W271 Group Lab 2

Co2 Analysis: from the Point of View of Present

Group Project (4 members)

Abstract

The study reevaluates our 1997 forecasts on atmospheric CO_2 trend and creates forecasts to year 2122 based on the best performing model from linear time trend model. Our final forecasts based on a SARIMA model are coherent with Keeling's curve predictions. Further studies are encouraged to evaluate the persistence or change in trends.

Contents

0.1 Task 0b: Introduction

Were models previously fit able to accurately predict for the future and are they still the best at capturing the behaviors of CO_2 concentration? From our previous report in 1997, we analyzed the atmospheric CO_2 evolution and highlighted both its seasonal and increasing trend behavior. We attempted to model the atmospheric CO_2 evolution with different models: linear at various degrees and ARIMA. At this stage, the quadratic seasonal fit was performing better by capturing both the trend and the seasonal pattern of the time series. For forecasting purposes, a second model was developed: ARIMA(0,1,1)(1,1,2). The forecasts predicted a strong C02 increase from approximately 360 ppm in 1997 to 420 ppm by 2031, 500 ppm by 2083, and eventually 525 ppm by 2100. Our report concluded that, considering the societal and environmental stakes, it would make sense to continue monitoring atmospheric CO_2. The additional data would allow us to assess models' performances and investigate any changes in the C02 behavior.

The current study aims at investigating if the data generation process has changed between 1997 and today. The forecasts of previously mentioned models (linear, ARIMA) are compared with realized atmospheric CO_2 levels. Other models, based on new CO_2 concentration observations, are developed and compared. The best model is used to perform forecasts up to 2122.

0.2 Task 1b: Create a modern data pipeline for Mona Loa CO_2 data.

Our data is collected by NOAA using infrared technology at the Mona Loa site. There they measure the mole fraction of CO_2 - the amount of molecules of carbon dioxide that are found in a given number of air molecules. By measuring the concentration at a higher altitude, in dryer air, measurements of differences in CO_2 concentration with less impact of environmental factors such as pressure, outside influences such as vegetation as well as local pollution are possible.

The data is a weekly data collected Sunday through Saturday since 1974. This 365 days span ignores the possibility of a leap year. it includes information on month, day, decimal, average, number ofdays (ndays), information on one year ago which is exactly 365 days before the collected observation, 10 years ago referring to exactly 10*365 days+3 days (for leap years) ago, and increases since 1800as a baseline. In the entirety of 1974, the "1 year ago" field is empty due to lack of recorded in the previous year. Likewise, the "10 years ago" field is empty from 1974 to May 1984 due to lack of recorded data in the 10 years preceding the relevant observations.

```
datasrc = "https://gml.noaa.gov/webdata/ccgg/trends/co2/co2_weekly_mlo.csv"
co2_present_raw <- read.csv(datasrc, header=TRUE,
                            stringsAsFactors=FALSE, comment.char="#")
```

This raw data includes 2,512 observations. Here we observe the missing values that are removed, represented as -1000 in average, X1.year.ago, and X10.years.ago variables. The decimal variable is a numerical representation of where in each year the atmospheric CO_2 levels are collected. The weekly CO_2 levels range from a minimum of 347 ppm to 422 ppm, with an average of 366 ppm and a median of 366 ppm. Further, the number of days between each collection is between 5 and 7 days, meaning the weekly collections are not always standardized at a 7 day week threshold. For our time series, the data we will be using are year, month, and day to represent our time in weeks as well as average to represent the CO_2 concentration.

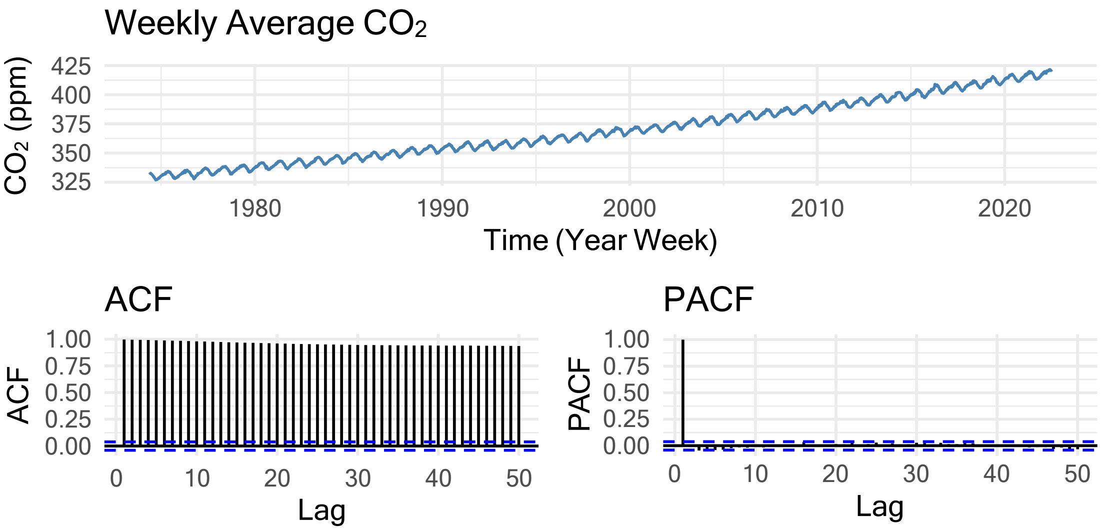


Figure 1: Initial EDA Plots: Time Plot, ACF, and PACF of the Monthly Average Atmospheric Concentrations of CO_2 (ppm) at Mauna Loa, Hawaii, from 1959 to present(2022)

The gaps are interpolated with the estimated values of an ARIMA model. This assumes no irregular or major events occurred during this time period that might later result in model underfitting. The resulting time series with the interpolated average values (renamed `values`) indexed on the date (`time_index`) is used to investigate the features of this time series - presence of seasonality, trends, and relationship between lagged values of the time series.

