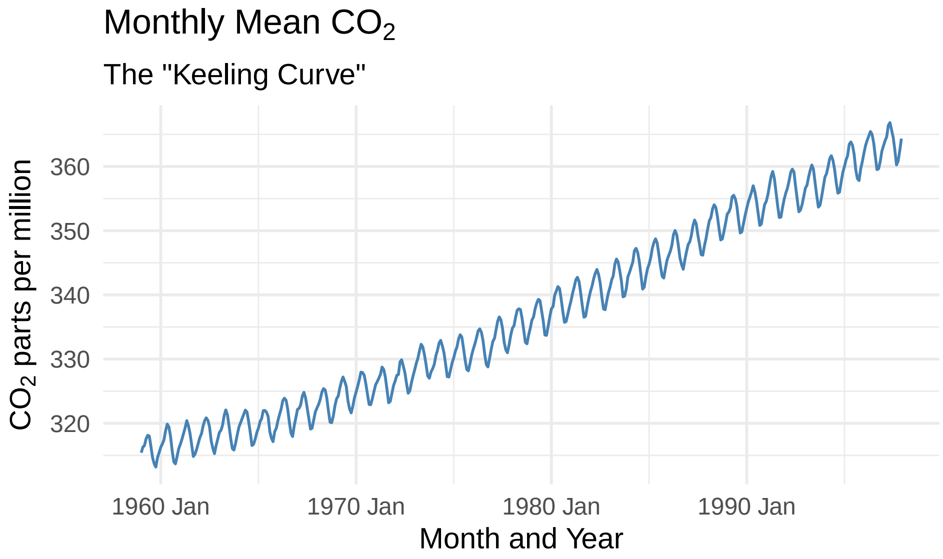
Forecast CO2 Emission Time Series (The "Keeling Curve")

Irene Na, William Lei, Cyrus Aghaee

Background

In the 1950s, the geochemist Charles David Keeling observed a seasonal pattern in the amount of carbon dioxide present in air samples collected over the course of several years. In 1958 Keeling began continuous monitoring of atmospheric carbon dioxide concentrations from the Mauna Loa Observatory in Hawaii and soon observed a trend increase carbon dioxide levels in addition to the seasonal cycle. He was able to attribute this trend increase to growth in global rates of fossil fuel combustion. This trend has continued to the present, and is known as the "Keeling Curve".

With that in mind, we decided to split our data into before (and include) 1997 Dec, and after 1997 Dec, conducts two parts of analysis. The first part serves as a training set where we experiment different modeling methods. The second part is used as a testing set where we evaluate the models with unseen data, and train new model incorporating the most recent data and evaluation observations.



Report from the Point of View of 1997

Part 0a: Introduction

The rising CO2 concentrations and their consequences, including global warming, ocean acidification, and rising sea levels are widely impacting human life around the globe. The repercussions of these impacts are extreme weather events, ecosystem degradation, agricultural disruption, and significant economic costs.

In light of the background, one research question comes to our mind is - *Is the growth of carbon dioxide (CO2) concentrations (partially or entirely) due to a deterministic trend, which can be associated with human* *activities such as fossil fuel combustion? Or is it mostly due to a stochastic process with a positive drift, thus harder to link to human activities?*

The question is critical, because it could indicate whether human intervention is going to be helpful or irrelevant in CO2 control. This analysis will be crucial in assessing future implications of rising CO2 levels and can guide preventative policies and initiatives.

More specifically, we will utilize the CO2 levels data collected at the Mauna Loa Observatory by Charles David Keeling from 1959. In this first part of analysis, we will focus on the data from 1959 Jan to 1997 Dec, which serves as the training data. We will explore different methods to fit models under different assumptions. Specifically, we will train a regression model under the assumption that CO2 growth is mostly due to deterministic trends. And we will look into ARIMA model, under the assumption that CO2 growth is a stochastic process. Then in the second part we will evaluate those models with unseen data post 1997.

Part 1a: CO2 data

First step, we examine the data structure and check missing values.

```
## tbl_ts [468 x 2] (S3: tbl_ts/tbl_df/tbl/data.frame)
##  $ index: mth [1:468] 1959 Jan, 1959 Feb, 1959 Mar, 1959 Apr, 1959 May, 1959 Jun...
##  $ value: num [1:468] 315 316 316 318 318 ...
##  - attr(*, "key")= tibble [1 x 1] (S3: tbl_df/tbl/data.frame)
##   ..$ .rows: list<int> [1:1]
##   .. ..$ : int [1:468] 1 2 3 4 5 6 7 8 9 10 ...
##   .. ..@ ptype: int(0)
##  - attr(*, "index")= chr "index"
##   ..- attr(*, "ordered")= logi TRUE
##  - attr(*, "index2")= chr "index"
##  - attr(*, "interval")= interval [1:1] 1M
##   ..@ .regular: logi TRUE
```

```
##       na_sum
## index      0
## value      0
```

```
## Head of data:
```

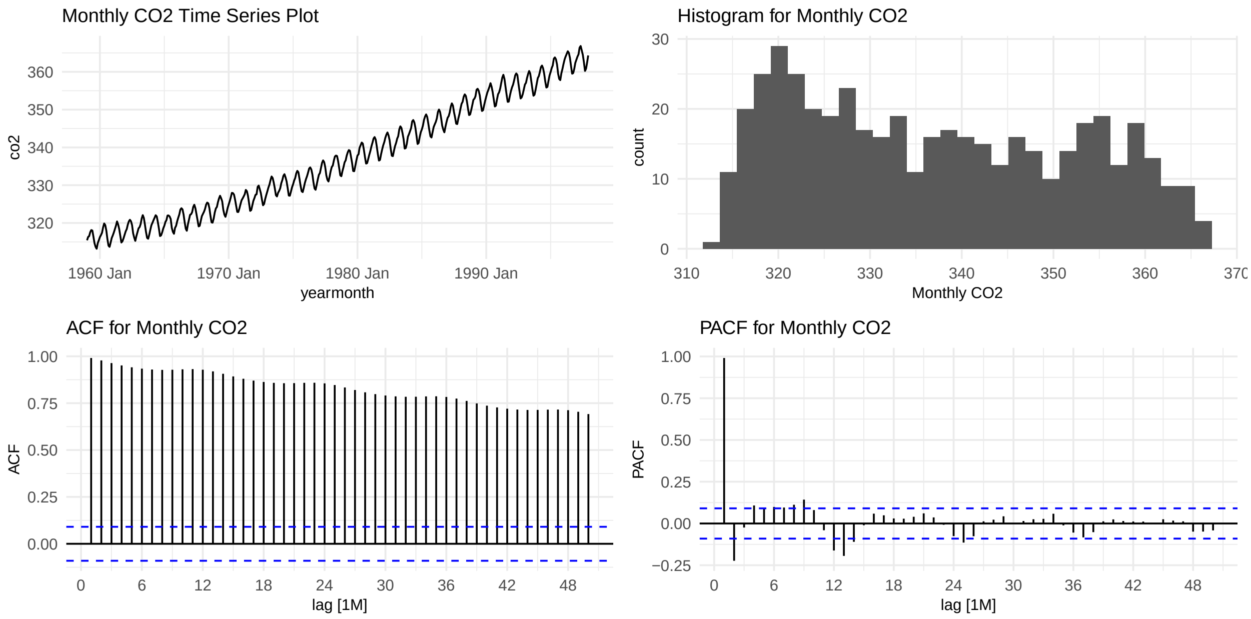
```
## # A tsibble: 6 x 2 [1M]
##      index value
##      <mth> <dbl>
## 1 1959 Jan  315.
## 2 1959 Feb  316.
## 3 1959 Mar  316.
## 4 1959 Apr  318.
## 5 1959 May  318.
## 6 1959 Jun  318
```

```
## Tail of data:
```

```
## # A tsibble: 6 x 2 [1M]
##      index value
##      <mth> <dbl>
## 1 1997 Jul  365.
## 2 1997 Aug  363.
## 3 1997 Sep  360.
## 4 1997 Oct  361.
## 5 1997 Nov  362.
## 6 1997 Dec  364.
```

Comments: we have 468x2 original data, and no N/A in data. The columns are index (year-month), and value of CO2 concentration, which has the unit of parts per million of CO2 (or ppm). For example 413 ppm of CO2 means in every million molecules of (dry) air there are on average 413 CO2 molecules. In this first part, we focus on the data from 1959 Jan to 1997 Dec.

Given this is a single time series data, we take a look at the time series, distribution, ACF and PACF charts for CO2 level as is. We are looking for patterns in data, and assessing if there is any feature engineering needed for further analysis.



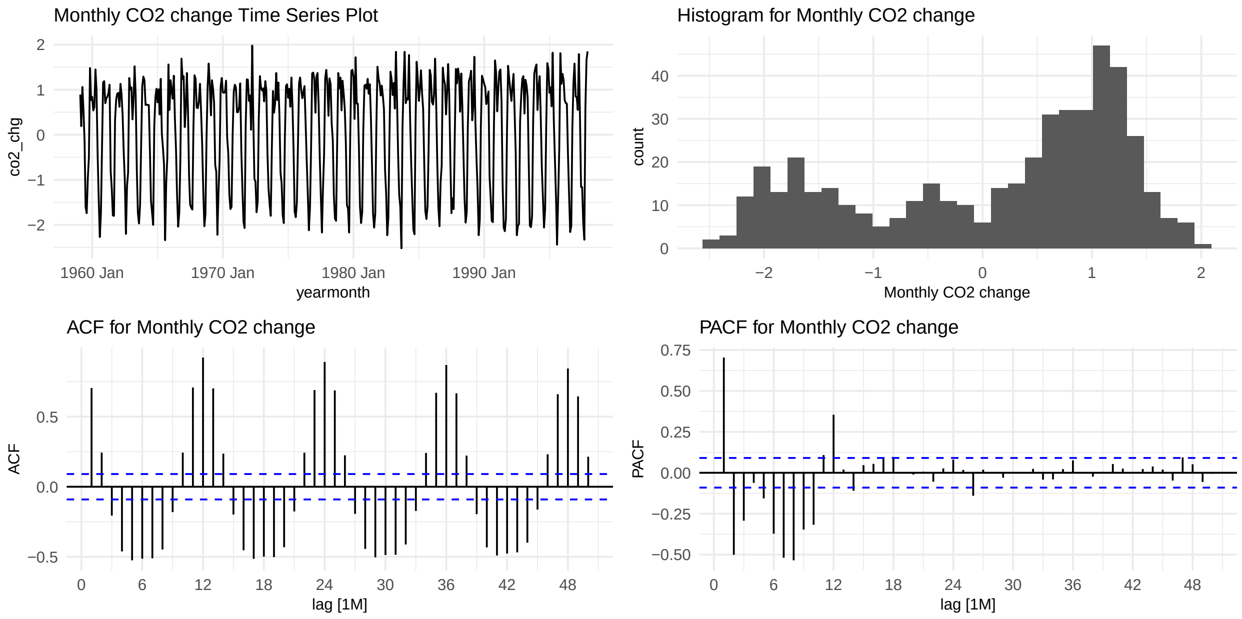
```
## Average annual growth rate is:

## [1] 15.15526
```

Comments: from the CO2 level EDA charts we can observe that:

1. The monthly CO2 concentration appears to have a deterministic trend, going up steadily with average annual growth of 15.15 ppm a year.
2. The monthly CO2 concentration presents a strong pattern of seasonality across months.
3. Persistent ACF with gradual decay and small fluctuation suggests the potential presence of AR process with seasonality. The slow decay suggests that the series is most likely not stationary.
4. Quickly dropping PACF with periodic significance suggests the potential existence of AR and MA processes.
5. The distribution chart is not too meaningful to read given the strong trends and seasonality.

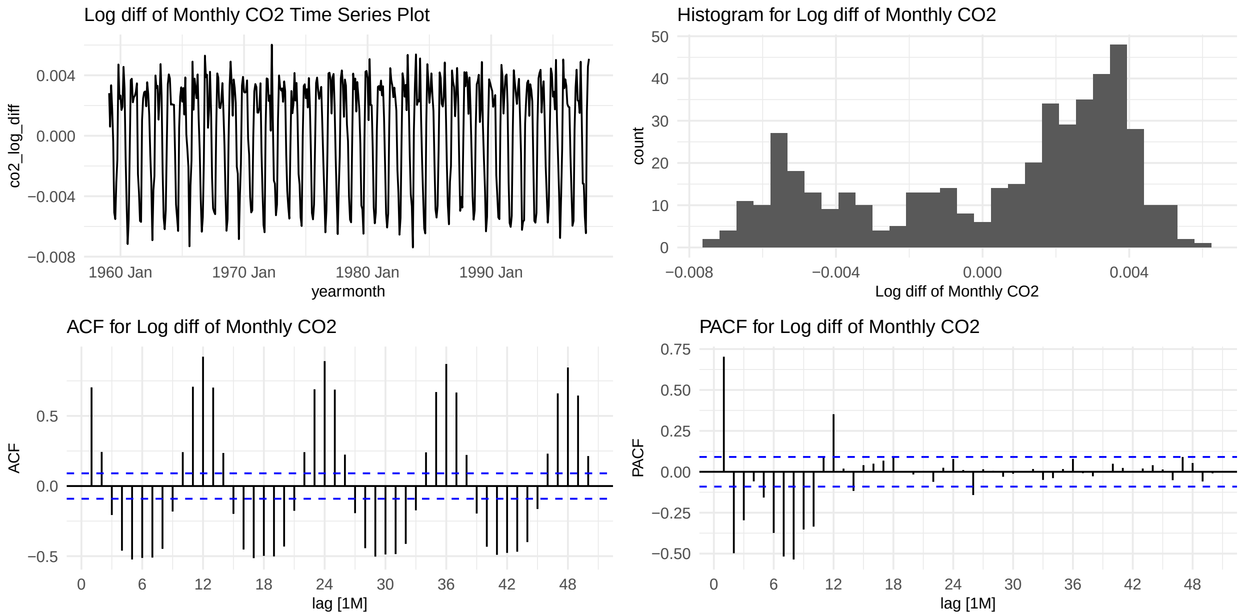
Given the level is not stationary, we would like to take a look at the differences of CO2 with the same charts:



Comments: from the monthly CO2 change EDA charts we can observe that:

1. The CO2 monthly change doesn't present trends anymore but still has strong seasonality. Further, the variance increases over time.
2. The patterned swing of ACF without decay provides more evidence of seasonality.
3. The persistent PACF without clear drop off suggests potential presence of MA process too.
4. The distribution chart is not too meaningful to read given the strong presence of seasonality.

