The Evolution of CO2 Levels from 1997 to Present

by Qiong Zhang, Ruiyu Zhou, Tejas Shirshikar, and Maria Lee

I. Background

A. The Keeling Curve

In the late 1950's, Charles Keeling initiated groundbreaking scientific endeavor, systematically measuring atmospheric carbon dioxide (CO_2) levels in Mauna Loa, Hawaii (1). This work unveiled a striking pattern in the CO_2 data, contradictory to the previous publications pointing to high variability (1)(2). Now known as the Keeling Curve, these findings have become an important reference in climate science, offering crucial evidence of the rapid changes occuring in our planet's atmosphere due to human activities.

B. Analyzing Trends in CO2 data: Significance and Evolution

Centering our analysis on the critical question of *What trends can be identified in the CO_2 data up to 1998, and what do they indicate about global environmental changes?* The investigation into seasonal fluctuations and the long-term increase in CO2 concentrations aimed to unravel insights into both the natural cycle affecting our climate system and the anthropogenic impacts stemming from increase fossil fuel combustion and agricultural activities(2).

The continuation of this analysis brings forth a pressing question: *How accurately have the models we utilized in the data up to 1997 predicted the CO_2 levels measured in the following years?* This inquiry probes not only the precision of our forecasts but also delves into whether any discrepancies highlight limitations within our models.

C. Technological and Methodological Advancements since 1997

Since 1997, significant advancements occurred in the data generating process for measuring CO_2 levels at the Mauna Loa Observatory. The adaptation of a new CO_2 analyzer, Cavity Ring-Down Spectroscopy (CRDS) technology, in April 2019 was a pivotal upgrade, replacing the previous infrared absorption-based analyzer. Calibration methods also evolved, with meticulous control of temperature, pressure, and flow rate, along with frequent calibrations using reference gas mixtures. Furthermore, detailed data selection criteria have been implemented to identify background air, which aimed to eliminate local influences on CO_2 measurements (3). In addition to the advancements, there was a disruption in measurements from November 2022 to July 2023 due to the eruption of the Mauna Loa Volcano, during which observations were conducted from the Maunakea Observatories approximately 21 miles north of the Mauna Loa Observatory. However, observations at Mauna Loa resumed in July 2023, ensuring continuity in the long-term CO_2 monitoring efforts (4).

D. Aims and Implications of Continued Analysis

As we extend our analysis into the present, our aim is not only to validate past predictions but also contribute to a deeper understanding of CO_2 impacts on Earth. This ongoing investigation serves as both a reflection on past observations and a forward-looking lens into future climate scenarios.

Reference Numbers (To move to the end later): (1) Autobiography of Keeling (2) First Publication (3) https://gml.noaa.gov/ccgg/about/co2_measurements.html (4) https://gml.noaa.gov/ccgg/trends/data.html

II. Measurement and Data

A. Measuring Atmospheric Carbon

This analysis begins with a meticulous measurement of atmospheric carbon dioxide levels, generated at the Mauna Loa Observatory. Located at a high elevation of 3,400 meters at near the summit of the Mauna Loa

volcano in Hawaii, this observatory is situated far from significant urban pollution sources and vegetative influences, providing an optimal setting for gathering representative samples of the global atmosphere. CO_2 measurements are collected by measuring the mole fraction of CO_2 in dry air and the new CRDS technology mentioned earlier, measures the rate of light absorption in an optical cavity rather than the magnitude of absorption, offering enhanced precision (3).

B. Historical Trends

In the left top panel of Figure 1, the time series plot of CO_2 concentrations up to 1997 clearly indicates a robust and consistent trend as well as seasonality within the monthly mean CO_2 data. This steady rise indicates the cumulative impact of human activities, primarily the burning of fossil fuels, which is corroborated by the histogram in the top-right panel. The histogram reveals the distribution of CO2 measurements, with most data points clustering around the higher end of the scale as time progresses, which supports the trend observed in the time series plot. The ACF plot in the bottom-left panel displays significant and sustained autocorrelation across numerous time lags, which denotes the presence of seasonality within the data. This seasonality is evidenced by the regular oscillations in CO_2 levels due to the cyclical nature of plant growth and decay, especially prominent in the northern hemisphere with its larger landmass and extensive vegetation (2). Lastly, the PACF plot in the bottom-right panel presents a first lag yielding a value of 1, indicating the present of a unit root within the series.

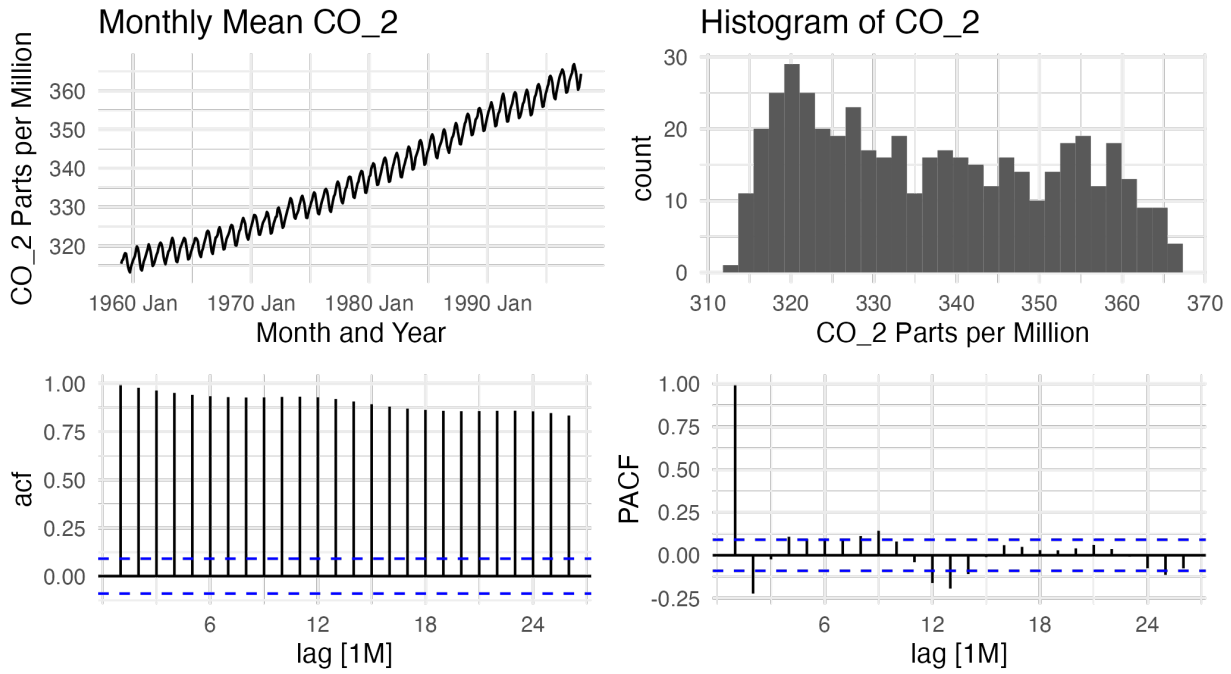


Figure 1: 1997 CO2

To further investigate this, first-differences CO_2 was analyzed to check the stationarity. In Figure 2, we see the elimination of the long-term trend, exposing the underlying seasonality more clearly. Which is further supported by periodic oscillations in the ACF plot. Significant lags for non-seasonal MA terms are observed, along with potential indications of seasonal MA terms. The PACF suggests the presence of non-seasonal and seasonal AR lags. To formally assess the stationarity of the differenced series, a KPSS root test was performed. The results from the KPSS test, with a p-value of 0.1 indicate that the null hypothesis of stationarity cannot be rejected, supporting the conclusion that the first differencing of the CO_2 series was necessary to achieve a stationary time series.

