W271 Group Lab 2 Team 2

Katt Painter, Bo He, Akanksha Chattopadhyay, Ian Vaimberg

POV: 1997

Context

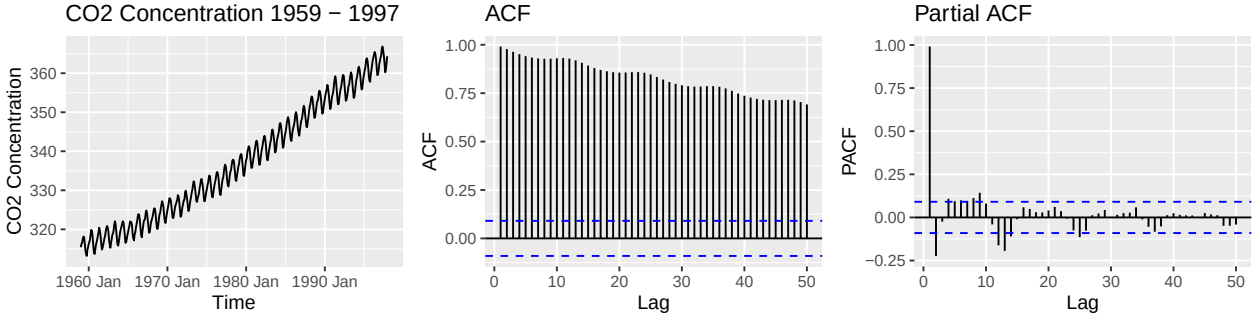
Climate science has emerged as a leading field of interest in the 20th century. Efforts to bring awareness to the impact of human intervention on the environment, such as the burning of fossil fuels, have proved fruitful. The Intergovernmental Panel on Climate Change (IPCC) has reinforced these efforts. In 1990 they released the First Climate Assessment Report stating that "human activities are substantially increasing the atmospheric concentration of greenhouse gases" (IPCC, 1990). Greenhouse gases are a group of gases, such as carbon dioxide, methane, and nitrous oxide, that when present in higher concentrations in the atmosphere, raise the surface temperate of the Earth. Carbon dioxide is the most abundant greenhouse gas that is produced from human activity, namely energy production via fossil fuel combustion. Since the industrial revolution, energy consumption from petroleum and natural gas sources has risen dramatically. This report aims to investigate the following questions: How have the levels of atmospheric CO2 changed over time? And, is there an identifiable pattern that will persist into the future? Forecasting atmospheric carbon dioxide concentrations allows scientists to measure the corresponding impact to the global environment and justify the need for human intervention in the opposite direction.

Data and Exploration

Charles Keeling was a research scientist who made it his life's work to survey the atmosphere in hopes of confirming Svante Arrhenius's theory that fossil fuel combustion is increasing the concentration of CO2 in the atmosphere. To this end, Keeling collected atmospheric CO2 concentration measurements at a number of sampling-stations including the Mauna Loa Observatory in Hawaii. These measurements were taken using a CO2 analyzer which detects the amount of infrared absorption present in a air sample and turns it into a mole fraction of CO2, defined as the total CO2 molecules divided by the total non-water vapor molecules in the air, measured in parts per million (ppm). This report uses the data collected at the Mauna Loa Observatory between January 1959 and December 1997. The dataset consists of 468 observations with each observation representing the monthly total atmospheric concentration of CO2 (ppm). The observations for February, March, and April 1964 were unavailable so the values in the dataset were generated via linear interpolation between the observations for January and May 1964.

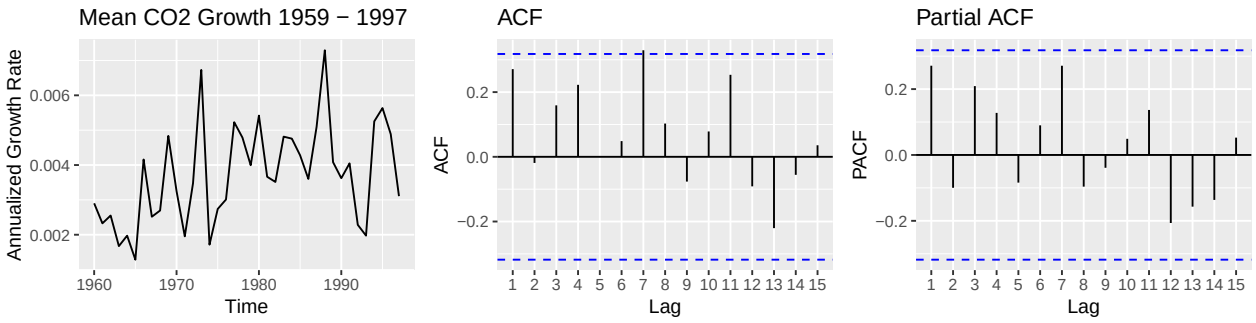
In order to better understand the characteristics of this time series, we conducted a exploratory analysis prior to modeling. Figure 1 shows the time series plot for CO2 concentration, its autocorrelation plot, and its partial autocorrelation plot. The time series plot shows a clear positive trend as well as the presence of seasonality. The autocorrelation plot provides evidence to support the presence of both trend and seasonality as it decays with increasing lags and shows a spike at about every twelfth lag, indicating a seasonal cycle. These characteristics were confirmed by conducting a STL decomposition which showed clear linear trend and seasonal oscillations.