In light of the increase of variance in CO2 change, we would like to take the log transformation of CO2 before taking the change in order to stabilize the variance.



Comments: from charts we can see that after the log transformation, the difference of log(CO2) has more stable variance over time. Similar to the difference of CO2, it no longer has visible trends, but possesses strong seasonality, which is visible from the swinging ACF.

Overall, from the comprehensive EDA on CO2 level, change and log CO2 change, we noted that in order to stabilize variance over time, log transformation on CO2 is recommended. Further, CO2 levels are not stationary. Besides trends and seasonality, ACF and PACF suggest potential existence of stochastic processes such as AR and MA. Note that log transformation will not change these observations on CO2 level.

Part 2a: Linear time trend model

Based on the EDA previously, it is believable that there exist deterministic trends and seasonality in CO2 level. Under the assumption that the trends and seasonality in CO2 are deterministic, we fit various regression models to CO2 (or log(CO2)), with time trend and/or seasonality as regressors. More specifically, we start with the simplest model, where we fit CO2 to a linear time trend. Then we gradually add model complexity by switching to fit log(CO2), and adding quadratic time trends, seasonality step by step. Then we examine the models with visual evaluations and in-sample comparison by RMSE.

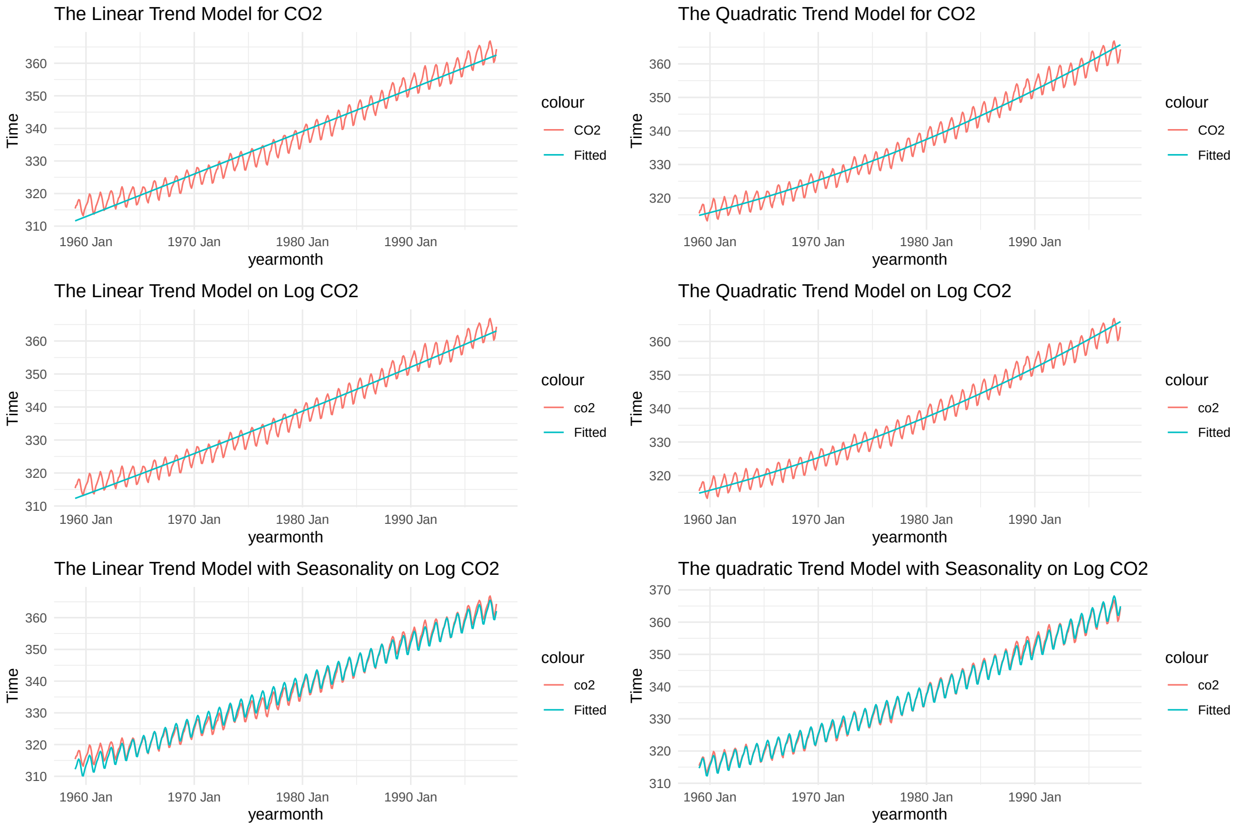
But before everything we would like to check the stationarity of the input - Log(CO2).

```
##
##  Augmented Dickey-Fuller Test
##
## data:  log(co2_ts_mthly_pri$co2)
## Dickey-Fuller = -3.2669, Lag order = 7, p-value = 0.0765
## alternative hypothesis: stationary
```

```
##
##  Phillips-Perron Unit Root Test
##
## data:  log(co2_ts_mthly_pri$co2)
## Dickey-Fuller Z(alpha) = -101.8, Truncation lag parameter = 5, p-value
## = 0.01
## alternative hypothesis: stationary
```

```
##
##  KPSS Test for Trend Stationarity
##
## data:  log(co2_ts_mthly_pri$co2)
## KPSS Trend = 0.86463, Truncation lag parameter = 5, p-value = 0.01
```

Comments: we note that two out of the three unit roots tests rejected the null of existence of unit root, leaving it somewhat inconclusive in terms of log(CO2)' stationarity. But it is encouraging to see that the p-value is overall less than 10%. Furthermore, if the assumption (under question) that the CO2 has deterministic trends is true, regressing log(CO2) on trends should result in a stationary process. Therefore, we continue with the regression analysis.



```
## # A tibble: 6 x 4
##   .model                                             .type     RMSE   MAE
##   <chr>                                              <chr>    <dbl> <dbl>
## 1 TSLM(co2 ~ trend())                                Training 2.61  2.15
## 2 TSLM(co2 ~ trend() + I(trend()^2))                 Training 2.18  1.84
## 3 TSLM(log(co2) ~ trend())                           Training 2.47  2.04
## 4 TSLM(log(co2) ~ trend() + I(trend()^2))            Training 2.19  1.85
## 5 TSLM(log(co2) ~ trend() + season())                Training 1.36  1.16
## 6 TSLM(log(co2) ~ trend() + I(trend()^2) + season()) Training 0.739 0.614
```

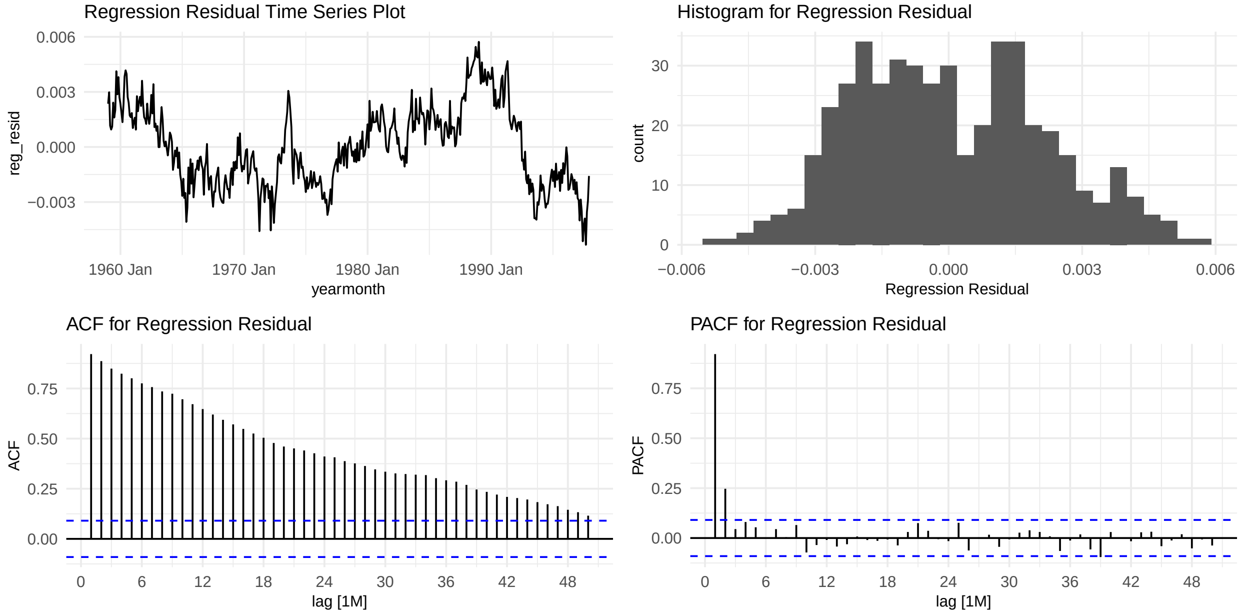
Comments: Judging by visual examination, the quadratic trend model with the seasonality modeled on log CO2 fits the in-sample data the best, which has the needed curvature and seasonality pattern. Further, judging by in-sample model evaluation, this model (the last model) has the lowest RMSE and MAE.

Therefore the chosen regression model with quadratic polynomial trends and seasonality has below model summary and mathematical form:

```
## Series: co2
## Model: TSLM
## Transformation: log(co2)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0053270 -0.0017362 -0.0001774  0.0015139  0.0057292
##
## Coefficients:
##                  Estimate Std. Error   t value Pr(>|t|)
## (Intercept)     5.751e+00  4.533e-04 12687.967  < 2e-16 ***
## trend()         2.223e-04  3.012e-06    73.817  < 2e-16 ***
## I(trend()^2)    2.150e-07  6.219e-09    34.566  < 2e-16 ***
## season()year2   1.969e-03  4.974e-04     3.959 8.73e-05 ***
## season()year3   4.163e-03  4.974e-04     8.371 7.16e-16 ***
## season()year4   7.498e-03  4.974e-04    15.075  < 2e-16 ***
## season()year5   8.911e-03  4.974e-04    17.916  < 2e-16 ***
## season()year6   6.965e-03  4.974e-04    14.004  < 2e-16 ***
## season()year7   2.480e-03  4.974e-04     4.986 8.78e-07 ***
## season()year8  -3.662e-03  4.974e-04    -7.362 8.61e-13 ***
## season()year9  -9.098e-03  4.974e-04   -18.290  < 2e-16 ***
## season()year10 -9.661e-03  4.974e-04   -19.423  < 2e-16 ***
## season()year11 -6.113e-03  4.974e-04   -12.290  < 2e-16 ***
## season()year12 -2.799e-03  4.974e-04    -5.627 3.21e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002196 on 454 degrees of freedom
## Multiple R-squared: 0.9976,  Adjusted R-squared: 0.9975
## F-statistic: 1.453e+04 on 13 and 454 DF, p-value: < 2.22e-16
```

$$
\begin{aligned}
log(CO2) = {} & 5.75 + 0.0002223t + 0.000000215t^2 + 0.001969m_2 \\
& + 0.004163m_3 + 0.007498m_4 + 0.008911m_5 + 0.006965m_6 \\
& + 0.00248m_7 - 0.003662m_8 - 0.009098m_9 - 0.009661m_{10} \\
& - 0.006113m_{11} - 0.002799m_{12}
\end{aligned}
$$

Next we examine the residuals of this model, and see if the residuals are close to white noise. If yes, we can believe that CO2 only has deterministic trends and seasonality components.



