point" indicator is the expected time. The "first" indicator is the month and year in which the upper bound of the 95% confidence interval crosses the PPM threshold, whereas the "last" indicator is the month and year in which the lower bound of the confidence interval crosses the PPM threshold. We can see that the SARIMA model crosses the threshold much slower, and there is a wider confidence interval.

When we look at year 2100, the SARIMA model projects that the CO_2 in the atmosphere will be close to 650PPM whereas the linear model projects there will be close to 700 PPM. There is, however, a wider range of results for the SARIMA model that can put the projections above and below the linear model and is illustrative of a higher level of model uncertainty. Ultimately, according to our calculations the Linear Model produces more stable results as the confidence interval is much narrower. The much broader confidence interval of the SARIMA model makes it less precise as a forecasting tool.

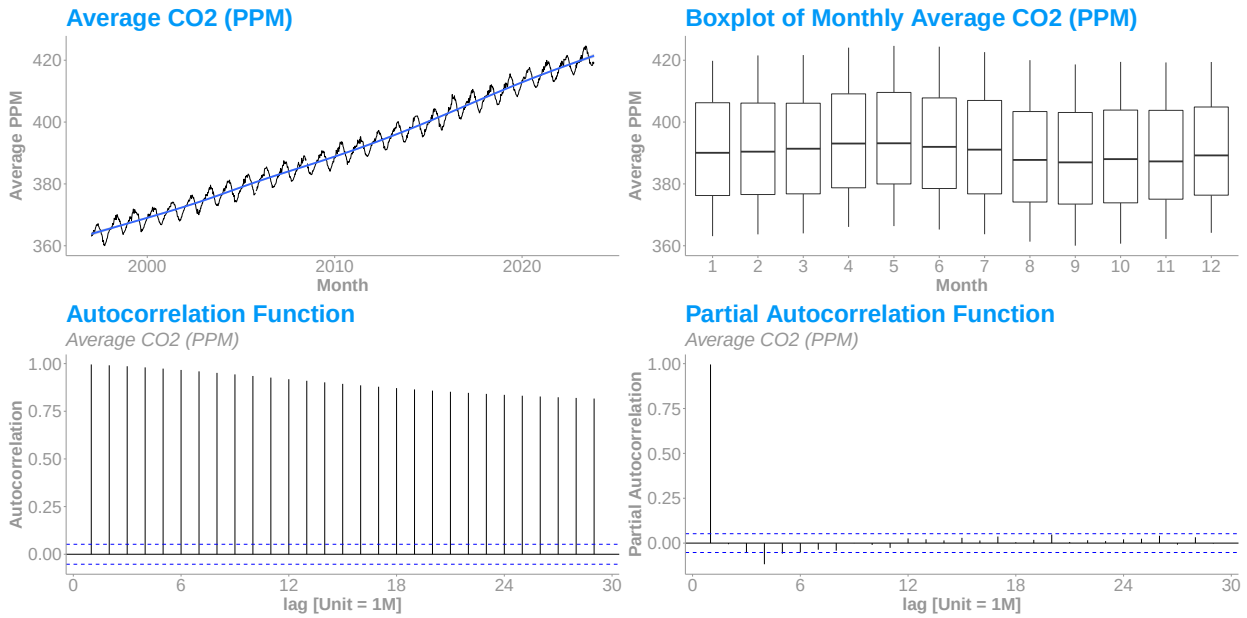
2 2023 Review of 1997 CO_2 Report

2.1 Introduction

In the following analysis, we would like to determine how well our best model trained on pre-1997 CO_2 concentration data collected at the Mauna Loa observatory performs against data collected post-1997. There were no major technical changes to the measurement from 1997 - 2019, but in that year a new CO_2 detector based upon Cavity Ring-Down Spectroscopy (CRDS) was installed [1]. While it is fundamentally different to the previous method of infrared absorption, robust calibration of the CRDS analyzer ensures that data from both instruments can be used in conjunction during modelling without issue.

2.2 Data Pipeline

A data pipeline was created to pull and clean atmospheric CO_2 concentration data hosted online by the NOAA Global Monitoring Laboratory.

