Study of Global CO_2 Emissions

Abstract

In this report, we study the concentrations of CO2 levels in the atmosphere. While early studies suggest that CO2 concentration levels are variable and hard to forecast, we will study datasets collected over years to determine if there are discernable patterns.Studying CO2 levels can help us assess health risks to the populations, wildlife, and ecosystem of different areas, and could also help us determine acitivities that adversely affect these levels.

Report From the Point of View of 1997

Introduction.

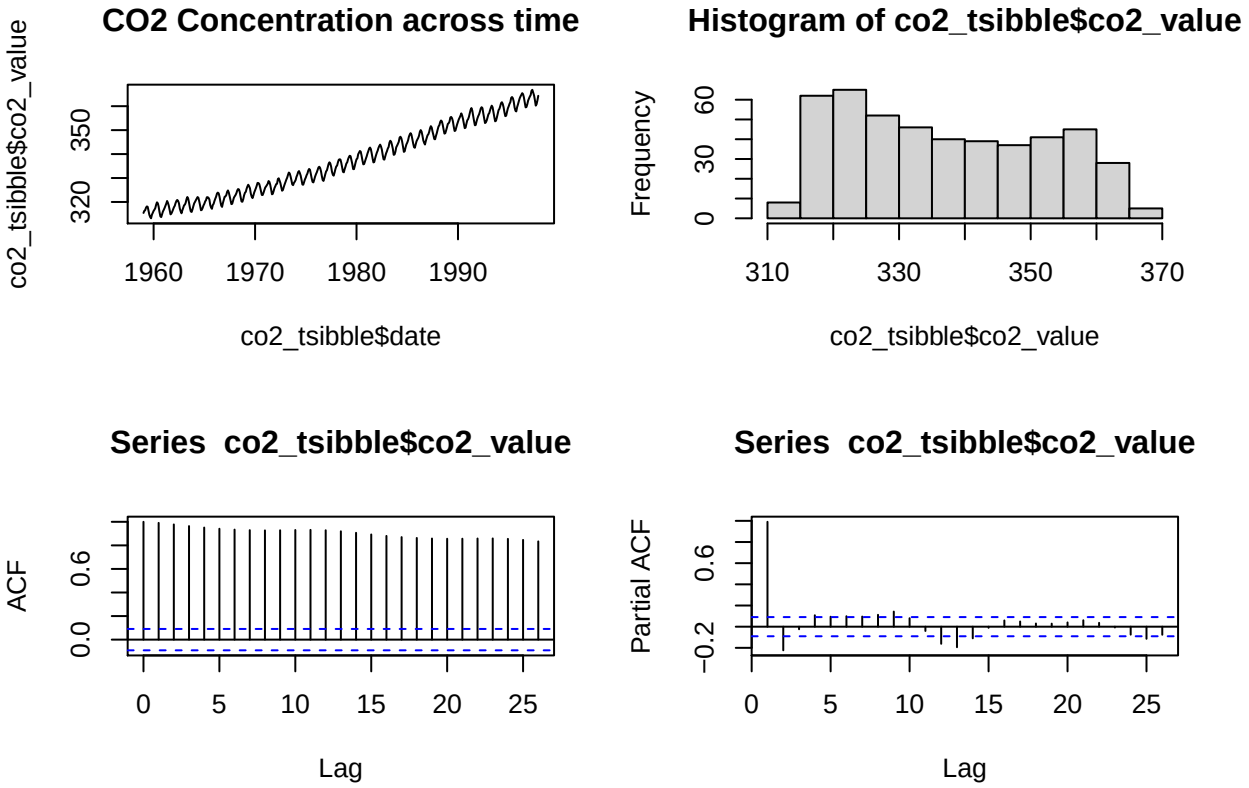
In this report we attempt to answer the following research question: Can we discover trend or seasonality factors that we can use to forecast global CO2 levels accurately?