Examine the stationarity of the regression residuals:

```
##
##  Augmented Dickey-Fuller Test
```

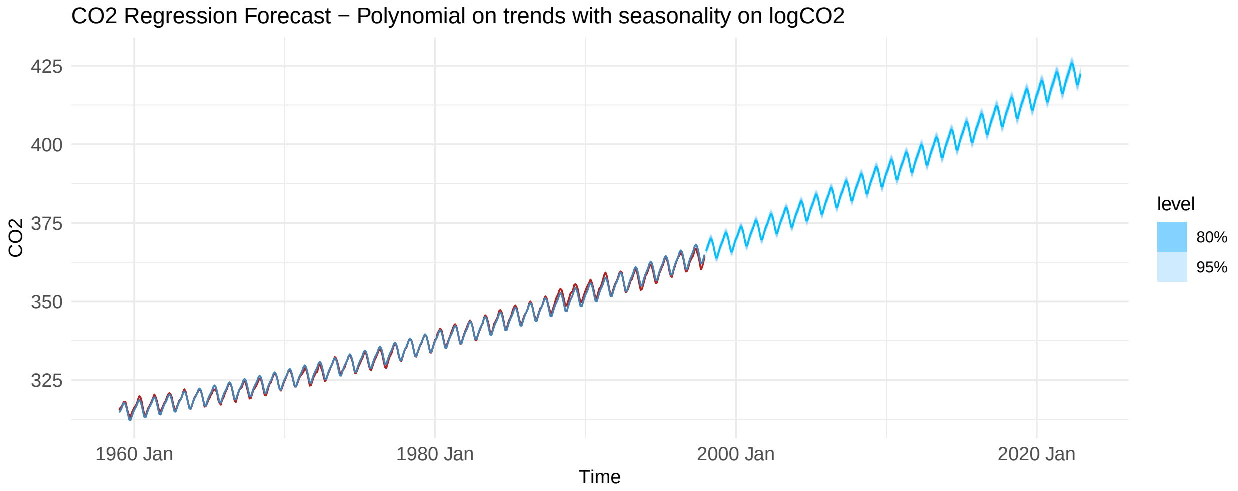
```
##
## data:  co2_ts_mthly_pri$reg_resid
## Dickey-Fuller = -2.1724, Lag order = 7, p-value = 0.505
## alternative hypothesis: stationary
```

```
##
##  Phillips-Perron Unit Root Test
##
## data:  co2_ts_mthly_pri$reg_resid
## Dickey-Fuller Z(alpha) = -24.253, Truncation lag parameter = 5, p-value
## = 0.02878
## alternative hypothesis: stationary
```

```
##
##  KPSS Test for Trend Stationarity
##
## data:  co2_ts_mthly_pri$reg_resid
## KPSS Trend = 0.63434, Truncation lag parameter = 5, p-value = 0.01
```

Comments: the residuals from the best regression model still present very persistent autocorrelation (ACF chart), which is decaying gradually. Combined with the abrupt dropping of PACF, it is very likely that there is some stationary AR process in the residuals, with or without some MA process. The residuals are most likely stationary by visual examination and unit root tests (with mixed results), but they are clearly not white noises. These results suggest that CO2 level does not only have deterministic trend and seasonality, but also stochastic components which are not captured by the regression technique.

Before we move to the ARIMA model, we forecasted CO2 using this regression model to 2022. Note that this forecast primarily focuses on the deterministic component of the CO2.



Comments: as expected, the best fit regression model provides a good in sample fit. It projects that the quadratic trends and seasonality will continue into the future in the forecast. The confidence intervals expand over time but are overall very narrow.

Part 3a: ARIMA times series model

To answer the other half of our research question - if CO2 level is mostly driven by a stochastic process, we then fit the training data into an ARIMA model, which focuses on exploring the stochastic process in the data.

First, we take notes from EDA and regression and adapt the log transformation of CO2 for modeling, to fulfill the constant variance model assumption. *Second*, we note that 1st order differencing is needed to remove the unit root from log(CO2), therefore d=1 is needed in the ARIMA modeling. *Third*, a seasonality component is needed to remove the seasonal patterns. The repeated pattern in ACF in log(CO2) differences suggest that a differencing, and / or a sMA process may be appropriate.

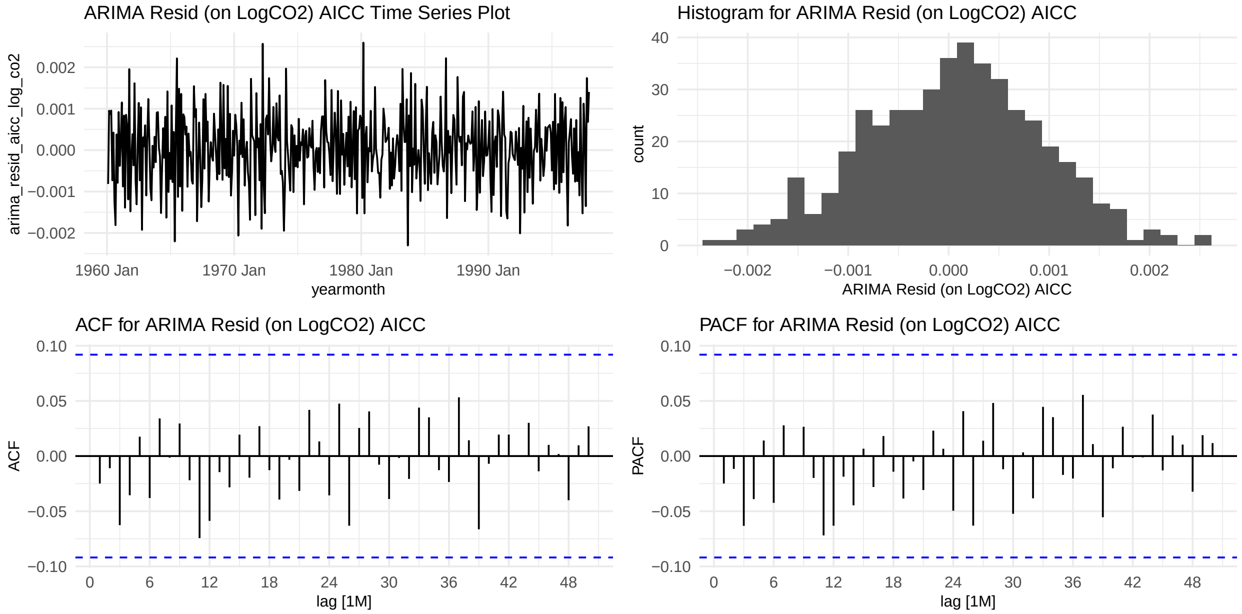
In order to find the most appropriate fit of ARIMA model, we went through a few iterations of modeling, checking residual diagnostics to see if the residuals resemble white noises, and readjust the model based on the observations. After a few iterations, we were able to find an SARIMA model with both seasonal and non-seasonal ARIMA components, which produces not only satisfactory model fit, white noises like residuals, but also is rather simple in parameters.

Specifically, the ARIMA model we found is ARIMA(0,1,0)(0,1,1)[12]. It means we take 1st order differencing in both seasonal and non-seasonal components, and include an additional SMA(1) to further account for the seasonality pattern. The model can be expressed in following mathematical form:

$$(1 - B^s)(1 - B)Log(CO2) = (1 + \phi B^s)w_t$$

We found the model specification overall consistent with our prior expectation, given it has first order differencing in non-seasonal components to remove the unit-root in Log(CO2), and includes both the first differencing and MA(1) in seasonal components to model the repeating patterns in seasonality.

```
## Series: co2
## Model: ARIMA(0,1,0)(0,1,1)[12]
## Transformation: log(co2)
##
## Coefficients:
##          sma1
##       -0.9415
## s.e.   0.0280
##
## sigma^2 estimated as 1.717e-06:  log likelihood=2544.61
## AIC=-5085.21   AICc=-5085.18   BIC=-5076.97
```



Ljung-box check on ARIMA model results

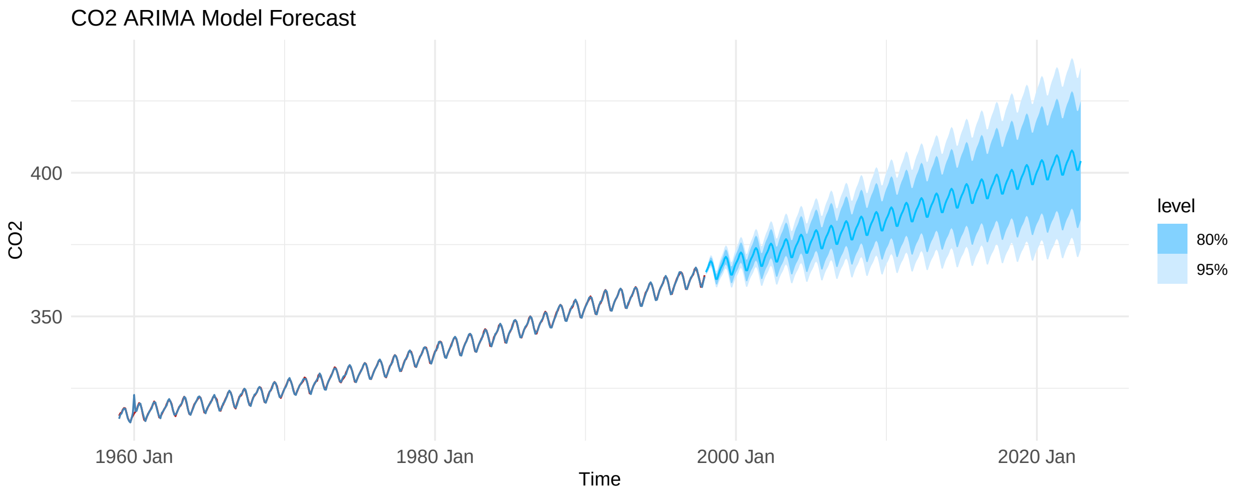
```
##
##  Box-Ljung test
##
## data:  co2_ts_mthly_pri$arima_resid_aicc_log_co2
## X-squared = 4.8413, df = 9, p-value = 0.8479
```

We ran the diagnostic charts on the ARIMA model residuals, and noted that ARIMA residuals hover around zero with relatively consistent variance over time. The distribution is overall symmetric optically.

The ACF and PACF are all well-within the 5% significance threshold, indicating the absence of obvious serial correlations. The Ljung-box test on the residuals (with lag 10 and degree of freedom 9) also fails to reject the Null that the residuals do not have serial correlation (p-value = 0.8479), supporting the visual diagnostic observations. All of the above suggest that the residuals resemble white-noise quite well.

Note that given the first order difference on both seasonal and non-seasonal components used up the first 13 observations to calculate the first model input, the model fit for those observations tend to be erratic. We therefore remove the first 13 residuals from visualization to avoid introducing confusion, as suggested by 'Time Series Analysis Forecasting and Control by Greta M. Ljung P321-324'.

Therefore, we believe the chosen ARIMA model produces satisfactory in-sample results. Next, We produced the CO2 level forecast and its confidence interval using the results of ARIMA from 1998 Jan to 2022 Dec for later model comparison in the second part.



Comments: Based on visual examination ARIMA model fits the in sample data reasonably well, and it forecasts the continuous growth and seasonality into the future. However, the growth trajectory in forecast is possessed with great uncertainty, evidenced by the wide and growing confidence interval around the mean, as forecast period increases. This is understandable given ARIMA assumes a stochastic process in essence.

Further, ARIMA model forecasts a CO2 level of 404.08ppm by Dec 2022, much lower than that's from the regression model (422.45ppm).

Notice that the two models - regression and ARIMA are generated with very different assumptions in mind. Which model performs better in-sample, the regression or the ARIMA? In order to answer this, we put the accuracy of the two models side by side and compare the in-sample results.

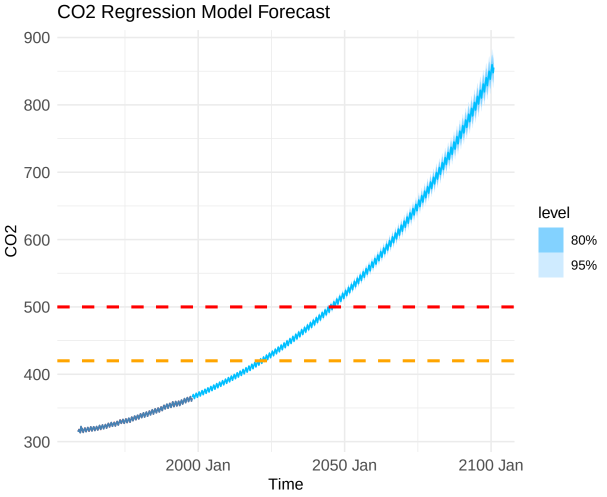
```
## # A tibble: 2 x 10
##   .model           .type      ME  RMSE   MAE      MPE   MAPE  MASE RMSSE    ACF1
## * <chr>            <chr>   <dbl> <dbl> <dbl>    <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1 Regression_model Trai~ 7.28e-4 0.739 0.614 -2.34e-4 0.181  0.485 0.525  0.924
## 2 ARIMA_model      Trai~ 5.71e-3 0.422 0.253  1.45e-3 0.0754 0.200 0.300 -0.0375
```

Comments: note that the ARIMA model produces lower RMSE and MAE model in-sample, suggesting a better in-sample fit.

Part 4a: Forecast atmospheric CO2 growth

Now we would like to check when the models predict the CO2 level reaches 420ppm and 500ppm level for the first and final times, and what the predicted CO2 level for year 2100 is.

We first examine the regression model, followed by the same examination of the ARIMA model.



