W271 Lab 2

The Keeling Curve

Kelly Short, Kent Bourgoing, Anshul Zutshi, Changhao Meng

Introduction

The Keeling Curve is the representation of rising atmospheric carbon dioxide (CO_2) levels, first measured continuously in 1958 by geochemist Charles David Keeling at the Mauna Loa Observatory in Hawaii. This curve represents two main insights: a regular seasonal cycle in CO_2 concentrations and a steady upward trend over decades. The seasonal pattern is primarily due to photosynthesis variation, as Earth's vegetation cover changes between the northern and southern hemispheres. The long-term increase, however, correlates strongly with human activities, especially fossil fuel combustion, which continually adds CO_2 to the atmosphere. Below is a visualization of the base keeling curve data used throughout this report:

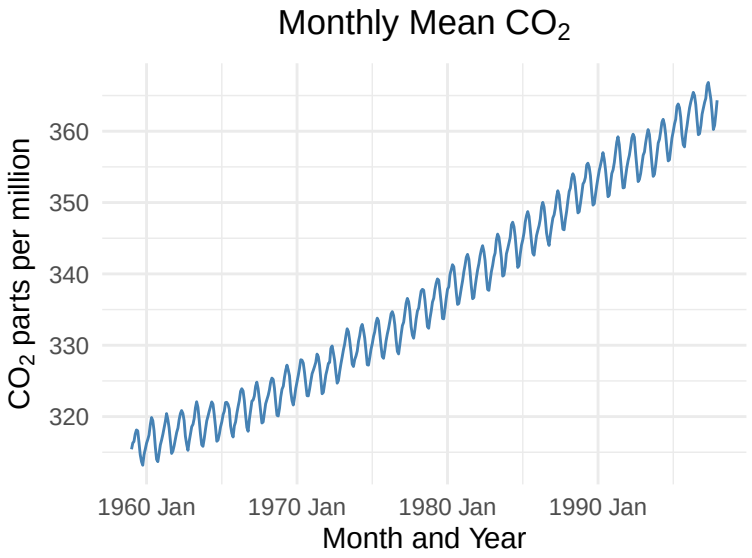
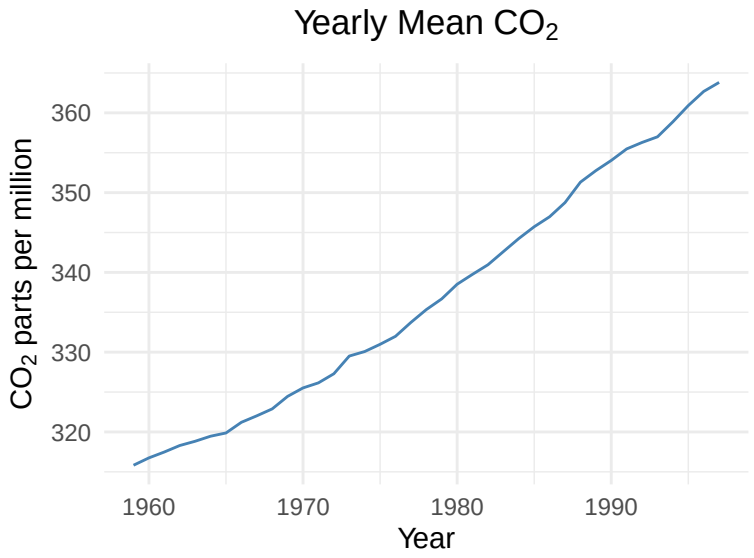


Figure 1: Keeling Curve from 1959 to 1997

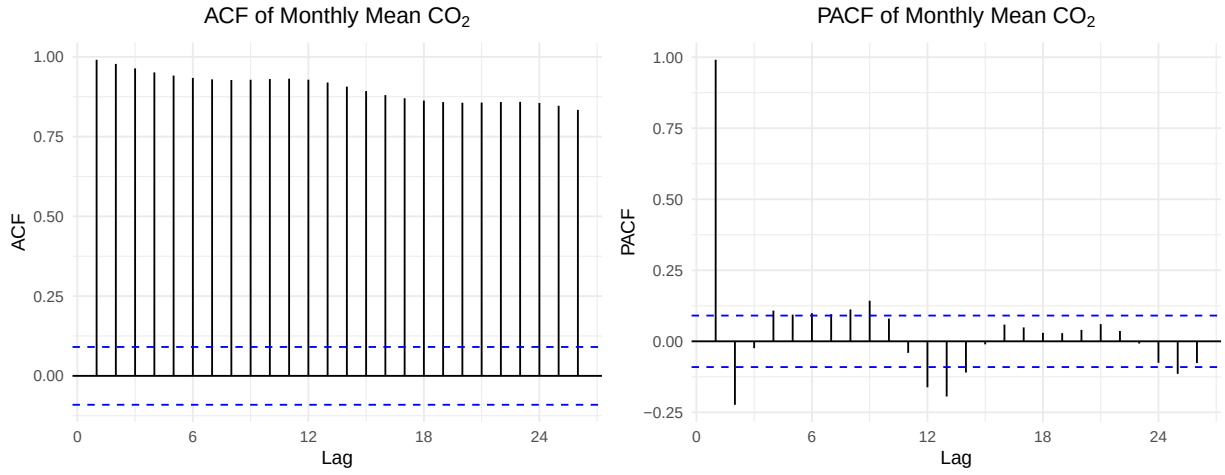
In this analysis, we will examine the characteristics of the Mauna Loa CO_2 data using the co2 dataset available in R, which provides monthly mean CO_2 levels from 1958 to 1997. The goal is to uncover insights into the seasonal and long-term trends in CO_2 concentrations, explore potential drivers, and consider the implications for understanding climate change. By examining this data from a 1997 perspective, we aim to build a foundation for assessing whether CO_2 levels pose an environmental challenge that requires further investigation. We want to better understand what trends and patterns in atmospheric carbon dioxide (CO_2) levels are evident from the Mauna Loa Observatory data, and what these trends suggest about the relationship between human activity and climate change.

EDA

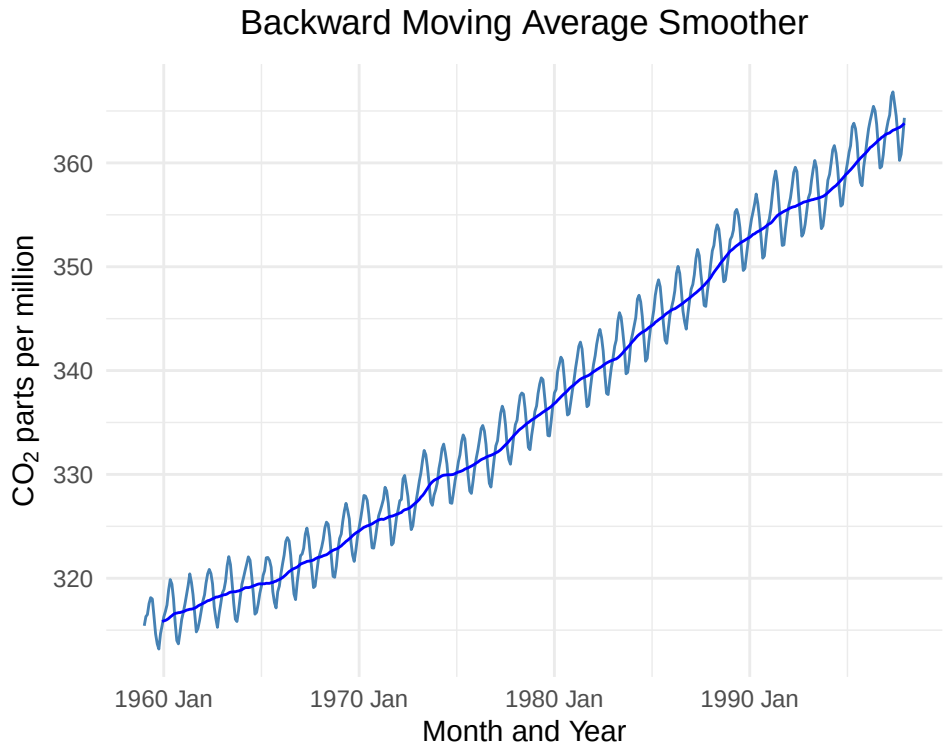
After loading the keeling data available in R from 1958 - 1997 We first assessed a few key charts to better understand how the data is behaving. Below you can see a plot of the Keeling curve yearly mean values.



After looking at the overall trend we then wanted to better understand the ACF and PACF plots. The ACF plot for the monthly CO_2 concentration shows a slow, gradual decay, indicating a strong persistence in the data, which suggests non-stationarity and the presence of a trend. Each monthly observation is highly correlated with prior months, supporting the idea of a consistent upward trend. The PACF plot, on the other hand, shows a significant spike at lag 1 and smaller spikes at seasonal lags (12, 24, etc.).

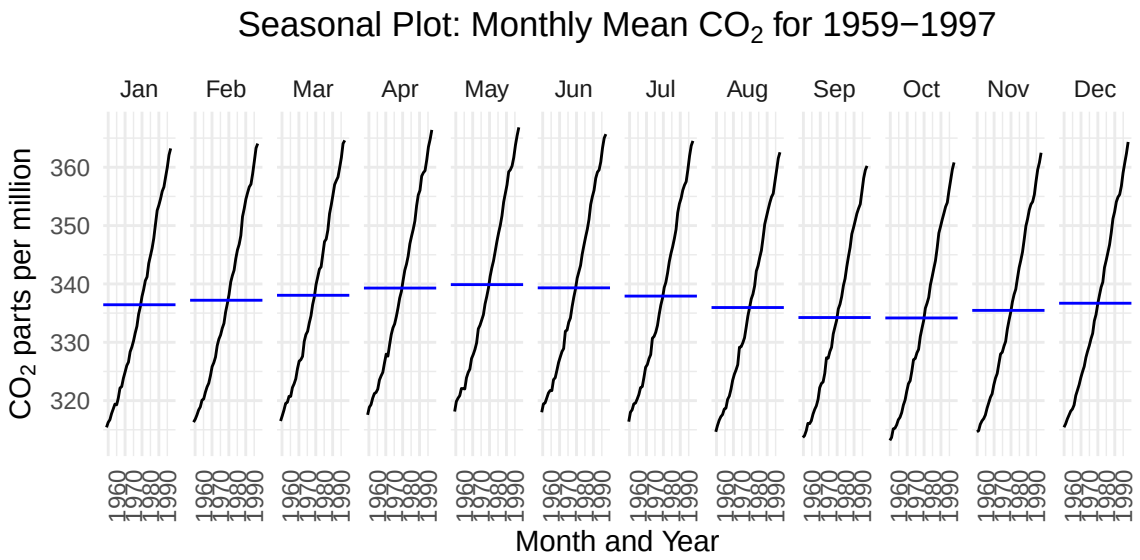


Given the nature of the initial yearly mean we decided to apply a smoothing technique. A backward moving average smoother is a technique used to smooth time series data by calculating the average of a specified number of past observations. For each point in the series, it takes the current value and a set number of previous values and averages them, giving a smoothed value for that point. For example, with a 12-month window, the smoothed value for each month is the average of that month's CO_2 level and the preceding 11 months. This method helps to reduce short-term fluctuations and highlights the longer-term trend in the data. The following plot displays the Backward Moving Average Smoother for Monthly Mean CO_2.