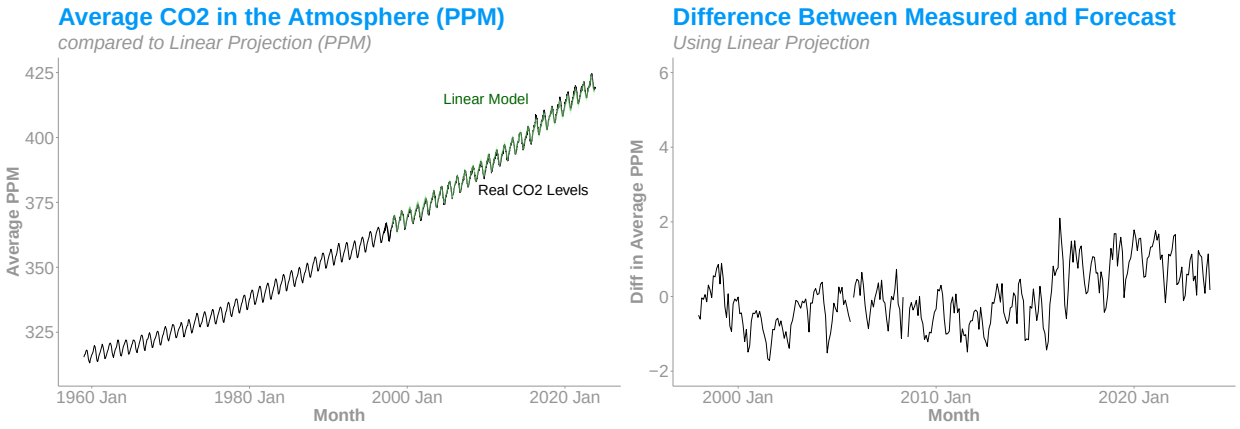


The EDA plots above suggest that the series post-1997 is similar to pre-1997 data, with a strong auto-regressive and seasonal component.

2.3 Compare Linear Model forecasts agaisnt realized CO_2

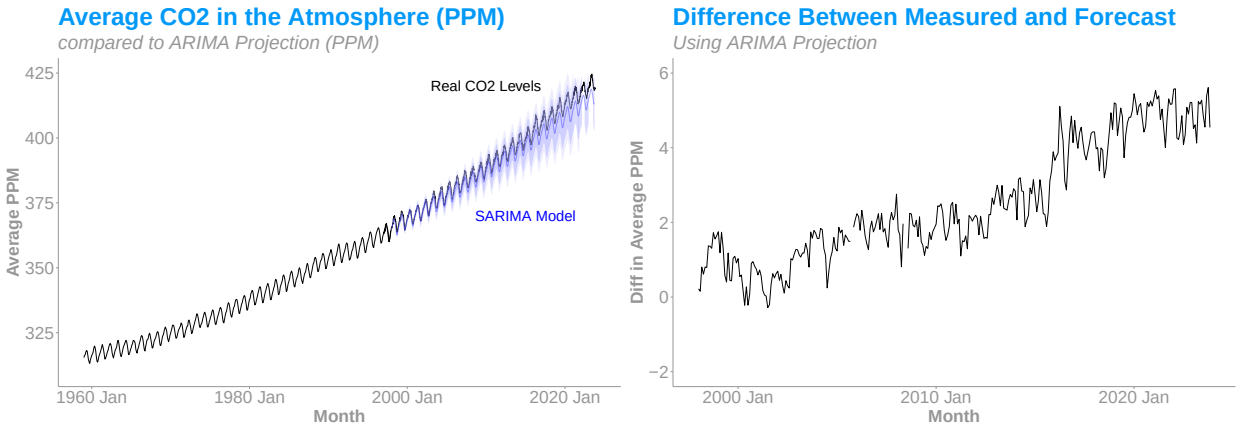
In the left plot below we show the forecast of the linear model (in green) against the true data (in black). For such a simple model the fit is surprisingly close. The right plot shows the difference between the model and data. It is noteworthy that the difference does not appear stationary and seems to increasingly under-estimate the measured atmospheric CO_2 concentrations as time passes.

[1] https://gml.noaa.gov/ccgg/about/co2_measurements.html



2.4 Compare ARIMA models forecasts against realized CO2

In the left plot below we show the forecast of the selected ARIMA model (in blue) against the true data (in black). The model fit is quite similar to the linear model, though the non-stationarity of the difference (right plot) is much stronger, more than twice as high in the ARMIA model versus the seasonal linear at the furthest forecast dates.