Figure 1 indicates the levels of CO_2 has significantly increased from the early 1980s to 2022, with an upward trend. Specifically, the monthly CO_2 concentration fluctuates over time and follows an increasing deterministic trend with seasonal pattern. The increasing mean of the time series indicates non-stationarity as well with regular seasonal trends observed each year - known as the Keeling curve, where the values fluctuate up and down regularly. In this seasonal variation CO_2, concentrations are highest in April to July, and at their lowest from August to October before picking back up.

The slow but consistent decay in the ACF plots shows a strong trend. The PACF abruptly drops after one significant lag. There are very small but somewhat significant sinusoidal fluctuations in each year observed, consistent with the Keeling curve seasonal variation patterns.

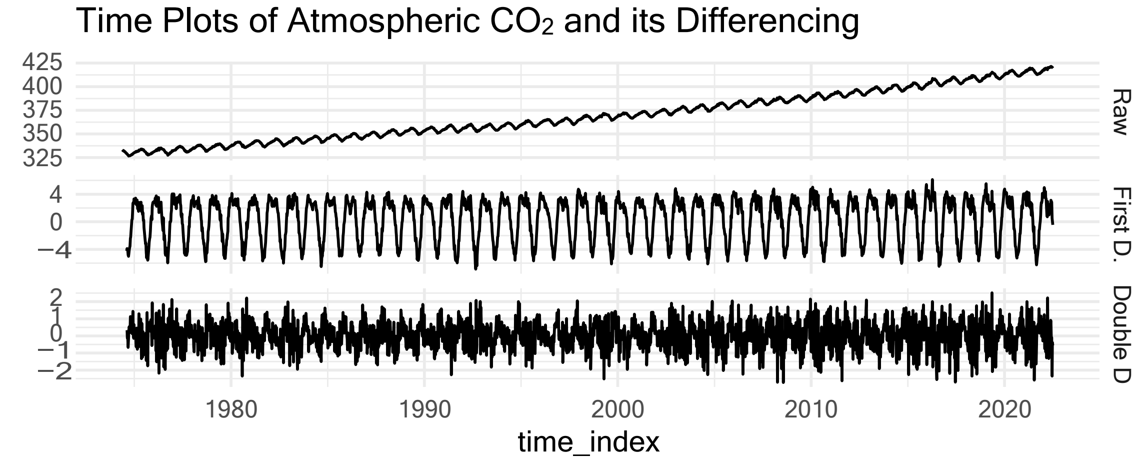


Figure 2: Time Plots of the Atmospheric CO_2 Time Series and its First and Second Order Differencing

In order to determine the most appropriate model of the time series, we will conduct further formal tests to further investigate its seasonality and trend. We will first investigate the nature of the seasonality in the data. Figure 2 suggests that the first differenced data visually seems to be stationary with non-fluctuating variance and constant mean but possess strong seasonality. The double differenced data seems to remove seasonality completely in addition of being stationary, making it looks a like white noise. Hence, the first differenced data is perhaps preferred over the double differenced data because it removes the trend and retains the seasonality of the original data.

Table 1: Unit Root Test Results of Double Diff CO_2 Time Series.

kpss_stat	kpss_pvalue
0.037	0.1

The unit root test reports whether a time series is non-stationary and possesses a unit root (Table

1). The null hypothesis is defined as the presence of a unit root whereas the alternative hypothesis is either stationarity, or trend stationarity. In this test, the p-value at 0.01 when less than 0.01 and as 0.1 in the case they are greater than 0.1 allows us to reject the null hypothesis. For the first differenced values, the test statistic is small and within expectations for stationary data. Here, with a p-value greater than 0.1, we can conclude the first differenced data appears stationary. The ACF and PACF (Figure 3) suggests that the series still contains a seasonal component and that we will need a model that adjusts for both trend and seasonality - a seasonal ARIMA model or a linear with both trend and seasonal adjustment are fitting candidates.

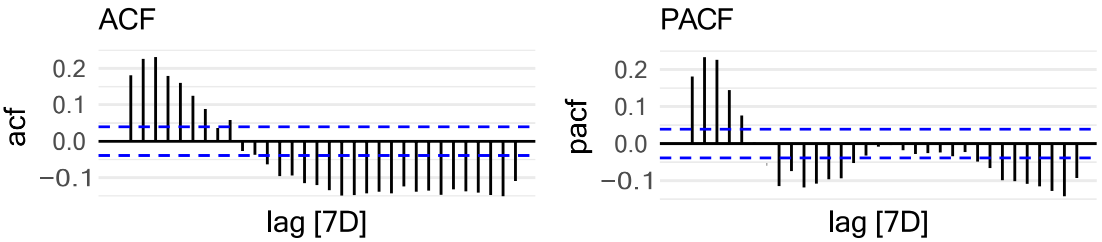


Figure 3: ACF and PACF of the First Differencing

0.3 Task 2b: Compare linear model forecasts against realized CO2

Before we create models on new data, we will assess previous models created in 1997 and their performance on the realized CO_2 levels (Figure 4).