The decomposition analysis in Figure 3, reflecting the persistent rise in CO_2 levels, alongside regular seasonal

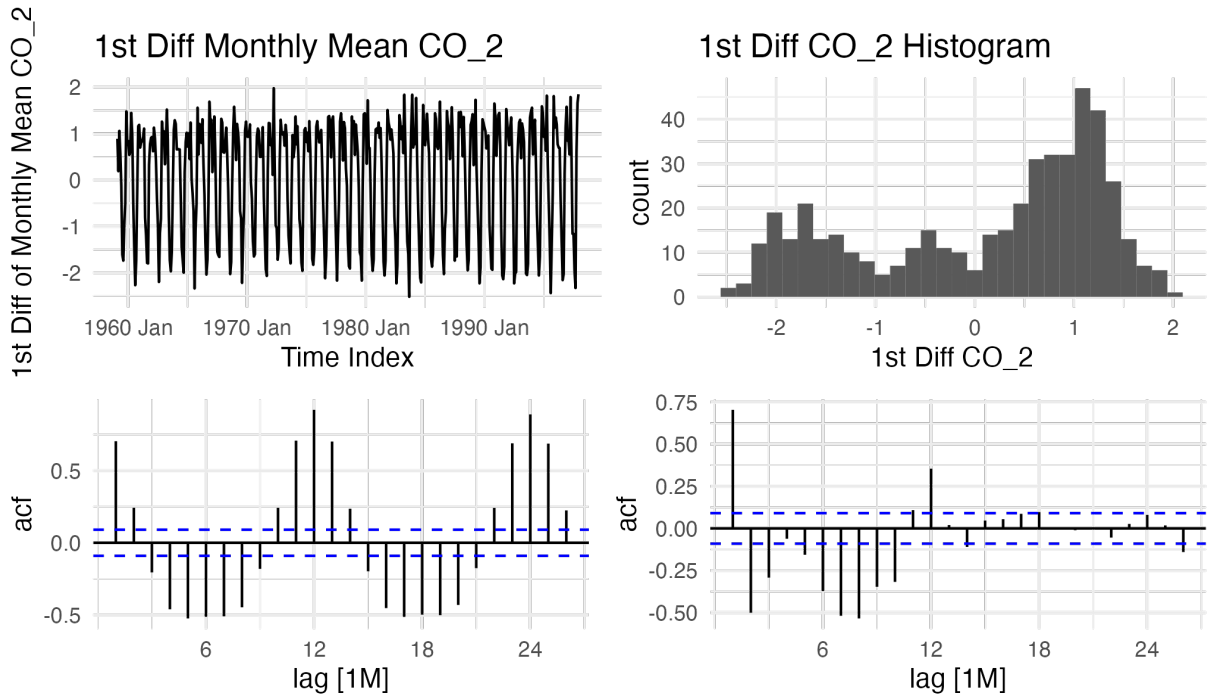


Figure 2: 1997 CO_2 First Difference.

patterns that show no change in magnitude over time. The first differencing of the data confirms these findings, indicating stable month-to-month variations once the trend is removed.

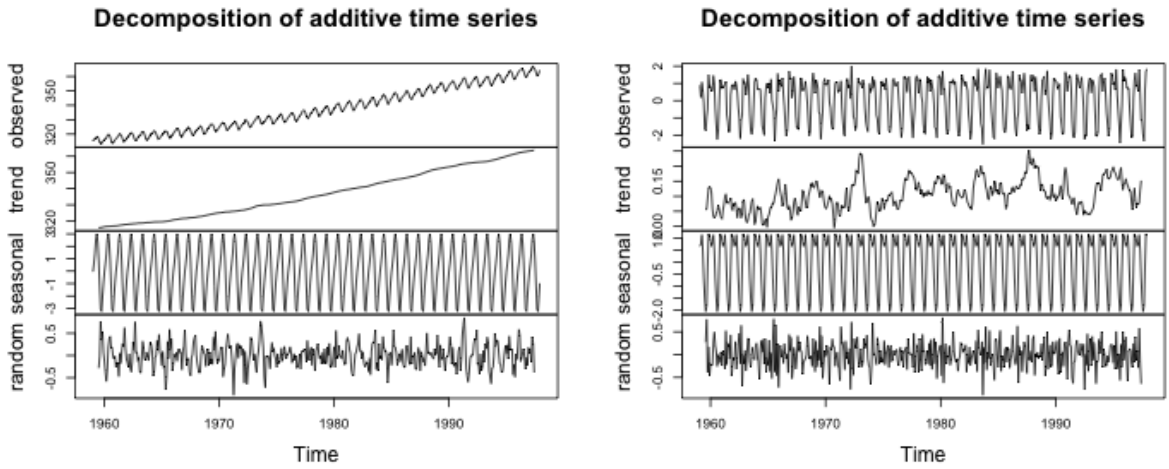


Figure 3: Decomposed CO2 and First Difference CO2 time series

Percentage growth rate assessment, indicates a steady increase in atmospheric concentrations, consistently below 0.8% annually. Despite appearing small, this percentage represents a meaningful and compounding rise in CO_2 levels, indicative of an accelerating trend with significant long-term implications.

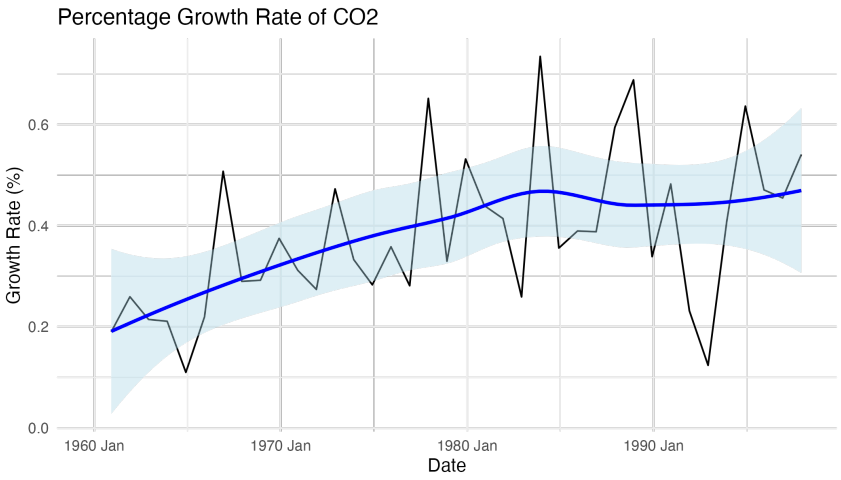


Figure 4: Percentage Growth Rate Assessment

C. Present Trends

When assessing the evolution of the Keeling Curve from 1997 to the present, we continue to observe a pronounced upward trend in CO2 levels, echoing the patterns identified in the earlier data. The ACF and PACF plots display sustained autocorrelation and the presence of a unit root, respectively, consistent with the historical analysis which necessitated differencing to achieve stationarity. The KPSS test corroborates this, with a p-value indicating stationarity in the differenced series. These findings confirm the ongoing influence of human activities on CO2 concentrations, with the percentage growth rate analysis further highlighting a compounding rise in levels, maintaining a rate below 0.8% annually. This persistence of trends and seasonality from the 1997 data to the present underscores the unrelenting trajectory of CO2 accumulation in the atmosphere.