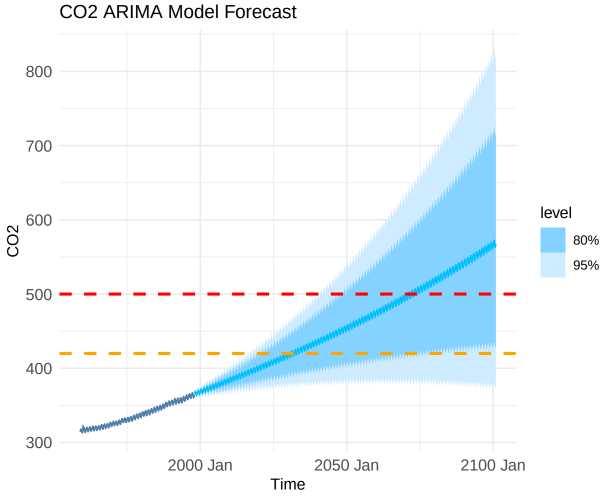


Table 1: CO2 Forecasts with ARIMA and Regression Models - 1997 Perspective

	Regression	ARIMA
First_time_reach_420	2020 May	2030 Apr
Last_time_stay_in_420_range	2022 Nov	2031 Jul
First_time_reach_500	2045 Mar	2070 May
Last_time_stay_in_500_range	2046 Oct	2074 Oct
First_time_reach_420_95%_IC	2020 Mar	2016 Apr
Last_time_stay_in_420_range_95%_IC	2023 Oct	> 2100 Dec
First_time_reach_500_95%_IC	2044 Mar	2040 Apr
Last_time_stay_in_500_range_95%_IC	2047 Oct	> 2100 Dec
2100_CO2_Forecast (ppm)	855.48	567.5

Comments: The point forecast as well as the 95% confidence interval forecast to reach critical CO2 levels (420, 500) for both models are presented in the above table. We also forecasted the CO2 level by 2100 Dec using both models. We can clearly observe that the regression model forecasts a much faster growth of CO2 in point estimate, and forecasts it with a much narrower band. On the other hand, ARIMA model models a slower point estimate growth and a much wider confidence band, reflecting its stochastic process nature assumed in the model.

Report from the Point of View of the Present

Part 0b: Introduction

Building on our original question from the 1997 analysis—*Is the growth of carbon dioxide (CO2) concentrations (partially or entirely) due to a deterministic trend, which can be associated with human activities such as fossil fuel combustion? Or is it mostly due to a stochastic process with a positive drift, thus harder to link to human activities?* —we now extend our analysis to the post-1997 period. This extension is essential to understand whether the observed trends in CO2 levels are consistent with our initial assumptions or if new patterns have emerged, suggesting changes in the underlying drivers of atmospheric CO2 growth.

In this updated analysis, we aim to address an even more pressing aspect of the question: *Are our models accurately forecasting the trend, and how severe could the increase in CO2 levels become?* Understanding whether the increase in CO2 is accelerating, decelerating, or continuing at a steady pace will help us project the potential future impact of rising CO2 levels on climate and guide the urgency of policy and intervention strategies.

We begin with an exploratory data analysis (EDA) on post-1997 data to examine how the Keeling Curve has evolved, comparing recent trends with those from pre-1997 data to assess if CO2 growth remains deterministic or shows signs of accelerating stochastically. Next, we revisit our 1997 models, comparing actual CO2 levels to our forecasts from a regression time model and ARIMA model. This comparison helps us evaluate model accuracy and detect any shifts in the data pattern, revealing whether the CO2 trend has worsened or remained stable. We'll also check how close our past prediction for CO2 crossing 420 ppm aligns with reality, providing insight into model reliability and current CO2 trajectory towards future thresholds. Finally, we'll project when CO2 levels are expected to reach critical thresholds of 420 ppm, 500 ppm, and beyond, extending forecasts to 2122. This analysis will help us understand the potential severity of CO2 increases under current or worsening trends.

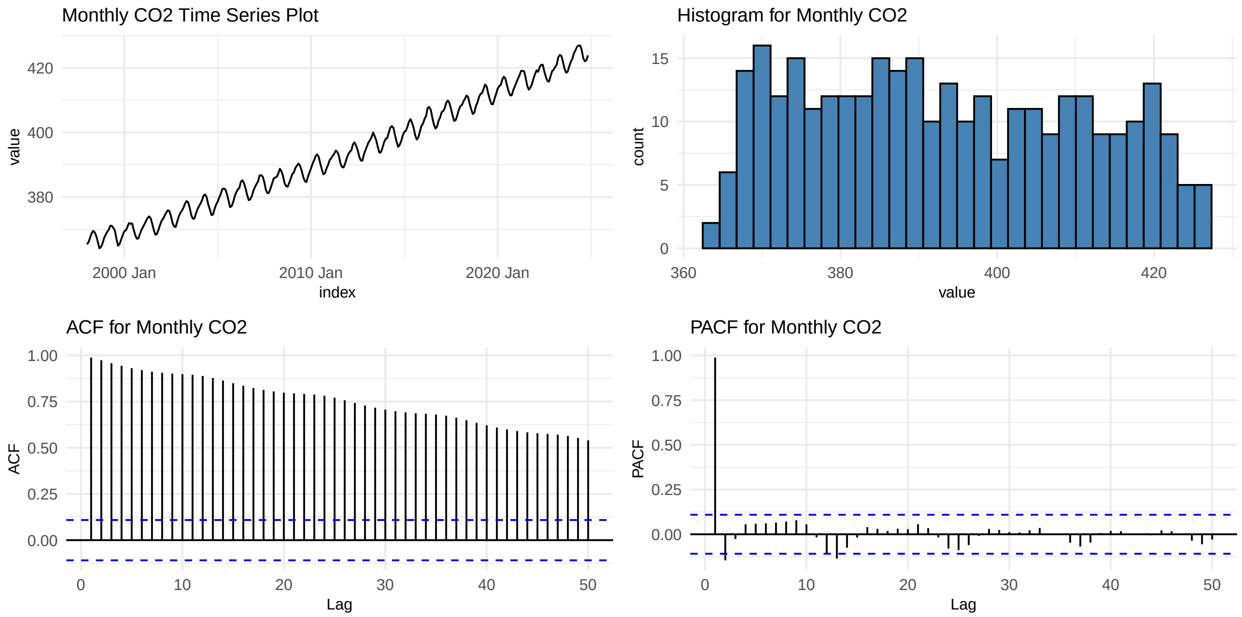
Ultimately, this analysis seeks to refine our understanding of whether the continued growth in CO2 concentrations is primarily driven by human activities (i.e an deterministic trend) or if it represents an intrinsic, accelerating trend that may be even harder to control. Insights from this updated analysis will be critical for guiding policies aimed at controlling atmospheric CO2 levels, understanding the potential severity of climate impacts, and assessing the urgency of actions required to mitigate climate change.

Part 1b: Create a modern data pipeline for Mona Loa CO2 data.

```
## # A tsibble: 323 x 2 [1M]
##       index value
##       <mth> <dbl>
##  1 1998 Jan  365.
##  2 1998 Feb  366.
##  3 1998 Mar  367.
##  4 1998 Apr  369.
##  5 1998 May  370.
##  6 1998 Jun  369.
##  7 1998 Jul  368.
##  8 1998 Aug  366.
##  9 1998 Sep  364.
## 10 1998 Oct  365.
## # i 313 more rows
```

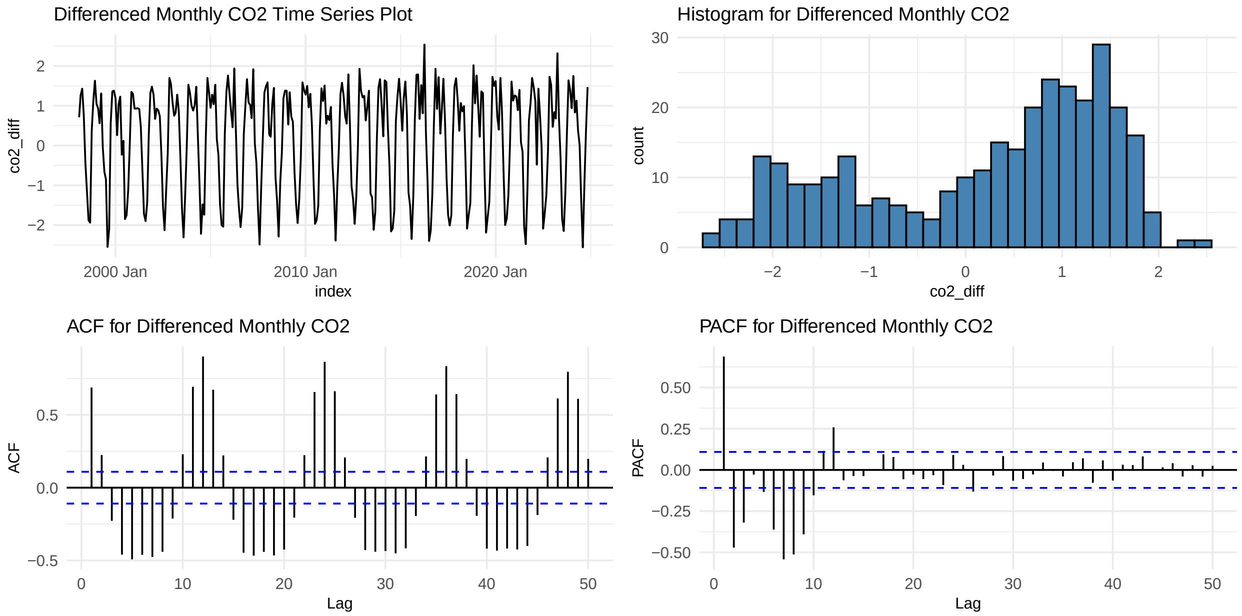
```
## [1] FALSE
```

Comments: There are no NA in the tsibble.



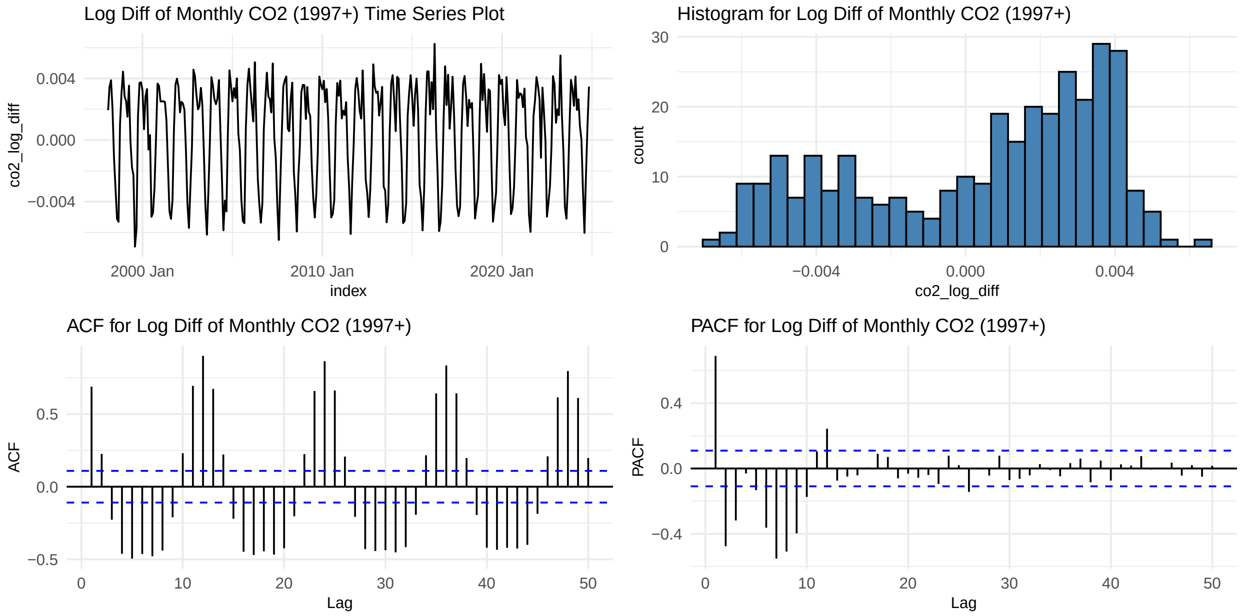
Comments: From the EDA charts for the present CO2 data, we observed that:

1. The Monthly CO2 Time Series Plot shows a clear upward trend from 1997 to present, with a consistent seasonal pattern indicated by the regular peaks and troughs.
2. The ACF for Monthly CO2 shows persistent autocorrelation, gradually decaying over lags, confirming non-stationarity and potential seasonality, likely an AR process.
3. The PACF for Monthly CO2 sharply drops after lag 1, suggesting a possible AR(1) component with some seasonal lags, indicating an ARIMA model might be suitable.
4. The histogram does not provide us with very meaningful insights to read given the strong trends and seasonality.



Comments: On the Differenced Monthly CO2 Change EDA plots:

1. The Time Series Plot displays stationary fluctuations around zero, indicating that the series is more stable after differencing, with no clear trend.
2. The ACF for Differenced Monthly CO2 shows significant autocorrelation at seasonal lags (e.g., 12 months), providing a strong evidence of seasonality.
3. The PACF for Differenced Monthly CO2 exhibits significant spikes at lag 1 and other seasonal lags, suggesting the potential presence of an AR(1) process along with seasonal components.
4. The histogram does not provide us with very meaningful insights to read given the strong trends and seasonality. But we do see more positive then negative.