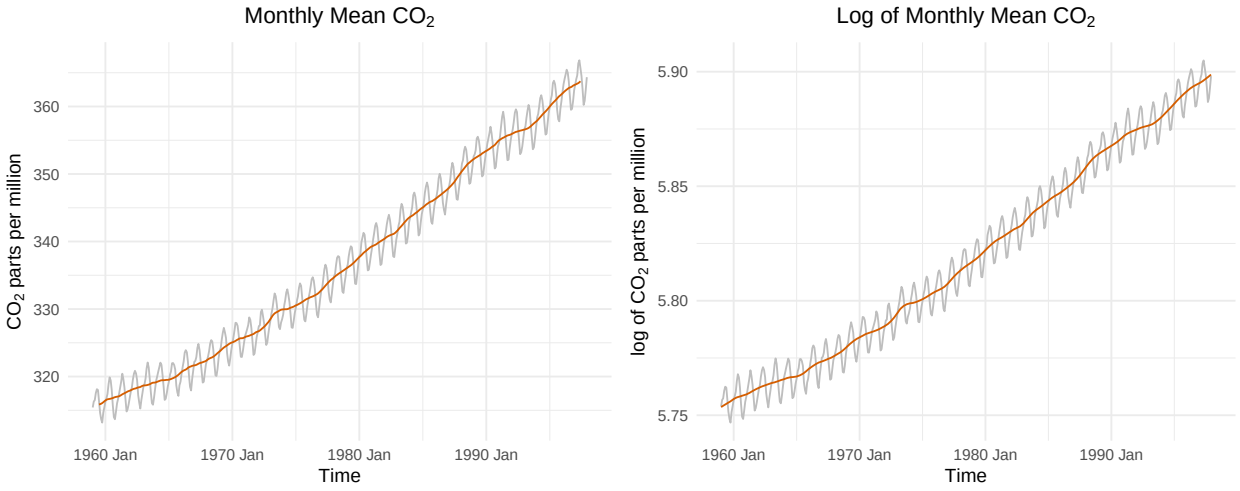


The moving average line highlights the long-term increase in CO_2 levels while filtering out short-term seasonal fluctuations, providing a clearer view of the upward trend. This smoothing approach helps reveal the consistent growth in atmospheric CO_2 over time.

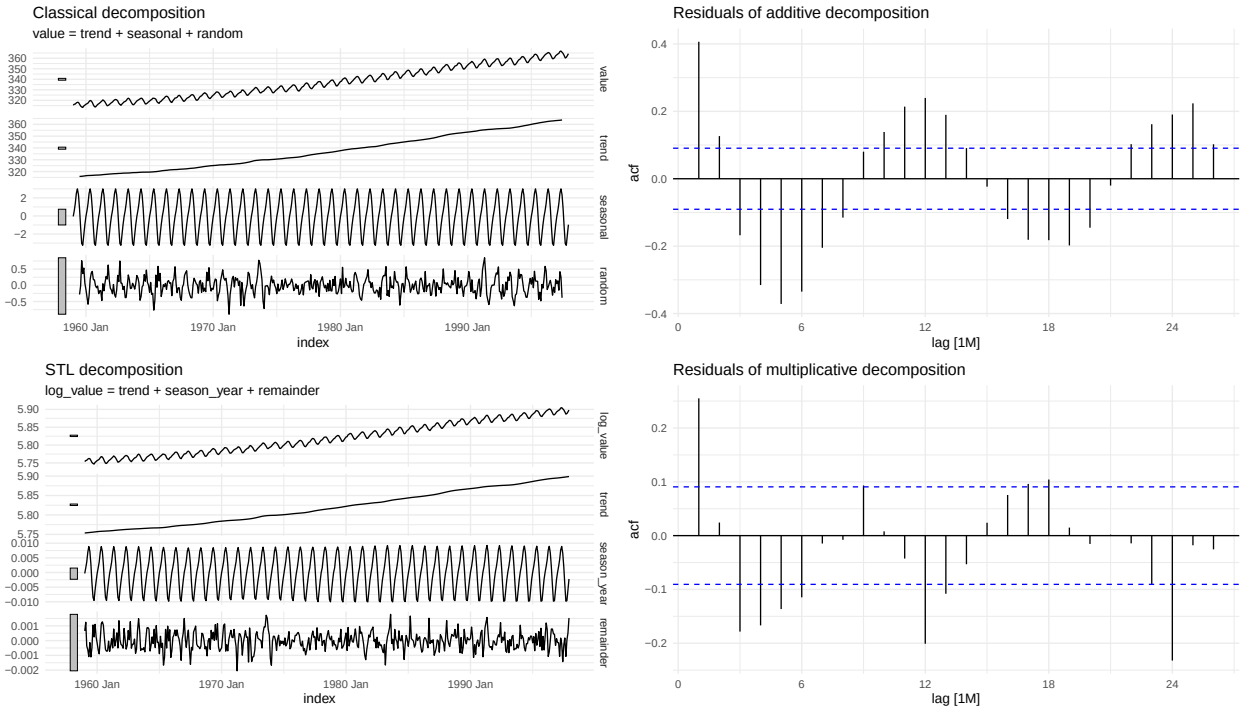
We next wanted to understand the seasonality trends in the data more.The seasonal plot below shows the monthly mean CO_2 levels from 1959 to 1997, arranged by month to highlight seasonal patterns over the years. Each line represents CO_2 levels for a particular year, with the months displayed along the x-axis. The plot reveals a clear seasonal cycle: CO_2 levels typically increase from January to May, peak around mid-year, and then decrease from July to October before rising again towards the end of the year. This pattern reflects the natural cycle of photosynthesis, where vegetation absorbs more CO_2 during the warmer months. There is also a steady upward trend in CO_2 levels across all months, indicating an overall increase in atmospheric CO_2 over time.



The next visualization we looked into shows two decompositions of monthly CO data: additive (left) and log-transformed multiplicative (right). The log transformation typically helps stabilize trends with increasing variance over time. However, both the additive and multiplicative plots display a rising trend with relatively constant seasonal fluctuations, so the log transformation may not significantly impact variance stabilization in this case.



To finish up our initial EDA we compared the classical additive decomposition (top) with the STL multiplicative decomposition (bottom) of CO_2 data. The classical model shows higher residual autocorrelation compared to the STL decomposition (with log transformation), suggesting that the STL model better captures the data's structure and reduces unexplained variation. However, notable autocorrelation remains at certain lags (e.g., 3, 4, 5) in the STL model.

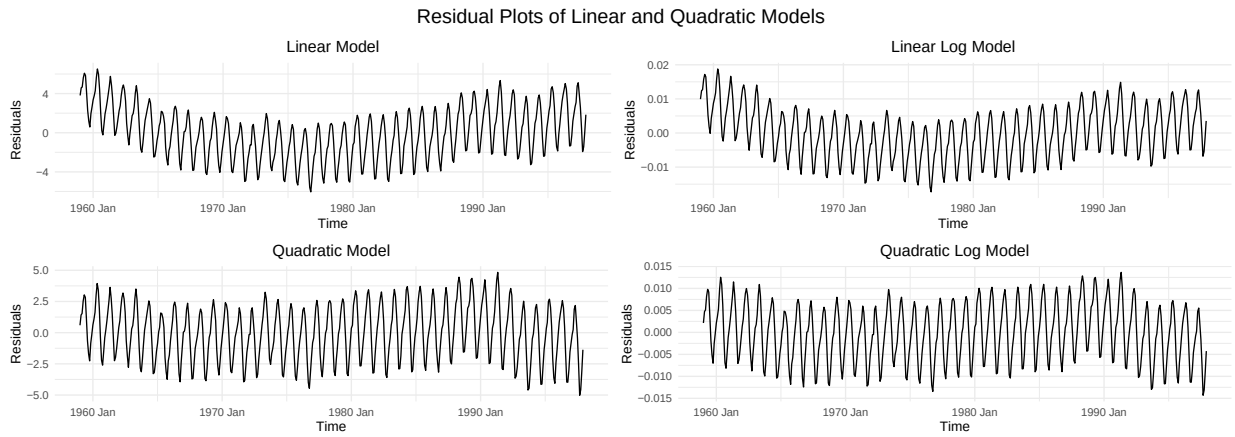


Model Development

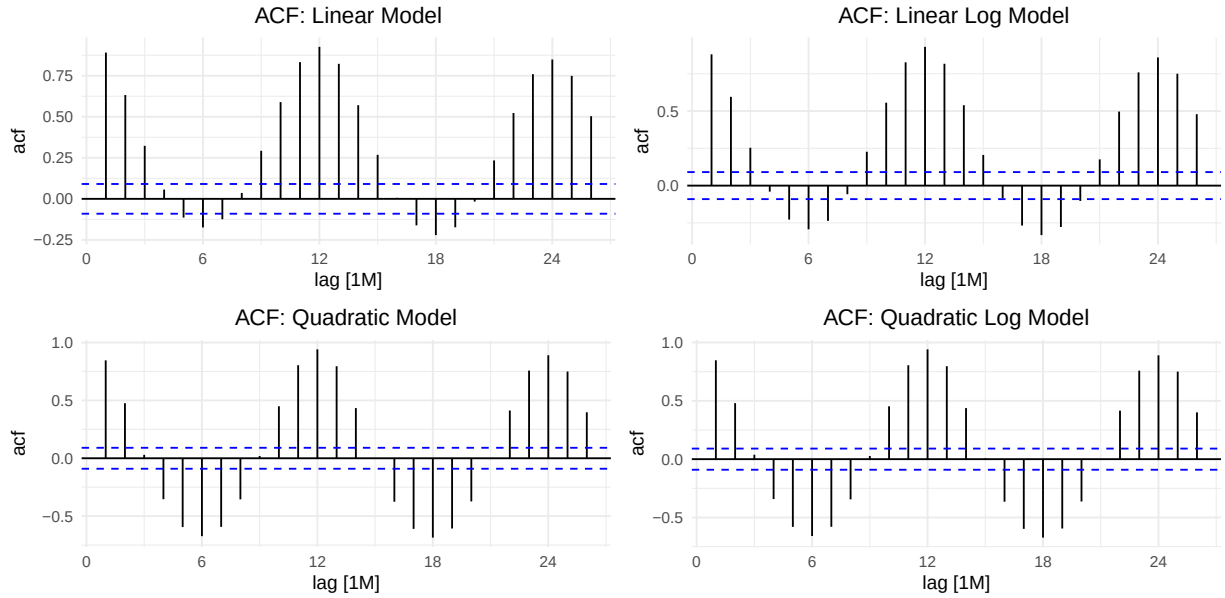
Linear and Quadratic Time Trend Models

After our initial EDA We looked at 4 different models to start: a linear time trend model, a quadratic time trend model, and then two models applying the log tranformation to the data for each model type. Below you can see the models being fit:

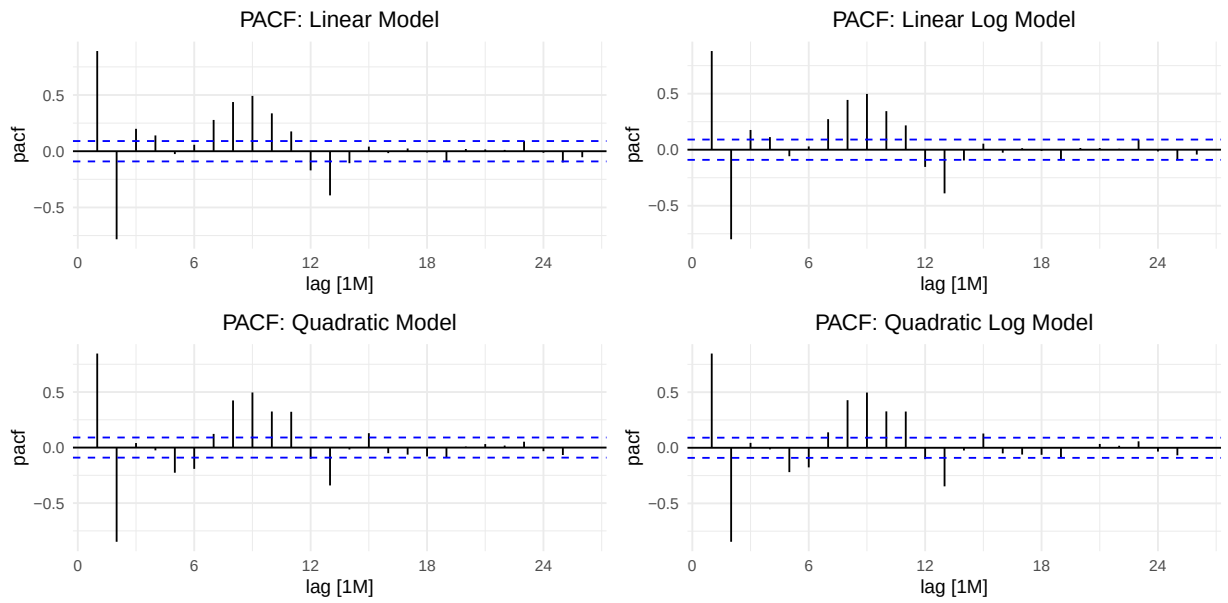
```
fit_linear <- co2_tsib %>%
  model(trend_model = TSLM(value ~ trend()))

fit_linear_log <- co2_tsib %>%
  model(trend_model = TSLM(log_value ~ trend()))

fit_quadratic <- co2_tsib %>%
   model(trend_model = TSLM(value ~ trend()+I(trend()^2)))

fit_quadratic_log <- co2_tsib %>%
   model(trend_model = TSLM(log_value ~ trend()+I(trend()^2)))
```



The figure above compares the residuals of linear and quadratic time trend models for CO_2 data, both with and without log transformation. The residuals in all models show strong seasonal patterns, indicating that neither the linear nor quadratic trends alone fully capture the seasonal component. However, the log-transformed models exhibit lower residual variation, suggesting improved variance stabilization. Additionally, the quadratic models show lower residual variation and capture the data better than the linear models.