Figure 1. Time Series, ACF, and PACF



Since the time series showed a clear positive linear trend, we also conducted exploratory analysis on the average annual growth rate of CO2 concentrations. Figure 2 indicates that the average annual growth rate of atmospheric CO2 concentration follows a white noise process. While there appears to be some acceleration in the growth rates, the ACF appears to not capture this trend. The annual growth of CO2 appears to be mean reverting, implying that the average long-run annual growth rate is expected to stay between 0.2% and 0.6% holding all else constant. If we were to ramp up activities that further increase CO2 emissions, we could expect a clearer linear trend to emerge.

Figure 2. Time Series, ACF, and PACF



Models and Forecasts

While the exploratory analysis allows us to infer the underlying process of the time series, we cannot confirm this process without empirical modeling. In this section we will produce multiple models from the following classes, linear models and ARIMA models. Following the modeling and model evaluation, we will use the models to produce forecasts for what the CO2 concentration will be over the next 20 years.

Linear Models

Three linear models were considered, these models are describe by the following set of equations:

$$y_t = t + \epsilon_t \tag{1}$$

$$y_t = t + t^2 + \epsilon_t \tag{2}$$

$$y_t = t + t^2 + d_{2,t} + d_{3,t} + ... + d_{11,t} + d_{12,t} + \epsilon_t \tag{3}$$

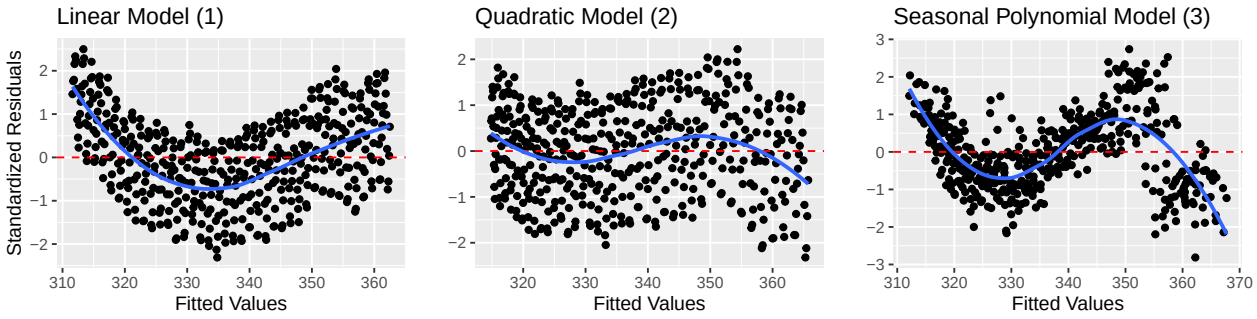
where t is the time index (i.e., year-month) represented as an integer and d is a dummy variable representing each month. For example d_2 would correspond to the month of February, d_3 would correspond to the month of March, and so on.

The following table reports the coefficients that produce the best fit for each of the previously defined models. It also reports several performance metrics such as R-squared, adjusted R-squared, F statistics, etc. The table suggests that each model explains a large amount of the variation in CO2 emissions with adjusted R-squared values ranging from 0.97 to 0.99. This indicates that these models appear to perform extremely well. The p-values associated with each predictor are also significant, indicating that each predictor increases the explanatory power of the model. However, in order to determine the appropriateness of using linear models to estimate CO2 concentrations we must examine the residuals of each model.

	Response: Atmospheric CO2 Concentration		
	(1)	(2)	(3)
Month Index	0.11^{***} (0.001)	0.07^{***} (0.003)	0.07^{***} (0.001)
Month Index^2		0.0001^{***} (0.0000)	0.0001^{***} (0.0000)
Feb			0.66^{***} (0.16)
March			1.41^{***} (0.16)
April			2.54^{***} (0.16)
May			3.02^{***} (0.16)
June			2.35^{***} (0.16)
July			0.83^{***} (0.16)
Aug			-1.23^{***} (0.16)
Sep			-3.06^{***} (0.16)
Oct			-3.24^{***} (0.16)
Nov			-2.05^{***} (0.16)
Dec			-0.94^{***} (0.16)
Constant	311.50^{***} (0.24)	314.76^{***} (0.30)	314.68^{***} (0.15)
Observations	468	468	468
R^2	0.97	0.98	1.00
Adjusted R^2	0.97	0.98	1.00
Residual Std. Error	2.62 (df = 466)	2.18 (df = 465)	0.72 (df = 454)
F Statistic	$14{,}794.94^{***}$ (df = 1; 466)	$10{,}749.90^{***}$ (df = 2; 465)	$15{,}314.77^{***}$ (df = 13; 454)
Note:			*p<0.05; **p<0.01; ***p<0.001