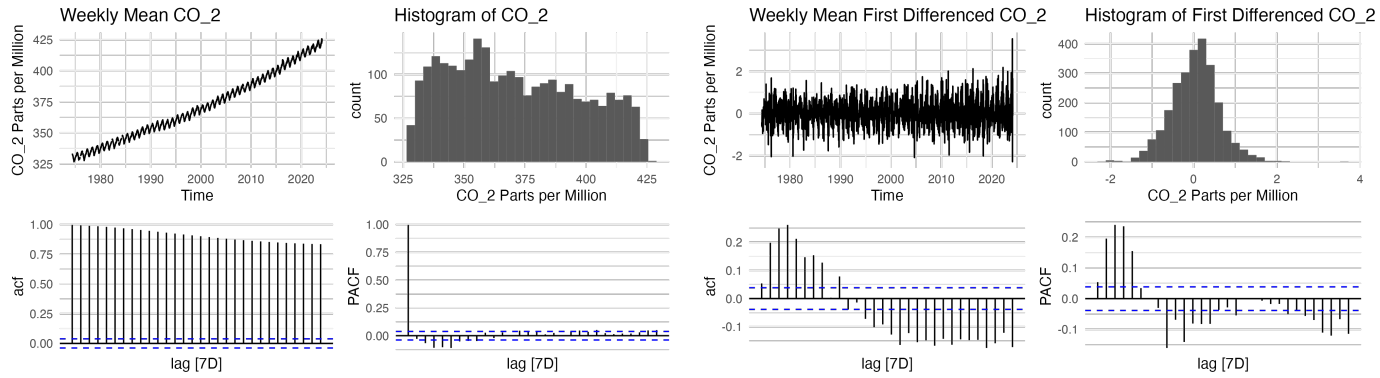


Figure 5: Present CO2 and First Differenced

The time series decomposition graphs illustrate the transformation from non-stationary to stationary data. The original dataset depicts an upward trend, signifying non-stationarity, with clear seasonality and considerable random fluctuations. Post differencing, the trend component is neutralized, evidencing stationarity with a consistent mean. The seasonal patterns remain unchanged, indicating their persistence regardless of stationarity. The random component, though still volatile, is now centered around zero without a discernible trend, characterizing the achieved stationarity.

III. Models and Forecasts

Diving into predictive analytics, we transitioned our focus from exploratory data analysis to the development and application of statistical models. A crucial step towards forecasting future levels of atmospheric CO_2 based on historical data trends.

A. Linear Model

Initiating our examination of the CO_2 series, we applied a linear model to uncover basic trends. When fitting a linear time trend model to the CO_2 data, the residuals displayed in Figure 5 exhibit systematic patterns, hinting at unaccounted seasonal effects. The ACF plot further confirms this with significant auto correlations at seasonal intervals. This suggests the linear model's limitations in capturing the CO_2 series' intricacies. Shifting to the quadratic model, we can see that the residuals in Figure 6 compared to the residuals in Figure 5 are more stationary. We can also see in the that the histogram in Figure 6 looks more normal than Figure 5. The quadratic model seems to fit the data better than the linear model. However, there is still seasonality in the ACF that needs to be addressed. We took a look at additive and multiplicative decomposition to see if using the log of CO_2 concentration will help in fitting the data.

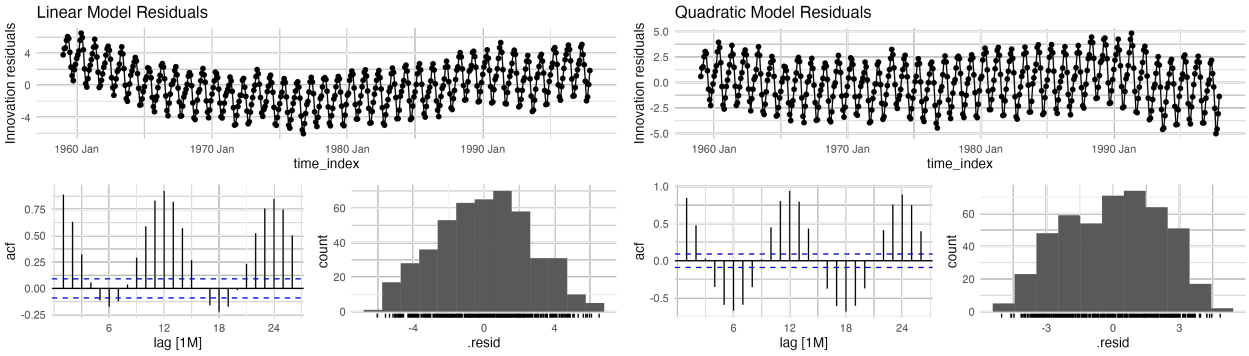


Figure 6: Linear and Quadratric Model Residuals

When comparing at the residuals in additive decomposition and multiplicative decomposition, we can looking at the ACFs that the residuals are less pronounced when we use multiplicative decomposition. Therefore by logging the values of CO2 we are able to better capture any non linear relationships that we have in our data. Next we will fit polynomial time trend model that incorporates seasonal dummy variables to capture the seasonality of the data. We will also be using the log of CO2 in our model.

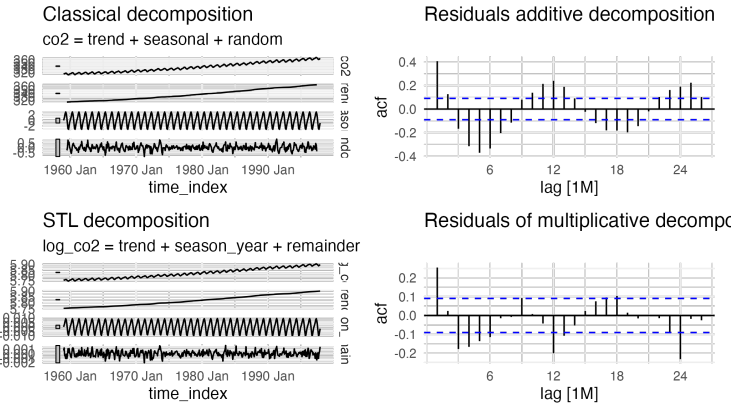


Figure 7: Additive and STL Decompositions

Based on Figure 7, we can see that the seasonal dummy variables have done a much better job capturing

the seasonality in our data than our previous models. The ACF shows no instance of seasonality. Based on the scale of the residuals, we can see that our data is better captured by the seasonal dummy variable, as the scales in the graph of the residuals is much smaller than the scale in Figures 3 and 4. Although we are capturing the data better, we do not have white noise as the correlations in the ACF as we increase the lag still seem to be significant. Next we will forecast using this model.