Understanding CO2 levels can he helpful in assessing health risks for a population, wildlife or ecosystem in a given area. Additionally, we could study CO2 levels in different areas and over different conditions to attempt to find the cause of changes in CO2 levels, potential activities such as burning fossil fuels, deforestation, among others.

It has been widely reported that the variability of CO2 levels make them hard to predict. In this report, we study measurements of CO2 levels in different areas, across years, and show that definitive patterns emerge that allow us to perform forecasts more accurate than previously conducted studies.

CO2 Data

The CO2 dataset used in this report consists of 468 observations, collected monthly from 1959 to 1997, of CO2 concentration expressed in parts per million (p.p.m.) in the preliminary 1997 SIO manometric mole fraction scale. The readings where collected in the Mauna Loa Observatory in Hawaii. The measurements were taken with a continuous gas analyzer, consisting of a thermostated cell, an optical system, and an electronic amplifier, manufactured by the Applied Physics Corporation. C.D. Keeling collected this data to determine whether we could detect patterns in CO2 concentration levels from potentially more accurate atmospheric measurements. Keeling was skeptical of previous research that used data collected with chemical measurements of CO2 observations to claim that CO2 levels are too variable to forecast, and decided to collect data with state-of-the art atmospheric measurement equipment.



We can see that the CO2 values in the dataset seem reasonable, from 313 to 367, and there are no missing values. The time series plot shows both a clear upwards trend and a level of seasonality very few months in terms of CO2 concentration levels. The histogram shows a slight right skew in the CO2 values. The ACF plot shows a high correlation for close lags and slow decay over time , similar to an AR process. But there's the oscillating behavior in the graph, corroborating the seasonality seen in the time series plot. The PACF plot shows a spike in lag 1, but a sinusoidal pattern with statistical significance at larger lags (2, 4, 12, 13, etc.). The initial exploration of the data suggests exploring an AR process, with at least one order difference to detrend, and a seasonal component.

Linear Time Trend Model

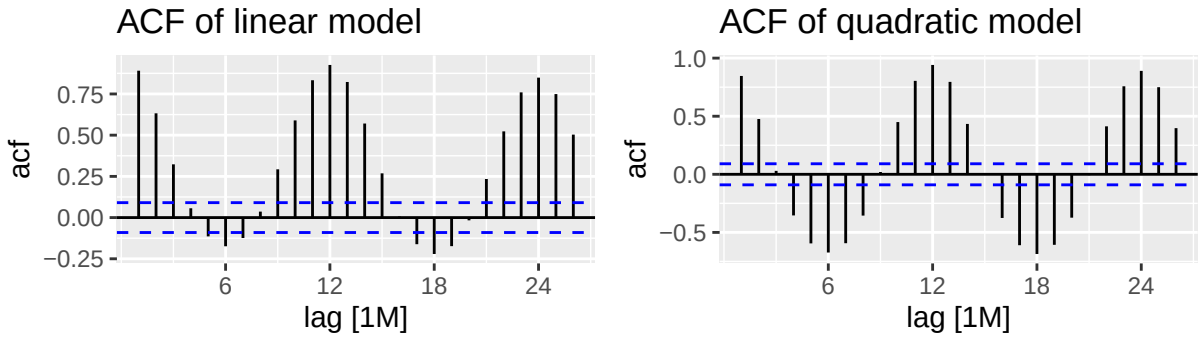
```
linear_model = co2_tsibble %>%
  model(trend_model = TSLM(co2_value ~ trend()))

linear_model %>% report()
```

```
## Series: co2_value
## Model: TSLM
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6.04  -1.95   0.00   1.91   6.51
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.12e+02   2.42e-01    1285   <2e-16 ***
## trend()     1.09e-01   8.96e-04     122   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.62 on 466 degrees of freedom
## Multiple R-squared: 0.969,   Adjusted R-squared: 0.969
## F-statistic: 1.48e+04 on 1 and 466 DF, p-value: <2e-16
```

```
quadratic_model = co2_tsibble %>%
  model(trend_model = TSLM(co2_value ~ I(trend()) + I(trend()^2)))

quadratic_model %>% report()
```

```
## Series: co2_value
## Model: TSLM
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5.02  -1.71   0.21   1.80   4.83
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.15e+02   3.04e-01  1035.7   <2e-16 ***
## I(trend())   6.74e-02   2.99e-03    22.5   <2e-16 ***
## I(trend()^2) 8.86e-05   6.18e-06    14.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.18 on 465 degrees of freedom
## Multiple R-squared: 0.979,   Adjusted R-squared: 0.979
## F-statistic: 1.07e+04 on 2 and 465 DF, p-value: <2e-16
```