Figure 3 shows the standardized residuals against the fitted values for each linear model. There are clear patterns present in each of these plots. The first model shows a U-shaped curve that indicates the model is under predicting CO2 concentrations that are in the mid-range and over predicting CO2 concentrations closer to the edges. Whereas the quadratic model appears to oscillate in whether it is over or under predicting CO2 concentrations. The polynomial model appears to also have a U-shaped pattern. However, compared to the first model, its errors appear to be larger, especially at edge values. Nevertheless, the residuals appear to be homoskedastic for all three models which suggests that no transformation (i.e., logarthmic) is needed. Overall, these plots suggest that linear models are not appropriate for this particular data set. They do not appear to adequately capture the underlying time series process. So we must turn to ARIMA models.

Figure 3. Linear Model Evaluation



ARIMA Models

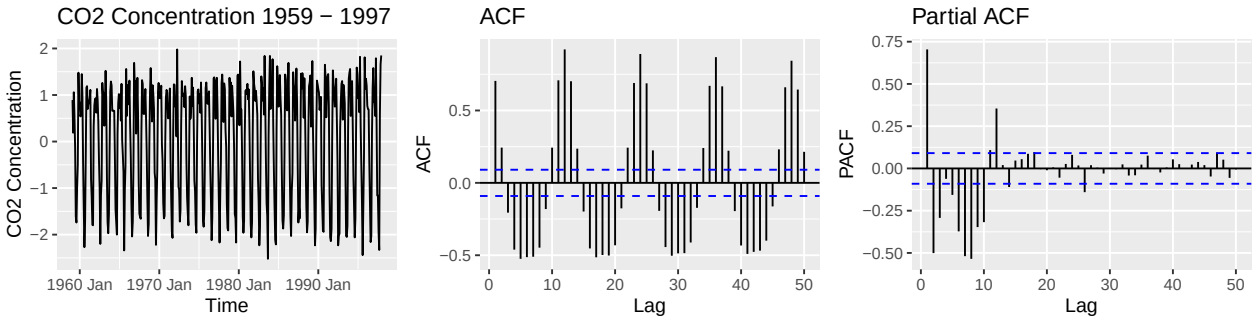
Due the strong presence of seasonality in the data, the next set of models are seasonal autoregressive integrated moving average models (SARIMA). The general notation (4) used to express these models is stated below as well as the short hand notation (5).

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D x_t = \mu + \theta(B)\Theta(B^s)\epsilon_t \tag{4}$$

$$SARIMA(p,d,q)(P,D,Q)_s \tag{5}$$

These models include moving average (MA) terms, autoregressive (AR) terms, and differencing (D) terms at non-seasonal and seasonal lags. This allows the model to more accurately capture and account for the seasonality present in the data. To help determine the seasonal MA, AR, and D terms we took a first difference of the data and re-plotted Figure 1 with the differenced data. Figure 4 shows this plot. The plot indicates that after one difference the data is near stationary suggesting that the differencing term should be 1. We performed ADF and KPSS test to confirm stationarity. Moreover, the PACF shows a large spike followed by oscillations between positive and negative correlations which appear to weaken as the lag value increases. This indicates the presence of a higher order MA term. The plots do not appear to support the presence of a AR term.

Figure 4. Time Series, ACF, and PACF After First Difference



Based on the assessment of Figure 3, we fit three SARIMA models of the form:

$$SARIMA(0,1,1)(0,1,1)_{12} \tag{6}$$

$$SARIMA(3,1,0)(3,1,0)_{12} \tag{7}$$

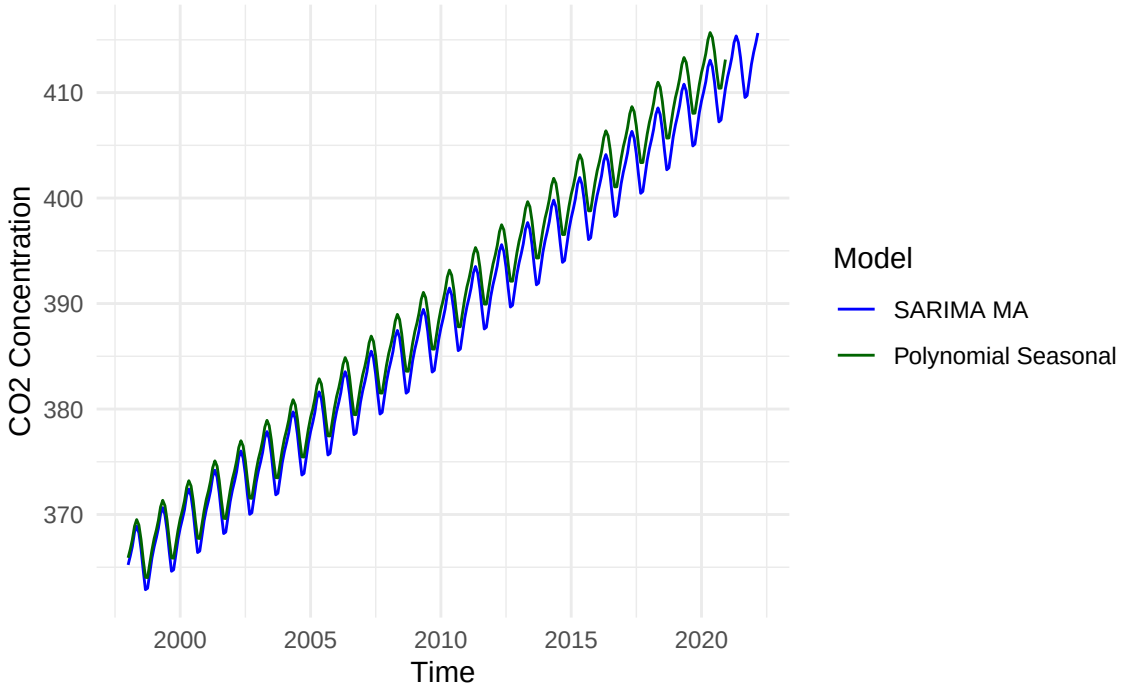
$$SARIMA(0,1,1)(1,1,2)_{12} \tag{8}$$

The final model was selected using the BIC scores with the ultimate winner being the $SARIMA(0,1,1)(0,1,1)_{12}$ model with a BIC value of 194.7. BIC was chosen due to its inherent penalty on adding additional terms leading to a well fitted and simpler model. Finding the ideal model is done through grid search of various parameters as we employed three simultaneous searches to see the most variety of different model specialties. All searches included non-zero difference terms based on the our EDA while one model focused on AR terms, another MA terms and one finally on both at the same time. To further assess the appropriateness of this model, we looked at a residual ACF plot which closely resembles that of a white noise process with lags rarely falling outside of the 95% confidence interval around 0 correlation. We also conducted a Box-Ljung test which had a p-value greater than 0.05, therefore the residuals do not exhibit significant autocorrelation giving greater confidence that the SARIMA model found is appropriate to model the CO2 dataset.

Forecasts

We used the models from equations 3 and 6 to generate forecasts of the expected atmospheric CO2 concentrations over the next 20 years. These forecasts are shown in Figure 5 below. The forecasts generated by each model appear to be similar in nature and both support that the trend and seasonal patterns present in the CO2 data will persist into the future.

Figure 5. Polynomial and SARIMA Model Forecasts



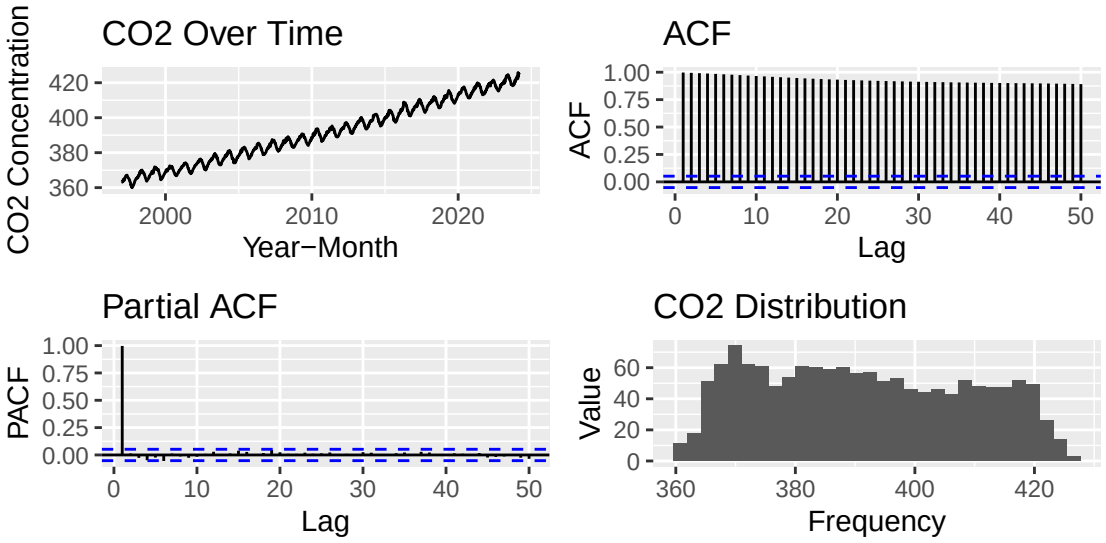
We can also use the forecasts to determine the year and month that CO2 concentrations will surpass specified thresholds. In particular, we were interested in identifying when the CO2 is expected to be at 420 ppm and 500 ppm, respectively. Climate scientists have indicated that these thresholds are significant due to the associated impact to the global climate such as further global warming, extreme weather, ocean acidification, and other detrimental effects. The SARIMA model estimates that the CO2 concentration will reach 420 ppm for the first time at May 2023 and for the last time at November 2025. A 95% confidence interval indicates that first time at which this threshold will be reached may vary between May 2017 to May 2036 and the last time will vary between December 2018 and November 2039. The model estimates that CO2 concentration will reach 500 ppm for the first time at March 2053. The 95% confidence interval for this estimate varies between April 2040 and May 2078. The model estimates that CO2 concentration will reach 500 ppm for the last time at December 2053. The 95% confidence interval for this estimate varies between October 2041 and November 2080. Lastly, the model estimates that at 2100, the atmospheric CO2 concentration will reach 674 ppm. Confidence is low in all of these predictions due to the widening variance of the predictions over time. Despite the strong cyclical patterns forecasting this far out in time is an incredibly hard task to do accurately.