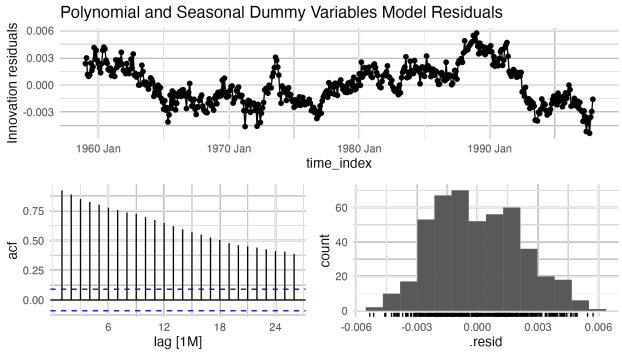


Figure 8: Polynomial and Seasonal Dummy Variables Model Residuals

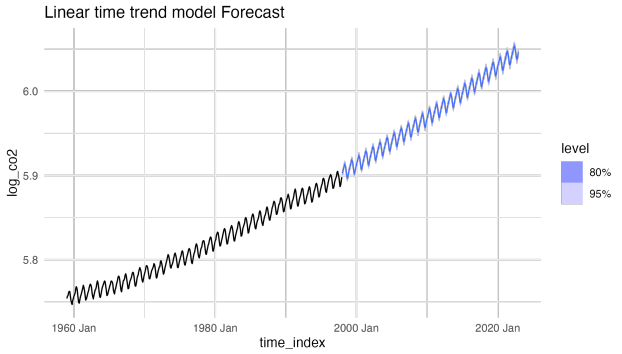


Figure 9: Linear time trend model Forecast

B. ARIMA Time Series Model

From above sections, we have proved that co2 has 1 unit root and lags in both non-seasonal and seasonal MA and AR models. So, our basic model will search MA(p:0-3) and AR(q: 0-3) with d=1. The EDA of the first differencing of CO_2 indicates that the first differenced CO_2 data has strong seasonality while doesn't have persistent and obvious trend. The ACF further verified the yearly seasonality as the autocorrelation peaks at lag 12 and 24. Thus, we tested different ARIMA models below. We will use ARIMA() to find out the exact number of lag by searching a set of different possible models, comparing AIC/BIC, and selecting the model with the lowest values. The optimal model based on pre-defined criteria is ARIMA(0,1,1)(1,1,2), which has non-seasonal and seasonal difference 1,seasonal lag for AR(1), non-seasonal and seasonal lag for MA(1,2) model, which is close to what we guess and observed before. We noticed that the model without intercept has much lower BIC(201.78) than the model with intercept(692.46).

```
## Series: co2
## Model: ARIMA(0,1,1)(1,1,2)[12]
##
## Coefficients:
##           ma1     sar1     sma1     sma2
##       -0.3482  -0.4986  -0.3155  -0.4641
## s.e.   0.0499   0.5284   0.5167   0.4369
```

```
##
## sigma^2 estimated as 0.08603:  log likelihood=-85.59
## AIC=181.18   AICc=181.32   BIC=201.78
```

```
## Series: co2
## Model: ARIMA(3,1,1)(0,0,2)[12] w/ drift
##
## Coefficients:
##          ar1      ar2      ar3      ma1    sma1    sma2  constant
##       1.1159  -0.1776  -0.3450  -0.9252  0.6573  0.3792    0.0427
## s.e.  0.0497   0.0738   0.0456   0.0167  0.0506  0.0417    0.0034
##
## sigma^2 estimated as 0.2311:  log likelihood=-321.64
## AIC=659.29   AICc=659.6   BIC=692.46
```

Then we use residual to check the model fitness. The Figure 9 shows that the histogram of residual close to normal distribution. The acf plot of residual shows most lags within the limit with only 2 significant lags. The Ljung Box test also proved that the we cannot reject the null hypothesis (p=0.1441) and the data are independently distributed and residual does not have serial correlation over time and stationary.

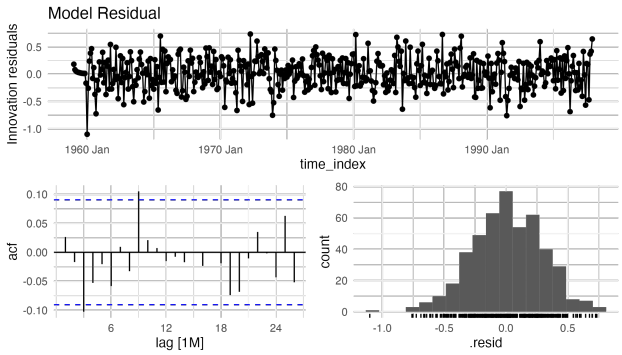


Figure 10: Model Residuals

In next section, we will demonstrate the predicted data to 2022 by using selected ARIMA model.

C. Atmospheric CO2 growth Forecast

In ARIMA forecast, it looks like the model captures the general increasing trend of co2 as well as the seasonality within entire predicting period.

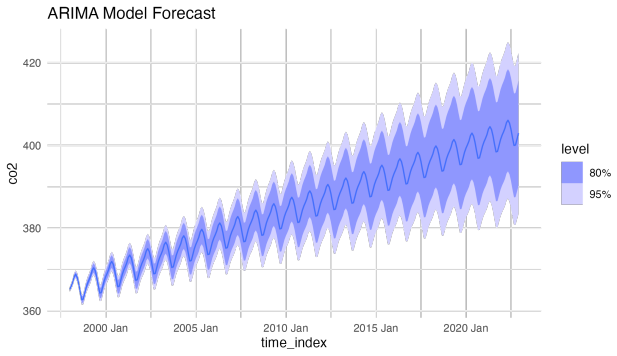


Figure 11: ARIMA Model Forecast

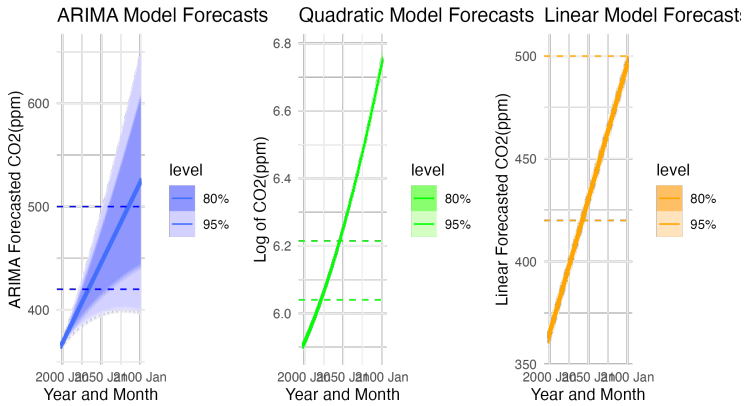


Figure 12: Three Model Forecasts

We calculate the difference between the point prediction (`.mean`) and 420, and extrac the row with lowest absolute difference. In this way, we can obtain the month and year of which CO_2 is closest to 420. Based on this approach, we found that in December 2033, the point prediction of CO_2 level is closest to 420, with variance equal to 242. We also calculate the 95% confidence interval, and we are 95% confident that the true CO_2 value is between 389.5 to 450.5.

We calculate the difference between the point prediction (`.mean`) and 500, and extract the row with lowest absolute difference. In this way, we can obtain the month and year of which CO_2 is closest to 500. Based on this approach, we found that in June 2083, the point prediction of CO_2 level is closest to 420, with variance equal to 242. We also calculate the 95% confidence interval, and we are 95% confident that the true CO_2 value is between 402.8 to 597.3.

The point prediction of CO_2 levels in year 2100 ranges from 521.1 to 517.2. Throughout the entire year of 2100, we are 95% confident that the CO_2 levels fall within then range of 400 to 600.

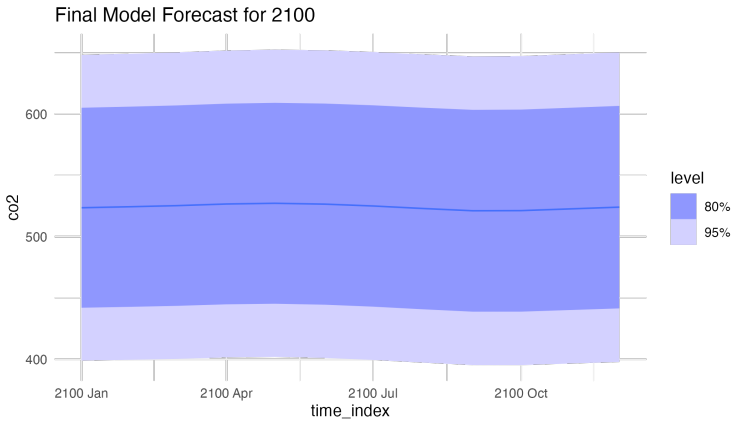


Figure 13: Final Model Forecast for 2100

D. Comparing models forecasts against realized CO2

In comparing the linear model forecasts against realized CO2, the actual monthly mean CO2 levels has a systematic increasing trend and regular fluctuations in fixed time period, indicating consistent growth and seasonality. The decomposition further proved its non-stationarity, increasing trend, and seasonality. In previous sections, we used linear model with quadratic term and season to capture the increasing rate and seasonality and predict the CO2 till 2022 Dec. The figure shows that the peak of 2020 is slightly below

6.05, while the peak of 2022 is over 6.05. In actual data plot, we can found that the peak of 2020 is around 6.03, while th peak of 2022 still below 6.05. Therefore, we can say the linear model was able to predict the seasonality in the predicted data but the systematic increasing trend was slightly over estimated, indicating higher slope coefficient in predicted trend than the actual trend of realized atmospheric CO2 data.