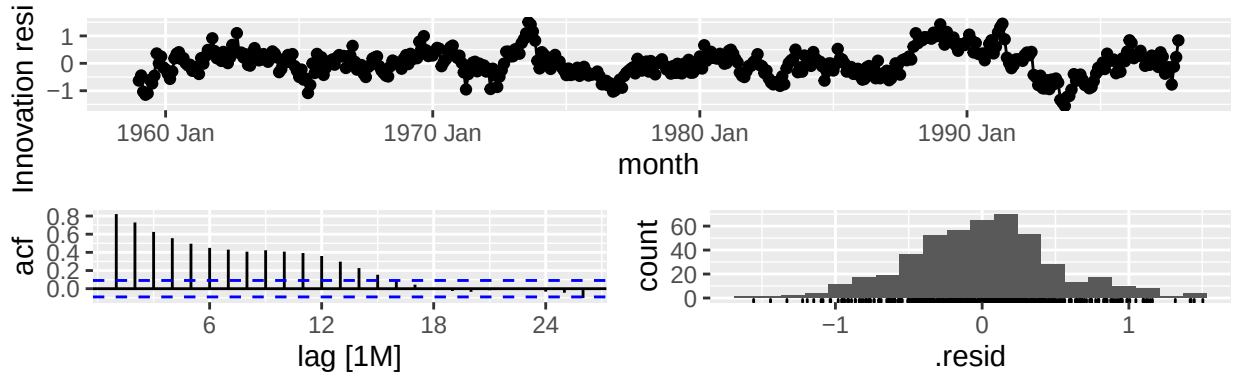
Fitting a linear model on the data and examining the residuals shows that the model is a poor fit for the dataset. There are clear autocorrelations in the ACF plot, with a seasonal pattern. The quadratic model is not a significant improvement over the linear model, apart from a slight increase in the adjusted R-square score (0.979 vs 0.969) and smaller residual standard error (2.18 vs 2.62), as we can see clear autocorrelations in the residuals in the ACF, still showing a seasonal pattern that's not modeled properly.

A logarithmic transformation of the data won't be of too much help in this case, because neither the data grows exponentially (the range is 315 to 364), nor the variance significantly changes over time, as we can see from the plot of the CO2 values over time.

We now explore the use of a polynomial model with seasonal dummy variables.