Conclusions

The primary questions posed at the beginning of this report were: How have the levels of atmospheric CO2 changed over time? And, is there an identifiable pattern that will persist into the future? The time series analysis conducted in this report addresses these questions. The levels of atmospheric CO2 have increased overtime. Moreover, there are seasonal and trend components in the data generation process that suggest the levels of atmospheric CO2 will continue to increase overtime at a relatively constant growth rate. It is important to understand that these conclusions are in the context of air samples taken at the Mauna Loa Observatory in Hawaii. To address this bias, it would be interesting to conduct the same analysis using data comprised of air samples from other land and sea stations around the world. Nevertheless, it is clear that there is a need for human intervention to decrease carbon dioxide emissions. Left unchecked carbon dioxide emissions will increase to a alarming point sometime between 2050 and 2100.

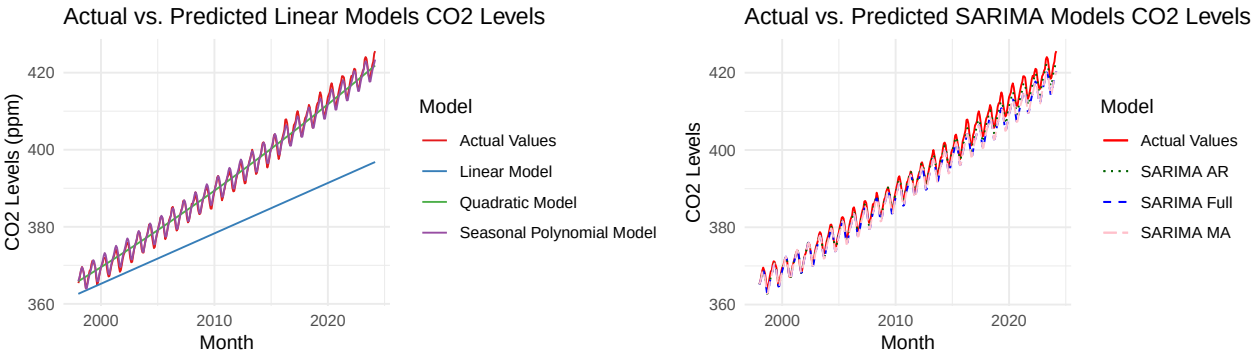
POV: Present

Moving forward in this report, we will delve deeper into the evolution of atmospheric CO2 levels and draw comparisons between the patterns projected in the previous sections and those present in the current dataset. This analysis will leverage data sourced from the United States' National Oceanic and Atmospheric Administration, collected at the Mauna Loa Observatory in Hawaii, consistent with our previous dataset. The data was compiled on a weekly basis, averaging CO2 values across days with valid data within each week. We retrieved the data by creating a data pipeline with the appropriate URL to the Global Monitoring Laboratory Website. Now we will examine the exploratory plots below to better understand the data.



The time series plot clearly shows an increasing trend in CO2 levels over time with regular seasonal fluctuations. This trend is consistent with the historical Keeling Curve data. The ACF plot reveals strong positive correlations at all lags, suggesting a very persistent and seasonal pattern in the CO2 data. In contrast, the PACF plot shows little to no significant correlations beyond the initial lag, implying that an autoregressive model may not be the best fit for this data. The distribution of CO2 concentrations indicates a multimodal distribution, suggesting the presence of different states in the data, potentially reflecting various environmental factors. With this understanding of our current data, we can now turn to compare it with the model forecasts from earlier sections of the report.

Comparision of Linear and SARIMA Models with Present data



In the graph comparison of linear models, the forecast from the seasonal polynomial model aligns closely with the actual CO2 values. The simple linear model consistently underestimates the CO2 levels. The quadratic model shows an improved fit in terms of the overall upward trend compared to the simple linear model, but it doesn't fully capture the periodic seasonal variations. Turning to the SARIMA models, forecasts from all variants—AR, MA, and Full—align closely with the actual CO2 values. The Keeling Curve from 1997 to the

present seems to have evolved pretty similarly to the historical curve.

