W271 Group Lab

Bike share demand

Annie DeForge, Hannah Abraham, Mariah Ehmke, Nora Povejsil

Contents

0.1 Introduction

The most recent UN Climate Conference known as COP28, resulted in international call to limit greenhouse gas emissions to limit overall planetary global warming to 1.5 degrees Celsius. The plans rely on reducing greenhouse gas emissions from anthropogenic activities and increasing the uptake of carbon dioxide (CO2) from the atmosphere into natural storage systems, such as forests and soils (United Nations Climate Conference, 2023). Global levels of CO2 have more than doubled around the world compared to pre-industrialization levels (Jiménez-de-la-Cuesta and Mauritsen, 2019). This doubling has occurred despite seasonal growth and contractions in CO2 levels, particularly in areas with forests.

Seasonal variation in CO2 levels reflects the terrestrial cycles related to plant growth and decomposition. Keeling (1960) showed atmospheric local seasonal CO2 levels in Mauna Loa, Hawaii peaked right before a new growing season, steadily decreased as plants absorbed CO2 during the growing season, and reached a low at the end of the growing season. Despite these seasonal adjustments, baseline CO2 levels in Hawaii and throughout the world have grown over time. The growth reflects increased emissions from industrial, agricultural, and transportation systems. The goal of the COP28 is to contain CO2 levels through improved forestry and biodiversity management to absorb more CO2 from the atmosphere.

The objective of this analysis is to measure the trend in atmospheric CO2 levels in Mauna Loa, Hawaii, controlling for seasonal fluctuations. We test the following null hypotheses:

H01: Atmospheric CO2 concentrations at Mauna Loa, Hawaii follow a linear trend from 1957 to 2020. HA1: Atmospheric CO2 concentrations at Mauna Loa, Hawaii follow a non-linear trend from 1957 to 2020.

If the null hypothesis holds, it will provide insight into the rate at which atmospheric CO2 levels would decrease given reductions in CO2 emissions from industrial, agricultural, and transportation technologies. If it the trend is linear, we would expect a proportional decrease in CO2 levels from emission reductions following a reverse linear trend. If it is not linear, then CO2 levels will respond differently to emission caps. We will explore possible growth paths if the null hypothesis is rejected.

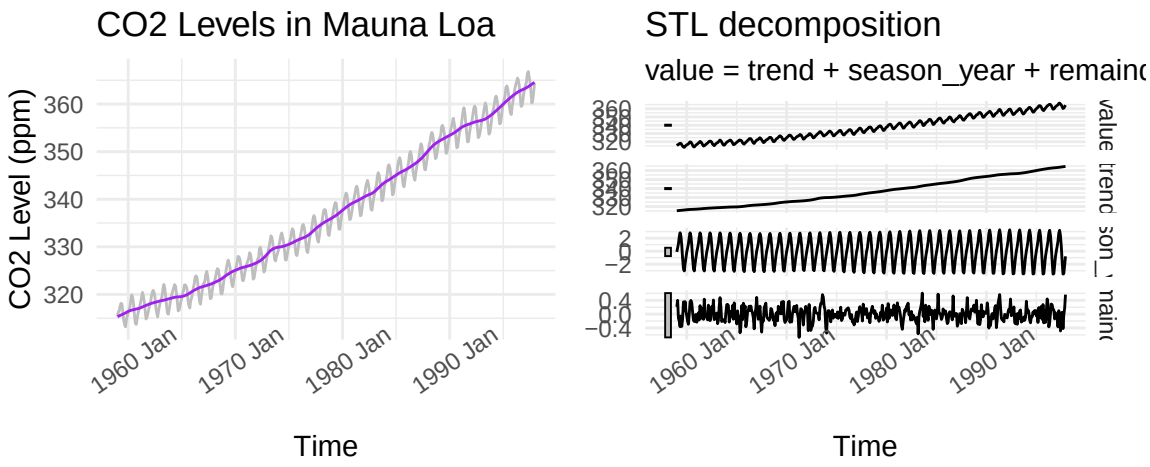
0.2 CO2 data

The data were collected at the Mauna Loa, Hawaii Observatory near the Mauna Loa Volcano in Hawaii. The amount of CO_2 is reported as the 'mole fraction' or number of carbon dioxide molecules present in a given number of molecules of air (National Oceanic and Atmospheric Administration, 2023). A CO_2 level of 400 indicates there are 400 parts per million (ppm) CO_2 molecules in every million molecules of dry air. The data collection began in 1957 by Dave Keeling (Keeling, 1960). The data are continuosly collected. We present data in part a that were monthly averages of CO_2 levels from 1957 to 1997. In part b, we extend our analysis using weekly averaged data from 1997 to the present.

0.2.1 Exploratory Data Analysis

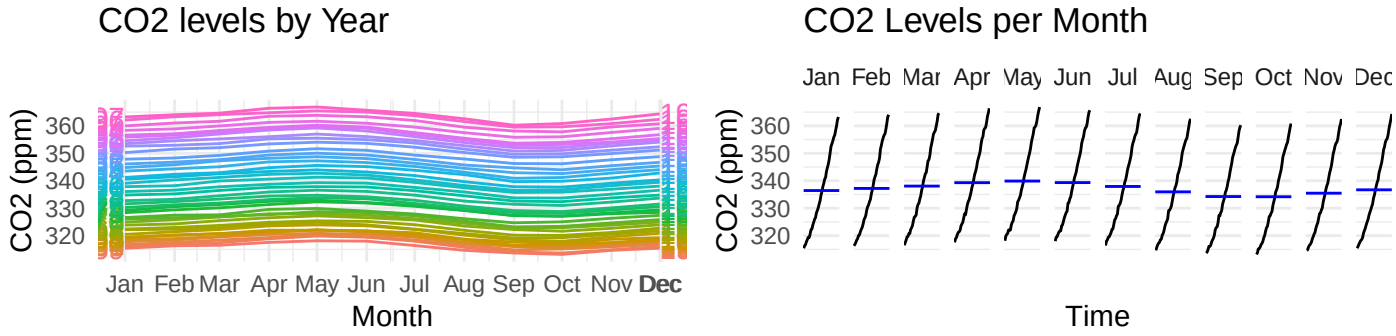
The exploratory data analysis includes a linear plot of the raw data and the trend-cycle components, and the additive components of the time series (i.e., trend, seasonality, and the remainder). According to Keeling (YEAR), the data follow a seasonal, cyclical pattern. Each year, CO2 levels are lowest at the end of the growing season following plant growth and CO2 absorption. Then, the CO2 levels peak at the start of the growing season, before plants have begun absorbing excess CO2 from the atmosphere. The raw data follow this pattern, but with an upward trend, illustrated by the purple line in Figure 1.