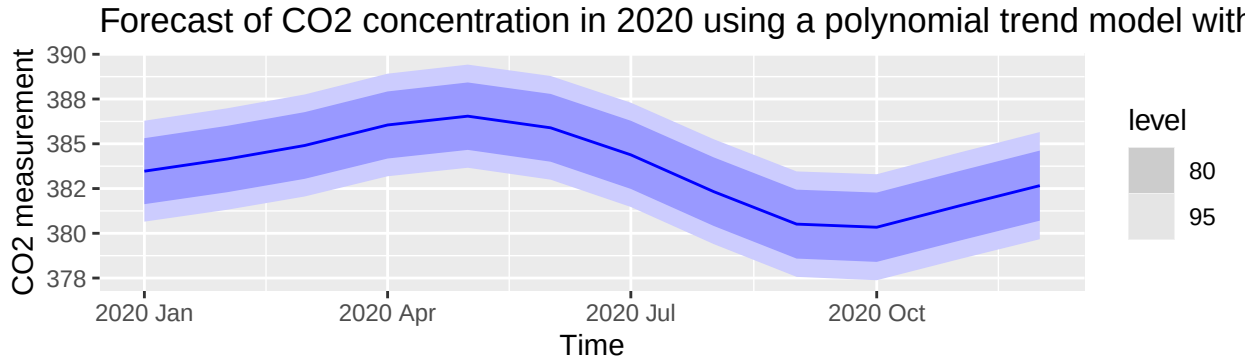
```r
polynomial_model = co2_tsibble %>%
  model(trend_model = TSLM(co2_value ~ I(trend()^1) + I(trend()^2) + I(trend()^3) +
                                       season()))

polynomial_model %>% report()
```

```
## Series: co2_value
## Model: TSLM
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.557 -0.331  0.001  0.288  1.504
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.16e+02   1.21e-01 2611.63  < 2e-16 ***
## I(trend()^1)    3.28e-02   1.74e-03   18.83  < 2e-16 ***
## I(trend()^2)    2.74e-04   8.61e-06   31.85  < 2e-16 ***
## I(trend()^3)   -2.64e-07   1.21e-08  -21.86  < 2e-16 ***
## season()year2   6.70e-01   1.14e-01    5.85  9.3e-09 ***
## season()year3   1.42e+00   1.14e-01   12.39  < 2e-16 ***
## season()year4   2.56e+00   1.14e-01   22.32  < 2e-16 ***
## season()year5   3.04e+00   1.15e-01   26.55  < 2e-16 ***
## season()year6   2.38e+00   1.15e-01   20.81  < 2e-16 ***
## season()year7   8.68e-01   1.15e-01    7.58  2.0e-13 ***
## season()year8  -1.19e+00   1.15e-01  -10.43  < 2e-16 ***
## season()year9  -3.01e+00   1.15e-01  -26.31  < 2e-16 ***
## season()year10 -3.19e+00   1.15e-01  -27.86  < 2e-16 ***
## season()year11 -2.00e+00   1.15e-01  -17.43  < 2e-16 ***
## season()year12 -8.74e-01   1.15e-01   -7.63  1.4e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.506 on 453 degrees of freedom
## Multiple R-squared: 0.999,   Adjusted R-squared: 0.999
## F-statistic: 2.92e+04 on 14 and 453 DF, p-value: <2e-16
```



The polynomial model with a trend degree equal to 3 and seasonal dummy variables improves the previous trend models in terms of residual standard errors (0.506 versus 2.62 and 2.18), and adjusted R-squared of 0.999. Larger polynomial degrees do not significantly improve these metrics, and all of its coefficients are statistically significant. The ACF plot shows lower autocorrelations between the residuals than in the linear

and quadratic models, but still large enough to state that the residuals do not behave like white noise. In comparison with the previous models, we have successfully modeled the seasonal pattern, and now there's only autocorrelations for low orders in the residuals.



This model suggests an uptrend in CO2 values after 1997. In particular, the point estimates for 2020 will be 383 in January 2020 and December 2020, with a range from 380 to 387 throughout the year.

ARIMA time series model

We explore different ARIMA models to predict levels of CO2 concentration. From the initial EDA, we can see that there are strong trend and seasonal patterns, indicating that we should consider differencing and seasonal adjustments in our ARIMA models, along with AR and MA orders.