2.5 Evaluate the performance of 1997 linear and ARIMA models

The first recorded time in the weekly data that CO2 crosses 420 ppm was in March of 2020. The seasonal linear model predicted May 2022, while the SARIMA model predicted April 2025. The models have 95% intervals of [411.9535, 415.4154] and [397.3001, 423.6106] respectively for March of 2020. (The linear model has a tighter interval and a closer point estimate, while the SARIMA model includes 420 in the CI but has a wider interval and a farther point estimate.) In this regard, the linear model proved to be more accurate as its forecast was only 26 months out vs. the 61 months for the SARIMA model.

From this information and the above graphs, we can see that neither of the models fully capture the rapid growth in level of atmospheric CO_2.

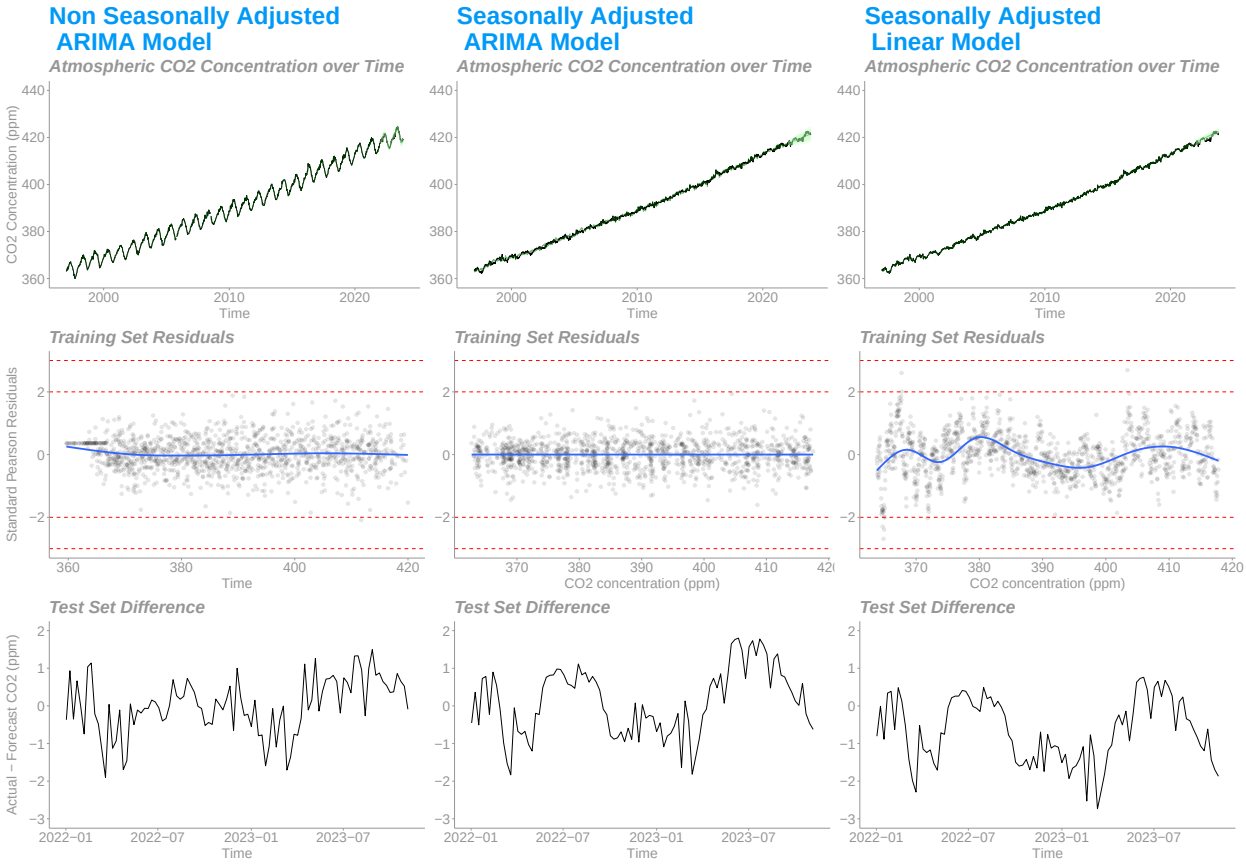
Comparing the root mean squared errors of the models, the linear model has an RMSE of 0.491 while the SARIMA model has an RMSE of 1.23. We therefore conclude that the linear model was

the more accurate of the two.