The ARIMA model we obtained in previous sections predicts that the CO2 will have a steady increasing trend with period fluctuation. However, contrary to linear model, the ARIMA model seems underestimate the CO2 increase by predicting the the peak of 2020 is slightly over 400, while the peak of 2022 is around 405. In actual data plot, we can found that the peak of 2020 is much over 410 even approaching 420 and the peak of 2022 is over 420. Therefore, we can say the ARIMA model was able to predict the seasonality and increasing trend in the predicted data but the trend is underestimate, indicating lower slope coefficient in predicted trend than the actual trend of realized atmospheric CO2 data.

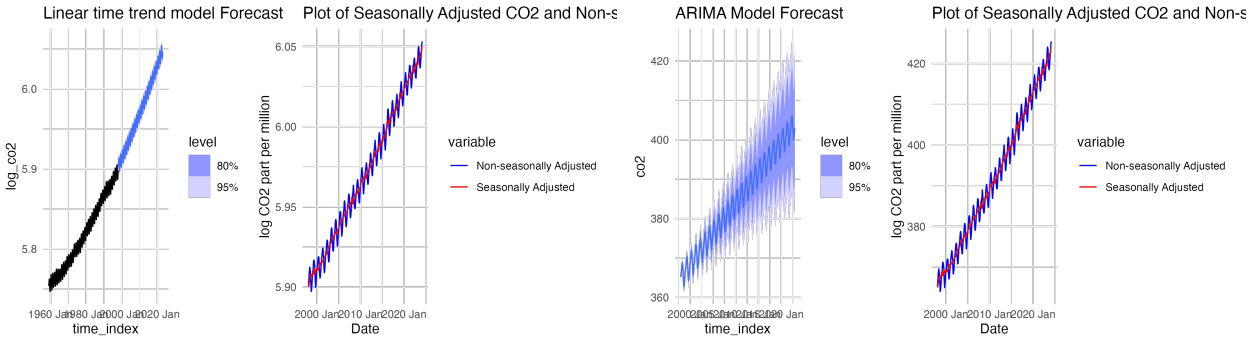


Figure 14: Linear and ARIMA Model Forecast and Plot of Seasonally Adjusted CO2 and Non-seasonally Adjusted CO2

Then we are able to evaluate the performance of 1997 linear and ARIMA models from earlier. In linear model, the first time that CO2 cross 420 ppm is 2022 Jan, while in actual data it is 2022 Apr. The ARIMA model did not predict CO2 could cross 420 ppm in 2022. Apparently, linear model is more close to the actual data. The time series plots are used to show the difference between actual minus predicted data. Consistent with above discussion, the linear model tends to over estimate the CO2 and the degree of overestimation tends to increase with the time evolve, while ARIMA model tends to underestimate the CO2 and the degree of underestimation tends to increase with the time evolve. However, ARIMA model has more smooth and consistent underestimation in residual, while linear model's prediction has bigger fluctuation in residual. From the histogram, we can tell that most residual in linear model lie between [-3,-1], while ARIMA model has a bigger range, [0.5,6]. Finally, we use accuracy function to test the gap between predicted and observed values. The results show that linear model has smaller gap (RMSE=2.14) than ARIMA model(RMSE=8.09), indicating better model fit. Need to choose which plot or output to place here from forecast evaluation

E. Training Models on Present Data

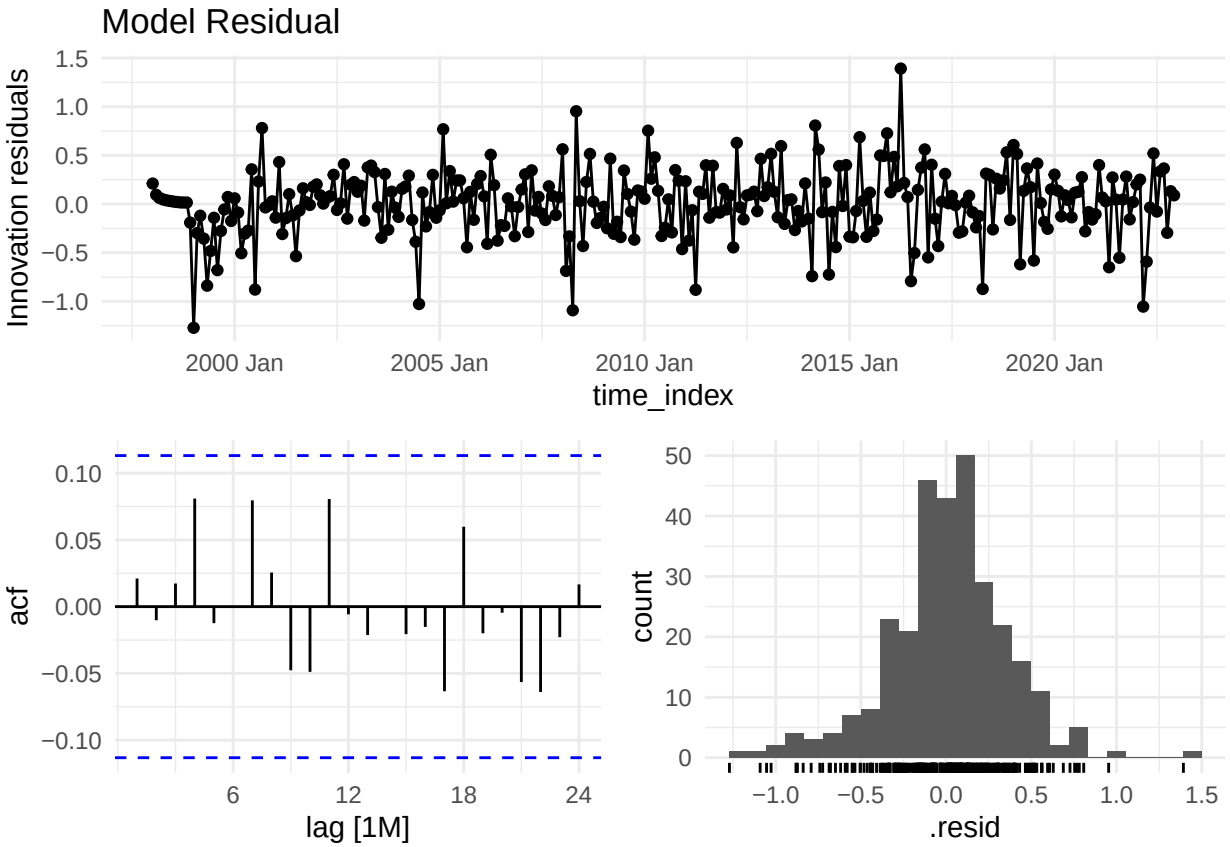
To train the models on present data, the weekly NOAA data was seasonally adjusted and split into both seasonally-adjusted (SA) and non-seasonally-adjusted (NSA) series into training and test sets, using the last two years of observations as the test sets. For both SA and NSA series, ARIMA models were fitted. Based on the EDA of first differencing data, we can observe that the mean of the first differenced CO_2 is fluctuated around zero. Thus, we set intercept to be equal 0 and parameter D to range from 1 to 2. The model with minimum BIC is `ARIMA(1,1,1)(2,1,1)[12]`. Although only `ma1` and `sma1` terms are sarcastically significant, based on the time plot and ACF, PACF plots in EDA section, we can observe strong and persistent non-seasonal and seasonal trend. The residual plots and KPSS test result suggest that the residuals of the model is stationary.