The ACF plots for the residuals show persistent autocorrelation across all models, indicating that none fully capture the data's structure, and the residuals do not resemble white noise. The log-transformed models (top right and bottom right) exhibit slightly reduced autocorrelation, suggesting an improved fit, but still show significant residual correlation, especially at seasonal lags (e.g., 12 and 24 months). This indicates that additional seasonal components or model refinements may be needed for a better fit.



The PACF plots of the residuals, similar to the ACF plots, indicate persistent partial autocorrelation across all models. This suggests that the models do not fully capture the underlying patterns in the data. While the log-transformed models (top right and bottom right) slightly reduce partial autocorrelation at lower lags, significant spikes remain, especially around seasonal lags like 12 months.

Table 1: AIC, AICc, and BIC Comparison for Models

Model	AIC	AICc	BIC
Linear	904.8343	904.8861	917.2798
Log-Linear	-4591.4423	-4591.3906	-4578.9969
Quadratic	735.4090	735.4954	752.0029
Log-Quadratic	-4710.0834	-4709.9970	-4693.4895

The table compares the AIC, AICc, and BIC values for four time trend models (Linear, Log-Linear, Quadratic, Log-Quadratic) fitted to CO_2 data. The Log-Quadratic model has the lowest values across AIC, AICc, and BIC, indicating the best fit among the models, followed closely by the Log-Linear model. The higher values for the Linear and Quadratic models suggest that they do not capture the data structure as effectively.

Based on the previous residual time series, ACF, and PACF analyses, as well as the information criteria values, a logarithmic transformation appears appropriate. It reduces the variability of the residuals, lowers autocorrelation and partial autocorrelation in the residuals, and results in lower information criteria values compared to the non-logarithmic models.

Next, we will fit logarithmic-transformed quadratic models of various orders with seasonal dummy variables to determine which best fits the data.

```
fit_poly_2_season <- co2_tsib %>%
  model(trend_model = TSLM(log_value ~ trend()+I(trend()^2)+ season()))

fit_poly_3_season <- co2_tsib %>%
  model(trend_model = TSLM(log_value ~ trend()+I(trend()^2)+ I(trend()^3) + season()))
```

```
fit_poly_4_season <- co2_tsib %>%
  model(trend_model = TSLM(log_value ~ trend()+I(trend()^2)+ I(trend()^3) + I(trend()^4) + season()))
```

Table 2: AIC, AICc, and BIC Comparison for Quadratic Models with Seasonal Component

Model	AIC	AICc	BIC
Log-Polynomial - 2nd order	-5713.443	-5712.381	-5651.216
Log-Polynomial - 3rd order	-6116.446	-6115.240	-6050.071
Log-Polynomial - 4th order	-6130.651	-6129.291	-6060.127

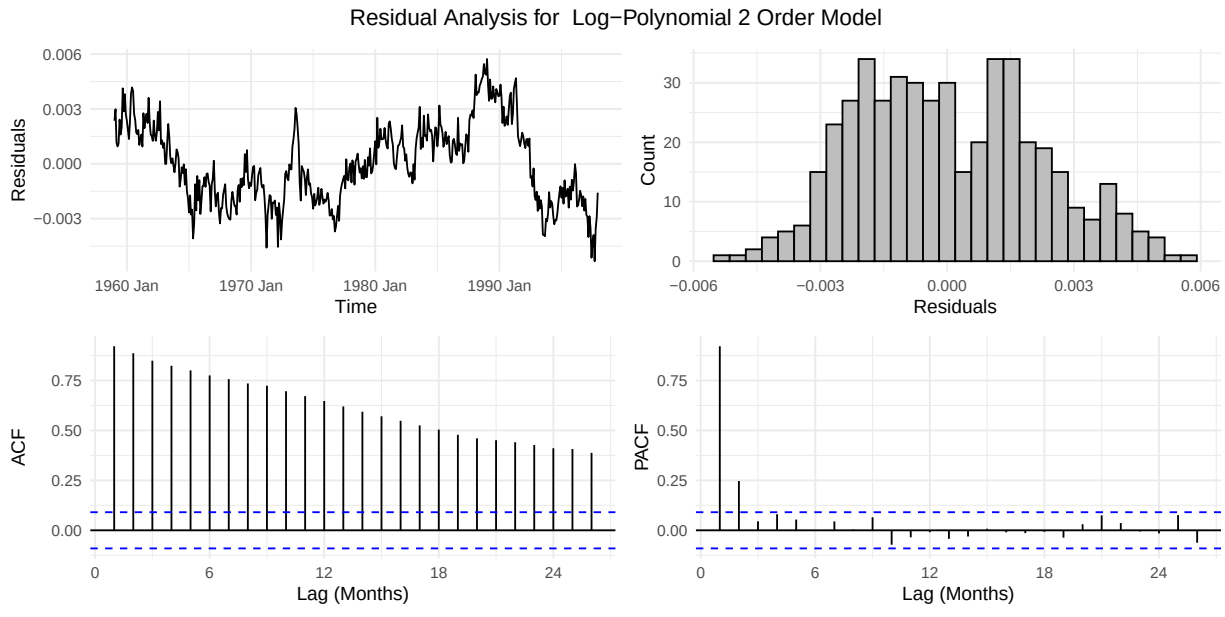
The table above compares AIC, AICc, and BIC values for Log-Quadratic models of different polynomial orders (2nd, 3rd, and 4th) fitted to CO_2 data. The information criteria values for the 3rd and 4th order models are approximately 400 units lower than those of the 2nd order model. While the 4th-order Log-Quadratic model has the lowest values, the differences in AIC, AICc, and BIC between the 3rd and 4th orders are minimal.

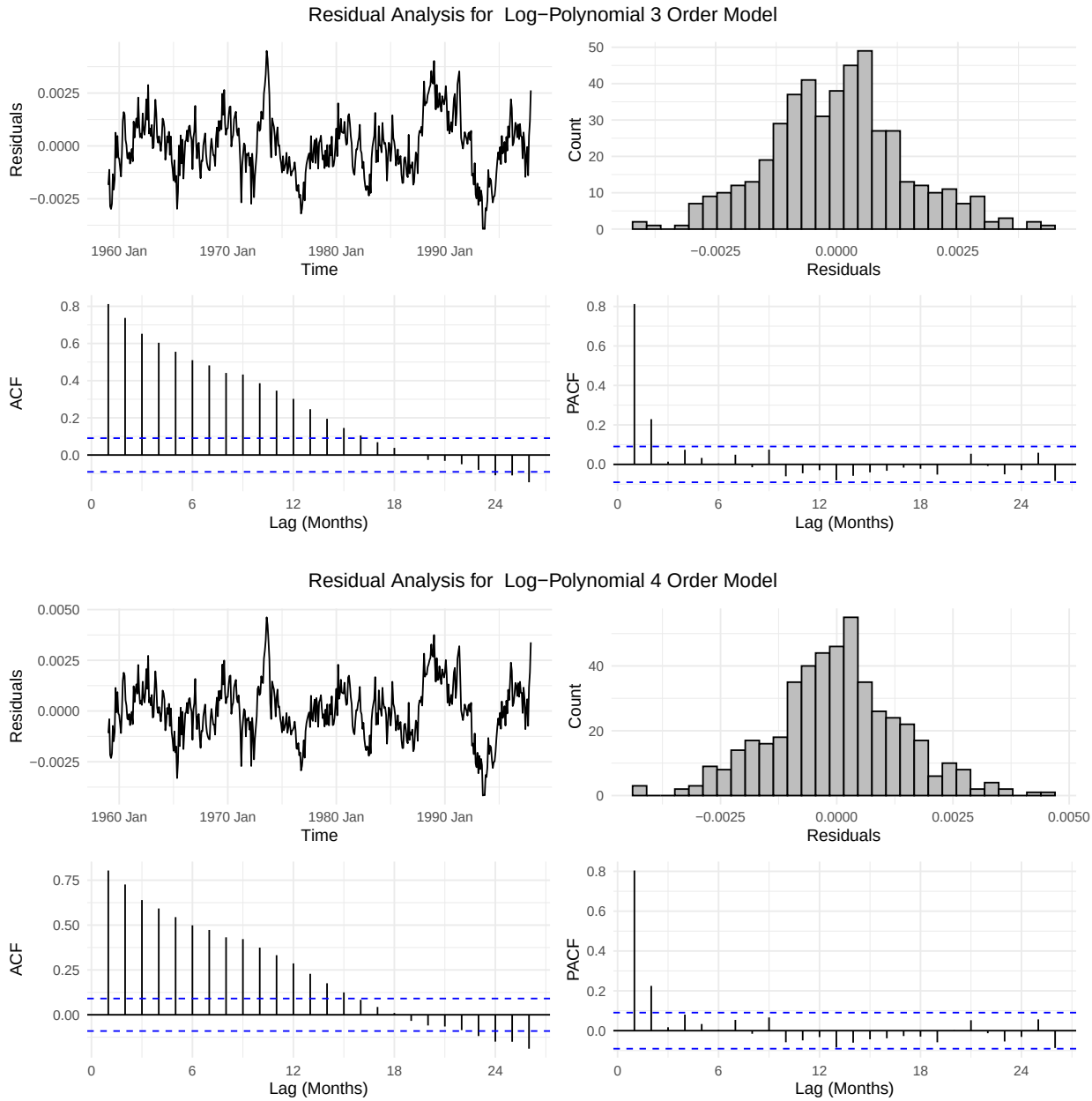
The Ljung-Box test is used to assess if residuals from a time series model are independently distributed, which is ideal for a good model fit. The null hypothesis (H_0) states that residuals are independent (white noise), while the alternative hypothesis (H_A) suggests they have autocorrelation. The test statistic (Q) considers the sample size, autocorrelation at each lag (ρ), and the number of lags tested (h), with Q following a chi-square distribution under H_0.

$$Q = n(n+2)\Sigma_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k}$$

If the null is rejected, it indicates serial correlation in the residuals, suggesting that the model may not adequately capture the data's structure.

We found that the Ljung-Box test statistic is high and the p-value is low for all models, so we reject the null hypothesis of no autocorrelation, concluding that the residuals are autocorrelated in each model.





In the residual plots displayed above, the three models show fluctuations over time, with the second-order model exhibiting more pronounced variations. As the polynomial order increases, residuals appear slightly more stabilized, especially in the 3rd and 4th order models.

All models show diminishing autocorrelation over time, with the 3rd and 4th order models having notably less significant autocorrelation, especially in the PACF plots. The ACF plot for the 2nd order model decreases more slowly than for the 3rd and 4th orders, suggesting that higher-order models reduce residual correlation.