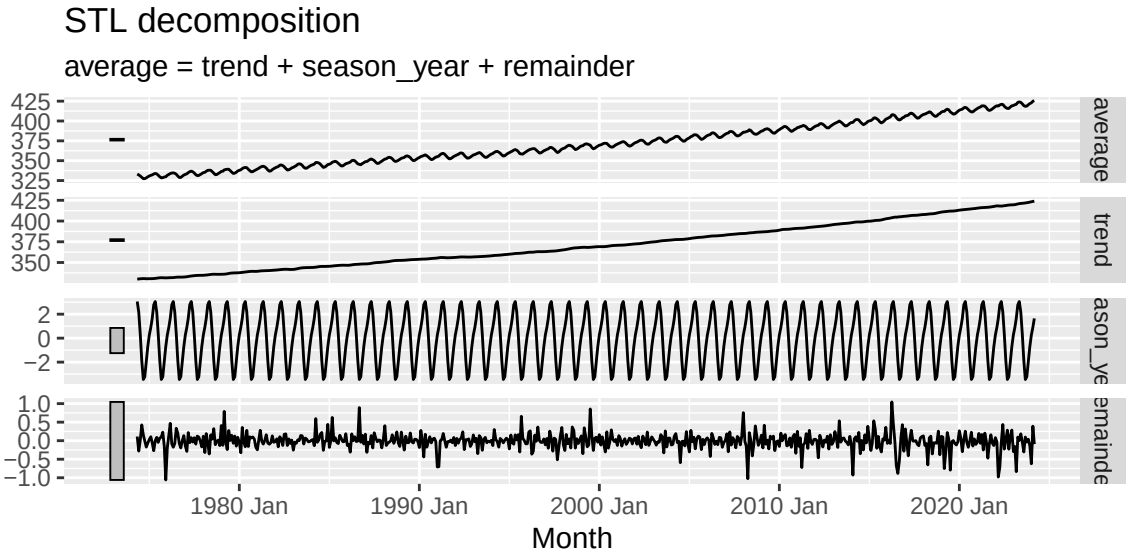
In 1997, our forecasts estimated when atmospheric CO2 levels might first surpass the 420 ppm threshold. The best-performing model, $SARIMA(0,1,1)(0,1,1)_{12}$, suggests that this milestone would be reached in different years under various scenarios: May 2017 was projected as a high extreme case, May 2023 as the model's central prediction, and May 2036 as a low extreme case. Actual measurements extracted from the Mauna Loa Observatory indicate that this level was first crossed in April 2022. This outcome demonstrates that the Seasonal ARIMA model provided a reasonable estimation, with the actual occurrence falling between the predicted central and high extreme case scenarios.

```
##                                ME    RMSE    MAE     MPE   MAPE   ACF1
## Linear Model               6.4116 10.3944 6.4116  1.5981 1.5981 0.9832
## Quadratic Model            0.0621  1.6161 0.9692  0.0117 0.2468 0.8338
## Seasonal Polynomial Model 0.0081  0.5776 0.3277 -0.0007 0.0828 0.8421
## SARIMA AR                  0.4229  0.7930 0.4554  0.1055 0.1140 0.8670
## SARIMA MA                  0.9848  1.6777 0.9929  0.2449 0.2471 0.9632
## SARIMA Full                0.9719  1.6585 0.9804  0.2417 0.2440 0.9627
```

Looking at the performance metrics of all three linear and SARIMA models, we see that the simple linear model shows high errors across all metrics. In contrast, both the quadratic and seasonal ploynomial models demonstrate lower error values, suggesting a more accurate representation of the data trends. The SARIMA models—SARIMA AR, SARIMA MA, and SARIMA Full—present moderate errors, with SARIMA MA and SARIMA Full yielding similar performance metrics. Among the models assessed, the seasonal polynomial model stands out with the lowest RMSE value at 0.5776, highlighting its accuracy in tracking the actual data points.

Models

As we have observed in our EDA section and observations from 1997 report, we believe the CO2 levels have both overall upward trend and very possible seasonal trend. We used STL decomposition to decompose the data into 3 components: 1) upward trend using 6 months window, 2) seasonal trend which is observed as yearly trend, and 3) remainder, which the mean is observed close to zero, and fluctuates around reasonably bounded variance. We will proceed with further investigation to check for stationary on the remainder.