Choosing which plots/outputs to show from the ARIMA Model

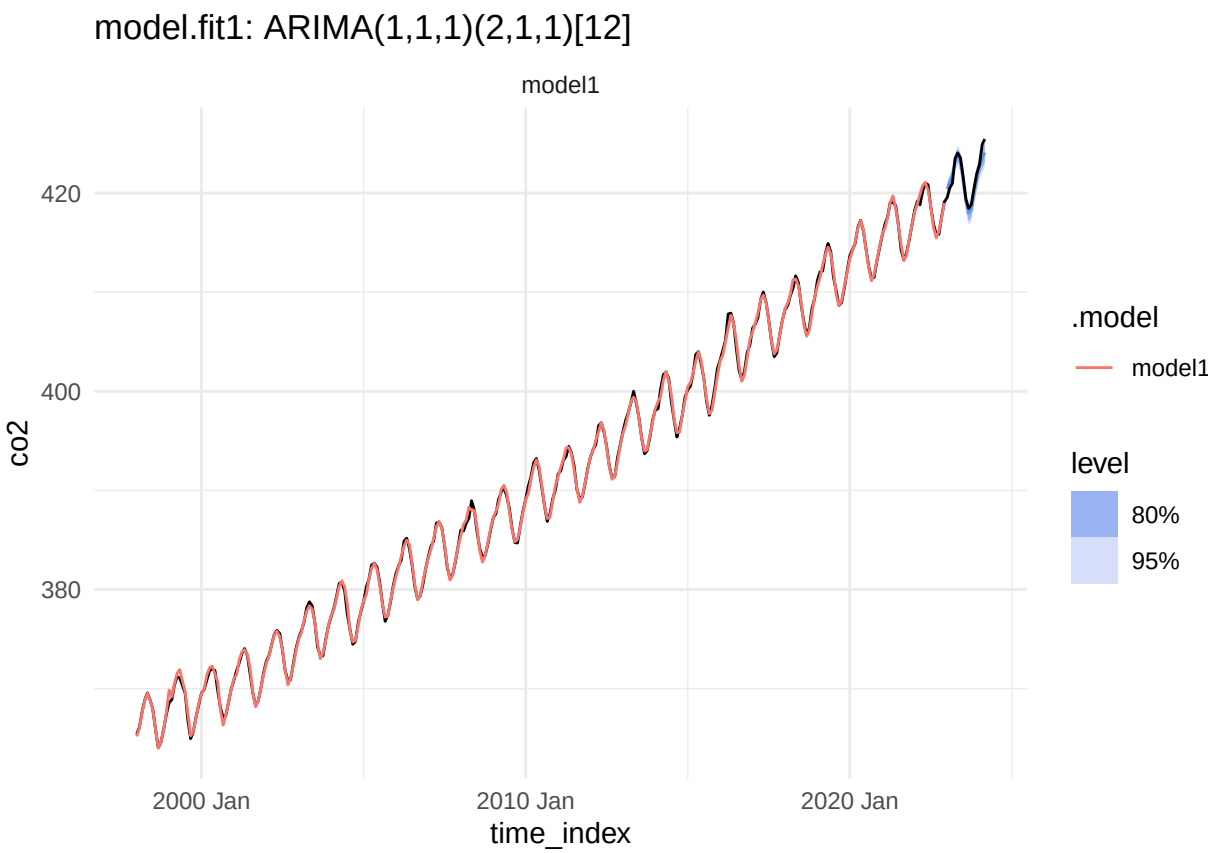
```
## Series: co2
```

```
## Model: ARIMA(1,1,1)(2,1,1)[12]
##
## Coefficients:
##          ar1      ma1    sar1    sar2     sma1
##       0.1676  -0.6189  0.0529  0.0062  -0.8614
## s.e.  0.1158   0.0939  0.0793  0.0757   0.0588
##
## sigma^2 estimated as 0.1302:  log likelihood=-112.93
## AIC=237.85   AICc=238.15   BIC=259.81
```

```
## # A tibble: 5 x 6
##   .model term  estimate std.error statistic  p.value
##   <chr>  <chr>    <dbl>     <dbl>     <dbl>    <dbl>
## 1 model1 ar1    0.168      0.116     1.45   1.49e- 1
## 2 model1 ma1   -0.619      0.0939   -6.59   2.06e-10
## 3 model1 sar1   0.0529     0.0793    0.668  5.05e- 1
## 4 model1 sar2   0.00616    0.0757    0.0814 9.35e- 1
## 5 model1 sma1  -0.861      0.0588  -14.6    1.21e-36
```




```
##
##  Box-Ljung test
##
## data:  .
## X-squared = 5.9277, df = 10, p-value = 0.8213
```

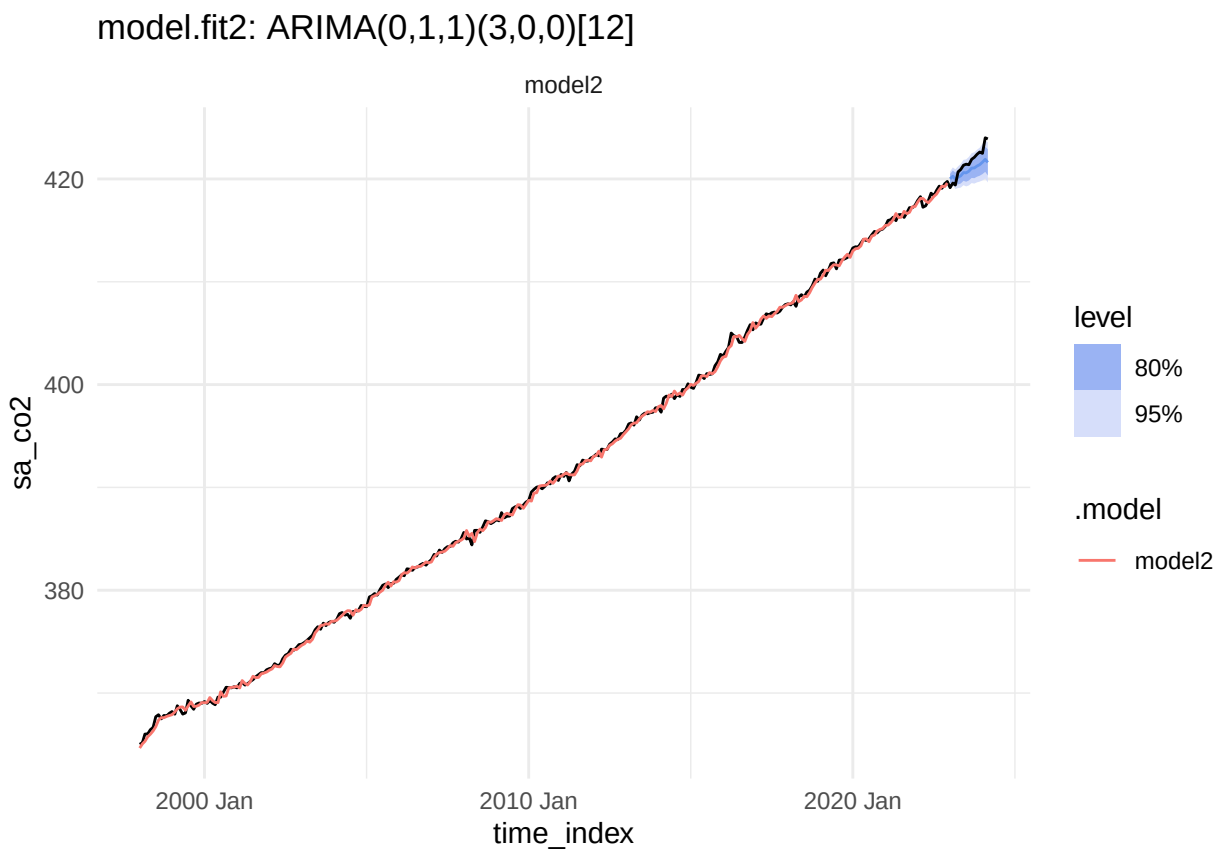



```
##                    ME      RMSE       MAE        MPE      MAPE
## Test set 0.2950163 0.7063206 0.5991691 0.06955459 0.1419224
```