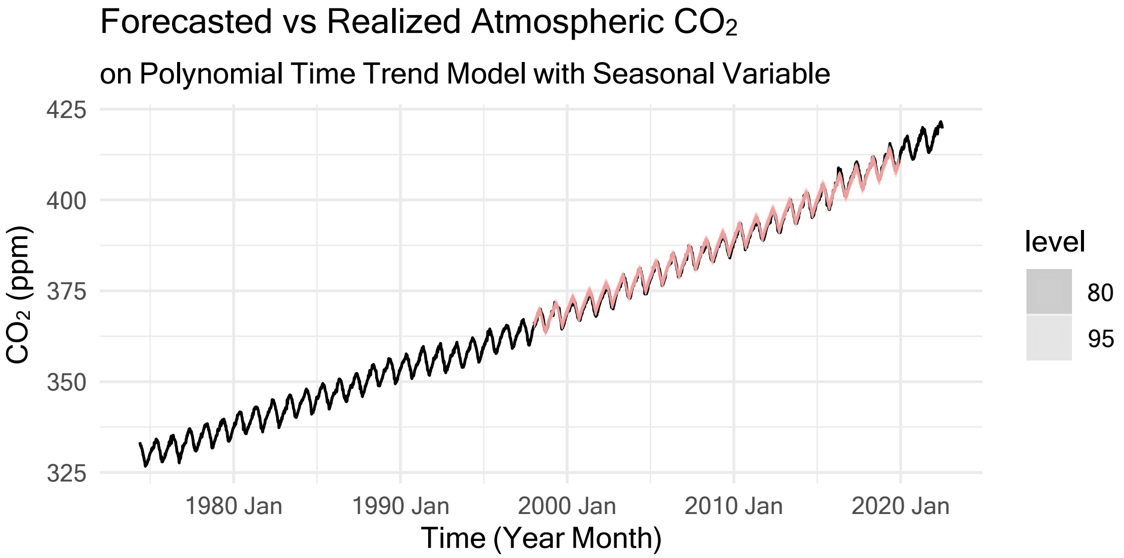


Figure 4: Comparison of 1997 linear model forecasted atmospheric CO_2 against realized CO_2

In our models from 1997, we created a linear model with a quadratic component for trend. Here it is clear to see that our forecasted data from this model captures both the seasonality and trend of the realized CO_2 levels well. We also observe that the linear model's forecasted data also contains a narrow confidence level. In effect of this or an increase in the steepness of the trend, some of the later realized data falls outside of the range of the predictions of this model.

0.4 Task 3b: Compare ARIMA models forecasts against realized CO_2

From our previous work, we found that our 1997 ARIMA forecast had the best performance, capturing trend and seasonality of the fitted data, with wider confidence intervals over time.

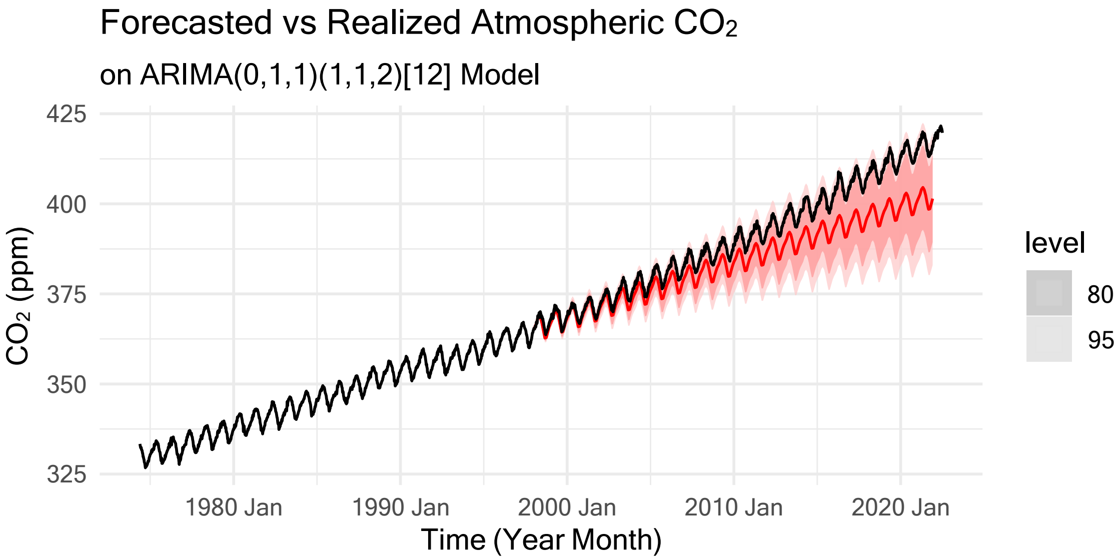


Figure 5: Comparison of 1997 ARIMA model forecasted atmospheric CO_2 against realized CO_2

Compared to realized values, the forecasted data from the ARIMA model diverges starting from the middle to the end of the forecasted range (Figure 5). Specifically, the forecasted data is consistent with the realized values from 1997 to about the mid-2000s. The discrepancy widens significantly thereafter, suggesting the ARIMA models may not best fit the model. However, it is to be noted that the seasonality patterns of the forecast are consistent with the actual data. We suspect a rapidly increasing trend on the emission of CO_2 from the mid-200s, which may have led to the underestimation in the forecast.

This graphical comparison is more subjective and we will further evaluate the models' performance using more objective measures.

0.5 Task 4b: Evaluate the performance of 1997 linear and ARIMAmodels

In 1997 we determined the ARIMA(0,1,1)(1,1,2)[12] model was the most performant evaluated by BIC and from it forecasted predictions for the first time that CO_2 would cross 420 ppm. Here we will evaluate how close our predictions were to the realized CO_2 levels.

In 1997, we predicted CO_2 ppm would first reach 420 in May 2031. In fact, we find this value to appear in April 2021, 11 years earlier than originally predicted. This reflects our graphical observations of the ARIMA model performance which illustrate our model to have predicted values with a trend that is slowly increasing when compared to the realized CO_2 concentration.

Following observations of our models from 1997, we continue to use the weekly data to generate a month-average series from 1997 to the present, and compare the overall forecasting performance of your models over the entire period.

```
train_co2 = co2_present %>%
           index_by(month_index = yearmonth(time_index)) %>%
           summarise(value = mean(value)) %>%
           filter_index( ~ "1997-12-31")

test_co2 = co2_present %>%
           index_by(month_index = yearmonth(time_index)) %>%
           summarise(value = mean(value)) %>%
           filter_index( "1998-01-01" ~ "2022-07-01")

model.comp <- train_co2 %>%
             model(quad = TSLM(value ~ trend() + I(trend()^2) + season()),
                  ARIMA = ARIMA(value ~ 0 + pdq(0,1,1) + PDQ(1,1,2)))

model.forecasts<-forecast(model.comp)
```

Table 2: Forecast Accuracy of the 1997 ARIMA and Linear Time Trend Models using Present Data

Model	RMSE	MAE
Quadratic Seasonal Model	1.08	0.979
ARIMA(0,1,1)(1,1,2)[12]	1.83	1.761

Table 2 shows that forecast accuracy, RMSE and MAE, of the 1997 models fitted on the all data prior to year 1998 and evaluated on the realized data from 1998 to present. As in the 1997 report, we are optimizing to minimize RMSE for forecasts of the mean. Recall in the 1997 report, the SARIMA model was more performant than the quadratic seasonal model with a lower RMSE. Interestingly, when evaluated against the observed values of the 25 years between 1997 and 2022, linear model has a lower value for both the RMSE and the MAE, and thus becomes the better preforming model than the SARIMA model.