We employ seasonal, trend, and remainder (STL) decomposition to parse out the role of seasonality, trends in the average across time, and remainder components on the time series in the four plots on the right-hand side of Figure 1. The trend is semi-linear and upward sloping. The yearly average is increasing over time. The seasonal variation appears to expand across time as the height and, especially the depth, of the seasonal line plot increases after the mid-1970s. The bottom right-hand line plot of remainder variation does not have a discernible visual pattern.

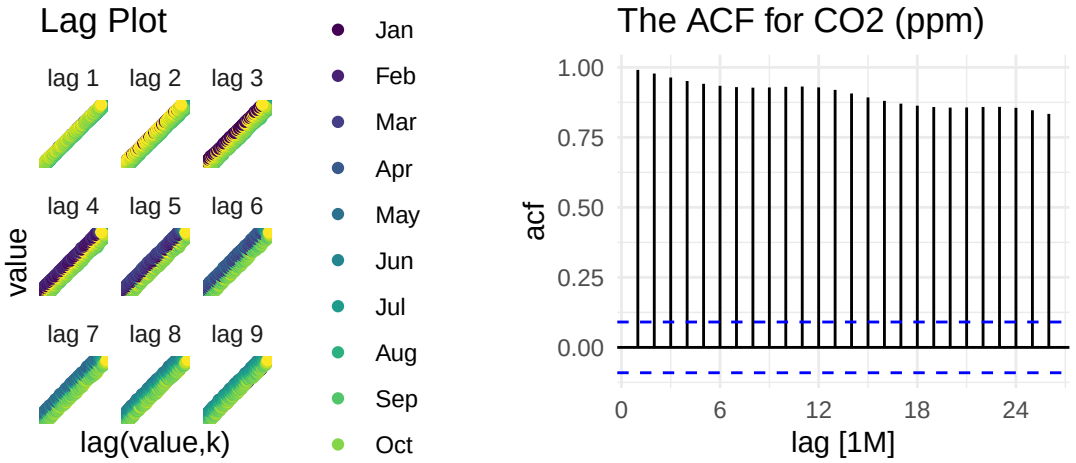


To analyze the trend year-by-year, Figure 2 provides another angle to view the upward trend in CO2 levels at Mauna Loa. The monthly CO2 (ppm) level was close to 315 ppm in 1959 and increased to upward of 365 ppm at the end of 1997. This is approximately a 16 percent increase in the concentration of CO2 at Mauna Loa across the nearly forty years of observations.

A break-down of the overall trend into time-trends in CO2 levels per month is presented in Figure 3. One will note CO2 levels appear to be lowest, on average in September and October and peak in May. This reflects the growing season patterns discussed by Keeling (1960). Although Keeling had fewer data points to consider, his observation transcends time. What is not clear is whether the upward trend in average CO2 levels is constant or increasing at a faster pace than in Keelings day. The trend line in Figure one appears to present as slightly convex or increasing at an increasing rate, especially after 1990. Also, the distance between lines in Figure 2 appears to increase in the 1990s as there is more white space between lines than previous decades, in particular when compared to the 1950s and 1960s.



Thus far, the data visualizations suggest a non-stationary process with strong seasonal trends. We now turn to lag and ACF plots to gauge possible auto-regressive tendencies in the data. We see strong correlation among the lag values in the lag-lag plot on the left-hand side of Figure 4. The first three month lags are nearly perfectly correlated with the time t CO_2 levels. From lags four to seven, there is more dispersion in lag correlates, but the linear, positive correlation remains strong. Finally, lags eight and nine realign with time t. The ACF supports a seasonal auto-regressive process. Lags decrease at a slow, but cyclical rate.



The EDA suggests we will need to consider a linear trend along with seasonal variation in the models of the carbon dioxide levels over time. Next, we build time-series models of carbon dioxide fluctuations to break down the seasonal variation and time trend to forecast future CO_2 accumulation at Mauna Loa.

0.3 Models

We test a series of models to determine the best model to explain carbon dioxide levels and accumulation path at Mauna Loa. The general formulation of the models follows a linear time trend (equation 1), quadratic transformation (equation 2), and polynomial time series model (equation 3).

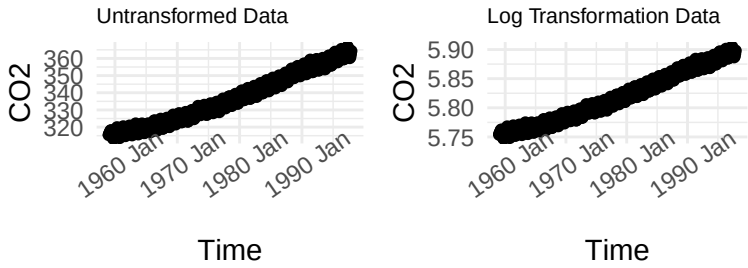
(Equation 1) $y_t = \beta_0 + \beta_1 t + \epsilon_t$
(Equation 2) $y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \epsilon_t$
(Equation 3) $y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \epsilon_t$

where y_t is the level of carbon dioxide in time t and ϵ_t is the white noise associated with the series.