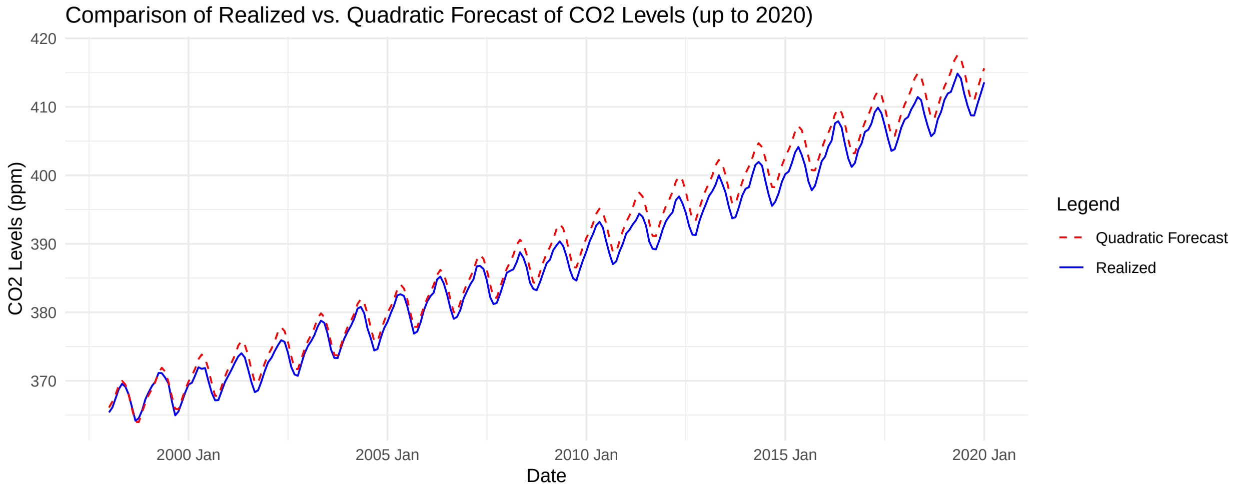


Comments: On the log-differenced plots, we noticed that:

1. The series appears to fluctuate around zero, indicating that the differencing has likely removed the trend, suggesting stationarity. The remaining variations reflect short-term dynamics and seasonality. The remaining variations reflect short-term dynamics and seasonality. The variance is more stable over time compared to the differenced CO2.
2. From the histogram, the distribution is centered around zero but appears slightly skewed to the right, with more frequent positive changes than negative. The distribution also shows some kurtosis with a peak around zero.
3. The ACF continues to exhibit noticeable seasonal peaks, indicating persistent seasonality even after log-differencing. This further strengthens our previous observations of seasonality.
4. The PACF spikes at lag 1 and other seasonal lags, suggesting potential AR terms in the model, particularly for capturing the seasonal structure.

Part 2b: Compare linear model forecasts against realized CO2

For this part of the analysis, we will compare the realized atmospheric CO2 levels with the forecasts produced by the regression model developed in 1997. This descriptive comparison will allow us to assess the accuracy and reliability of our previous model. By examining how well the predicted CO2 concentrations align with the actual observed data, we can gain insights into the model's performance and evaluate whether the underlying assumptions still hold true for the post-1997 period. This step is crucial for understanding the model's predictive limitations and will inform adjustments for future forecasting models if necessary.

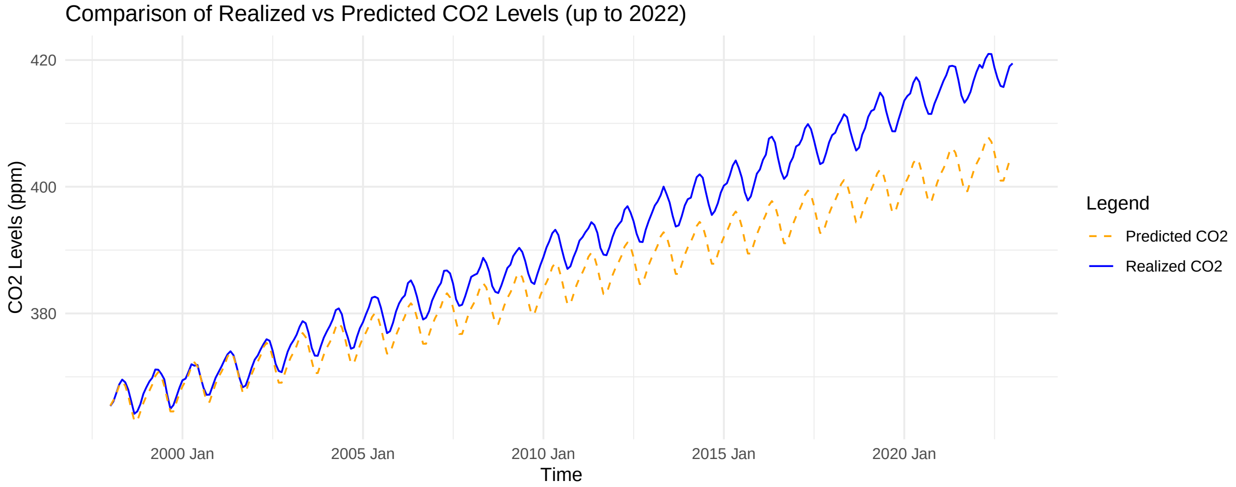


Comments: The predicted quadratic regression model with seasonality and realized CO2 levels track closely together which indicates that our model fits the data well even beyond 1997. Both the trend and the seasonal pattern (e.g., periodic fluctuations) are well-captured by the model, as evidenced by the overlap of the two lines.

Overall, the forecast aligns well with the actual CO2 levels from 1997 onwards, indicating that the quadratic model with seasonality provides a good approximation of atmospheric CO2 behavior from 1997 onward.

Part 3b: Compare ARIMA models forecasts against realized CO2

In this section, we will perform another comparison between the realized atmospheric CO2 levels and the forecasts generated by the ARIMA model fitted in 1997. This analysis will provide insights into how well the model's projections align with actual observations over the years. This comparison is essential for understanding the model's performance in capturing the long-term growth and seasonal fluctuations in CO2 levels, which will help in assessing the model's predictive accuracy and identifying any shifts in the underlying CO2 dynamics.



Comments: The plot compares predicted and realized atmospheric CO2 levels from 1998 to 2020. The

ARIMA model effectively captures both the upward trend and the seasonal fluctuations of the Keeling Curve. However, it consistently underpredicts actual CO2 levels, particularly in later years.

A potential reason for this underprediction could be accelerated emissions from increased global industrial activity or other unforeseen factors, such as deforestation or changes in ocean absorption rates, which the model may not have fully incorporated. This suggests the model may need further adjustment to account for these evolving environmental dynamics. This very observation could be due to either CO2 growth is really not a stochastic process, or the model is forecasting too far in the future with dated information, which naturally hinders the performance.

Part 4b: Evaluate the performance of 1997 linear and ARIMA models

In this section, we aim to examine the predictive performance of our regression and ARIMA models regarding a critical milestone: the first instance when atmospheric CO2 levels exceeded 420 ppm. Specifically, we will compare the predicted crossing points from both models with the actual observed data. This analysis will help us evaluate the accuracy of our models' projections and assess how closely the model predictions align with the realized trend in CO2 growth over time.

Table 2: First Crossing of 420 ppm for CO2 Levels

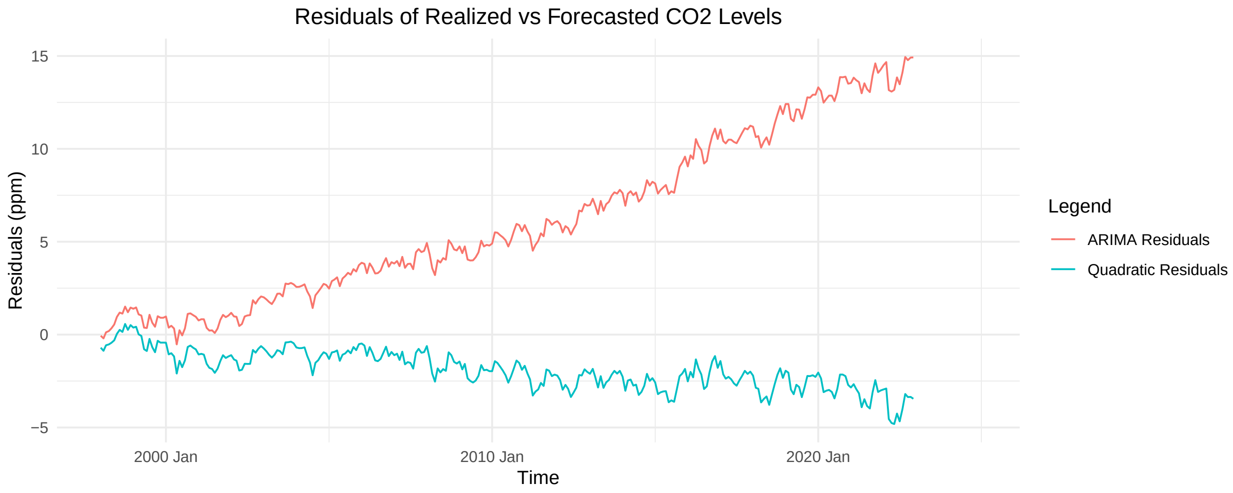
Model	First_Crossing_420ppm
Regression Model	2020 May
ARIMA Model	NA
Realized CO2	2022 Apr

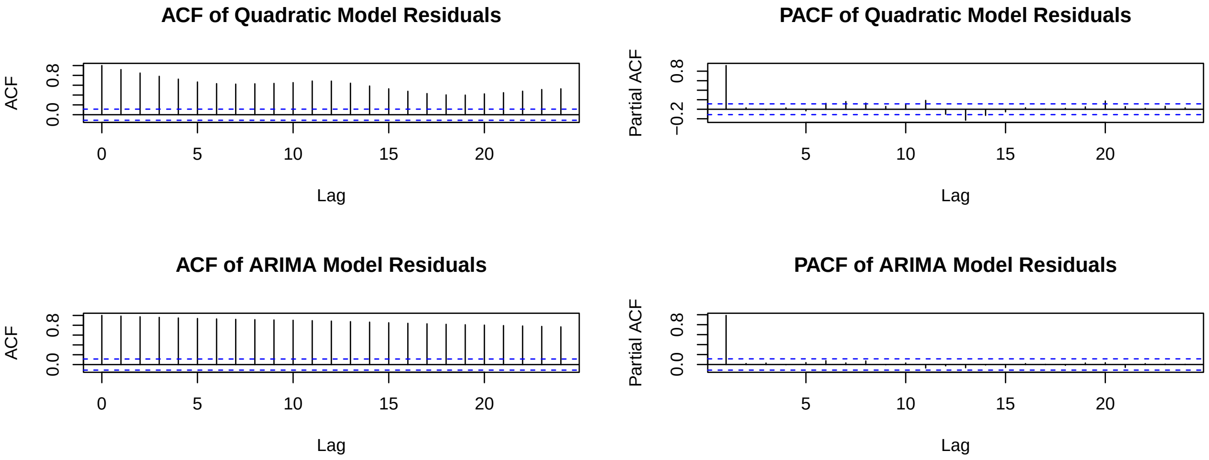
Comments:

The quadratic regression model predicts that CO2 levels would cross 420 ppm in May 2020, while the realized crossing actually occurred in April 2022. This suggests that the quadratic model underestimated the time it would take to reach this threshold by nearly two years (i.e. the quadratic model overestimated the CO2 growth rate, by around 2 years).

On the other hand, the ARIMA model did not predict a crossing of 420 ppm within the available forecast horizon, indicating that it underestimated the CO2 growth rate very significantly. This suggests that the ARIMA model is less aggressive in projecting the upward trend of CO2 levels compared to both the quadratic model and actual observed data.

In this part, we continue to examine the forecasting accuracy of quadratic regression and ARIMA models by generating a monthly average CO2 series by the weekly data from 1997 onward. We assess each model's performance over the entire period to determine which best captures the underlying trend and will apply formal tests to validate model performance.





```
##
##  Box-Ljung test
##
## data:  residuals_quadratic
## X-squared = 2494.9, df = 20, p-value < 2.2e-16
```

```
##
##  Box-Ljung test
##
## data:  residuals_arima
## X-squared = 4990.8, df = 20, p-value < 2.2e-16
```

```
## RMSE for Quadratic Regression Model: 2.141397
```

```
## RMSE for ARIMA Model: 7.532728
```

Comments:

This analysis reveals contrasting performance between the quadratic regression and ARIMA models in forecasting CO2 levels from 1997 onward. The regression model exhibits relatively stable residuals close to zero, indicating that it aligns more closely with the actual CO2 levels over time. In contrast, the ARIMA model shows a clear upward trend in residuals, suggesting that it consistently underpredicts CO2 levels as time progresses.

The Ljung-Box test results for both models indicate significant autocorrelation in the residuals, with p-values below 2.2e-16. This implies that neither model fully captures all the patterns in the data, leaving room for further improvements or the inclusion of additional variables.