0.6 Task 5b: Train best models on present data

From our EDA, we found that seasonal adjustment would be an important component of our final model. Here we begin with an STL model to remove seasonality from the observed data.

The STL decomposition with 52-week seasonal period (there are 52 weeks in year) in Figure 6 suggests that a trend that has been seasonally adjusted with a stationary remainder that resembles white noise. This seasonally adjusted trend data will be used for forecasting in our models.

To evaluate the performance of models chosen, we split our data into a training and a test set where training is all data up until 2020 and the test set represents the last two years.

```
co2_present_training <- co2_seasonal %>% filter_index( ~ "2020-01-01")
co2_present_test <- co2_seasonal %>% filter_index("2020-01-01" ~ "2022-07-01")
```

We first begin the analysis with non-seasonal ARIMA model. As in the 1997 report, the EDA implies strong yearly seasonality pattern, hence the seasonal components with 12-period interval is added

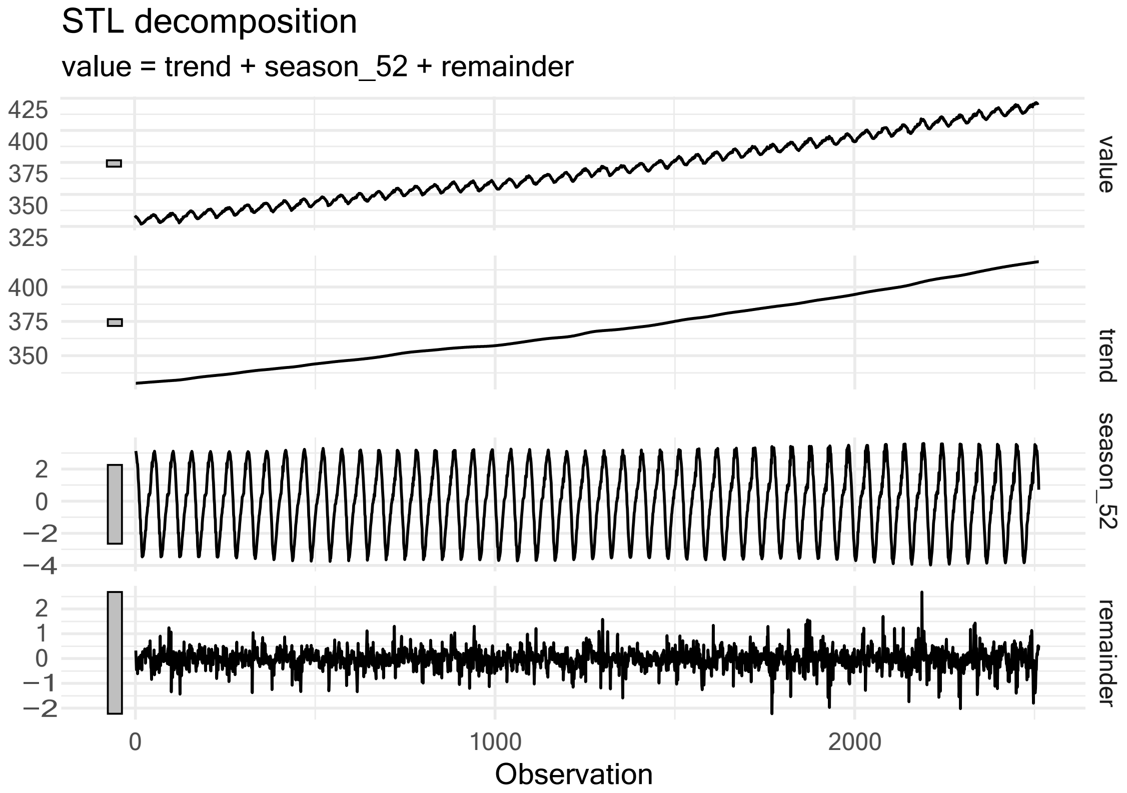


Figure 6: STL Decomposition of the Presetn CO_2 Data

the search space of the ARIMA model. The search space of AR and MA terms for both non-seasonal and seasonal components are limited to less than 3 as higher order models is discouraged by R and takes too much time and computing resources for an initial investigation. The differencing order for both non-seasonal and seasonal components is limited to less than 2 as difference order beyond 2 is difficult to interpret.

```
nsa.arima.fit <- co2_present_training %>%
  model(
    ARIMA(value ~ 0 + pdq(0:3, 0:2, 0:3, p_init=0, q_init=0),
               order_constraint=TRUE, ic="bic", stepwise=T, greedy=F)
  )
```

```
## Series: value
## Model: ARIMA(2,1,3)
##
## Coefficients:
##          ar1      ar2      ma1     ma2      ma3
##       1.8537  -0.896  -2.064  1.4047  -0.2665
## s.e.  0.0186   0.018   0.028  0.0461   0.0242
##
## sigma^2 estimated as 0.2457:  log likelihood=-1705
## AIC=3421   AICc=3421   BIC=3456
```

After iterating through the space, the model with the lowest BIC is an ARIMA(2,1,3). To evaluate the model, we start by observing its residuals.