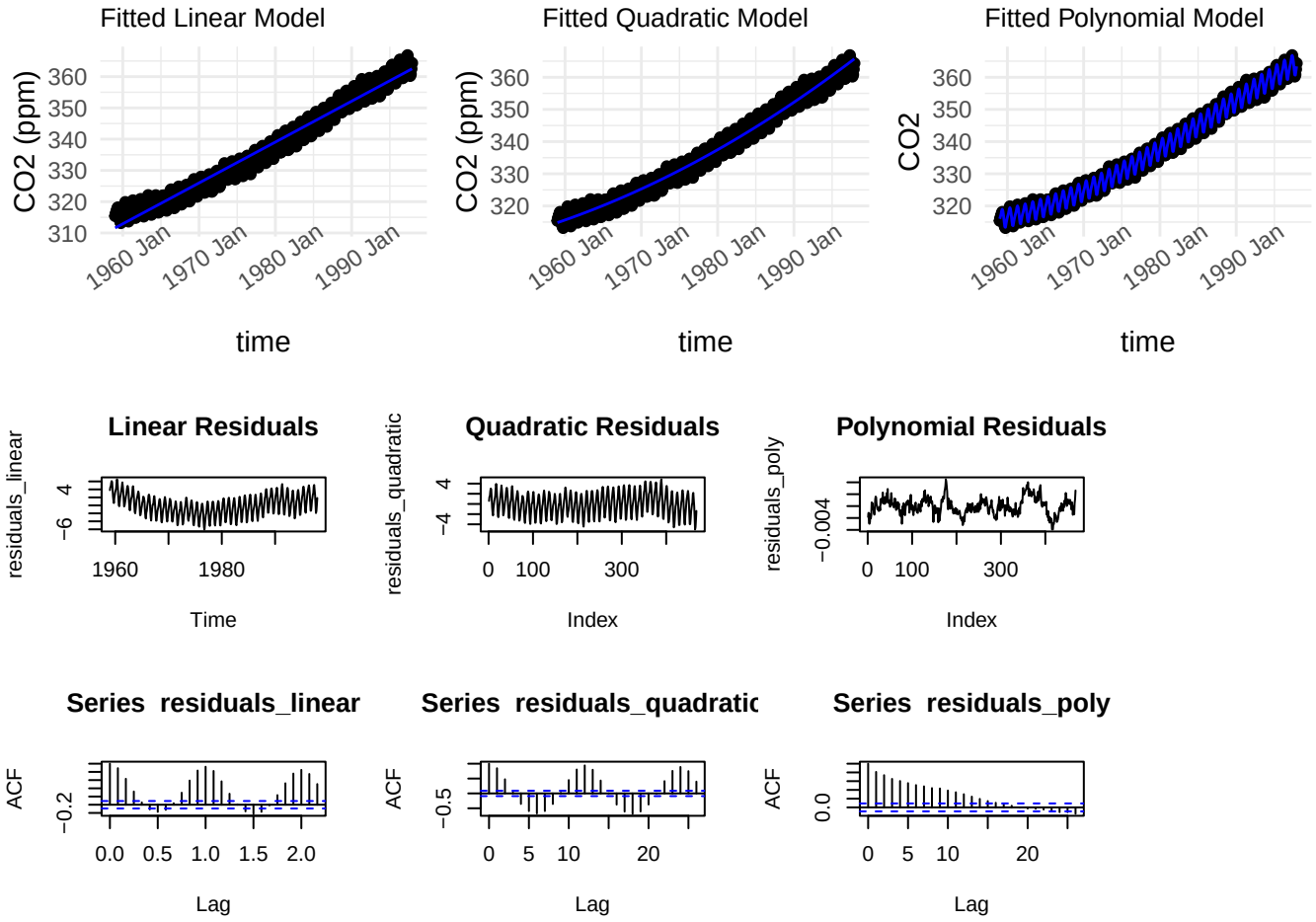
0.4 Linear time trend model

0.4.1 Log-Linear Model consideration

The data has a pretty linear trend with equal variance throughout the data. From the plot, you can see that a log transformation does not help with linearizing the data. Therefore, there is no transformation that is required for this data



The time trend plots for the linear and quadratic models appear smooth compared to the polynomial model. Both the linear and quadratic model have patterns in the residuals and significant lags in the ACF plot; all of this indicates that seasonality has not yet been accounted for. The polynomial residuals appear more like a white noise process, but there are significant lags in ACF which potentially indicate that the series needs to be detrended.

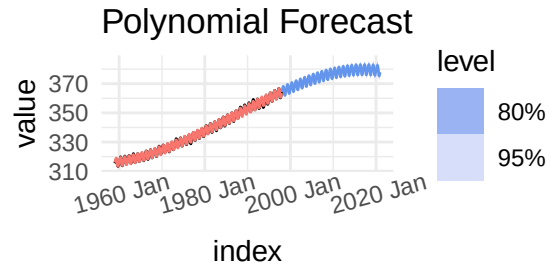


The AIC and BIC measurements form the three fitted models indicate the polynomial model has the best fit or makes the best use of the parameters because it has both the lowest AIC and BIC.The plot of the polynomial model results are in the right-side panel of the three panel diagram above. The residual plots below indicate the residuals from the polynomial plot are distributed more precisely around zero. The ACF of residuals was not stationary, but not the seasonal variation of the linear and quadratic models.