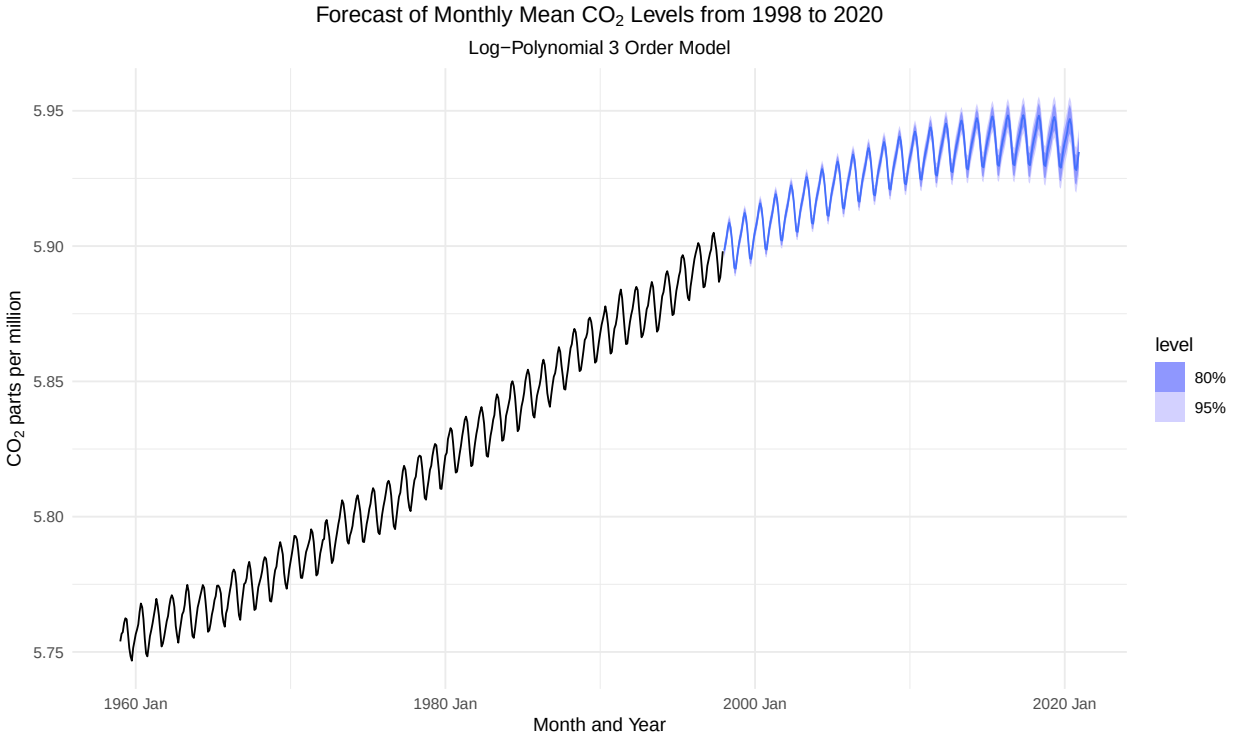
Residual distributions across all three models are approximately normal, though slightly more centered in the 3rd and 4th order models.

Given the minimal differences in AIC, AICc, and BIC between the 3rd and 4th order models, both of which are significantly lower than the 2nd order, and considering that the residual plots for the 3rd and 4th orders exhibit lower variance and reduced residual correlation at higher lags, the 3rd-order polynomial model is preferred for forecasting.

$$\log(\text{CO}_2) = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot t^2 + \beta_3 \cdot t^3 + \sum_{i=1}^{11} \gamma_i \cdot \text{Season}_i + \epsilon$$

This choice balances model fit with simplicity, avoiding unnecessary complexity and potential overfitting associated. It's important to note that our model violates the assumption that the residuals are white noise, which may impact the accuracy of our forecast.

Now that we have selected a model we want to visualize the forecast trend.



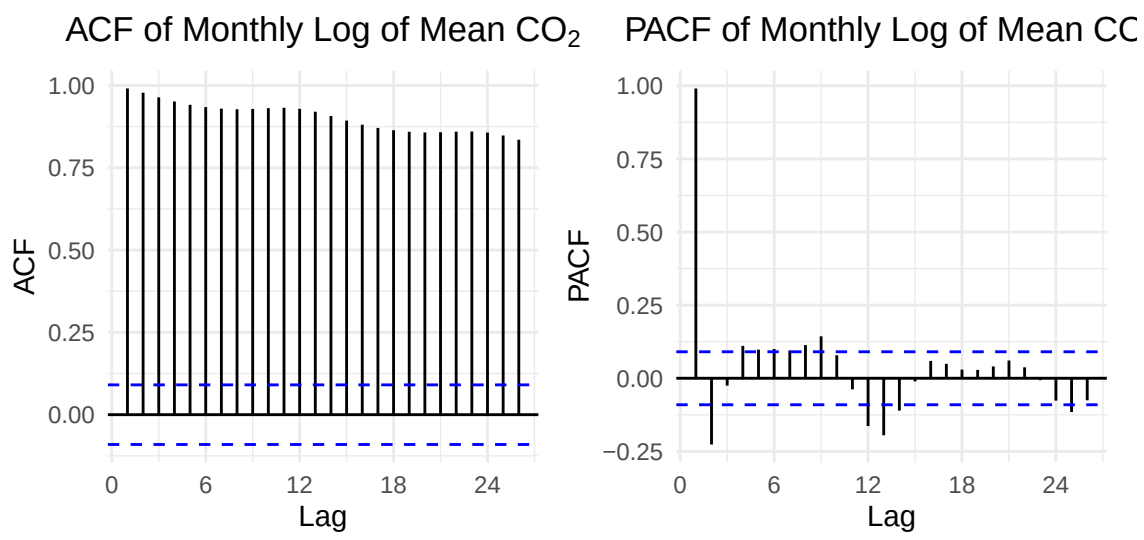
This plot shows the forecasted CO_2 levels from 1998 to 2020 based on a 3nd-order polynomial model. The prediction follows the historical trend, with seasonal fluctuations and an upward trend in CO_2 levels. The shaded areas represent 80% and 95% confidence intervals, which widen over time, indicating increased uncertainty in longer-term predictions. This widening suggests that while the model anticipates a continued upward trend with seasonal patterns, there is greater variability in potential outcomes as we approach 2020.

ARIMA Times Series Models

In order to fit an ARIMA model there is some additional EDA needed. Let's firts perform a unit root test on the log-transformed series to assess if differencing is necessary to achieve stationarity.

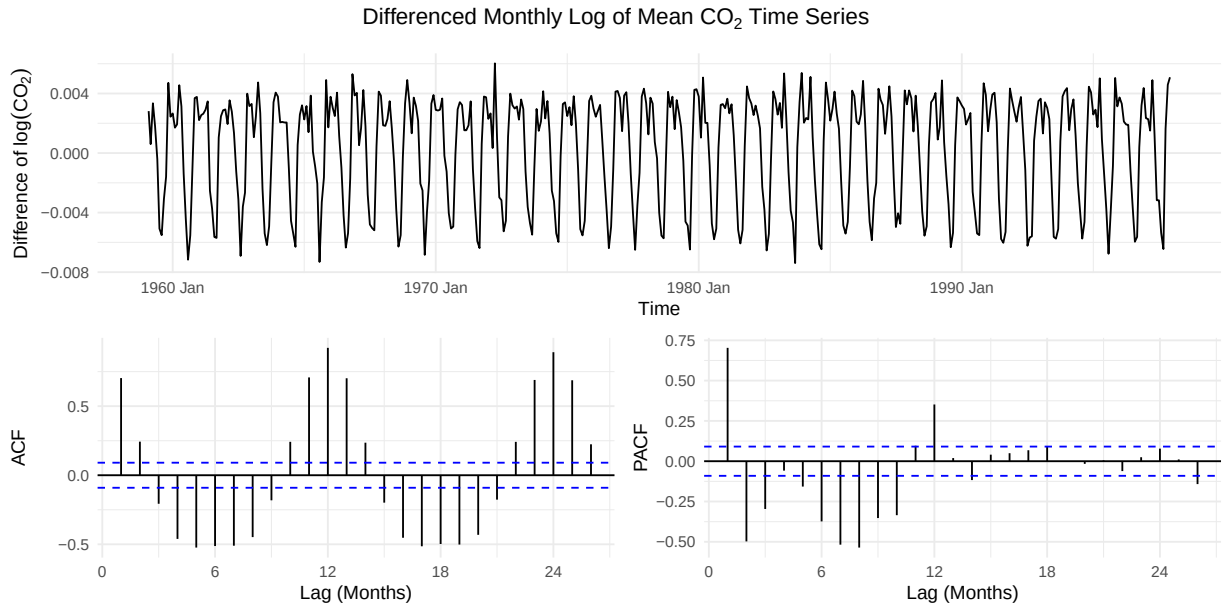
```
##
##  Augmented Dickey-Fuller Test
##
## data:  co2_tsib$log_value
## Dickey-Fuller = -3.2669, Lag order = 7, p-value = 0.0765
## alternative hypothesis: stationary
```

Since the p-value is greater than the conventional significance level of 0.05, we fail to reject the null hypothesis that the series is non-stationary. This suggests that the series may exhibit non-stationary behavior, indicating that additional differencing may be needed to achieve stationarity.



The ACF plot of the monthly log-transformed mean CO_2 shows a slow decay over time, suggesting the presence of a unit root and potential non-stationarity in the series.

Given the non-stationarity indicated by the previous ADF test and ACF plot of the log-transformed series, we will difference the series to transform it into a stationary series.



This plot displays the differenced monthly log of mean CO_2, with the time series plot in the top panel and the ACF and PACF plots below. The differenced series shows reduced trend and variability. The ACF plot exhibits significant spikes at seasonal lags (e.g., 12 months), while the PACF plot has a few notable lags, especially around seasonal intervals. This pattern suggests that the differenced series retains some seasonal correlation, but is closer to meeting stationarity assumptions.

The ACF plot above displays a strong seasonal pattern with slowly decaying lags, suggesting the possible presence of a unit root at the seasonal frequency, which may necessitate seasonal differencing. This strong seasonal pattern may also indicate the presence of at least one AR component. Additionally, it is challenging to clearly identify non-seasonal AR and MA components. The slow decay in the PACF plot hints at at least

one MA component. Meanwhile, the ACF shows significance at the first two lags, possibly indicating up to two AR lags, though this could be influenced by seasonal effects. It is possible that the model may have up to two AR lags or potentially no AR lags at all.

To build an ARIMA model, we will fit four different ARIMA models and compare their performance. The first three models will be based on initial assumptions with increasing complexity, while the fourth model will be automatically selected by R.

The first model includes only non-seasonal and seasonal differencing components set to order 1, specified as ARIMA(0,1,0)(0,1,0). The second model adds both a non-seasonal and a seasonal MA term at order 1, while maintaining non-seasonal and seasonal differencing at order 1, specified as ARIMA(0,1,1)(1,1,0). The third model also sets non-seasonal and seasonal differencing to order 1 but adds a non-seasonal AR term of order 1, a non-seasonal MA term of order 1, and a seasonal AR term of order 1, specified as ARIMA(1,1,1)(1,1,0).

```
# Set a random seed for reproducibility
set.seed(123)

arima_fit <- co2_tsib %>%
          model(
          model_1 = ARIMA(log_value~0+pdq(0,1,0)+PDQ(0,1,0)),
          model_2 = ARIMA(log_value~0+pdq(0,1,1)+PDQ(1,1,0)),
          model_3 = ARIMA(log_value~0+pdq(1,1,1)+PDQ(1,1,0)),
          model_4 = ARIMA(log_value),
          best_model = ARIMA(log_value~0+pdq(0,1,0)+PDQ(0,1,1))
          )
```

```
## Series: log_value
## Model: ARIMA(0,1,0)(0,1,0)[12]
##
## sigma^2 estimated as 2.318e-06:  log likelihood=2425.3
## AIC=-4848.59   AICc=-4848.58   BIC=-4844.47
```

```
## Series: log_value
## Model: ARIMA(0,1,1)(1,1,0)[12]
##
## Coefficients:
##           ma1     sar1
##       -0.3715  -0.4516
## s.e.   0.0476   0.0428
##
## sigma^2 estimated as 1.939e-06:  log likelihood=2499.41
## AIC=-4992.82   AICc=-4992.77   BIC=-4980.46
```

```
## Series: log_value
## Model: ARIMA(1,1,1)(1,1,0)[12]
##
## Coefficients:
##          ar1      ma1     sar1
##       0.1668  -0.5225  -0.4482
## s.e.  0.1342   0.1186   0.0430
##
## sigma^2 estimated as 1.94e-06:  log likelihood=2500.05
## AIC=-4992.1   AICc=-4992.01   BIC=-4975.62
```

```
## Series: log_value
## Model: ARIMA(0,1,3)(1,1,0)[12]
##
## Coefficients:
##           ma1     ma2      ma3     sar1
##       -0.3512  0.0026  -0.1136  -0.4434
## s.e.   0.0469  0.0476   0.0470   0.0431
##
## sigma^2 estimated as 1.934e-06:  log likelihood=2502.62
## AIC=-4995.24   AICc=-4995.11   BIC=-4974.64
```

The ARIMA(0,1,0)(0,1,0) model is the simplest and performs the worst, with the highest AIC and BIC values. The ARIMA(0,1,1)(1,1,0) model shows improvement, with AIC and BIC values approximately 144 points lower than those of model 1. The ARIMA(1,1,1)(1,1,0) model has nearly identical AIC and BIC values to model 2, indicating that the non-seasonal AR(1) term does not significantly improve the fit. The ARIMA(0,1,3)(1,1,0) model has the lowest AIC and BIC values, indicating a slightly better fit than the second and third model. However, the improvement over model 2 and 3 is minimal.