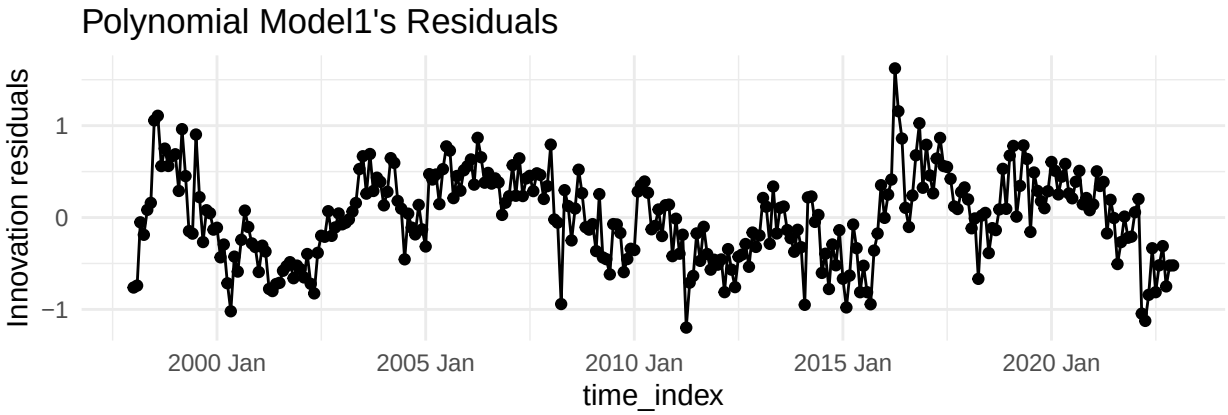
COMMENT: compare to models without PDQ termss

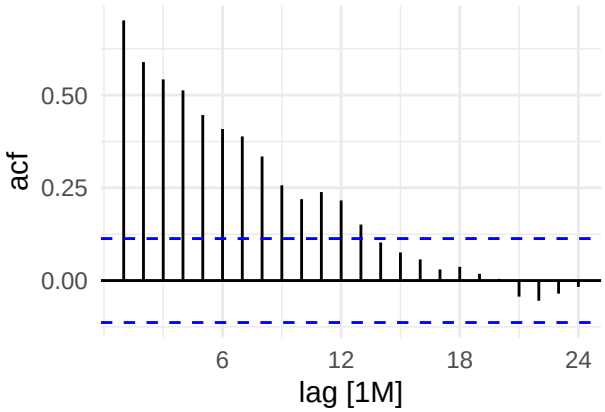
```
## Series: sa_co2
## Model: ARIMA(0,1,1)(3,0,0)[12]
##
## Coefficients:
##           ma1    sar1    sar2    sar3
##       -0.3794  0.2871  0.2240  0.2150
## s.e.   0.0639  0.0590  0.0621  0.0627
##
## sigma^2 estimated as 0.135:  log likelihood=-126.15
## AIC=262.29   AICc=262.5   BIC=280.8
```

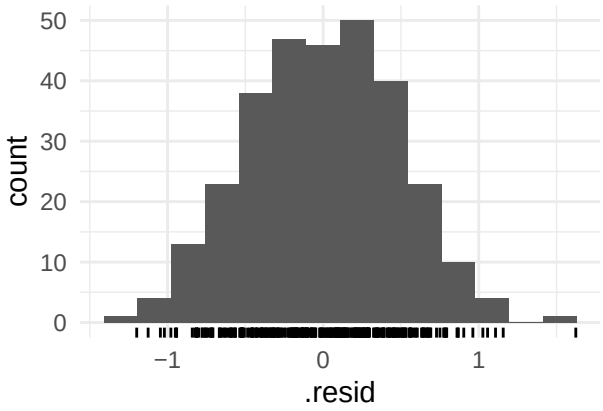
```
## # A tibble: 4 x 6
##   .model term  estimate std.error statistic       p.value
##   <chr>  <chr>    <dbl>     <dbl>     <dbl>         <dbl>
## 1 model2 ma1     -0.379    0.0639     -5.94 0.00000000788
## 2 model2 sar1     0.287    0.0590      4.86 0.00000187
## 3 model2 sar2     0.224    0.0621      3.61 0.000364
## 4 model2 sar3     0.215    0.0627      3.43 0.000684
```

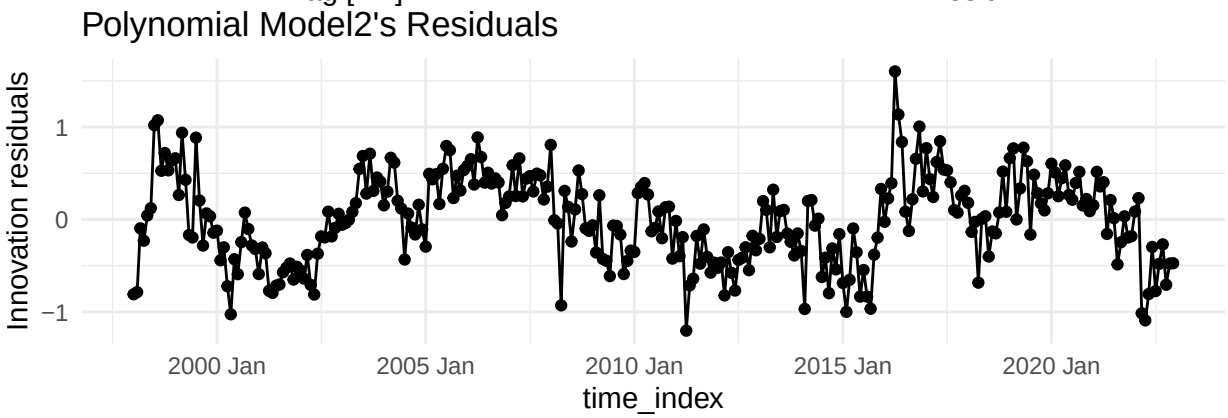


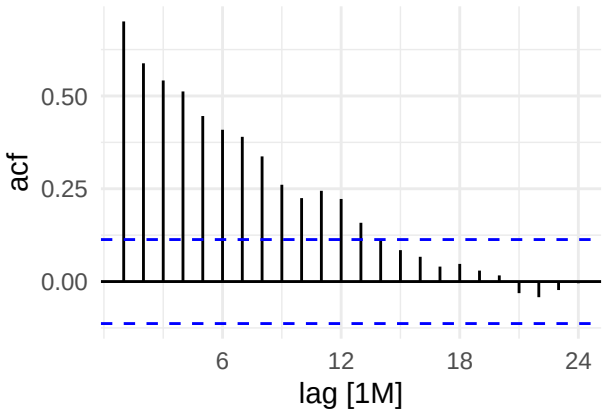

```
##                    ME     RMSE       MAE       MPE      MAPE
## Test set 0.7141089 1.117187 0.9947379 0.1687327 0.2356507
```