Compared to the linear and quadratic models, the polynomial had the strongest performance in both, with significantly lower AIC and BIC

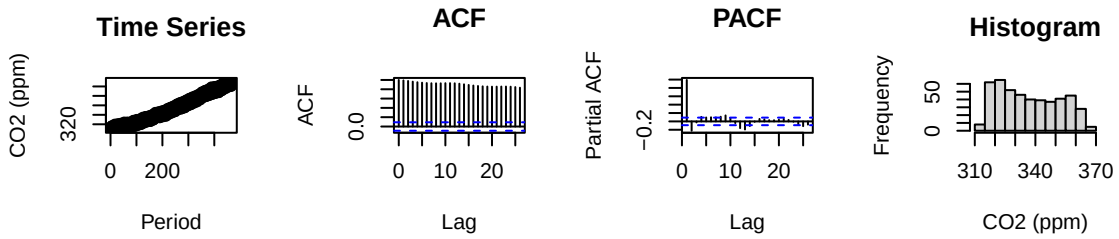
```
##      mod_names    mod_aic             mod_bic
## [1,] "linear"     "2232.96081427566"  "2245.40621916342"
## [2,] "quadratic"  "735.409043710632"  "752.002916894303"
## [3,] "polynomial" "-6116.44609583328" "-6050.0706030986"
```

The forecast until 2020 using the polynomial model had a similar seasonal pattern, with a decreasing growth rate as time went on.



0.5 (3 points) Task 3a: ARIMA times series model

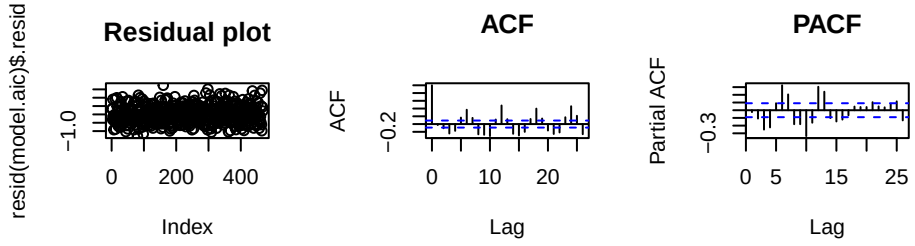
Following all appropriate steps, choose an ARIMA model to fit to the series. Discuss the characteristics of your model and how you selected between alternative ARIMA specifications. Use your model (or models) to generate forecasts to the year 2022.



The time series plot shows an increasing average as time increases. The histogram of Co2 values also shows an non-normal distribution. Both of these point to an increasing trend in the data. The ACF plot shows persistent significant lags, even for very large lag values, this is an attribute of an AR process. Additionally, the PACF plot quickly dies, but maintain an oscillating pattern, which is an attribute of an MA process. All of these factors together point towards an ARIMA process that underlies the Co2 time series that we wish to model.

```
## Series: value
## Model: ARIMA(2,1,4) w/ drift
##
## Coefficients:
##          ar1      ar2      ma1     ma2     ma3     ma4  constant
##       1.6886  -0.9587  -1.3228  0.1540  0.1374  0.1909    0.0286
## s.e.  0.0137   0.0134   0.0481  0.0749  0.0902  0.0563    0.0039
##
## sigma^2 estimated as 0.2901:  log likelihood=-373.34
## AIC=762.68   AICc=762.99   BIC=795.85
```

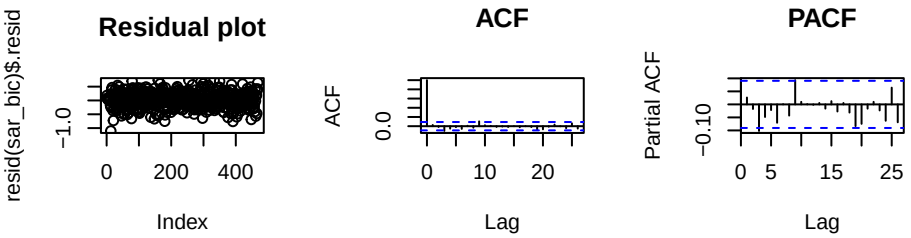
Model selection across AIC, AICc, and BIC all chose an ARIMA model with 2 AR parameters, 4 MA parameters and linear differencing. However, from looking at the ACF and PACF plots, below there are signs that there are still unaccounted seasonality trends that could be incorporated into the model.



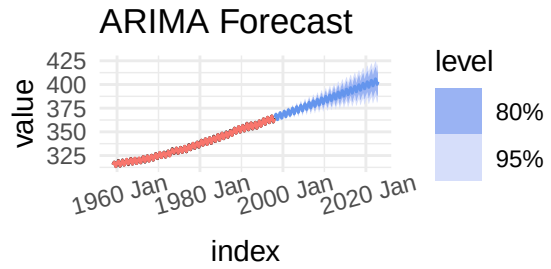
We built a SARIMA model and did parameter selection using BIC, because BIC is favored for larger datasets, and the dataset that we are training had nearly 500 observations. Using BIC as the model selection, the model included a main and seasonal linear trend, 1 main MA, 1 seasonal MA, and 2 seasonal AR. The seasonal components tended to have high errors compared to the coefficient estimates, while the main MA effect had much higher confidence.

```
## Series: value
## Model: ARIMA(0,1,1)(1,1,2)[12]
##
## Coefficients:
##           ma1     sar1     sma1     sma2
##       -0.3482  -0.4986  -0.3155  -0.4641
## s.e.   0.0499   0.5281   0.5164   0.4366
##
## sigma^2 estimated as 0.08603:  log likelihood=-85.59
## AIC=181.18   AICc=181.32   BIC=201.78
```

Adding in seasonal factors creates a residual plot that resembles a white noise process and an ACF and PACF with no significant lags, indicating that this model is a good fit.

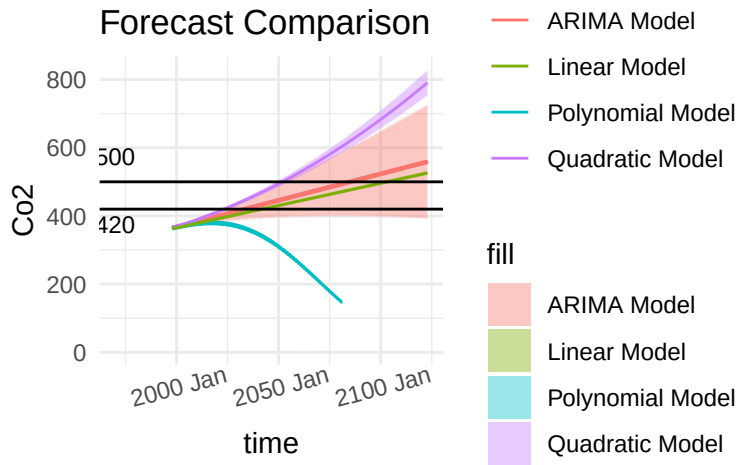


The forecast until 2022 using the ARIMA model continues the seasonal patterns with a linear growth rate.