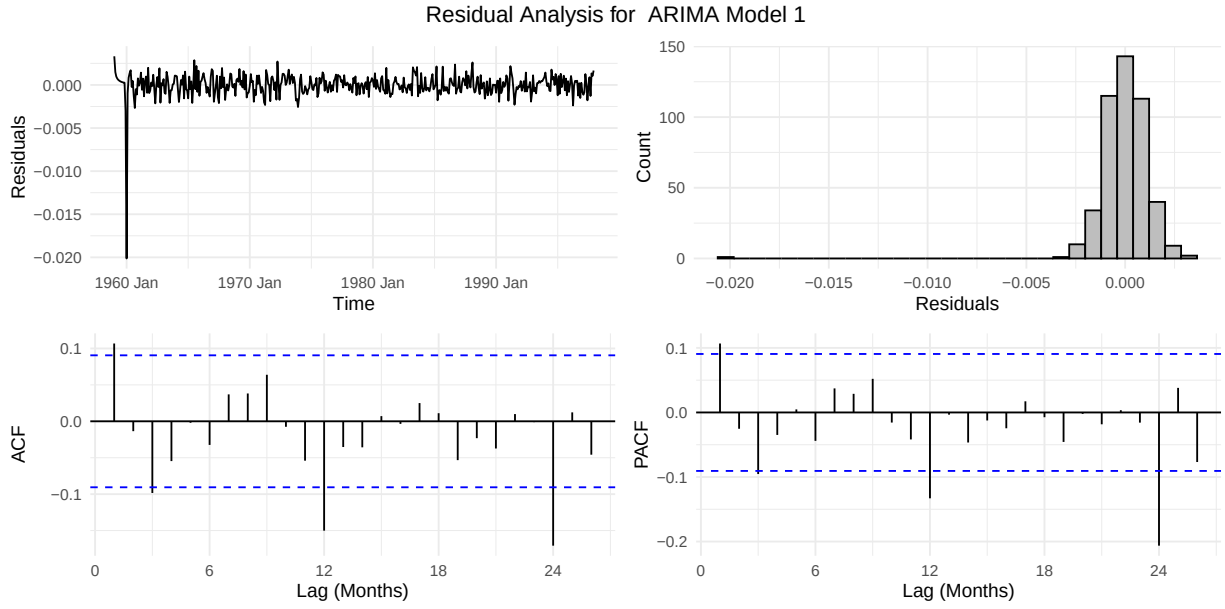
Let's perform a Ljung-Box tests to assess if residuals from a the time series models are independently distributed

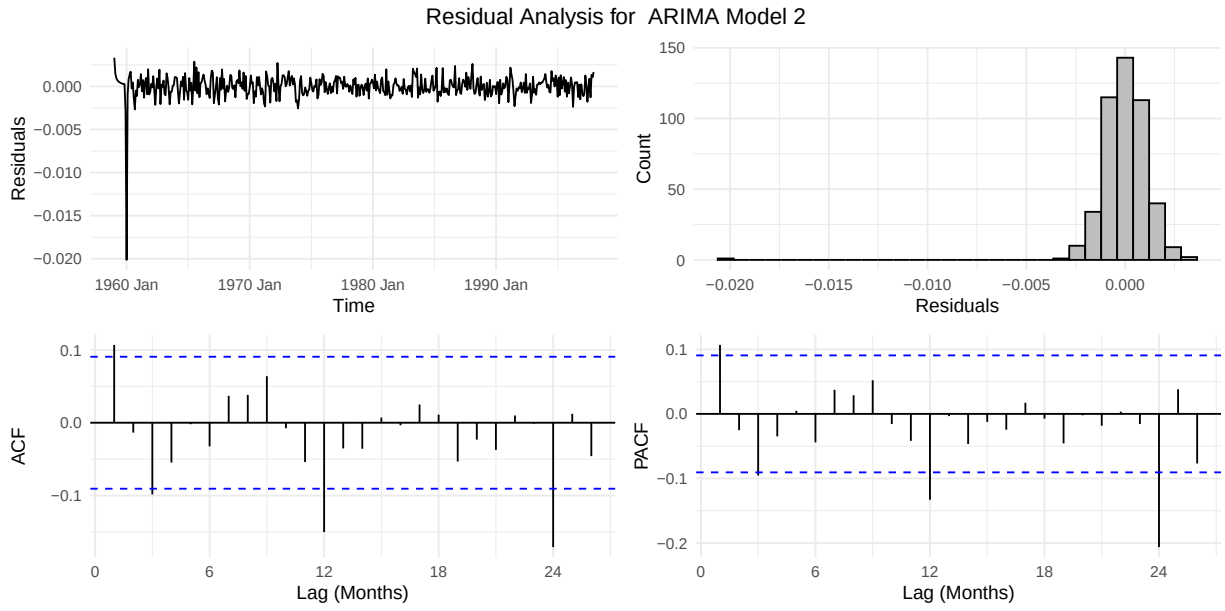
Table 3: Ljung-Box Test Results for ARIMA Models

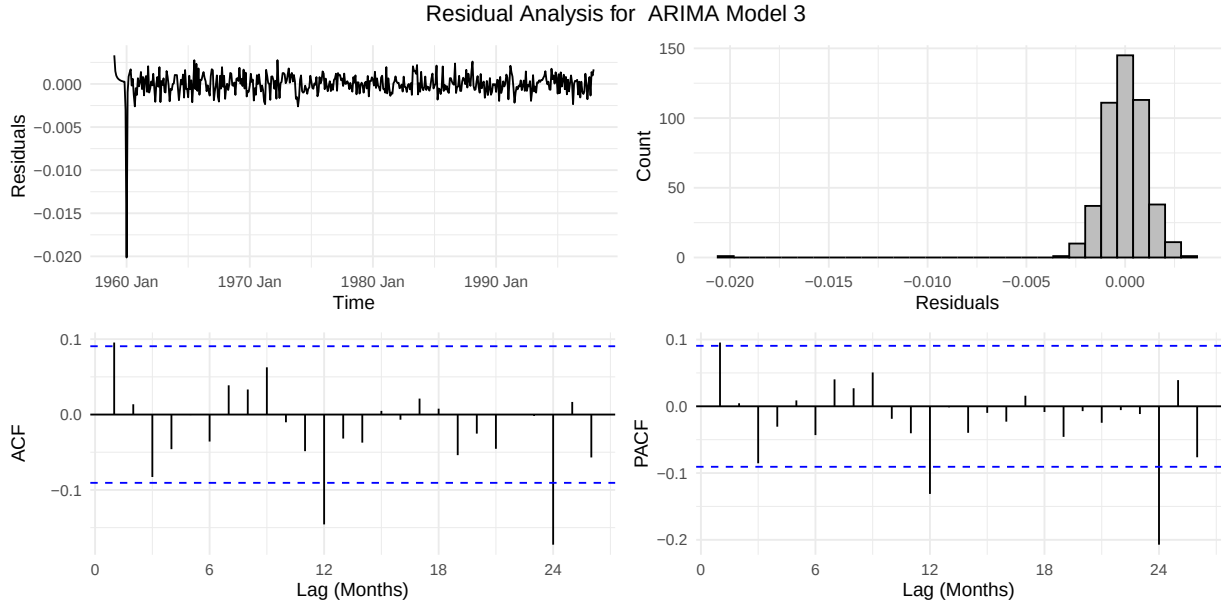
Ljung-Box p-value	Model
0.0000022	ARIMA Model 1
0.0019529	ARIMA Model 2
0.0033722	ARIMA Model 3
0.0066913	ARIMA Model 4

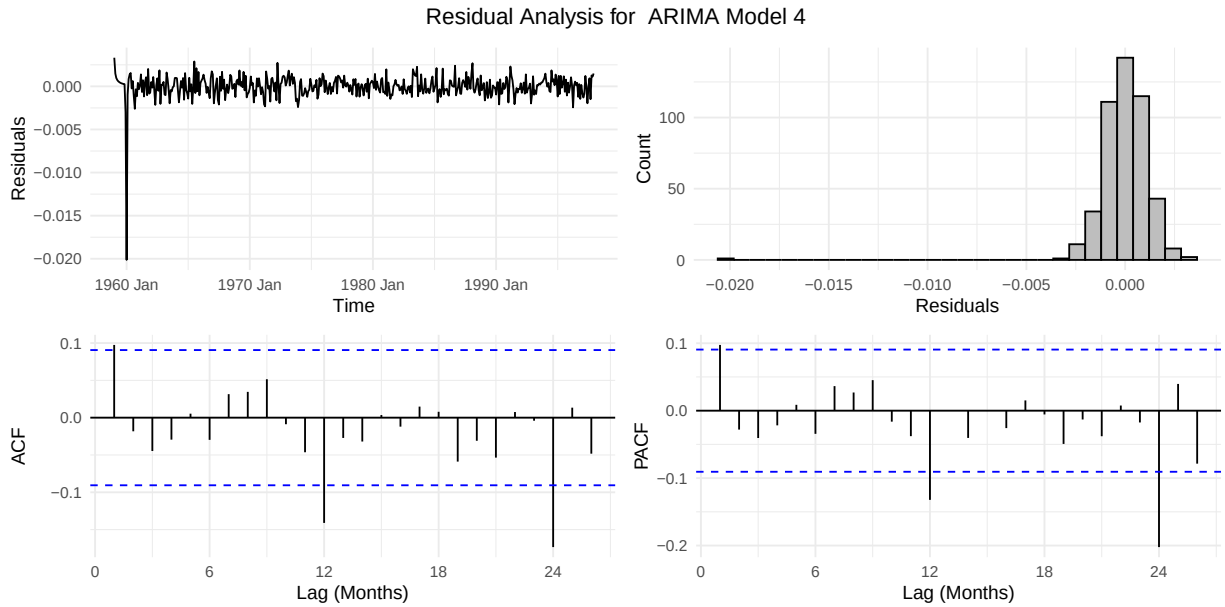
Based on the Ljung-Box test, the p-value is less than 5% for all models, and we reject the null hypothesis of no serial correlation.

The residual plots for each model were examined.




Residual Analysis for ARIMA Model 2
Residuals
0.000
−0.005
−0.010
−0.015
−0.020
1960 Jan
1970 Jan
1980 Jan
1990 Jan
Time
Count
150
100
50
0
−0.020
−0.015
−0.010
−0.005
0.000
Residuals
ACF
0.1
0.0
−0.1
0
6
12
18
24
Lag (Months)
PACF
0.1
0.0
−0.1
−0.2
0
6
12
18
24
Lag (Months)


Residual Analysis for ARIMA Model 3
Residuals
0.000
−0.005
−0.010
−0.015
−0.020
1960 Jan
1970 Jan
1980 Jan
1990 Jan
Time
Count
150
100
50
0
−0.020
−0.015
−0.010
−0.005
0.000
Residuals
ACF
0.1
0.0
−0.1
0
6
12
18
24
Lag (Months)
PACF
0.1
0.0
−0.1
−0.2
0
6
12
18
24
Lag (Months)



Based on the plots, there is some type of outlier present in the residuals for all models in January 1960. The residual on January 1960 has a value of '-2.013132e-02,' approximately '-0.02011184' lower than the residual mean. Adjusting parameters by adding AR and MA terms did not remove this anomaly. Examination of both raw and transformed data revealed no errors. This high residual may indicate a unique shift in the CO_2 pattern in January 1960 that the models fail to capture.

Since the Ljung-Box test and residual plots indicated that the residuals do not resemble white noise, we need to adjust the ARIMA model to include some MA terms to include this serial correlation in our model.