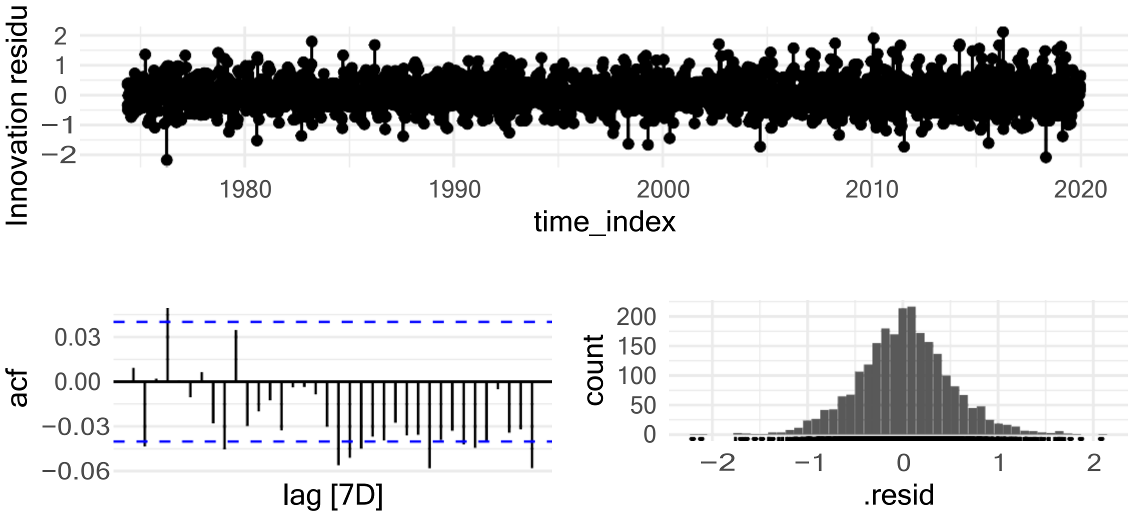


Figure 7: Diagnostic Plots of the Residuals of ARIMA(2,1,3) Model

The visual residual diagnostic of the ARIMA(2,1,3) model (Figure 7) implies that the time series to resemble that of white noise data, centered around 0 and constant fluctuation. The ACF plot of the residuals also shows significant values, especially for lags greater than 10. Finally, we observe a distribution that resembles that of a normal distribution. While there are signs that the residuals resemble white noise, we perform the Ljung-Box test to support whether these are in fact independently distributed.

Table 3: Ljung-Box test for model: ARIMA(2,1,3)

df	pvalue
1	0.645
10	0.023

From results of the Ljung-Box Test for seasonally-adjusted ARIMA model (Table 3) show a statistically significant p-value at lag 10 and reject the null hypothesis for the Ljung-Box test - suggesting that the data is not independently distributed and distinguishable from white noise. Because the lags exhibit dependence, we might expect this model to not perform well in forecasting future data.

Now we model for a seasonally-adjusted ARIMA model.

After iterating through the space, the model with the lowest BIC is an ARIMA(0,1,2)(0,0,2)[12].

```
## Series: season_adjust
## Model: ARIMA(0,1,2)(0,0,2)[12]
##
## Coefficients:
##           ma1      ma2    sma1    sma2
##       -0.5791  -0.0728  0.0566  0.0795
## s.e.   0.0209   0.0196  0.0210  0.0196
```

```
##
## sigma^2 estimated as 0.161:  log likelihood=-1202
## AIC=2414   AICc=2414   BIC=2442
```

Compared to the seasonally adjusted , the AIC and AICc of 2414 as well as the BIC of 2442 score much lower, and therefore better, than in the non-seasonally adjusted model. Consequently, we expect this model to be better at forecasting for future data.

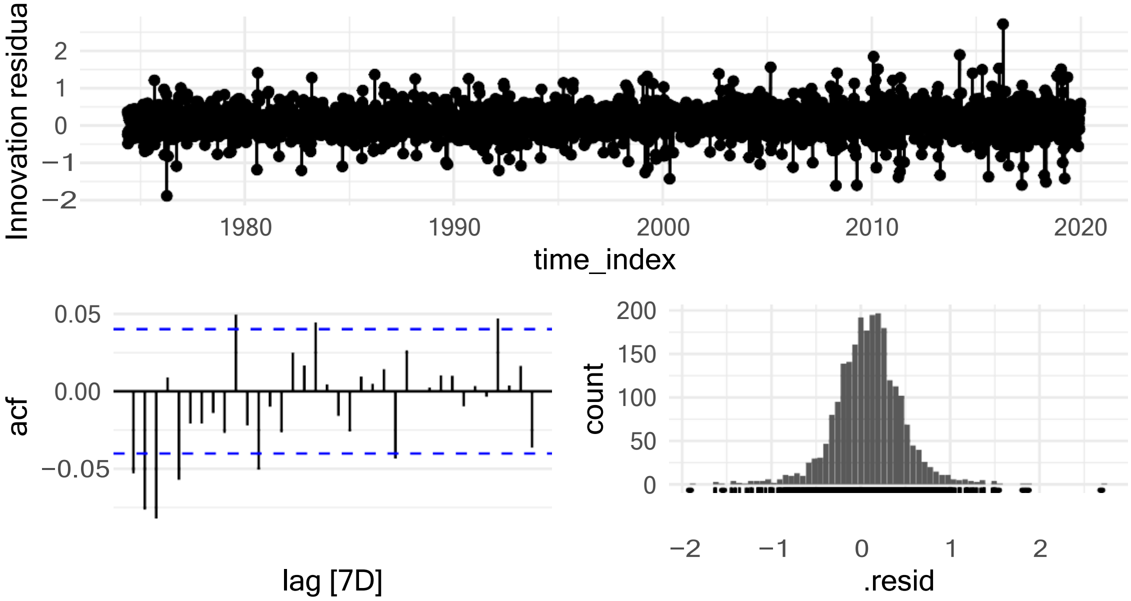


Figure 8: Diagnostic Plots of the Residuals of Seasonal Adjusted ARIMA(0,1,2)(0,0,2)[12] Model

Observing the residuals of our ARIMA(0,1,2)(0,0,2)[12] model (Figure 8), we find our time series to resemble that of white noise data, centered around 0. The ACF plot of the residuals also shows significant values, and oscillate from negative to positive. Finally, we observe a distribution that resembles that of a normal distribution. As before, we perform the Ljung-Box test to support whether these are in fact independently distributed.