When comparing RMSE, the regression model achieves a lower RMSE of 1.96 compared to the ARIMA model's RMSE of 7.95, suggesting better predictive accuracy. While the ARIMA model may have performed well on historical data, its forecast accuracy appears weaker, highlighting its limitations in predicting future values accurately. In contrast, the quadratic model demonstrates stronger forecasting performance in this scenario, despite being simpler. Overall, the quadratic model currently serves as a more reliable forecasting approach for CO2 levels. This also indicates that the reality about CO2 growth is likely closer to the assumption behind using the regression model - CO2 growth is primarily driven by deterministic trends, which could be human activities.

Part 5b: Train best models on present data

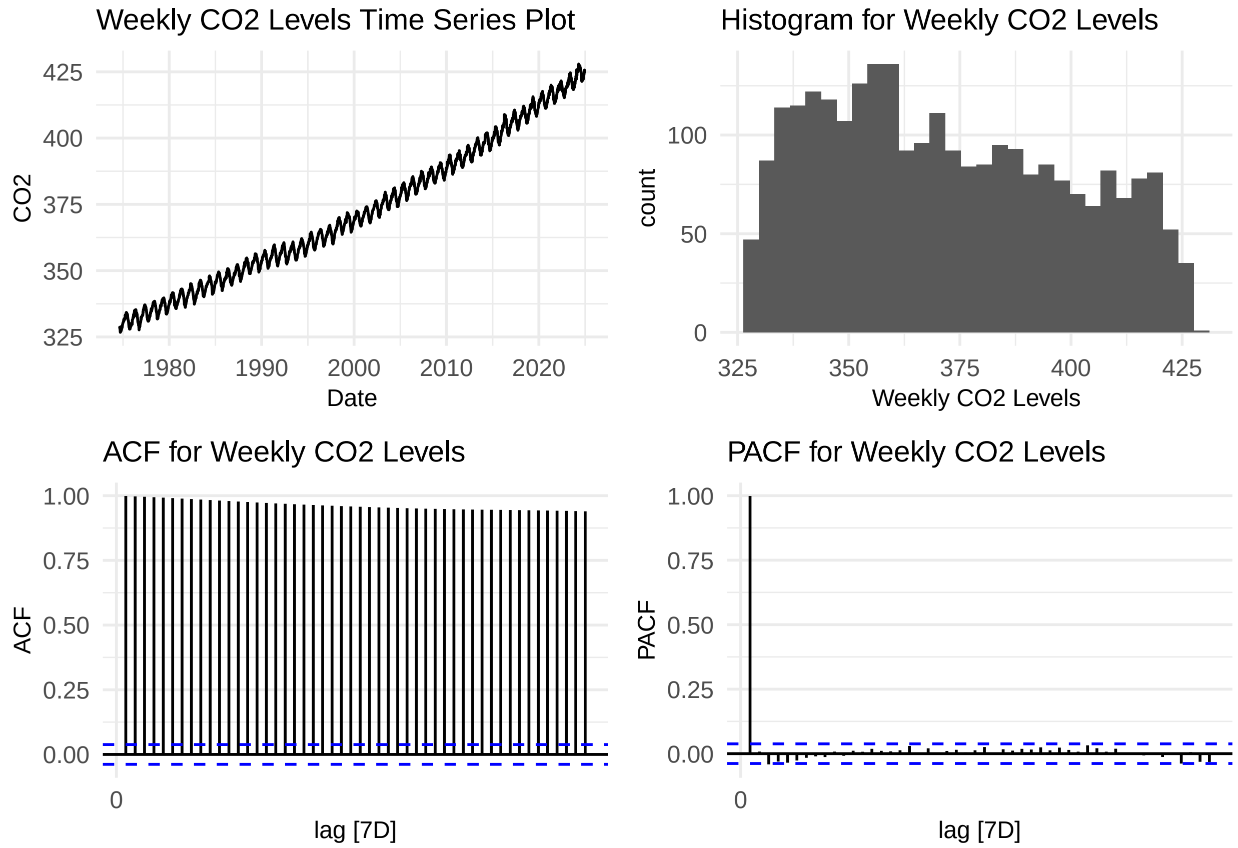
In this section, we use weekly CO2 data to train ARIMA models (both seasonally adjusted (SA) and non-seasonally adjusted (NSA)) and a polynomial regression model on the SA series. Training data is used up to two years before the most recent data, with the final two years reserved for testing. This setup enables us to evaluate how each model performs in predicting recent CO2 trends and compare their forecasting accuracy against each other.

```
## # A tsibble: 6 x 2 [7D]
##   Date         CO2
##   <date>     <dbl>
## 1 1974-08-11  329.
## 2 1974-08-18  329.
## 3 1974-08-25  328.
## 4 1974-09-01  328.
## 5 1974-09-08  328.
## 6 1974-09-15  327.
```

To prepare for the weekly data, we replaced any missing values (-999.99) with NA, and then filled those NA values by carrying forward the last observed value to create a clean time series. Afterwards, we applied seasonal adjustment to the data accordingly. Let's take a look at the training and testing sets.

```
## # A tsibble: 6 x 3 [7D]
##   Date         CO2 CO2_SA
##   <date>     <dbl>  <dbl>
## 1 1974-08-11  329.   328.
## 2 1974-08-18  329.   329.
## 3 1974-08-25  328.   329.
## 4 1974-09-01  328.   329.
## 5 1974-09-08  328.   329.
## 6 1974-09-15  327.   328.
```

```
## # A tsibble: 6 x 3 [7D]
##   Date         CO2 CO2_SA
##   <date>     <dbl>  <dbl>
## 1 2022-10-23  416.   418.
## 2 2022-10-30  416.   418.
## 3 2022-11-06  417.   418.
## 4 2022-11-13  417.   418.
## 5 2022-11-20  418.   419.
## 6 2022-11-27  418.   418.
```



Comments: From the CO2 level EDA charts we can observe that:

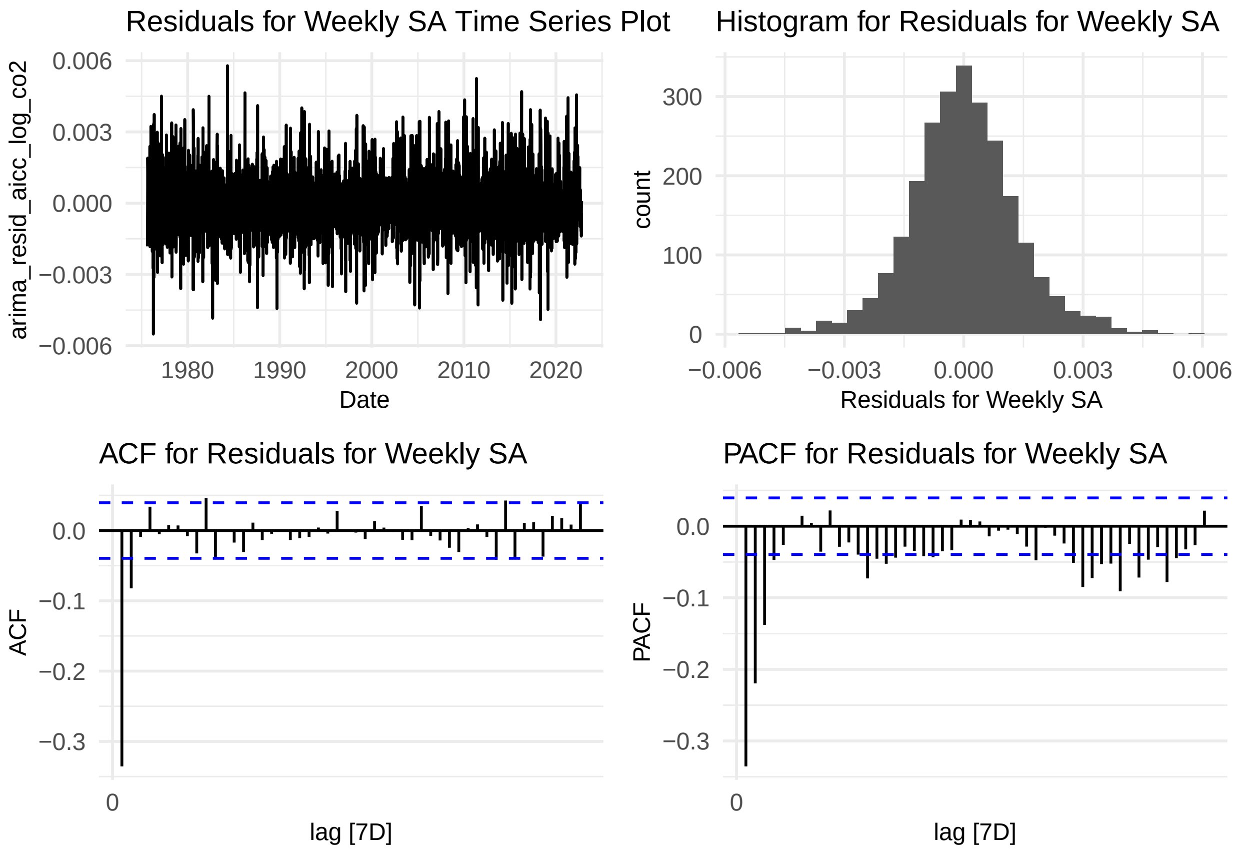
1. The weekly CO2 concentration presents a strong pattern of seasonality across months.
2. Persistent ACF with very slow decay and small fluctuation suggests the potential presence of AR process with seasonality. The slow decay suggests that the series is most likely not stationary.
3. Quickly dropping PACF with periodic significance suggests the potential existence of AR and MA processes.

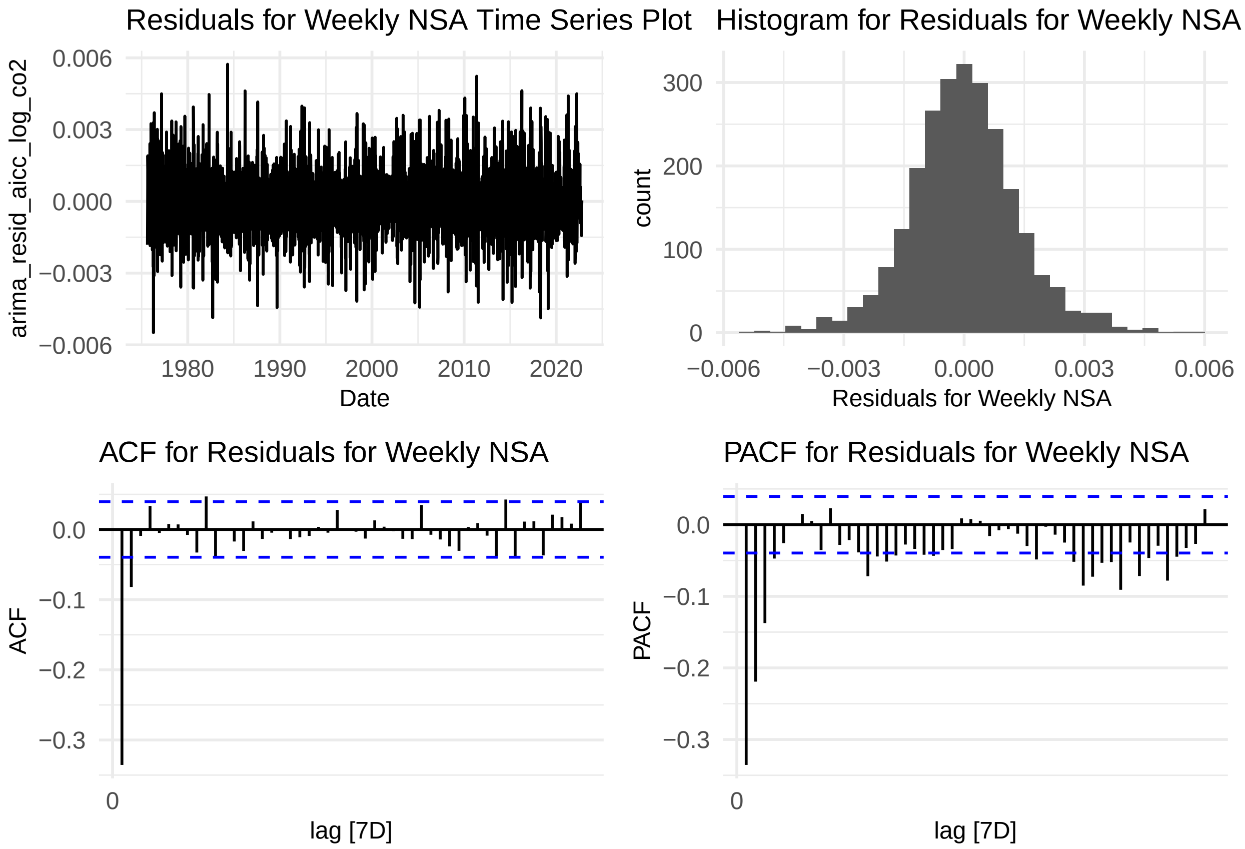
Given the similarity between the weekly and the monthly data, we decided to apply the same ARIMA model parameters to the weekly CO2 data. We begin by taking the log difference of CO2, setting d=1 in the ARIMA model, to address trends and ensure stationarity. Ultimately, if the initial model provides satisfactory residual diagnostics, we keep it unchanged, ensuring a consistent and parsimonious approach across both monthly and weekly analyses.