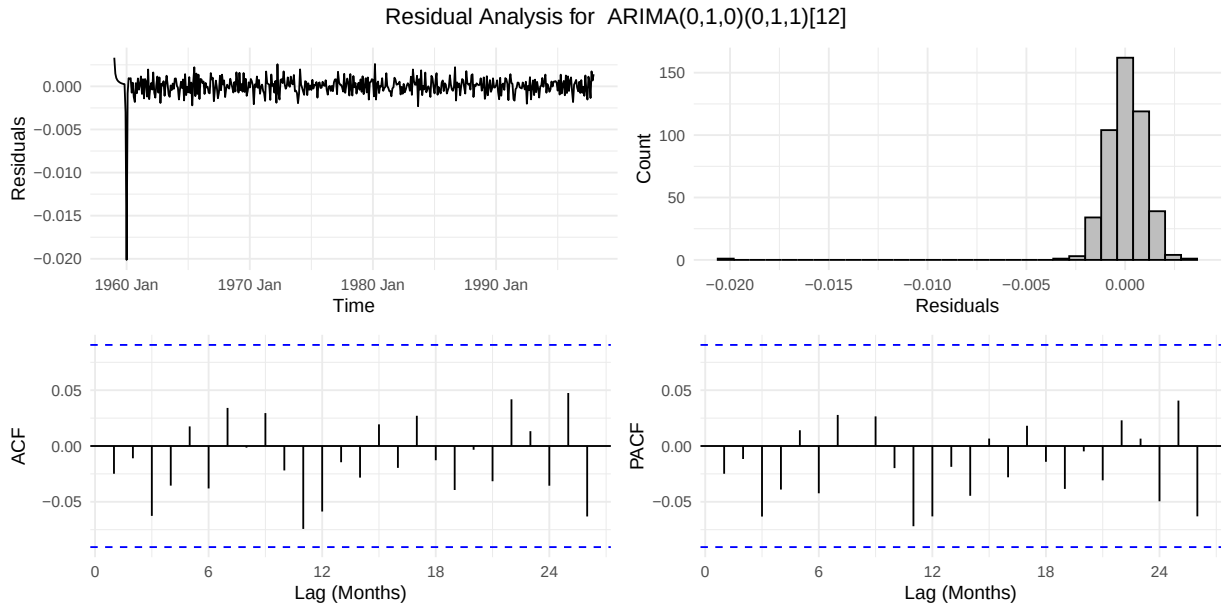
After testing various parameters, the best ARIMA model, with the fewest parameters, lower AIC and BIC values than previous models, and residuals resembling white noise, is:

$$\text{ARIMA}(0,1,0) \times (0,1,1)_{12}$$

```
## Series: log_value
## Model: ARIMA(0,1,0)(0,1,1)[12]
##
## 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
```

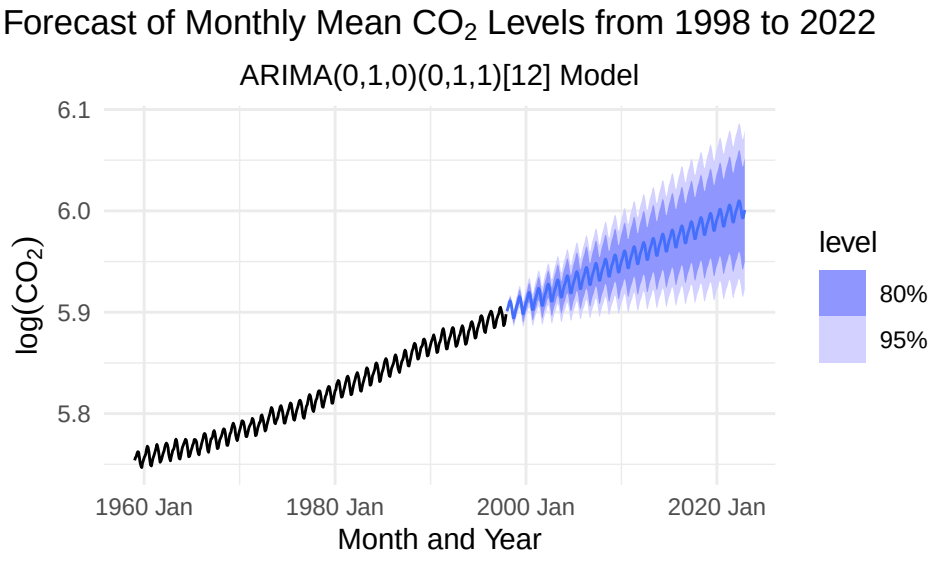
```
## # A tibble: 1 x 2
##   lb_pvalue Model
##       <dbl> <chr>
## 1     0.945 Best ARIMA Model
```

Based on the Ljung-Box test, the p-value is greater than 5%, and we fail to reject the null hypothesis of no serial correlation.



The residual analysis for the ARIMA(0,1,0)(0,1,1)[12] model indicates that the residuals are close to white noise. The time plot shows residuals centered around zero, with no obvious patterns, suggesting that the model captures most of the trend and seasonality in the data. The histogram of residuals is roughly symmetric, indicating a normal distribution. The ACF and PACF plots reveal no significant autocorrelation at different lags, with all values remaining within the significance bounds.

The following plot displays the forecasted monthly mean CO_2 levels from 1998 to 2022 using an ARIMA(0,1,0)(0,1,1)[12] model.



The forecast follows the observed data trend, capturing both the seasonal fluctuations and the general upward trajectory in CO_2 concentration. The shaded areas represent the 80% and 95% confidence intervals, which widen over time, indicating increasing uncertainty in the forecast as it extends further into the future. This suggests that while the model is effective at capturing the trend and seasonality, the precision of long-term predictions decreases.

Forecasting Atmospheric CO2 Growth

Using our ARIMA model, we can generate predictions for when atmospheric CO_2 is expected to be at 420 ppm and 500 ppm levels. To estimate the earliest date (first time) these thresholds are reached, we examine the upper bound of the 95% confidence interval. Likewise, we can use the the lower bound of the 95% confidence interval to find the latest date when CO_2 levels reach 420 ppm and 500 ppm.Additionally, we can pinpoint the first date when the predicted average atmospheric CO_2 level is expected to reach 420 ppm and 500 ppm.

The table below provides the earliest forecasted dates, or "first times," when atmospheric CO_2 levels are projected to reach 420 ppm and 500 ppm. The estimates include the mean forecast and the 95% confidence interval. The model indicates, with 95% confidence, that the earliest date when the CO_2 levels are likely to reach 420 ppm is by April 2016, while for 500 ppm, it is anticipated around April 2040.

Table 4: Forecast Summary with 95% Confidence Intervals

Level	First_Time	First_Mean_Estimate	First_Lower_95	First_Upper_95
420 ppm	2016 Apr	396.8926	374.2075	420.9529
500 ppm	2040 Apr	438.0234	382.7225	501.3149

The table below presents the the latest date, or "last times" when atmospheric CO_2 levels are expected to exceed 420 ppm and 500 ppm. However, due to increasing variance over time, the 95% lower confidence interval begins to decline, making it impossible to provide a definitive "last time" within the forecasted period. This uncertainty indicates that while levels are projected to cross 420 ppm and 500 ppm, the predictions suggest that these levels might not reach those thresholds until sometime after 2100.

Table 5: Forecast Summary with 95% Confidence Intervals for Last Observations

Level	Last_Time	Last_Mean_Estimate	Last_Lower_95	Last_Upper_95
420 ppm	>2100	NA	NA	NA
500 ppm	>2100	NA	NA	NA

The table below provides the projected dates when the mean atmospheric CO_2 levels are expected to reach 420 ppm and 500 ppm. These estimates include the mean forecasted value along with the 95% confidence interval. According to the model, CO_2 levels are likely to reach 420 ppm by April 2030, with a 95% confidence interval ranging from 379.91 to 465.19 ppm. Levels are expected to reach 500 ppm by May 2072, with a wider confidence interval ranging from 385.46 to 649.65 ppm due to increased uncertainty over the longer forecasting horizon.

Table 6: Forecast Summary with 95% Confidence Intervals for Mean Estimates

Level	Time	Mean_Estimate	Lower_95	Upper_95
420 ppm	2030 Apr	420.39	379.91	465.19
500 ppm	2072 May	500.41	385.46	649.65

We can also project CO_2 levels for the year 2100. The table below displays the estimated average CO_2 concentration in ppm, along with the corresponding 95% confidence interval.

Table 7: Forecast for CO2 Levels in 2100 with 95% Confidence Intervals

Mean Estimate (ppm)	Lower 95% CI (ppm)	Upper 95% CI (ppm)
556.73	376.58	820

The estimated mean CO_2 concentration for the year 2100. The 95% confidence interval ranges from 376.58 ppm to 820.00 ppm. This wide interval highlights the increasing variability and uncertainty in projections as we move further into the future. For that reason, we are uncertain about the accuracy of previous predictions due to the long time range involved. However, if our predictions were constrained to a shorter timeframe (such as within a year or two after the last observed data point in 1997) the model would likely yield more accurate and reliable forecasts, as shorter-term predictions reduce the accumulation of uncertainties and better reflect the observed data trends.

Shifting to the Present Point of View

Now, we turn our attention to the modern Mauna Loa CO_2 dataset, and attempt to complete the following:

1. Evaluate the performance of our best-performing linear and ARIMA models trained on 1997 data to determine accuracy and generalizability of time series models on Mauna Loa CO_2 data.
2. Train and evaluate new models using a longer period of training data (up until 2022), and determine if a better model can be fit with the additional information.
3. Perform new predictions on when we expect to see CO_2 levels of 420 ppm and 500 ppm for the first and last time, and generate a long-term prediction for CO_2 levels in 2122.

Initial EDA of more recent data

In this section of the report, we aim to evaluate the performance of our linear and ARIMA models derived from 1997 data on known, observed data up until 2024. Then, with the new data period, we again attempt to find and fit the best model for predicting CO_2 in the future. Since 1997, in addition to sporadic data outages due to equipment malfunction or minor data adjustments due to data collection errors, two major scale changes were implemented - first in 2017 then again in 2021 to the latest X2019 scale. These scale changes are retroactively applied to all data, meaning that pre-1997 data is also changed to a varying degree. Finally, the modern CO_2 data used in this section has weekly granularity, as opposed to the monthly cadence from before.

Create a modern data pipeline

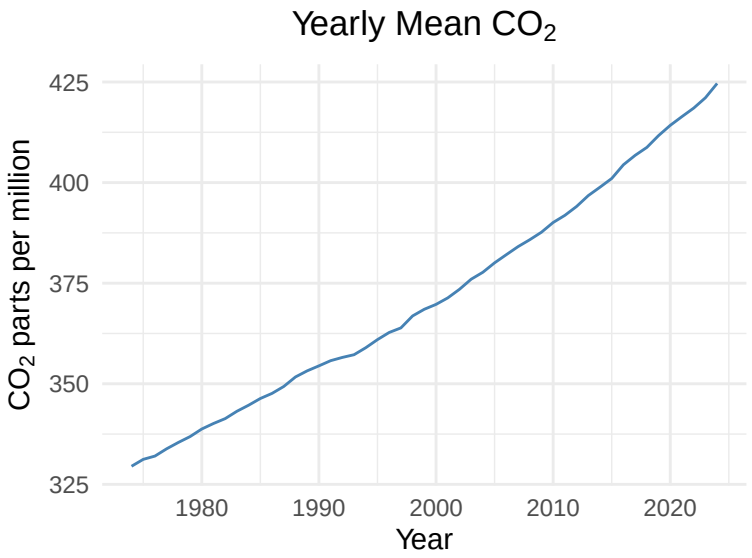
Given the improved capabilities of recording CO_2 data we are now able to obtain the latest CO_2 data from the NOAA Global Monitoring Library, we directly imported from noaa.gov using a permalink to its latest weekly Mauna Loa CO_2 data release.