2.6 Train best models on present data

Three models were fit against data from 1997 up to 2022 and tested against the data from 2022 onward. The first was a SARIMA model trained on the raw weekly data, the second was a SARIMA model trained on the seasonally corrected data (only the trend and white noise components of the data), and the final was a polynomial time-trend model based on the same seasonally corrected data.

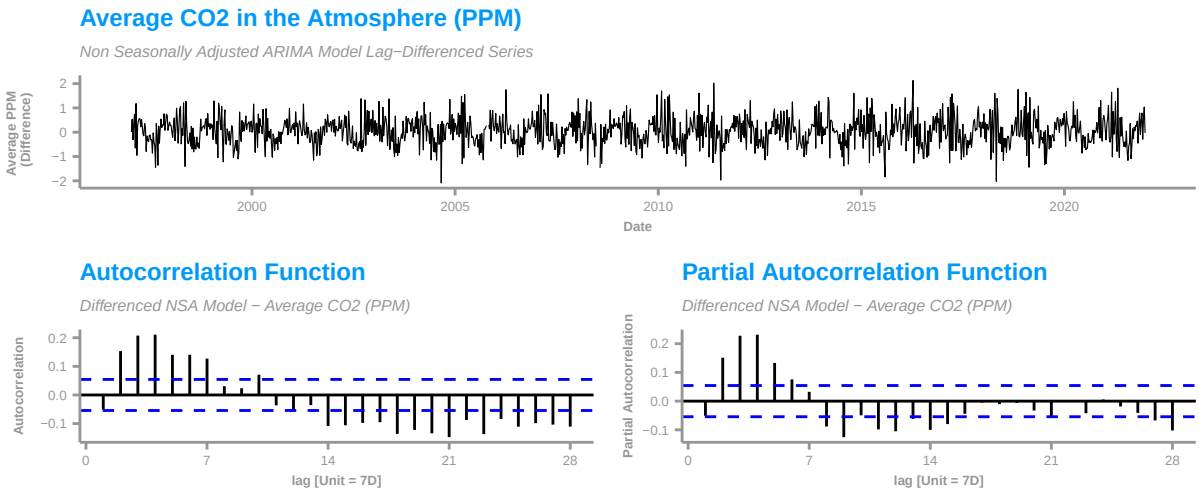
These are summarized graphically in the following plot and discussed in detail in the sections below.



2.6.1 Non Seasonally Adjusted ARIMA Model

$$ARIMA(1,0,2)(1,1,0)[52] \text{ w/ drift}$$

Similar to the 1977 report, differencing of the time series of the model is required to make the data stationary. This is validated in the unit root test, where we get a p-value of 0.1 after a first difference. This was the motivation to force a difference term for the seasonal component of the SARIMA search. Again, similarly to the optimized model in 1977, we can see an oscillation in the ACF which is indicative of a ARIMA model. There is also still some oscillation in the PACF for this model that are likely due to the difficulty in paramter search for a series with a relatively large period of 52.



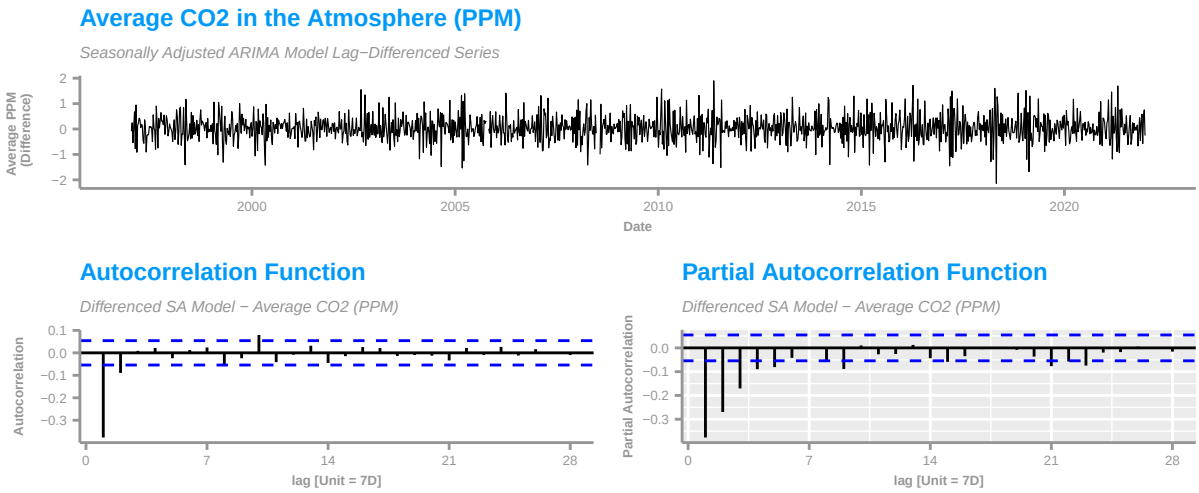
When we review the residuals of the model we can see that they are closely centered around zero, and the projection of the model track well against the test set data. We also ran the Ljung-Box test and we failed to reject the null hypothesis that all auto-correlations are zero in the time series at lag 1 (p=0.419), but can reject at lag 10 (p=0.0153), suggesting some residual auto-correlation that is not captured in the seasonal model.