0.6 (3 points) Task 4a: Forecast atmospheric CO2 growth

The ARIMA model had an AIC of 181 (see model output in section 3a), which was not as low as the polynomial model, but lower than both the linear and quadratic models, which could not account for seasonality. However, the polynomial model does poorly out of sample, as it predicts CO2 levels to peak and then steeply drop off towards negative predictions. Thus the best model for long-term forecasts is the ARIMA model



0.6.1 Point Estimate and CI for 420 ppm

The ARIMA model predicts that CO2 levels will pass 420 ppm for the first time in May of 2031, and for the last time in October of 2035. As seen in the plot above, the confidence interval grows extremely large and the lower bound of the prediction interval does not exceed 420 even in 2100, so the upper bound of the confidence interval was not findable using the ARIMA model over the time period that we forecasted over. The linear model prediction confidence interval does not have much overlap with the ARIMA CI, but the quadratic model has a decent amount of overlap. Because of the strength of the ARIMA AIC and the fact that there are similar predictions between different models, we can be fairly confident that the real time in which CO2 is within the 95% confidence interval.

```
##                 420: Predicted first time first time 95% CI
```

```
## Linear Model    "2041 Dec"                "( 2037 Dec , 2046 Jan )"
## Quadratic Model "2023 Jun"                "( 2021 Apr , 2025 Oct )"
## ARIMA Model     "2031 May"                "( 2020 May , N/A )"
##                 420: Predicted last last last time 95% CI
## Linear Model    "2042 Aug"                "( 2038 Aug , 2046 Oct )"
## Quadratic Model "2023 Oct"                "( 2021 Aug , 2026 Feb )"
## ARIMA Model     "2035 Oct"                "( 2022 Nov , N/A )"
```

0.6.2 Point Estimate and CI for 500 ppm

The ARIMA model predicts that the CO2 levels will pass 500 ppm for the first time in April of 2083 and for the last time in September of 2087. We can be much less confident in this prediction because the confidence interval is so wide at this point that the margin of error is almost 3 decades, thus we are much less confident in this estimate.

```
##                 500: Predicted first time 95% CI
## Linear Model    "2103 Feb"                "( 2098 Oct , 2107 Aug )"
## Quadratic Model "2051 Nov"                "( 2048 Nov , 2055 Apr )"
## ARIMA Model     "2083 Apr"                "( 2051 Mar , N/A )"
##                 500: Predicted last time 95% CI
## Linear Model    "2103 Nov"                "( 2099 Jun , 2108 Apr )"
## Quadratic Model "2052 Feb"                "( 2049 Jan , 2055 Jul )"
## ARIMA Model     "2087 Sep"                "( 2052 Jan , N/A )"
```

0.6.3 Predictions for 2100 January

```
## [1] "Point Estimate:  523.54"
```

```
## [1] "95% CI: ( 398.83 , 648.24 )"
```

We estimated the predicted CO2 levels in January of 2100. The ARIMA model predicted 524 ppm and we are 95% confident that the true value is between 399 and 648 ppm. This is very far in the future and the confidence interval is very large so it is hard to have much confidence in these predictions.

1 Report from the Point of View of the Present

1.1 (1 point) Task 0b: Introduction

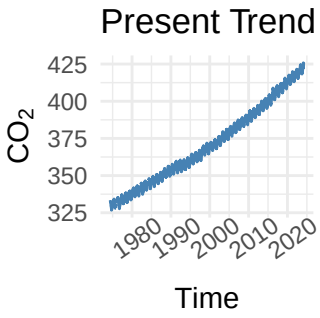
In 2024, we now have access to an additional 26 years of co2 concentration data. The data from the 1997 report was collected by the Scripps Institution at the Mauna Loa Observatory, Hawaii. This report will use co2 concentration data collected by NOAA from 1997 to 2024. The data presented in this report and the 1997 report were both collected at 3400m in Mauna Loa with samples from stations in the northern subtropics. The collection of the data is relatively consistent over the 66 years of observations. The rest of this report will analyze if forecasts of co2 concentration were accurate based on the data available in 1997, and if error is due to the models or other factors. Over the past 20 years, the world population has increased by 2 billion, and we've been burning more fossil fuels than ever to keep up with energy demands. These circumstances that aren't accounted for in the 1997 model might explain some of the error we will see in previous forecasts.

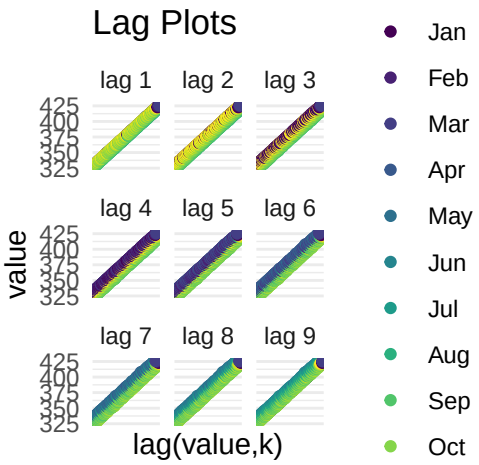
1.2 (3 points) Task 1b: Create a modern data pipeline for Mona Loa CO2 data.