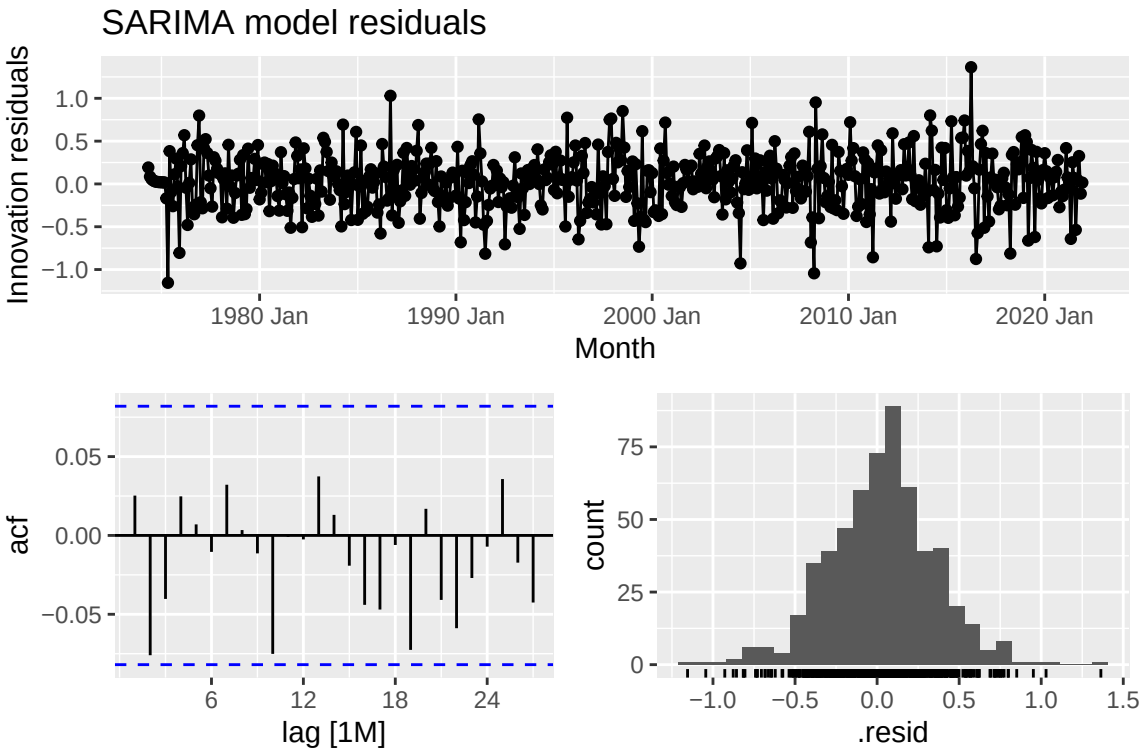


From the residual plots and the Box-Ljung test results in p-value of $0.04 < 0.05$, although the residuals of the decomposition appeared to be stationary, they do not appear to be completely white noise. This means that while the decomposition method eliminates the deterministic components from this specific time series, there are some correlation remains in the data. Combined with the knowledge that we learned from 1997

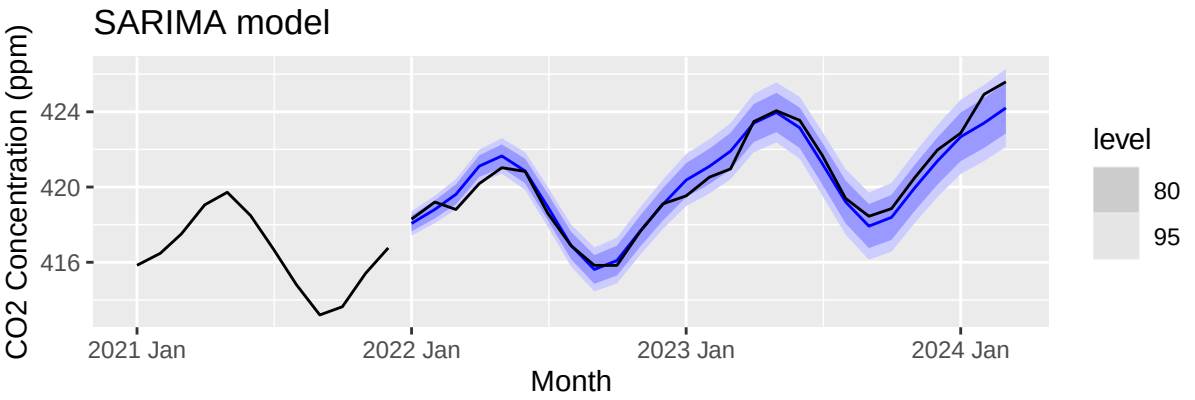
report, we decided to fit our data using SARIMA and polynomial time models with our new data.