Here we are training the same ARIMA model on both seasonally adjusted (SA) and non-seasonally adjusted (NSA) and a polynomial regression model on the SA series. And here are the residuals diagnosis for both SA and NSA ARIMA models:

```
## Series: CO2_SA
## Model: ARIMA(0,1,0)(0,1,1)[52]
## Transformation: log(CO2_SA)
##
## Coefficients:
##          sma1
##       -0.9225
## s.e.   0.0099
##
## sigma^2 estimated as 2.524e-06:  log likelihood=12739.63
## AIC=-25475.26   AICc=-25475.26   BIC=-25463.65

## Series: CO2
## Model: ARIMA(0,1,0)(0,1,1)[52]
## Transformation: log(CO2)
##
## Coefficients:
##          sma1
##       -0.9225
## s.e.   0.0099
##
## sigma^2 estimated as 2.52e-06:  log likelihood=12742.58
## AIC=-25481.16   AICc=-25481.16   BIC=-25469.55
```




Residuals for Weekly NSA Time Series Plot

Histogram for Residuals for Weekly NSA

ACF for Residuals for Weekly NSA

PACF for Residuals for Weekly NSA

```
##
##  Box-Ljung test
##
## data:  residuals_sa_arima
## X-squared = 115.54, df = 10, p-value < 2.2e-16
```

```
##
##  Box-Ljung test
##
## data:  residuals_nsa_arima
## X-squared = 115.36, df = 10, p-value < 2.2e-16
```

Comments: The Box-Ljung tests show a very low p-value suggesting to reject the null hypothesis that there is no serial correlation on the residuals for both SA and NSA ARIMA models. Additionally, based on the ACF and PACF plots of the residuals, we observe signs of serial correlation in the residuals, which indicates that the current model might not fully capture all the underlying patterns in the data. Since we're using a previous monthly-based model on weekly data, adjusting the model by increasing the MA (moving average) component to 4 is a reasonable approach. This adjustment aligns with the fact that there are approximately 4 weeks in a month. By setting q=4 for the MA component, we account for potential correlations across these weekly periods within each month. This adjustment should help reduce residual autocorrelation and improve the model's fit to the weekly structure of the data. Below are the adjusted models:

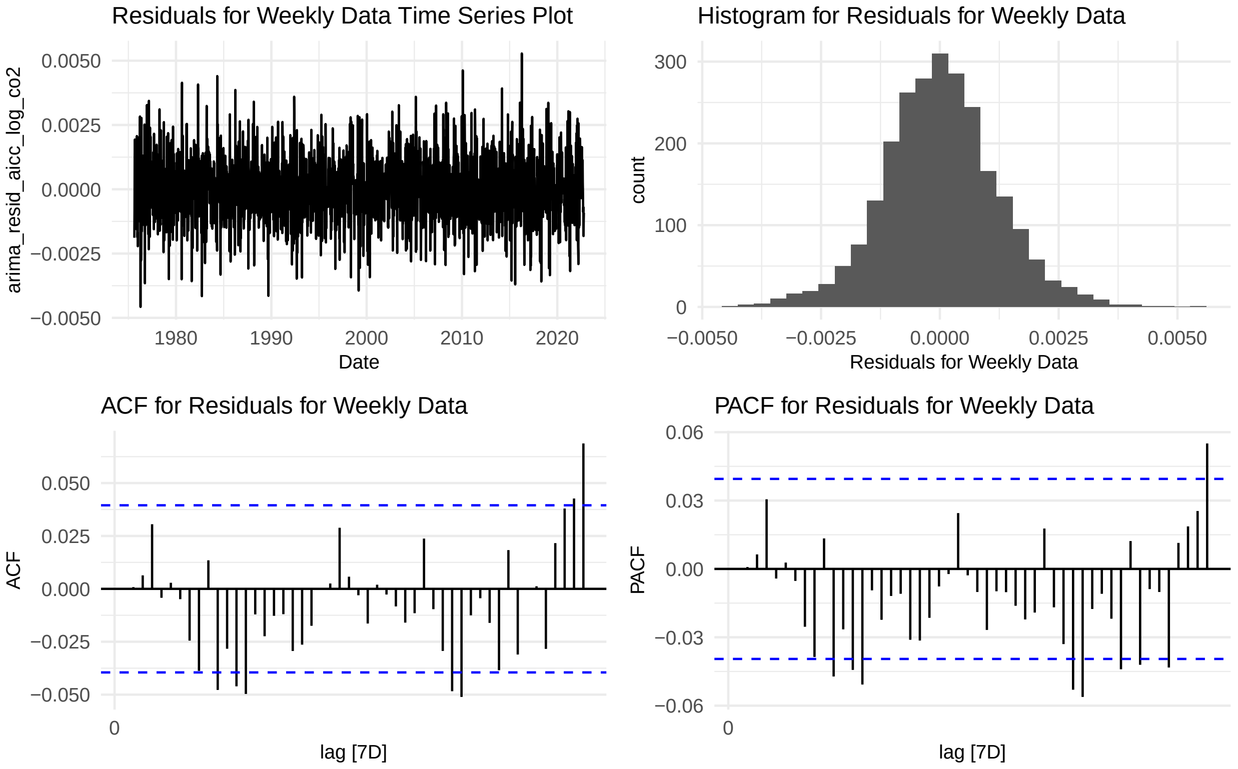
```
## Series: CO2_SA
## Model: ARIMA(0,1,4)(0,1,1)[52]
## Transformation: log(CO2_SA)
```

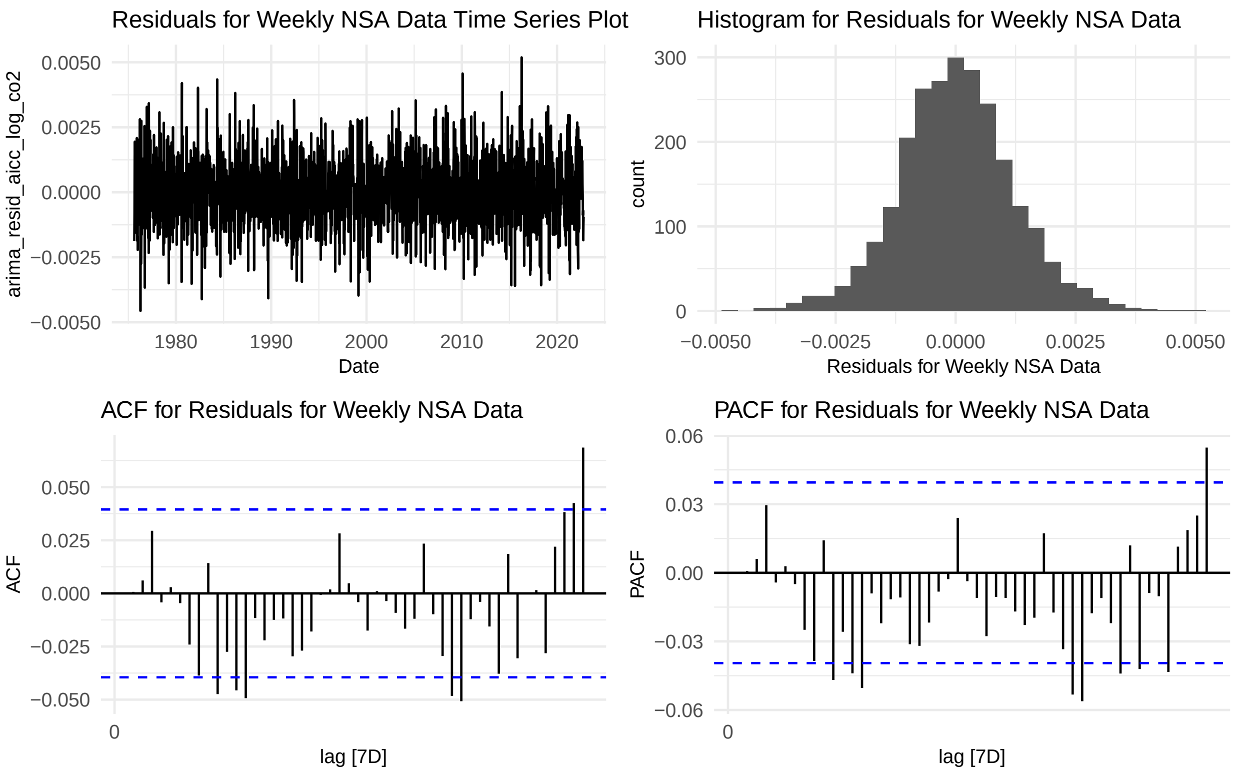
```
##
## Coefficients:
##           ma1      ma2      ma3      ma4     sma1
##       -0.5483  -0.1363  -0.0250  -0.0034  -0.8043
## s.e.   0.0212   0.0228   0.0231   0.0213   0.0127
##
## sigma^2 estimated as 2.166e-06:  log likelihood=13038.95
## AIC=-26065.91   AICc=-26065.87   BIC=-26031.06
```

And then we fit the non-seasonally adjusted data to the ARIMA model.

```
## Series: CO2
## Model: ARIMA(0,1,4)(0,1,1)[52]
## Transformation: log(CO2)
##
## Coefficients:
##           ma1      ma2      ma3      ma4     sma1
##       -0.5477  -0.1357  -0.0248  -0.0026  -0.8044
## s.e.   0.0212   0.0228   0.0231   0.0213   0.0127
##
## sigma^2 estimated as 2.163e-06:  log likelihood=13041.29
## AIC=-26070.58   AICc=-26070.54   BIC=-26035.72
```

Now we want to examine the residuals for the new ARIMAs on both SA and NSA to check for serial correlations.



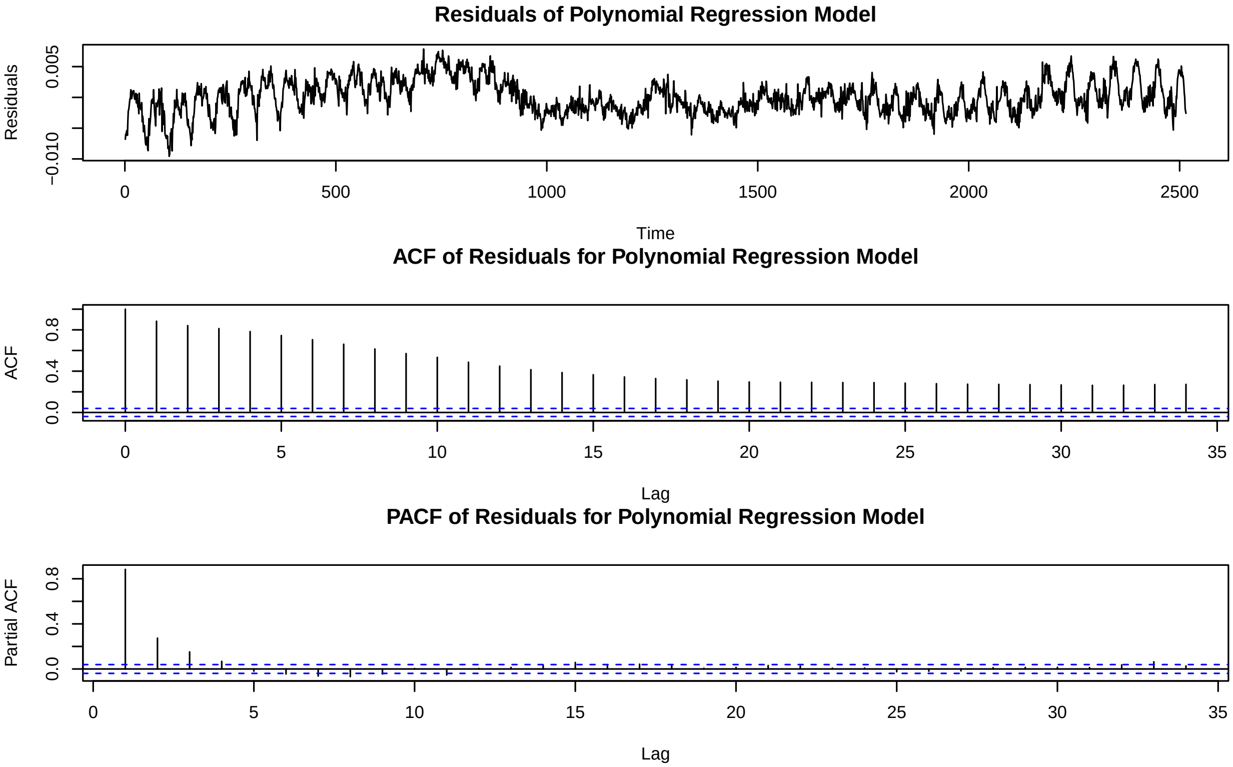


```
##
##  Box-Ljung test
##
## data:  residuals_sa_arima_411
## X-squared = 8.7632, df = 10, p-value = 0.5547
```

```
##
##  Box-Ljung test
##
## data:  residuals_nsa_arima_411
## X-squared = 8.7506, df = 10, p-value = 0.5559
```