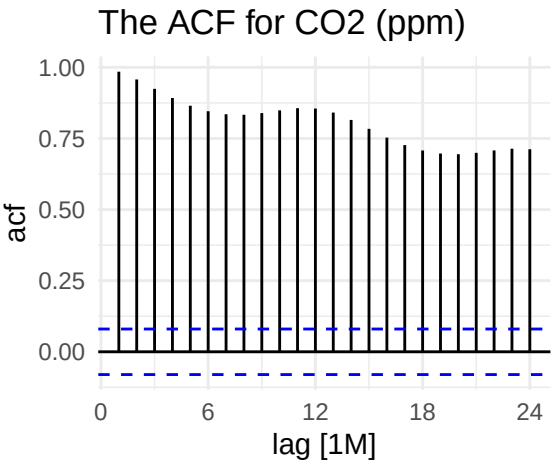
We created a pipeline which connected to the NOAA data portal and pulled down the data and locally saved the file with more recent data.

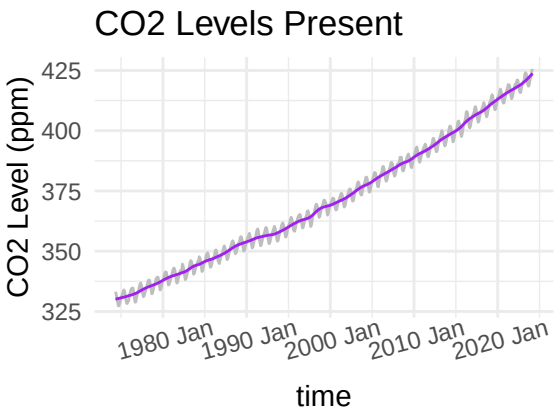
Compared to the keeling curve in 1997, the NOAA data shows a similar consistent seasonal trend, but the values are increasing rapidly. The NOAA data shows a steeper incline in co2 concentration. The Keeling curve from 1958-1997 (39 years) shows a roughly 40 ppm increase overall, while the Keeling curve from 1997-2023 (26 years) shows around a 90 ppm increase.The ACF, STL decomposition, and lag plots look similar to the 1958-1997 observations. The ACF's persistent significant lags and the oscillation in the PACF plot continue to show traits of both AR and MA processes, just as it did in 1997. These

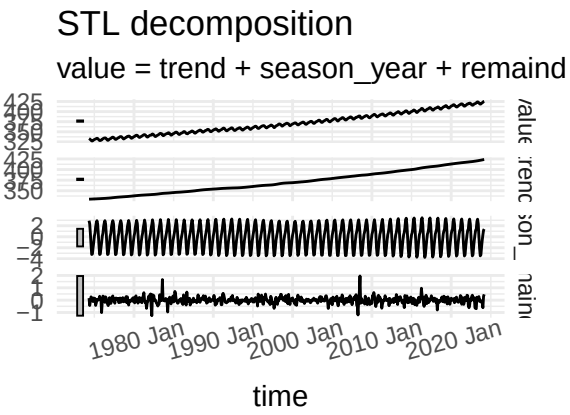
comparison of the present data timeplot and the 1997 timeplot reveal a possible non-linear trend in co2 which could mean our linear model from 1997 will perform poorly.

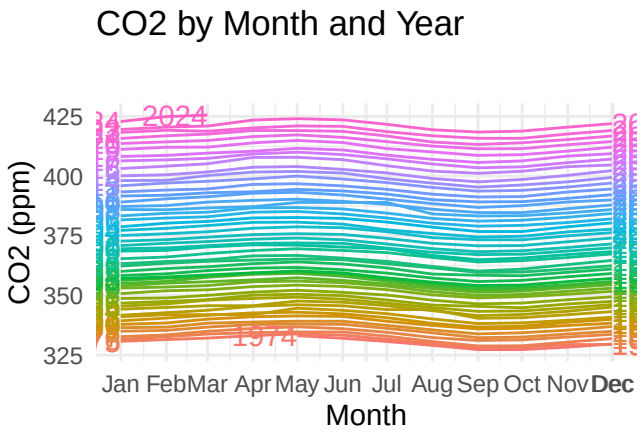


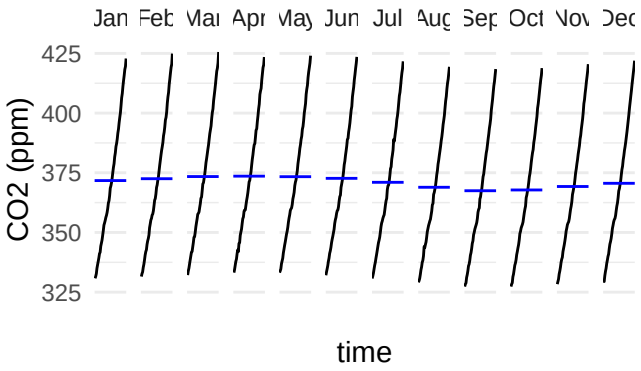






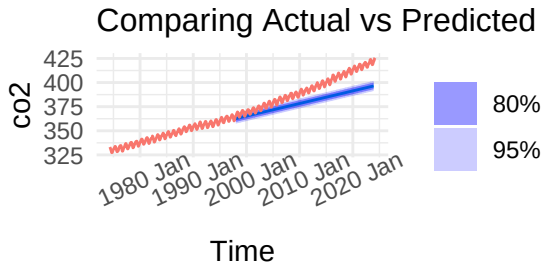






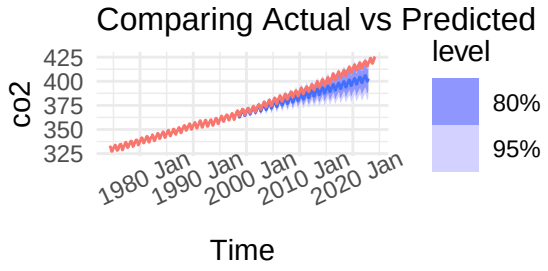
1.3 (1 point) Task 2b: Compare linear model forecasts against realized CO2

The graphs show that the actual co2 concentration levels are significantly higher than what was forecasted in 1997. According to the linear model, 2023 was forecasted to see levels around 390 ppm, but the actual ppm recorded was about 420. From the residuals of the linear model in the 1997 report we already saw that this model wasn't a great fit, and this graph confirms it.