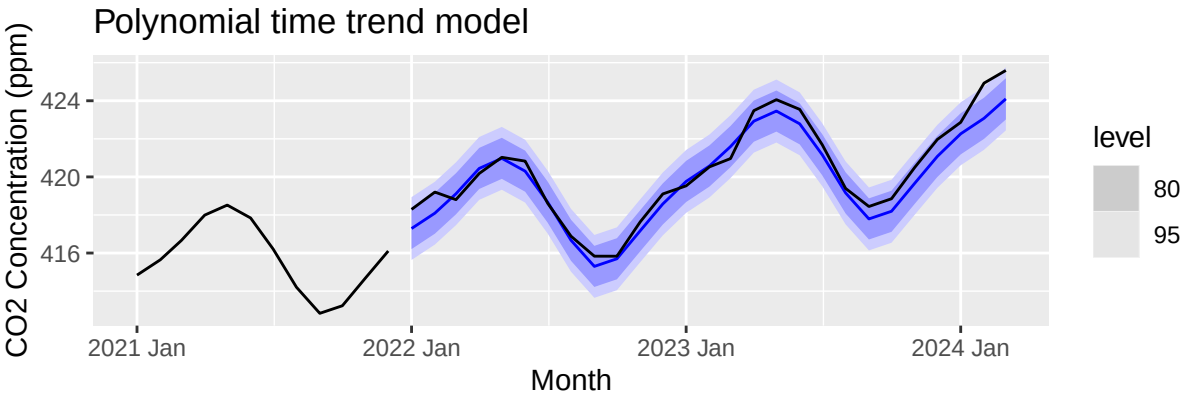
Prior to model building, we have divided up our data to use data points prior to 2022 as our training dataset, and post 2022 as our testing dataset. Then, we start fitting SARIMA and polynomial time models using our training data and test the accuracy using the testing data.

SARIMA Model



The residuals of the $SARIMA(0,1,1)(2,1,2)_{12}$ model appear to be close to white noise. There is no significant lags, but with some seasonal pattern that suggests we should run a statistical test on the residuals from the models to see if they are randomly distributed i.e. are white noise, which is what we want for a good model fit, or if they appear to have some serial correlation over time and violate the assumptions for a stationary time series fit. The Box-Ljung test result p-value of $0.53 > 0.05$ confirmed that the residuals of the SARIMA model are stationary and appeared to be white noise. And we performed the similar steps and test on Polynomial model confirmed the same, so we decided to use both models for forecasting and accuracy comparison.

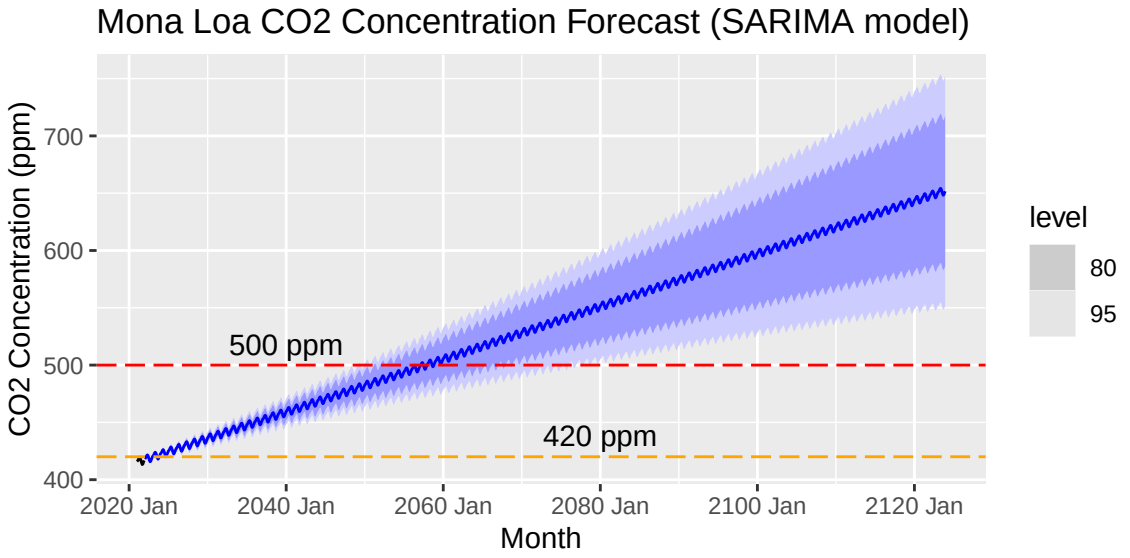





```
## # A tibble: 2 x 10
##   .model      .type     ME  RMSE   MAE    MPE  MAPE  MASE RMSSE  ACF1
##   <chr>       <chr>  <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 .model      Test  0.0718 0.611 0.469 0.0167 0.111   NaN   NaN 0.574
## 2 trend_model Test  0.470  0.718 0.584 0.111  0.139   NaN   NaN 0.437
```

From the above accuracy table, it is evident that SARIMA model has outperformed polynomial time-trend model in terms of minimize the forecast error in the test dataset. As a result, we have decided to use SARIMA model to forecast out fo 2122.

Forecast to 2122



We can see that using our SARIMA model the forecasts also have a trend and seasonal movement and fluctuations increase overtime. But these forecasts will get very inaccurate as we move beyond at best 5 year forecasts of the model, as the confidence intervals of the model prediction begin to open up very widely due to the inherent model restriction and the unpredictability of the future. We can be pretty confident to say that we will hit 420 ppm CO2 level in 2023-2025, less confident about 500 ppm CO2 level as the model suggest it can be as early as 2045 or as late as 2060 or beyond.