The new ARIMA model has done a much betetr job in terms of residuals resembling white noise, as visually supported by the plots as well as the much higher p-value from Box-Ljung test. Lastly, we fit the seasonally adjusted data to our polynomial model.

```
## Series: CO2_SA
## Model: TSLM
## Transformation: log(CO2_SA)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0095794 -0.0018983 -0.0002202  0.0018512  0.0078633
##
## Coefficients:
##                   Estimate Std. Error   t value Pr(>|t|)
## (Intercept)      5.802e+00  4.224e-04 13737.419  < 2e-16 ***
## trend()           7.102e-05  2.996e-07   237.077  < 2e-16 ***
## I(trend()^2)      9.195e-09  1.153e-10    79.755  < 2e-16 ***
## season_factor2   -1.874e-04  5.563e-04    -0.337 0.736228    
## season_factor3   -3.353e-04  5.563e-04    -0.603 0.546788    
## season_factor4   -2.919e-04  5.563e-04    -0.525 0.599847    
## season_factor5   -1.991e-04  5.563e-04    -0.358 0.720465    
## season_factor6   -6.079e-04  5.563e-04    -1.093 0.274587    
## season_factor7   -9.388e-04  5.563e-04    -1.688 0.091616 .  
## season_factor8   -7.844e-04  5.563e-04    -1.410 0.158658    
## season_factor9   -6.870e-04  5.563e-04    -1.235 0.216969    
## season_factor10 -1.127e-03  5.563e-04    -2.026 0.042842 *  
## season_factor11 -7.850e-04  5.563e-04    -1.411 0.158363    
## season_factor12 -5.596e-04  5.563e-04    -1.006 0.314542    
## season_factor13 -1.943e-04  5.563e-04    -0.349 0.726927    
## season_factor14 -2.758e-05  5.563e-04    -0.050 0.960457    
## season_factor15  2.066e-04  5.563e-04     0.371 0.710380    
## season_factor16  2.373e-04  5.563e-04     0.427 0.669748    
## season_factor17  5.245e-04  5.563e-04     0.943 0.345895    
## season_factor18  1.866e-04  5.563e-04     0.336 0.737272    
## season_factor19  1.394e-04  5.563e-04     0.251 0.802171    
## season_factor20  6.302e-04  5.563e-04     1.133 0.257409    
## season_factor21  6.517e-04  5.563e-04     1.171 0.241515    
## season_factor22  3.221e-04  5.563e-04     0.579 0.562700    
## season_factor23  4.474e-06  5.563e-04     0.008 0.993583    
## season_factor24 -8.632e-05  5.563e-04    -0.155 0.876703    
## season_factor25 -2.148e-04  5.563e-04    -0.386 0.699400    
## season_factor26 -3.758e-04  5.563e-04    -0.676 0.499410    
## season_factor27 -5.905e-04  5.563e-04    -1.061 0.288571    
## season_factor28 -6.840e-04  5.563e-04    -1.229 0.219004    
## season_factor29 -8.800e-04  5.563e-04    -1.582 0.113824    
## season_factor30 -9.365e-04  5.563e-04    -1.684 0.092404 .  
## season_factor31 -1.246e-03  5.563e-04    -2.240 0.025192 *  
## season_factor32 -1.331e-03  5.535e-04    -2.405 0.016245 *  
## season_factor33 -1.351e-03  5.535e-04    -2.440 0.014752 *  
## season_factor34 -1.752e-03  5.535e-04    -3.166 0.001566 ** 
## season_factor35 -1.971e-03  5.535e-04    -3.561 0.000377 ***
## season_factor36 -2.163e-03  5.535e-04    -3.907 9.58e-05 ***
## season_factor37 -2.106e-03  5.535e-04    -3.805 0.000145 ***
## season_factor38 -2.367e-03  5.535e-04    -4.276 1.97e-05 ***
## season_factor39 -2.302e-03  5.535e-04    -4.159 3.31e-05 ***
## season_factor40 -1.891e-03  5.535e-04    -3.416 0.000645 ***
## season_factor41 -1.739e-03  5.535e-04    -3.142 0.001701 ** 
## season_factor42 -1.301e-03  5.535e-04    -2.351 0.018786 *  
## season_factor43 -9.892e-04  5.563e-04    -1.778 0.075507 .  
## season_factor44 -7.651e-04  5.563e-04    -1.375 0.169136    
## season_factor45 -3.477e-04  5.563e-04    -0.625 0.531961    
## season_factor46 -1.482e-04  5.563e-04    -0.266 0.789967    
## season_factor47 -1.086e-04  5.563e-04    -0.195 0.845192    
## season_factor48 -1.767e-04  5.563e-04    -0.318 0.750787    
## season_factor49 -1.346e-04  5.563e-04    -0.242 0.808768    
## season_factor50 -4.926e-05  5.563e-04    -0.089 0.929443    
## season_factor51 -1.655e-04  5.563e-04    -0.298 0.766109    
## season_factor52 -1.199e-04  5.563e-04    -0.215 0.829429    
## season_factor53 -2.015e-04  1.041e-03    -0.194 0.846517    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002725 on 2460 degrees of freedom
## Multiple R-squared: 0.9985,  Adjusted R-squared: 0.9984
## F-statistic: 2.943e+04 on 54 and 2460 DF, p-value: < 2.22e-16
```



Comments: The SA and NSA ARIMA models both exhibit residuals centered around zero with random fluctuations, indicating they effectively capture the primary data patterns and are well-suited for short-term forecasting. The white noise nature of these residuals confirms that both ARIMA models accurately capture recent fluctuations without significant autocorrelation. Conversely, the polynomial regression model's residuals display a more structured, non-random pattern, suggesting it is better suited for capturing long-term trends but less effective for short-term fluctuations.

In summary, the ARIMA models excel in short-term forecasting due to their white noise residuals, while the polynomial regression model is more reliable for long-term trend analysis. Below is the in-sample and out-of-sample evaluation for the SA ARIMA, NSA ARIMA, and SA Polynomial models based on the forecasted out-of-sample data:

Table 3: Model Evaluation Comparison

Metric	SA.ARIMA	NSA.ARIMA	SA.Polynomial.Regression
In-Sample RMSE	364.7181	364.7081	364.7161
Out-of-Sample RMSE	415.7433	415.7714	416.0884
In-Sample AIC	-26065.9079	-26070.5757	-29646.6213
In-Sample AICc	-26065.8737	-26070.5415	-29644.0240
In-Sample BIC	-26031.0555	-26035.7234	-29320.1397

Comments: In conclusion, the analysis highlights the complementary strengths of the SA ARIMA and polynomial regression models in forecasting CO2 levels. The SA ARIMA model demonstrates a stronger capacity for short-term prediction accuracy, as evidenced by its lowest out-of-sample RMSE among the models. This suggests that the seasonally adjusted ARIMA model effectively captures recent fluctuations in CO2 levels, making it well-suited for short-term forecasts. On the other hand, the Polynomial Regression model, while showing a slightly higher out-of-sample RMSE, has significantly lower in-sample AIC, AICc, and BIC values. This indicates a better in-sample fit, suggesting that the polynomial model captures the long-term underlying trend in CO2 levels more effectively. This is consistent with our prior findings, where the ARIMA model was observed to underperform over extended periods, potentially due to its focus on short-term seasonal patterns. In contrast, the Polynomial Regression model's structure inherently allows it to model long-term deterministic trends, reflecting continuous influences on CO2 growth.

Given these insights, a combined approach could be advantageous: leveraging the SA ARIMA model for accurate short-term predictions and utilizing the polynomial regression model for reliable long-term trend analysis. This hybrid strategy could enhance forecasting accuracy and robustness, especially in applications where both short-term variability and long-term patterns are important.

Part 6b: How bad could it get?

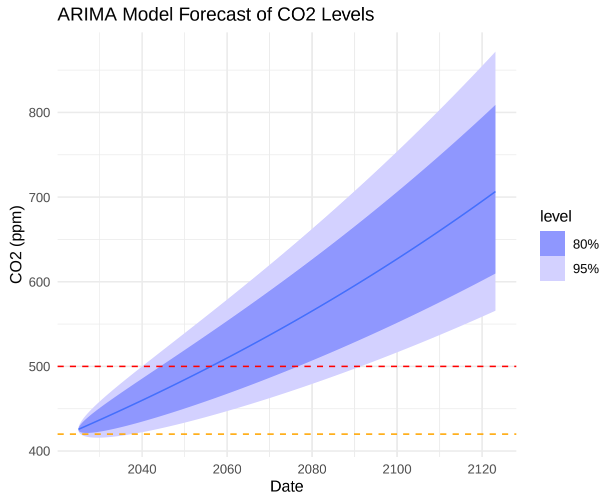
In this section, we will estimate what dates atmospheric CO2 is expected to be at 420 ppm and 500 ppm levels after fitting our models on the non-seasonally adjusted data series. After, we will aim to estimate CO2 levels in the year 2122.

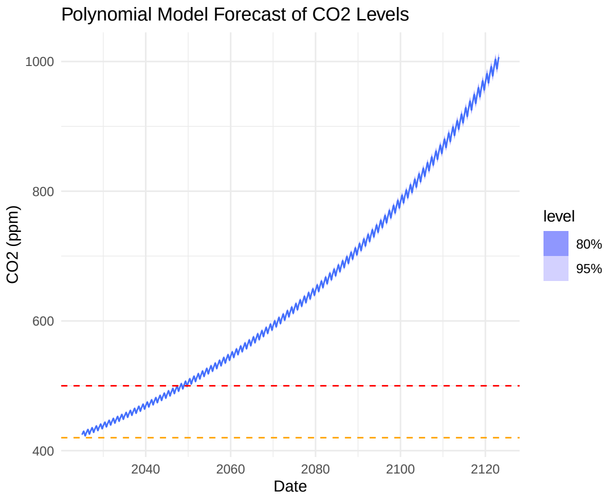
Table 4: Estimated Dates that Target CO2 Levels are crossed

Model	Target_Level	Lower_Bound	Mean	Upper_Bound
Polynomial	420	2024-12-29	2024-12-29	2024-12-29
Polynomial	421	2024-12-29	2024-12-29	2024-12-29
Polynomial	500	2048-04-19	2048-03-15	2047-04-07
Polynomial	501	2049-02-28	2048-03-29	2047-04-21
ARIMA	420	2024-12-29	2024-12-29	2024-12-29
ARIMA	421	2024-12-29	2024-12-29	2024-12-29
ARIMA	500	2091-06-03	2056-03-12	2040-01-01
ARIMA	501	2091-12-16	2056-07-30	2040-04-01

Table 5: 2122 CO2 Level Predictions with 95% Confidence Prediction Intervals

Model	Lower_Bound	Mean	Upper_Bound
ARIMA	564.29	704.19	868.25
Polynomial	987.74	995.94	1004.20





Comments:

Firstly, we note that both forecast models are expected to hit the 420 ppm and 421 ppm values immediately, as current CO2 levels (as of 10-20-2024) have already surpassed 421 ppm. Therefore, we focus the target level commentary on the dates predicted for reaching 500 ppm.

When examining the mean predictions, the polynomial regression model forecasts that CO2 levels will reach 500 ppm around almost 10 years earlier than the ARIMA model (2048-03-15 vs. 2057-10-14) and will reach 501 ppm in just two weeks following the 500 ppm mark, compared to nearly five months for the ARIMA model. Similarly, in our 2122 CO2 level predictions, the polynomial regression model projects levels to be 993 ppm, which is approximately 300 ppm higher than the ARIMA model's prediction of 694 ppm.

However, we also observe that the ARIMA model has a much wider 95% confidence interval compared to the polynomial regression model. For instance, when reaching 500 ppm, ARIMA's prediction interval spans over 54 years, while the polynomial regression model's interval spans only 1-2 years. This pattern is also reflected in our 2122 CO2 level predictions, where the ARIMA model's 95% confidence interval covers approximately 300 ppm versus about 130 ppm for the polynomial model.

When considering the upper bounds of the ARIMA model as opposed to its mean prediction, we see that the upper bound actually predicts reaching 500 ppm faster than the polynomial regression model's mean prediction. This underscores the stochastic nature of the ARIMA model, leading to increased uncertainty as forecast horizons extend, in contrast to the deterministic stability of the polynomial regression model.

Based on these results, if the trend continues until 2122, we expect CO2 levels to be between the ARIMA and polynomial regression model predictions, likely closer to the polynomial model due to its deterministic nature. At the current CO2 level of 422 ppm, the 693 ppm and 993 ppm levels predicted by our models represent 1.6 to 2.3 times the present level, posing serious environmental and societal risks for humanity.

Conclusion:

In this analysis, we start with the research question - *Is the growth of carbon dioxide (CO2) concentrations (partially or entirely) due to a deterministic trend, which can be associated with human activities such as fossil fuel combustion? Or is it mostly due to a stochastic process with a positive drift, thus harder to link to human activities?*

Overall, our analysis as of 1997 perspective and as of current perspective indicate that although with continuously updated information and model retraining, ARIMA model could do a good job in predicting CO2 concentration in short term future, polynomial regression, which inherently assumes a deterministic trend, does a much better job forecasting CO2 concentration even with dated information. Deterministic trends model outperformed stochastic trends model in out-of-sample meaningfully.

These results suggest that the CO2 series are more deterministic than stochastic in nature, which could relate to human activities. However, this also showcases that we have the potential to create change in this growing issue. The results of this analysis highlight an urgency for policy makers and individuals to take initiative in mitigating and reducing CO2 emissions.