Table 8: Weekly CO2 Levels at Mauna Loa

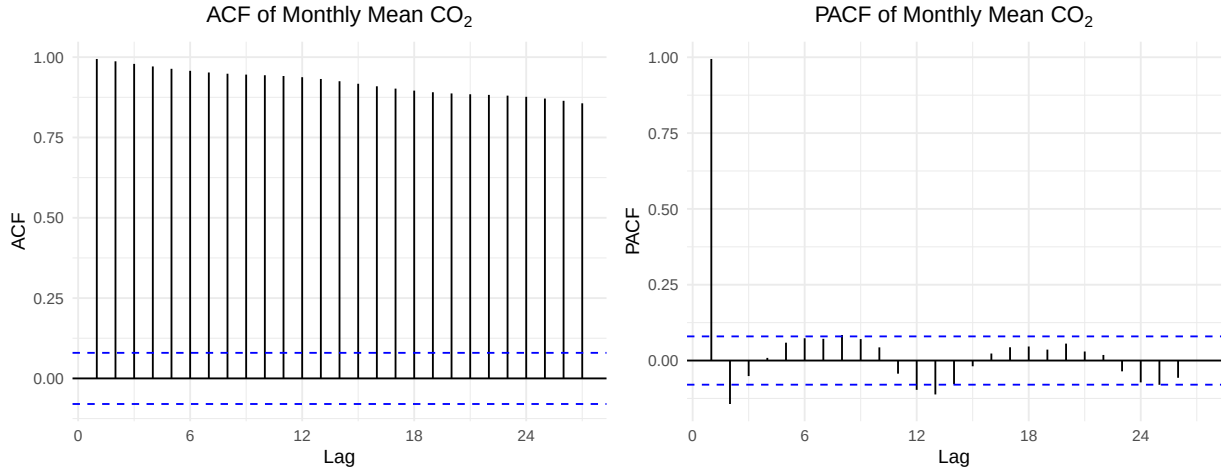
year	month	day	decimal	average	ndays	1 year ago	10 years ago	increase since 1800	date
1974	5	19	1974.380	333.37	5	-999.99	-999.99	50.40	1974-05-19
1974	5	26	1974.399	332.95	6	-999.99	-999.99	50.06	1974-05-26
1974	6	2	1974.418	332.35	5	-999.99	-999.99	49.60	1974-06-02

year	month	day	decimal	average	ndays	1 year ago	10 years ago	increase since 1800	date
1974	6	9	1974.437	332.20	7	-999.99	-999.99	49.65	1974-06-09
1974	6	16	1974.456	332.37	7	-999.99	-999.99	50.06	1974-06-16
1974	6	23	1974.475	331.73	5	-999.99	-999.99	49.72	1974-06-23

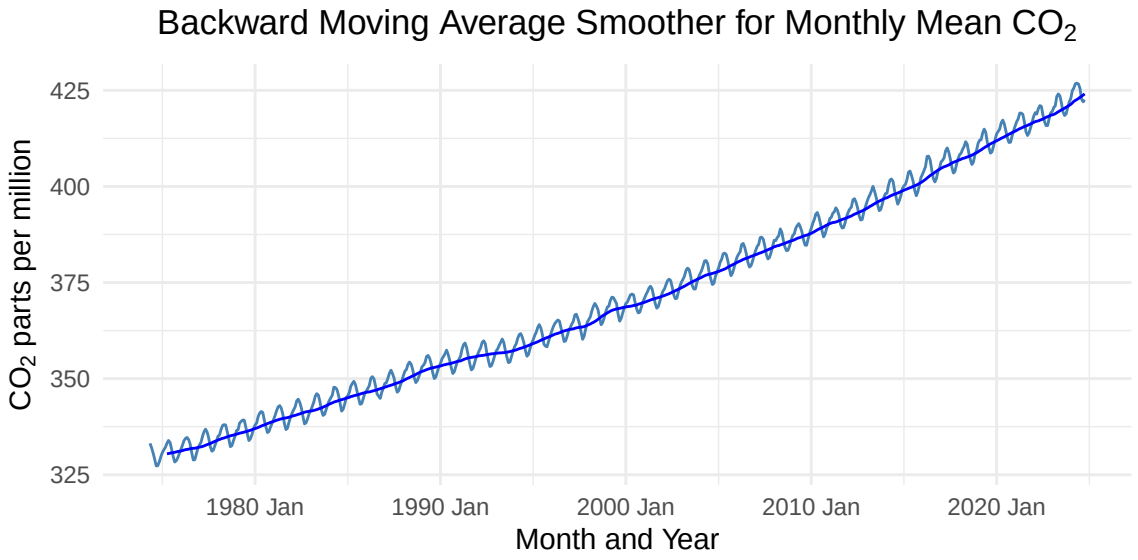
We conducted a similar EDA on this updated data and found the following trends:



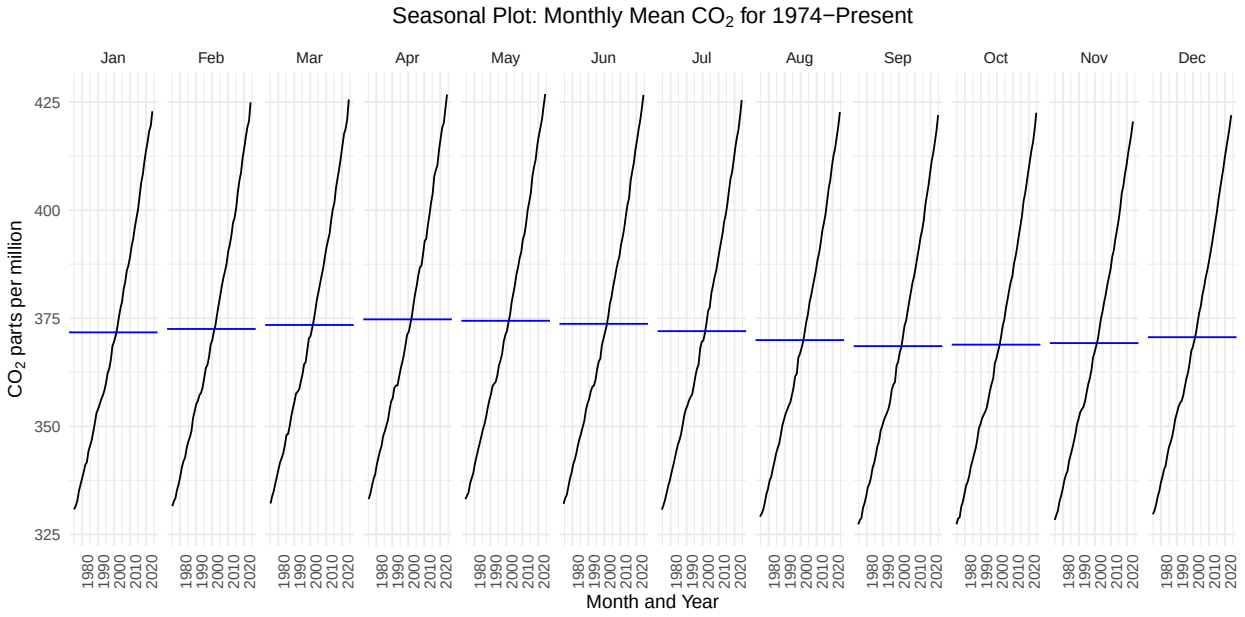
From a macro perspective, it appears that the yearly average CO_2 is continuing its upward trend, with the rate of growth slowly increasing as well in recent years.



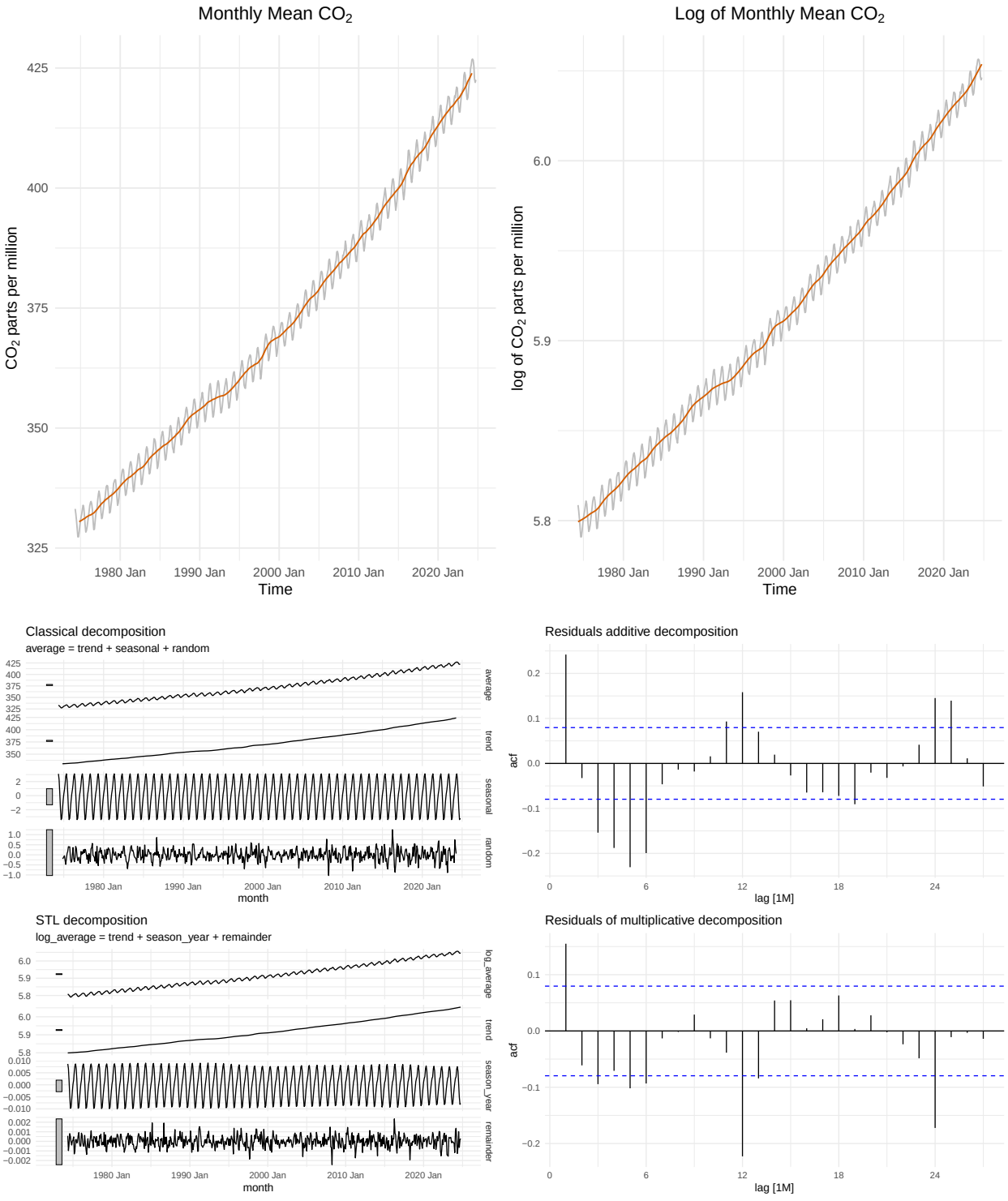
The ACF and PACF charts here reaffirm our findings in the first section of the report, again depicting strong evidence for a trend in the data and an autoregressive process.



The moving average line, coupled with the time series plot, demonstrate a continuing trend from previous data in a cyclical CO_2 pattern throughout the year, with a steady underlying increasing trend that appears to depict a slowly rising slope.



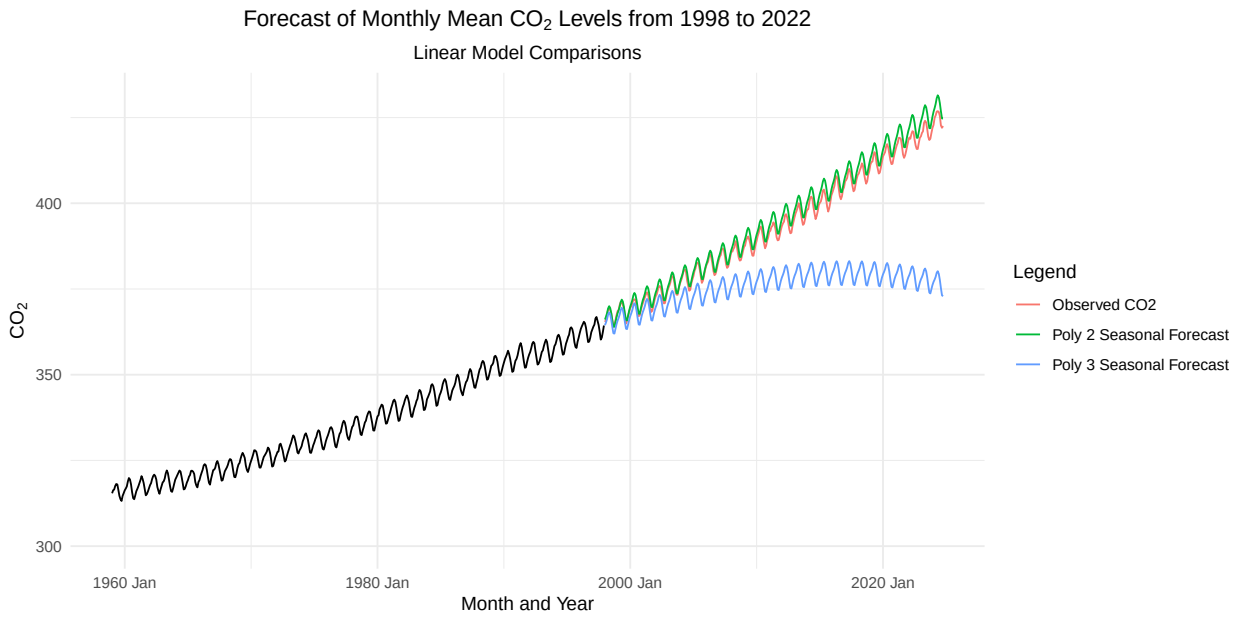
Isolating months of the year to account for seasonality, we can clearly see the aforementioned increasing trend that has continued to persist from the previous section, and the slowly increasing slope.