Table 4: Ljung-Box test for model: ARIMA(0,1,2)(0,0,2)[12]

df	pvalue
1	0.01
10	0.00

From results of the Ljung-Box Test for seasonally-adjusted ARIMA model (Table 4), there are statistically significant p-values at both lag 1 and lag 10 and so we reject the null hypothesis for the Ljung-Box test - suggesting that the data is not independently distributed and distinguishable from white noise. Again, because the lags exhibit dependence, we might expect this model to not perform well in forecasting future data.

In addition to our non-seasonal and seasonal ARIMA models, we will also fit a seasonal polynomial model. As we observed in our EDA, there is a curve in the CO_2 suggesting a non-linear term be included as well as strong seasonality and so we will fit against our seasonally-adjusted data. To determine the appropriate degree for the polynomial model, defined as the model with the lowest information criteria measures, multiple polynomial models of degree 2 to 4 are fitted.

```
poly.season.models <- co2_present_training %>%
  model(linear = TSLM(season_adjust ~ trend()),
        quad   = TSLM(season_adjust ~ trend()+I(trend()^2)),
        cubic  = TSLM(season_adjust ~ trend()+I(trend()^2)+I(trend()^3) ),
        quart  = TSLM(season_adjust ~ trend()+I(trend()^2)+I(trend()^3) +
                        I(trend()^4) ))
```

Table 5: Model Fit Measurements of Linear, Quadratic, Polynomial (with Seasonal Variable) Time Trend Models

Model	AIC	AICc	BIC
Linear Seasonal Model	3782	3782	3799
Quadratic Seasonal Model	-769	-769	-746
Cubic Seasonal Model	-2088	-2088	-2059
Quartic Seasonal Model	-2187	-2187	-2152

The measures in Table 5 illustrate polynomial models greater than 1-degree with the seasonal component all have negative information criteria measurements, and fall in the range of -2100s to 3700s. Here, the greatest improvement of the information criteria measures occurs when the polynomial degree moves from 1 to 2 as well as from 2 to 3 while a degree of 4 model contains the lowest value for BIC. From this, we narrow our potential choices to degrees 2 to 4. To incorporate a holistic approach for our choice, we will rely on the error from the test set and we will also compare our non-seasonal ARIMA model and our seasonal ARIMA model.

```
poly.comp <- co2_present_training %>%
             model(polynomial2 = TSLM(season_adjust ~ trend() + I(trend()^2)),
                   polynomial3 = TSLM(season_adjust ~ trend() + I(trend()^2) +
                                        I(trend()^3)),
                   polynomial4 = TSLM(season_adjust ~ trend() + I(trend()^2) +
                                        I(trend()^3) + I(trend()^4))
                   )

poly.forecasts<-forecast(poly.comp)

model.comp<-co2_present_training %>%
  model(SA.ARIMA = ARIMA(value ~ 0 + pdq(0,1,2) + PDQ(0,0,2), ic="bic"),
        ARIMA = ARIMA(value ~ pdq(2,1,2))
        )

model.forecasts<-forecast(model.comp)
```

Comparing our different chosen models, Table 6 contains the two most commonly used forecast accuracy metrics: root mean squared error(RSME) and mean absolute error(MAE). MAE is popular when comparing forecasts based on the same time series because of its simplicity in computation and interpretation. However, a forecast method minimizing MAE would lead to forecasts of the median whereas those minimizing RMSE would lead to forecasts of the mean. Hence, RMSE is also a popular choice of forecast accuracy metrics. From Table 6, we find the polynomials of degree 3 and 4 have the smallest values for RMSE and MAE. As RMSE will lead to forecasts of the mean we use this in our final decision leading the polynomial of degree 3 to be the preferred model for the analysis onward.

Table 6: Forecast Accuracy Evaluation of the ARIMA and Seasonal Polynomial Time Trend Models

Model	RMSE	MAE
Seasonal Polynomial (deg 3)	0.297	0.229
Seasonal Polynomial (deg 4)	0.307	0.227
Non-Seasonal ARIMA	0.346	0.344
Seasonally-Adjusted ARIMA	0.422	0.399
Seasonal Polynomial (deg 2)	1.304	1.284

```r
poly.fit <- co2_present_training %>%
            model(poly3 = TSLM(season_adjust ~ trend() + I(trend()^2) + I(trend()^3)))
```

To illustrate the performance of the models across the ARIMA, seasonally-adjusted ARIMA, and polynomial degree 3 models, we plot the forecasts against the realized CO_2 levels of the test set.

In our short term comparison, we see that the ARIMA non-seasonal forecast model runs through the trend of our test data and fully contain the seasonality at 95% confidence level. For the ARIMA based on seasonally-adjusted data, we in fact find this to not follow the trend and to only contain values at the beginning of the test data. Last, for the polynomial model based on seasonally-adjusted data, we find this data to follow the trend running running along the data as most observations seem to fall with the 95% confidence interval until a dip in later 2022. Here, two models stand out as prospective: the non-seasonally-adjusted ARIMA model and the seasonally-adjusted polynomial model. To more objectively choose one for predictions, we may rely on the results of Table 6 which suggests that the polynomial model of degree 3 is preferred and we choose this model to form predictions.