1.4 (1 point) Task 3b: Compare ARIMA models forecasts against realized CO2

The ARIMA model did a better job at forecasting than the linear model, but it still has a tendency to underestimate the co2 levels. The correct co2 concentrations are included in the upper end of the 95% confidence interval of the ARIMA forecast. The Keeling curve had a greater rate of change than anticipated.



1.5 (3 points) Task 4b: Evaluate the performance of 1997 linear and ARIMA models

1.5.1 Linear Model Eval

```
##                         ME      RMSE       MAE          MPE      MAPE      MASE
## Training set 4.373498e-15  2.612462  2.146882 -0.005352854 0.6392463  1.994736
## Test set     1.269036e+01 14.526501 12.690362  3.166701389 3.1667014 11.791014
##                    ACF1
## Training set 0.8910172
## Test set            NA
```

1.5.2 ARIMA Model Eval

```
##                ME     RMSE      MAE      MPE     MAPE
## Test set 6.979727 8.573007 6.979727 1.730336 1.730336
```

It was forecasted by the ARIMA model that we would reach 420 ppm in May 2031, and the linear model forecasted we'd reach 420 ppm in December 2041. However we crossed the 420 ppm threshold in February 2023. Our models seemed to be far from the truth since they didn't account for the quickly increasing rate of change of the Keeling curve.

From looking at the linear model residual plot from the 1997 report, we see that the residuals aren't white noise. This tells us that the model wasn't a good fit for the data. From the accuracy test we see that the RMSE is roughly 14.6 which shows there is substantial error.

The ARIMA model has a better RMSE than the linear model (8.6), and the residual plot from the 1997 report showed that the residuals were white noise. This means the ARIMA model is a better fit for representing present co2 concentration levels. The plot in 3b also shows that the correct co2 concentrations were included in the 95% confidence interval of the ARIMA forecast, which shows that the ARIMA was close to learning the correct trend and seasonality.

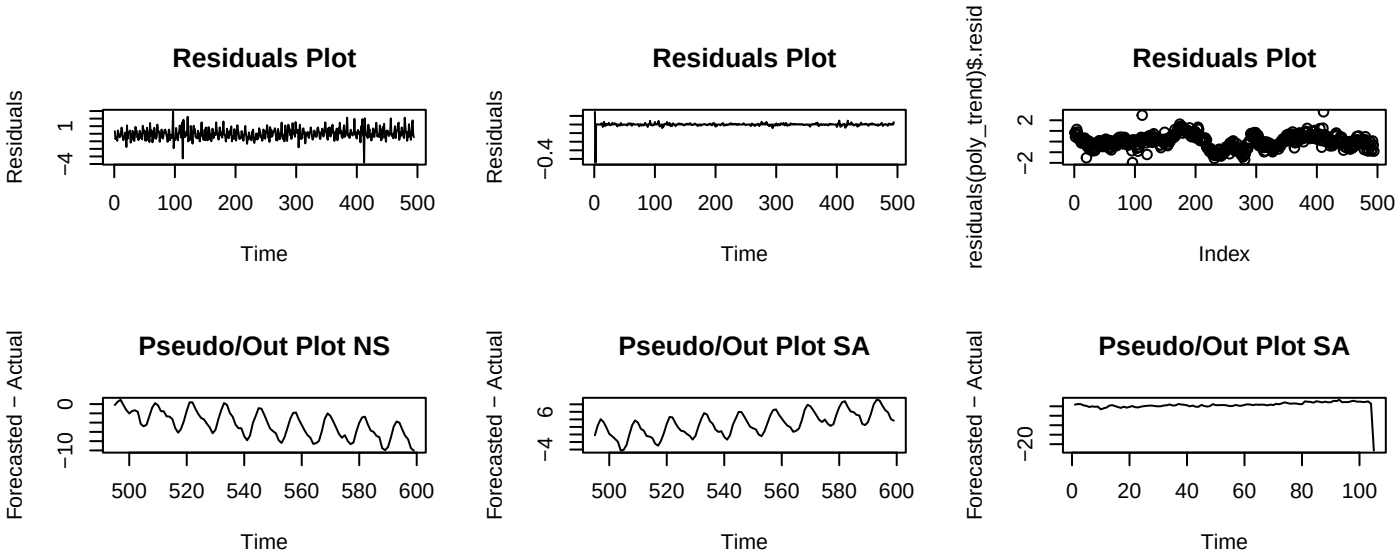
1.6 (4 points) Task 5b: Train best models on present data

The residuals plots show that the ARIMA model trained on the seasonally adjusted CO2 data frame performs better than the non-seasonally adjusted data frame (the in-sample residuals stay closer to 0), as it had the lowest AIC compared to the other models (see table below). The polynomial AIC was also good compared to the ARIMA NAS.