The decomposition charts clearly show the cyclical seasonality and increasing trend over time of CO_2 data.

Compare linear model forecasts against realized CO2

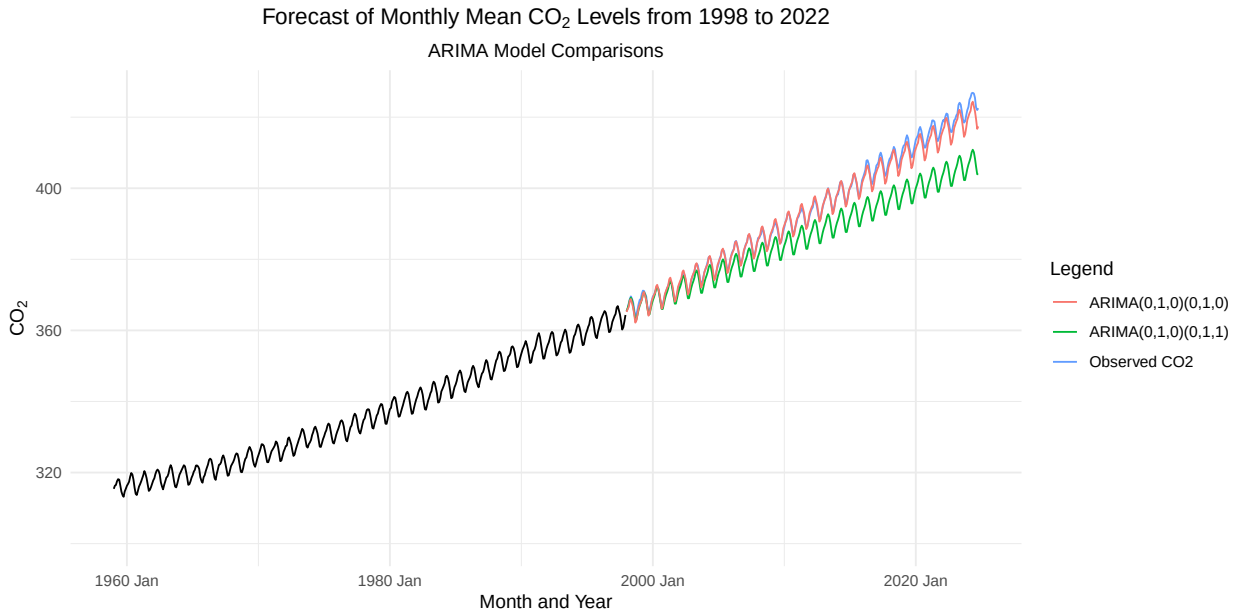
First, we evaluate the simple linear and polynomial models trained using 1997 data on present observed data to evaluate their efficacy.



Comparing the linear models that we attempted in the first section, we observe that the Poly 2 Seasonal model actually performs the best when evaluated against our true, observed data from 1998 onwards. Previously, we found that the Poly 3 Seasonal model had the best performance when only observing data up until 1998. This shows that the alarming growth of CO_2 is outpacing expectations by even the best predictions in 1998.

Compare ARIMA models forecasts against realized CO2

Next, we consider the ARIMA models trained on 1997 data.



Again, we observe that the best model selected in the previous section of the report is not the best-performing model when compared against true observed data. Here, the $ARIMA(0,1,0)(0,1,0)$ model actually outperforms the more complex $ARIMA(0,1,0)(0,1,1)$ model. The discrepancy highlights that simpler structures may better capture underlying trends in this case.

Evaluate the performance of 1997 linear and ARIMA models

We previously predicted that CO_2 levels would surpass 420ppm in 2030, on average, with it occuring as soon as 2016 or as late as beyond 2100 around 95% of the time. In reality, we see that CO_2 levels have already been observed to surpass 420 ppm in 2022, towards the early end of that scale.

Table 9: Accuracy Comparison of Linear Models

Model	ME	RMSE	MAE	MPE	MAPE	ACF1
Linear	13.1411177	15.0957095	13.1411177	3.2671290	3.2671290	0.9776995
Log-Linear	0.0268173	0.0306366	0.0268290	0.4472125	0.4474115	0.9688562
Quadratic	0.1652185	2.3079722	1.9530049	0.0331849	0.4962282	0.8400328
Log-Quadratic	-0.0046458	0.0076443	0.0061562	-0.0776005	0.1029947	0.8470648
Poly 2 with Season	-0.0049783	0.0056780	0.0050267	-0.0830763	0.0838963	0.9065832
Poly 3 with Season	0.0451137	0.0567314	0.0451137	0.7511693	0.7511693	0.9886045
Poly 4 with Season	0.0735580	0.0950903	0.0735580	1.2243355	1.2243355	0.9888532

Table 10: Accuracy Comparison of ARIMA Models

Model	ME	RMSE	MAE	MPE	MAPE	ACF1
ARIMA(0,1,0)(0,1,1)	0.0173361	0.0211135	0.0173489	0.2887905	0.2890069	0.9839047
ARIMA(0,1,0)(0,1,0)	0.0017088	0.0034547	0.0025353	0.0283778	0.0422602	0.8418235
ARIMA(0,1,1)(1,1,0)	0.0108738	0.0132751	0.0108845	0.1811571	0.1813383	0.9694969
ARIMA(1,1,1)(1,1,0)	0.0126900	0.0153098	0.0126907	0.2114428	0.2114540	0.9735963
ARIMA(0,1,3)(1,1,0)	0.0141811	0.0169895	0.0141815	0.2363050	0.2363128	0.9760387

Here, using the accuracy chart, we can see a more concrete representation of the best linear and ARIMA models. For linear models, the Poly 2 model performed the best (as opposed to the Poly 3 model in the previous section). For ARIMA models, the $ARIMA(0,1,0)(0,1,0)$ model performed the best (as opposed to the $ARIMA(0,1,0)(0,1,1)$ model in the previous section).

Train best models on present data

We then proceed to create two versions of our modern CO_2 data - one adjusted for seasonality and one without the adjustment. For both versions of the data, we perform a train/test split and train new ARIMA models.

Table 11: Optimal ARIMA Models

Model	Parameters
ARIMA (NSA)	<ARIMA(0,1,4) w/ drift>
ARIMA (SA)	<ARIMA(1,1,5)>

Table 12: Model Accuracy Comparison

Model	ME	RMSE	MAE	MPE	MAPE	ACF1
ARIMA (NSA, In-Sample)	0.0000008	0.0068748	0.0024066	-0.0001020	0.0408001	0.0019715
ARIMA (SA, In-Sample)	0.0006486	0.0070459	0.0027668	0.0108393	0.0469349	-0.0091240

Model	ME	RMSE	MAE	MPE	MAPE	ACF1
ARIMA (NSA, Out-of-Sample)	0.0097573	0.0111894	0.0097573	0.1613020	0.1613020	0.9208802
ARIMA (SA, Out-of-Sample)	0.0125555	0.0141959	0.0125859	0.2077661	0.2082709	0.8997722
Polynomial Trend (In-Sample)	0.0000000	0.0097243	0.0060202	-0.0002736	0.1020717	0.6849691
Polynomial Trend (Out-of-Sample)	0.0004418	0.0055330	0.0046832	0.0072279	0.0775276	0.8658350

The learned best parameters are ARIMA(0,1,4) w/ drift for the non-seasonally adjusted data and ARIMA(1,1,5) for the seasonally adjusted data. The polynomial trend model appears to be relatively equal in performance to the ARIMA models.

How bad could it get?

Armed with the models trained on the latest-available data, we again perform the prediction tasks to determine when we expect CO_2 to surpass 420 ppm and 500 ppm for the first time, and when we expect to last see a CO_2 level under 420 ppm and 500 ppm respectively.

Table 13: CO2 Threshold Forecasts

Level	First Time Date	Last Time Date
420 ppm	2022-10-30	2092-07-13
500 ppm	2034-07-09	2122-10-25

Here, we can see that the expected time for CO_2 levels to hit 420ppm is late 2022, which is a lot closer to reality. All other dates appear to have pushed forward from the 1997 estimate as well, with the date that we last observe a CO_2 level under 420ppm already within scope in 2092. However, since the Last Time Date for 500ppm above is the last date of our predicted range, we can say that the date where we last observe a CO_2 level under 500ppm is still further out than 2122.

Finally, examining the longer-term predictions for the year 2122, one hundred years out from the last date in the training data in our model, we obtain the following conclusions:

Table 14: Forecasted CO2 Levels in 2122

Week	Mean Estimate (ppm)	Lower 95% CI (ppm)	Upper 95% CI (ppm)
2122-01-04	658.6951	453.1971	957.3743
2122-01-11	658.7535	453.1971	957.3743
2122-01-18	658.8119	453.1971	957.3743
2122-01-25	658.8704	453.1971	957.3743
2122-02-01	658.9288	453.1971	957.3743
2122-02-08	658.9873	453.1971	957.3743
2122-02-15	659.0457	453.1971	957.3743
2122-02-22	659.1042	453.1971	957.3743
2122-03-01	659.1626	453.1971	957.3743
2122-03-08	659.2211	453.1971	957.3743
2122-03-15	659.2795	453.1971	957.3743
2122-03-22	659.3380	453.1971	957.3743
2122-03-29	659.3965	453.1971	957.3743
2122-04-05	659.4550	453.1971	957.3743
2122-04-12	659.5135	453.1971	957.3743
2122-04-19	659.5720	453.1971	957.3743

Week	Mean Estimate (ppm)	Lower 95% CI (ppm)	Upper 95% CI (ppm)
2122-04-26	659.6305	453.1971	957.3743
2122-05-03	659.6890	453.1971	957.3743
2122-05-10	659.7475	453.1971	957.3743
2122-05-17	659.8060	453.1971	957.3743
2122-05-24	659.8645	453.1971	957.3743
2122-05-31	659.9230	453.1971	957.3743
2122-06-07	659.9816	453.1971	957.3743
2122-06-14	660.0401	453.1971	957.3743
2122-06-21	660.0987	453.1971	957.3743
2122-06-28	660.1572	453.1971	957.3743
2122-07-05	660.2158	453.1971	957.3743
2122-07-12	660.2743	453.1971	957.3743
2122-07-19	660.3329	453.1971	957.3743
2122-07-26	660.3914	453.1971	957.3743
2122-08-02	660.4500	453.1971	957.3743
2122-08-09	660.5086	453.1971	957.3743
2122-08-16	660.5672	453.1971	957.3743
2122-08-23	660.6258	453.1971	957.3743
2122-08-30	660.6844	453.1971	957.3743
2122-09-06	660.7429	453.1971	957.3743
2122-09-13	660.8016	453.1971	957.3743
2122-09-20	660.8602	453.1971	957.3743
2122-09-27	660.9188	453.1971	957.3743
2122-10-04	660.9774	453.1971	957.3743
2122-10-11	661.0360	453.1971	957.3743
2122-10-18	661.0946	453.1971	957.3743
2122-10-25	661.1533	453.1971	957.3743

Table 15: Mean CO2 Forecast for 2122

Mean Estimate (ppm)	Lower 95% CI (ppm)	Upper 95% CI (ppm)
659.923	453.1971	957.3743

It is difficult to have high confidence in these predictions due to a couple of reasons. Firstly, due to the significant time gap between our last observed data and the predicted time range, there could be a lot of potential deviation, trends, outliers, or other external influencers that can cause a deviation that our current model cannot capture. As we have seen in the past, conflicting interests between increasing global economic output at the expense of increased CO_2 emissions, and the various efforts to curb these emissions and protect the environment can cause significant changes to established trends. Secondly, also due to the significant time gap, the confidence interval is signficantly expanded, creating a large range of approximately 400pm and thus diminishing the tangible value of the predicted range results. Furthermore, it is difficult to factor in the improvement in technology that would take place in the next 100 years, as technology is currently rapidly evolving requiring higher resources like in the case with GPUs.