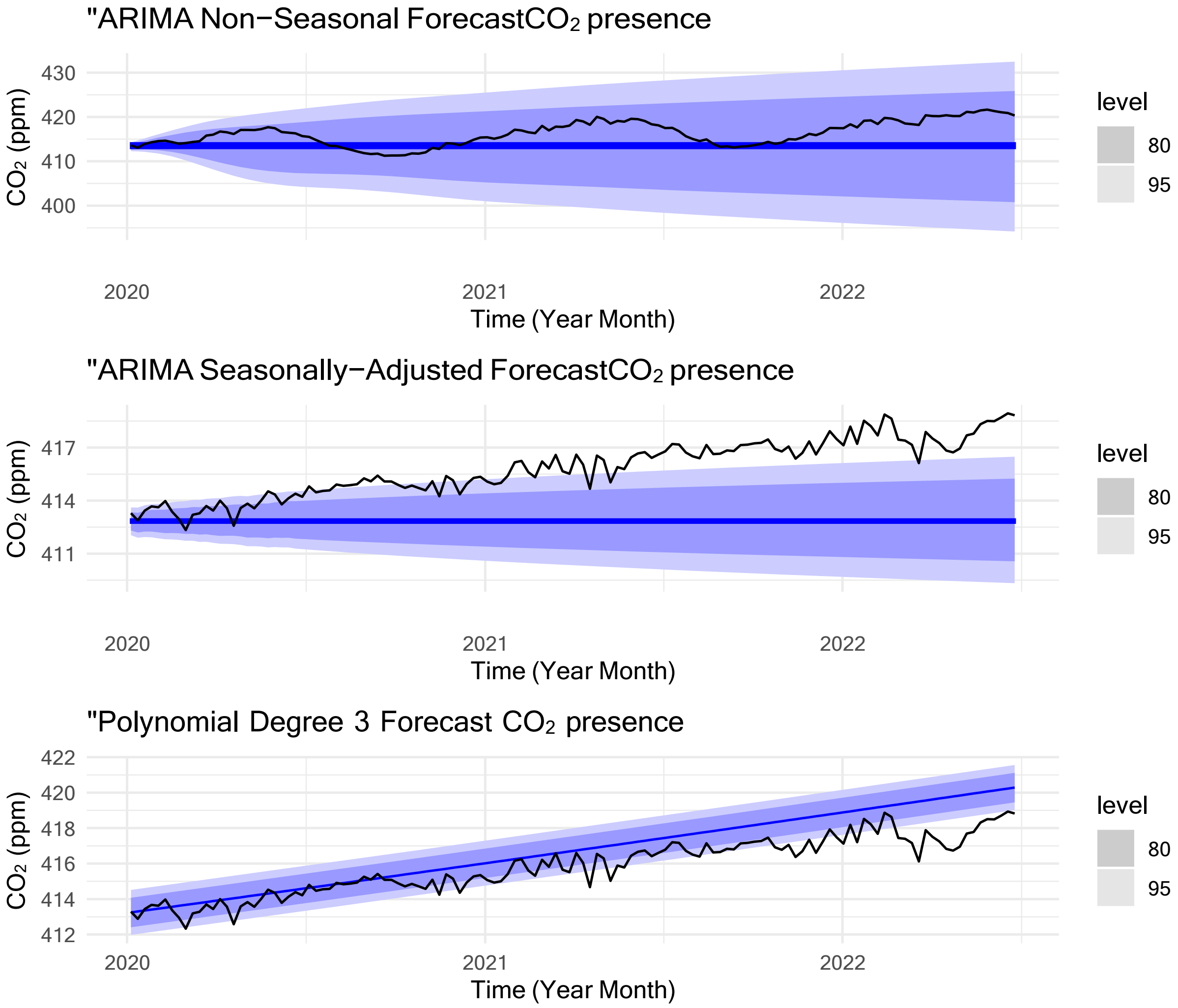
0.7 Task Part 6b: How bad could it get?

As in 1997, we generated predictions atmospheric CO_2 to estimate when it we might expect it to be at 420 ppm and 500 ppm levels for the first and final times as well as for atmospheric CO_2levels in the year 2122.

```r
poly_forecast <- forecast(poly.fit, level = c(95), h = (2200-1998)*12 + 6)
```

From our time series plot of historic and predicted values, we observe that our model's forecasted values continue to curve similar to the historic data and, as our 1997 polynomial model's forecasts, contains a narrow confidence interval.

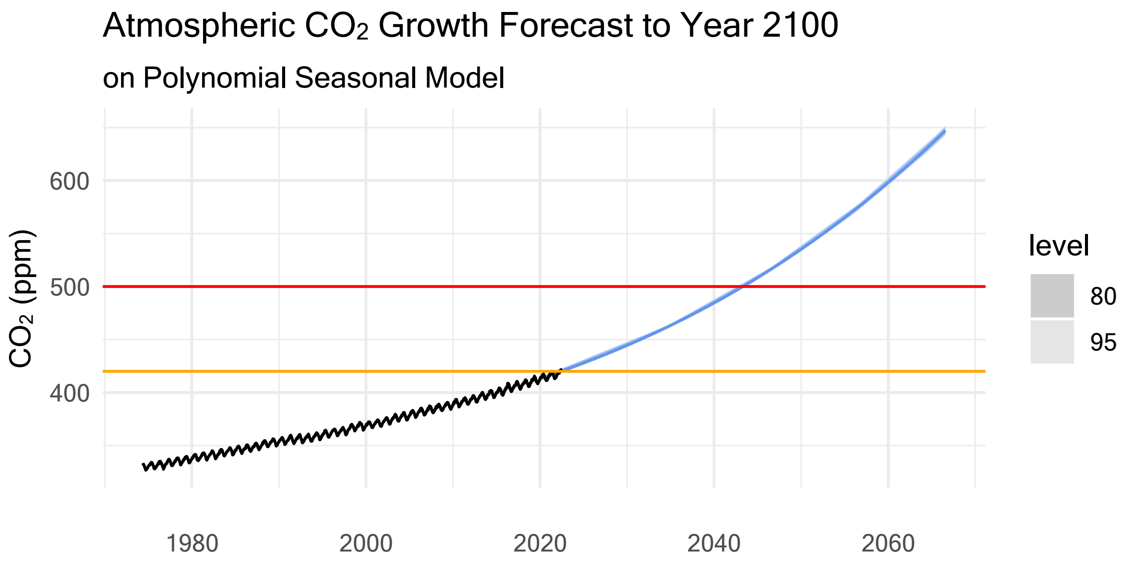
Table 7 specifically shows the month and year of the first and final times when the forecasted atmospheric CO_2 level reaches 420 ppm and 500 ppm respectively, in terms of the estimated mean, 80% confidence intervals, and 95% confidence intervals.