A RMSE of 0.734 was calculated for this model against the test set.

2.6.2 Seasonally Adjusted ARIMA Model

$$ARIMA(0, 1, 2) \text{ w/ drift}$$

Similar to the non-seasonally adjusted model, these data were suspected to be non-stationary and therefore a differencing function needed to be added. This is validated in the unit root test, where we also get a p-value of 0.1 after a first difference. This was the motivation to force a difference term in the arima search.



There are some key differences vs. our non-seasonally adjusted model. The first is that there is less uniformity in the time series and the residual difference looks more like white noise. Also, while

there is still an oscillation in the ACF, fewer of the spikes are significant. The PACF also shows a more sensible exponential fall-off. Running the Box-Leung test, we fail to reject the null hypothesis for lags of both 1 and 10 suggesting that the model has captured sufficient autocorrelative behaviour in the series.

The calculated RMSE for this model against test data is 0.620, which shows a slight improvement against the non-seasonally adjusted SARIMA model.

2.6.3 Seasonally Adjusted Polynomial Linear Model

$$3.64 \times 10^2 + lag \times 3.26 \times 10^{-2} + lag^2 \times 6.76 \times 10^{-6}$$

In the 1977 report, the best performing model was the linear model. We therefore also trialed optimizing a linear model with a quadratic elements on the training data.

When we review the projection and residuals for this model, the mean of the residuals doesn't follow as straight of a line as for the ARIMA models. Similar to the pre-1997 data modelling, it had the best performance on the test data with an RMSE of 0.426.

2.7 How bad could it get?

Estimates are provided in the Tables below for when the atmospheric concentration of CO_2 are expected to breach 420PPM and 500PPM. While the seasonally adjusted and non-adjusted ARIMA models are fairly consistent in estimates for each, the polynomial model predicts much higher values. The University Corporation for Atmospheric Research [2] suggests that the climate sensitivity parameter, or the rise in mean global temperature per doubling of atmospheric CO_2 values, is approximately 3C. Both of the ARIMA models are similar in their predictions, but worryingly, the historically more accurate quadratic model is predicting that an atmospheric concentration will be reached over a decade earlier, and that by 2122 the global mean temperature will be significantly (>1.5C) warmer on average than predicted by either of the ARIMA models.

Model	420 (low)	420 (mean)	420 (high)	500 (low)	500 (mean)	500 ppm (high)
Linear	2022 Mar	2022 May	2023 Apr	2033 May	2051 Apr	2052 Apr
SARIMA	2019 Apr	2025 Apr	2035 Apr	2044 May	2056 Mar	2076 Apr
SARIMA (New)	2022-04-24	2022-03-20	2022-01-30	2063-01-28	2058-03-10	2055-02-07
ARIMA (New)	2025-03-30	2023-03-05	2022-05-15	2067-07-24	2059-11-23	2053-07-06
Linear (New)	2023-05-07	2022-10-30	2022-04-17	2048-12-13	2048-05-31	2047-12-01

Type	2122 Low	2122 Mean	2122 High
Linear	775.6497	787.6282	799.6066
SARIMA	469.4515	477.1600	484.8685
SARIMA (New)	624.2220	638.6720	653.1220
ARIMA (New)	610.5660	635.2813	659.9966
Linear (New)	853.9854	864.5171	875.0487

[2] https://scied.ucar.edu/interactive/climate-sensitivity-calculator