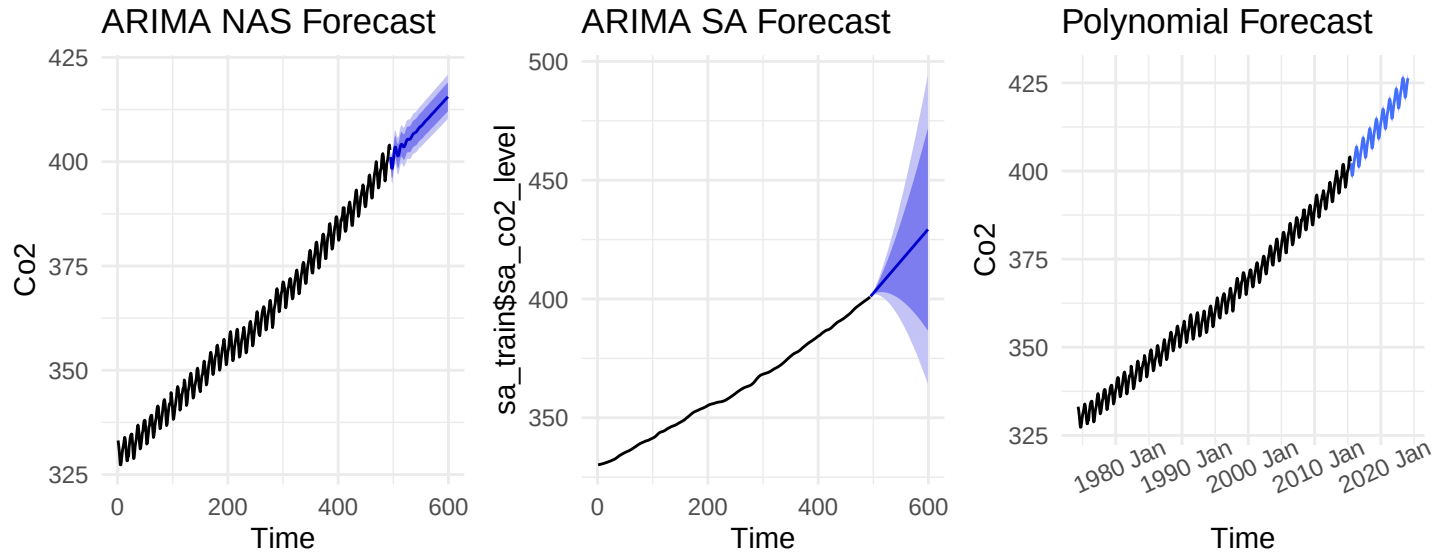
Analyzing the pseudo/out prediction values, we can see that the seasonally adjusted ARIMA model tends to overestimate CO2 values, while the non-seasonally adjusted model underestimates CO2 values.

From the pseudo/out graph for the seasonally adjusted polynomial model, we see that for the first set time steps are forecasted fairly accurately, but further into the future the model severely underestimates the Co2 concentration with high error.

The three forecasts for non-seasonally adjusted ARIMA, seasonally adjusted ARIMA, and polynomial model (same model as used in question 4b) show a difference in scale and trend for future predictions. SA ARIMA and NSA ARIMA models are similar, but SA overestimates in comparison to NSA. The polynomial model underestimates in comparison to the other two, and follows a nonlinear trend with significant dips.




```
## Plot variable not specified, automatically selected `.vars = value`
```




```
##      models          aic
## [1,] "ARIMA NSA"     "1101.77468825552"
## [2,] "ARIMA SA"      "-3111.476092516"
## [3,] "Polynomial SA" "-429.227754466318"
```

1.7 (3 points) Task Part 6b: How bad could it get?

Using the non-seasonally adjusted data series, the first time CO2 is expected to be at 420 ppm is October 2026 (with the interval October 2023 - March 2030) and to be 500 ppm in January 2073 for the first time (with the interval April 2068-March 2078). In the year 2122, CO2 is expected to be 747.21 from the seasonally adjusted ARIMA model predictions with the 95% confidence interval -2058.8 to 3553.26. The polynomial model forecasts that in the year 2122 we will reach 1406.12 ppm with the 95% confidence interval 1319.66 to 1492.58. We are not entirely confident in these predictions being accurate, since there is variability in the models, as well as many factors that may speed up or slow down CO2 concentration trends in the next 100 years besides seasonal patterns. It also seems that the confidence intervals are very large for co2 concentration predictions, which might indicate that it is hard to accurately forecast based on the current data.

1.7.1 Point Estimate and CI for 420 ppm

```
##                   Predicted first time 95% CI
## ARIMA (NSA)      "2026 Oct"           "( 2023 Oct , 2030 Mar )"
## ARIMA (SA)       "2021 May"           "( 2018 Feb , N/A )"
## Polynomial Model "2021 May"           "( 2022 Jan , 2023 Jan )"
##                   Predicted last time 95% CI
## ARIMA (NSA)      "2027 Apr"          "( 2024 Mar , 2030 Sep )"
## ARIMA (SA)       "2021 Aug"          "( 2018 Mar , N/A )"
## Polynomial Model "2023 Sep"          "( 2022 Aug , 2023 Aug )"
```

1.7.2 Point Estimate and CI for 500 ppm

```
##                   Predicted first time 95% CI
## ARIMA (NSA)      "2073 Jan"           "( 2068 Apr , 2078 Mar )"
## ARIMA (SA)       "2046 Jan"           "( 2024 Aug , N/A )"
## Polynomial Model "2042 May"           "( 2042 Feb , 2043 May )"
##                   Predicted last time 95% CI
## ARIMA (NSA)      "2073 Jul"          "( 2068 Sep , 2078 Sep )"
## ARIMA (SA)       "2046 Mar"          "( 2024 Aug , N/A )"
## Polynomial Model "2043 Oct"          "( 2042 Oct , 2044 Oct )"
```

1.7.3 Predictions for 2122 January

```
##                   Point Estimate     95% confidence Interval
## ARIMA Model      "747.206691477701" "( -2058.84513453578 , 3553.25851749118 )"
## Polynomial Model "1406.12040895587" "( 1319.66371514165 , 1492.57710277009 )"
```

2 References

Jiménez-de-la-Cuest, Diego and Thorsten Mauritsen. 2019. Emergent constraints on Earth's transient and equilibrium response to doubled CO2 from post-1970s global warming. Nature Geosciences 12: 902-905.

Keeling, Charles D. 1960. The Concentration and Isotopic Abundances of Carbon Dioxide in the Atmosphere. Tellus 12(2)

United Nations Climate Conference. 2023. Linking Nature Conservation with Climate Action. Retrieved from https://unfccc.int/cop28/5-key-takeaways#protecting-nature