"ARIMA Non−Seasonal ForecastCO₂ presence

"ARIMA Seasonally−Adjusted ForecastCO₂ presence

"Polynomial Degree 3 Forecast CO₂ presence

Figure 9: Monthly CO_2 Concentration at Mauna Loa: Fitted vs Actual values


Atmospheric CO₂ Growth Forecast to Year 2100

on Polynomial Seasonal Model

Time (Year Month)

Figure 10: Atmospheric CO_2 Growth Forecast to Year 2100 based on Seasonal Polynomial Model.

Table 7: Forecasted First and Final Times of When Atmospheric CO_2 Reaching 420 ppm and 500pm on the Estimated Mean, 80% Confidence Intervals and 95% Confidence Intervals

Theshold Value	Estimated Mean	X80 Lower	X80 Upper	X95 Lower	X95 Upper
420ppm First Time	2022-03-20	2022-07-03	2021-12-12	2022-08-28	2021-10-17
420ppm Final Time	2022-07-17	2022-10-30	2022-04-03	2022-12-25	2022-02-13
500ppm First Time	2043-03-15	2043-06-07	2042-12-14	2043-07-26	2042-11-02
500ppm Final Time	2043-05-24	2043-08-16	2043-02-22	2043-10-04	2043-01-11

We based our forecasts on the estimated mean of the polynomial model and 95% confidence interval, including the 80% bounds for outer limits. First, the estimated atmospheric CO_2 concentration is expected to reach 420 ppm for the first time in March 2022 and for the final time in July 2022 or, at the 95% confidence interval as early as October 2021 until as late as December 2022. Second, we estimate that we will reach 500 ppm for the first time in March 2043 and for the last time May 2043 or, at the the 95% confidence interval in November 2042 until October 2043.

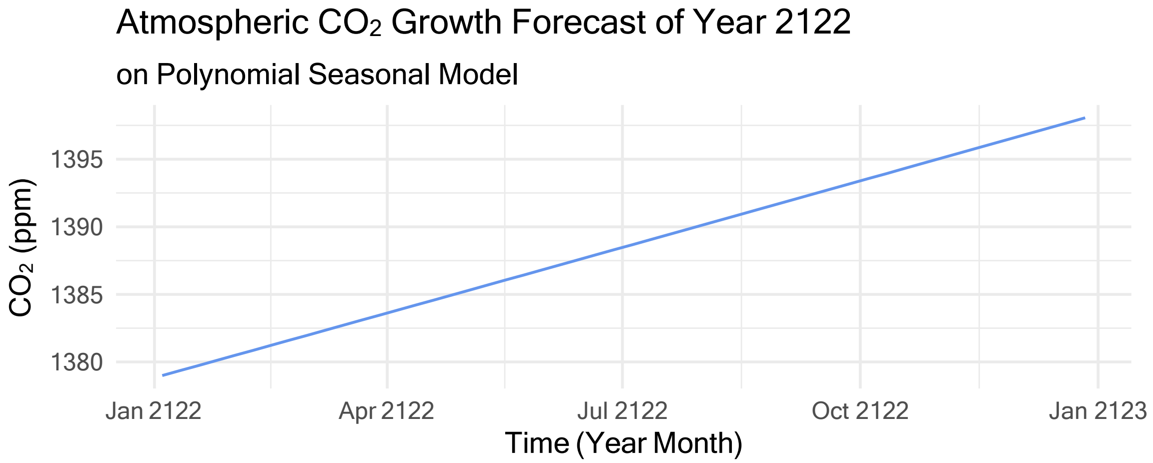


Figure 11: Atmospheric CO_2 Growth Forecast of Year 2100 based on the Seasonal Polynomial Model.

Finally, Figure 11 illustrates the projected values for atmospheric CO_2 over the year of 2122. In prior forecasts, we expect 420 ppm to last about four months in observations and 500 ppm would last about two months. For year 2122, atmospheric CO_2 are predicted to be well above 1000 ppm, ranging from 1379 ppm to 1398 ppm and increasing by 19 ppm in the single year.

0.8 Conclusion:

In this investigation, we sought to answer whether models previously fit able to accurately predict for the future and are they still the best at capturing the behaviors of CO_2 concentration. Here, we assessed models created in 1997 given realized observations of the last 25 years as well as forecasts generated from them. We also fitted a new model given this new data. From our investigation of the time series of $C0_2$ presence in the atmosphere, we found that the time series exhibits a curved trend with regular seasonality. In assessing 1997 models it was found that, counter to prior findings, the polynomial model was more performant than the ARIMA model on the realized data. This illustrated how the ARIMA would be more successful on the fitted data but less successful on newly realized data than the polynomial model. In general, the ARIMA under fit and predicted lower values, not exhibiting a distinct curve present in the long term data. Having a trend and seasonal factors incorporated as well as containing a trend of degree 2 allowed the polynomial model to fit the long term data well. This same result appeared again in comparing the second degree linear model to the two ARIMA models. This time, we leveraged RMSE for selecting one of several polynomial models of different degrees as well, selecting a linear model with degree 3 this time.

Because of these features, in comparison to observed values, we found that our 1997 ARIMA model produced estimates that were too conservative and unable to accurately predict for the future. Our previous estimates for CO_2 level reaching 420 ppm for the first time was 2031 yet the value was met 11 years earlier than predicted. The atmospheric CO_2 was at 313 ppm in 1958 and has recently reached 420 ppm in the current years - the highest in human history. Moreover, the polynomial model predicts a more aggressive upward trend in atmospheric CO_2 growth than did our 1997 ARIMA model and was better fit to future data. We see this trend reflected in our projection of the year 2122 with a projected increase of 19ppm for the year. The study suggests continuing atmospheric CO_2 monitoring to continue to compare forecasted and realized values to help assess the model's quality and observe whether these observed trends persist or change over time.