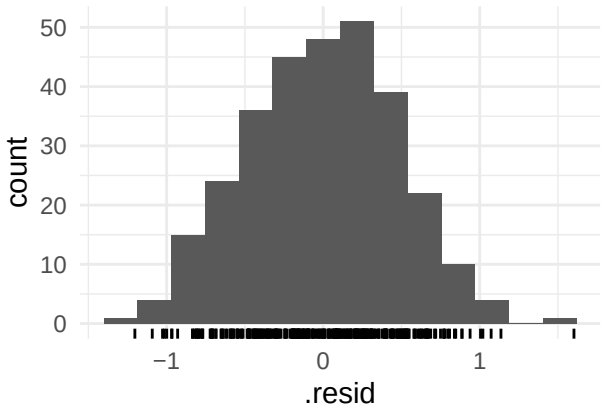










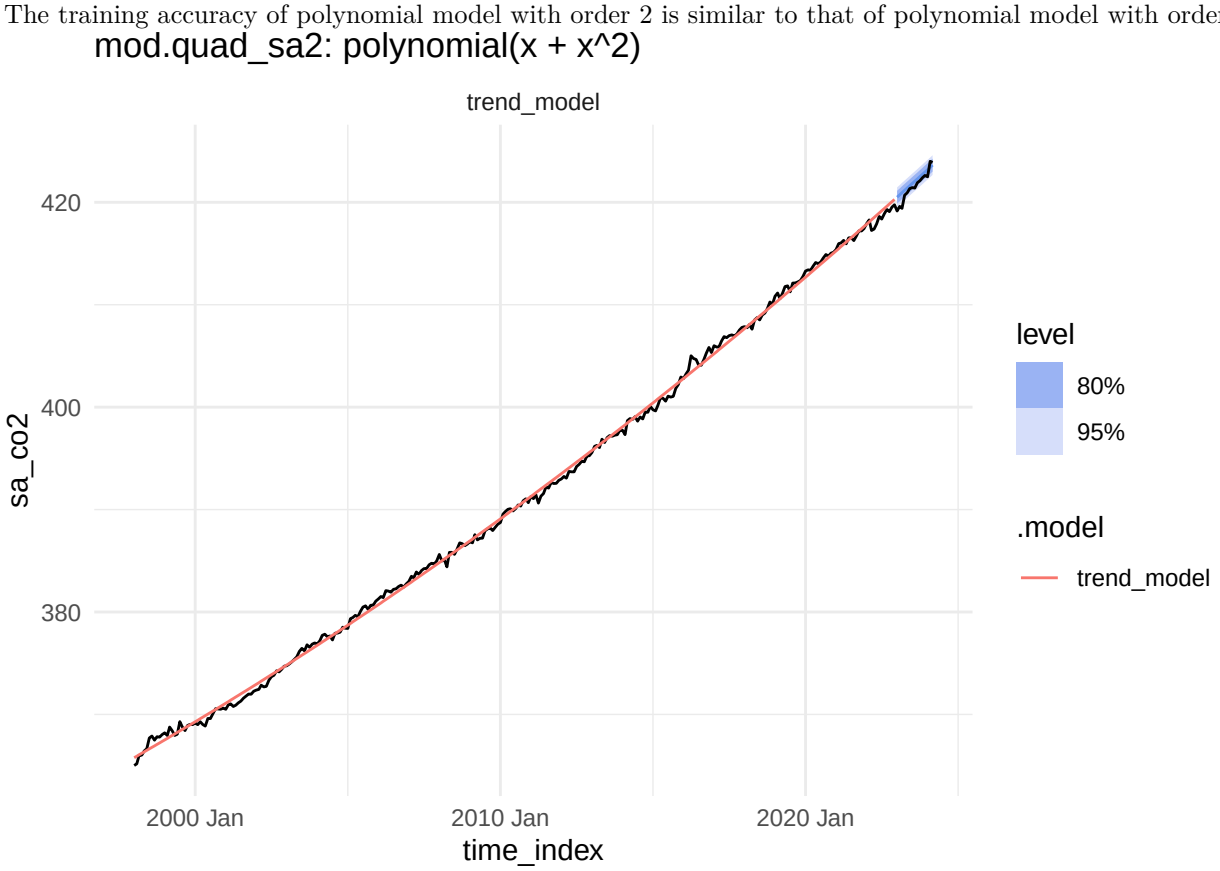



```
## # A tibble: 1 x 10
```

```
##   .model      .type          ME  RMSE   MAE       MPE  MAPE  MASE RMSSE  ACF1
##   <chr>       <chr>       <dbl> <dbl> <dbl>     <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 trend_model Training 8.39e-16 0.477 0.391 -0.000150 0.100 0.182 0.213 0.701
```

```
## # A tibble: 1 x 10
##   .model      .type          ME  RMSE   MAE       MPE  MAPE  MASE RMSSE  ACF1
##   <chr>       <chr>       <dbl> <dbl> <dbl>     <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 trend_model Training -7.78e-16 0.477 0.392 -0.000150 0.100 0.182 0.213 0.701
```

The training accuracy of polynomial model with order 2 is similar to that of polynomial model with order 3.




```
##                  ME      RMSE       MAE        MPE      MAPE
## Test set -0.5059376 0.7308907 0.6200114 -0.1203975 0.1473037
```

Polynomial model generates more accurate forecasting results for seasonally adjusted CO_2 trend compared to ARIMA.

```
##                        ME      RMSE       MAE         MPE      MAPE
## NSA ARIMA       0.2950163 0.7063206 0.5991691  0.06955459 0.1419224
## SA ARIMA        0.7141089 1.1171874 0.9947379  0.16873269 0.2356507
## SA Polynomial -0.5059376 0.7308907 0.6200114 -0.12039749 0.1473037
```

F. Atmospheric CO2 Predictions

Based on the non-seasonally adjusted data series, we used the non-seasonally adjusted ARIMA model to generate predictions for atmospheric CO2 levels until the year 2122.

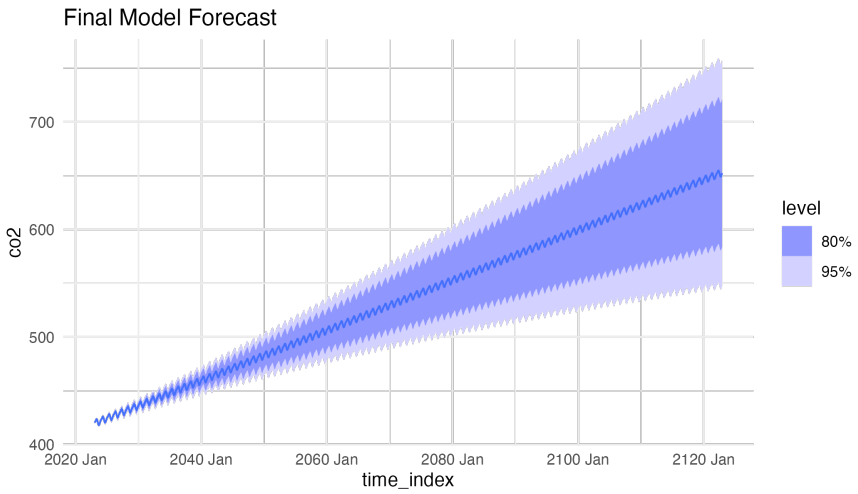


Figure 15: Final Model Forecast

Using a 95% confidence intervals we noted the first and last years that we would expect the atmospheric CO2 levels to be at 420 ppm and 500 ppm. With 95% confidence, we expect atmospheric CO2 levels to first exceed 420 ppm in February 2023, with the upper confidence interval remaining below this threshold until October 2023. For the 500 ppm mark, the lower confidence interval is projected to surpass this level in April 2074, while the upper confidence interval is expected to remain below 500 ppm until November 2049.

IV. Conclusion