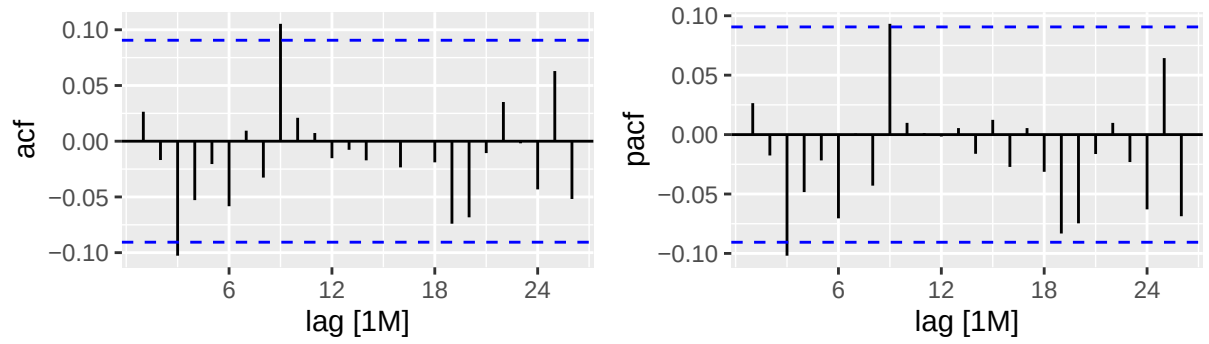
```
automated.arima.model.bic<-co2_tsibble %>%
  model(ARIMA(co2_value ~ 0 + pdq(0:10,0:2,0:10) + PDQ(0:10,0:2,0:10), ic="bic", stepwise=F, greedy=F))
```

We invoked the ARIMA automated model selection mechanism with the BIC metric as the criteria, and obtained the model shown below. As expected, there is a differencing component included in the model, to apply detrending to the upward trend of CO2 concentration levels. Also, there are seasonal orders for AR, MA and differencing, to account for the sinusoidal behavior that we observed in the time series.

```
arima.model.bic<-co2_tsibble %>%
  model(ARIMA(co2_value ~ pdq(0,1,1) + PDQ(1,1,02), ic="bic", stepwise=F, greedy=F))

arima.model.bic %>% report()
```

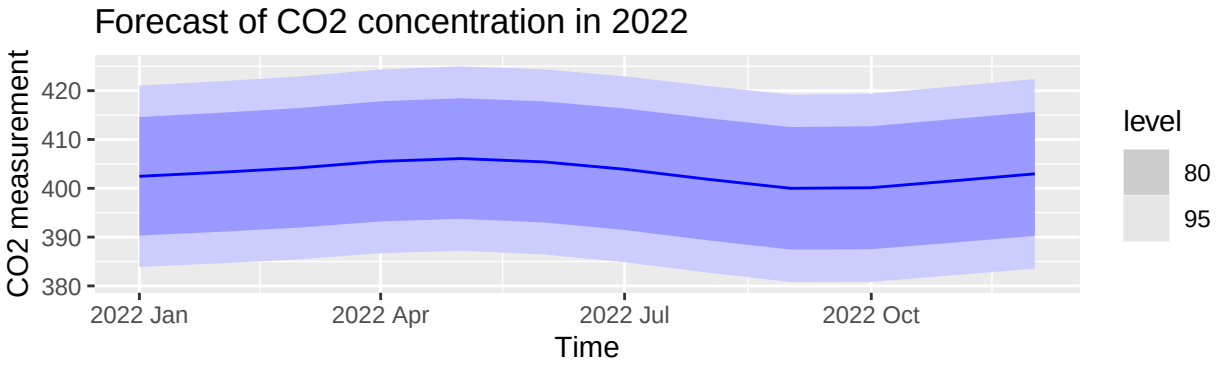
```
## Series: co2_value
## Model: ARIMA(0,1,1)(1,1,2)[12]
##
## Coefficients:
##           ma1    sar1    sma1    sma2
##       -0.3482  -0.499  -0.316  -0.464
## s.e.   0.0499   0.528   0.516   0.437
##
## sigma^2 estimated as 0.08603:  log likelihood=-85.6
## AIC=181   AICc=181   BIC=202
```

We now examine the residuals to determine if the model is a proper fit.



The ACF and PACF plots show that the residuals bahve roughly like white noise, as all but two of the autocorrelations are barely statistically significant. With a 95% confidence, we would expect up to 5% of the autocorrelations to be statistically significant to consider the residuals independent.

We have run and passed Ljung-Box tests for lags 1-10, results are not shown, providing further formal evidence that the residuals behave like white noise.



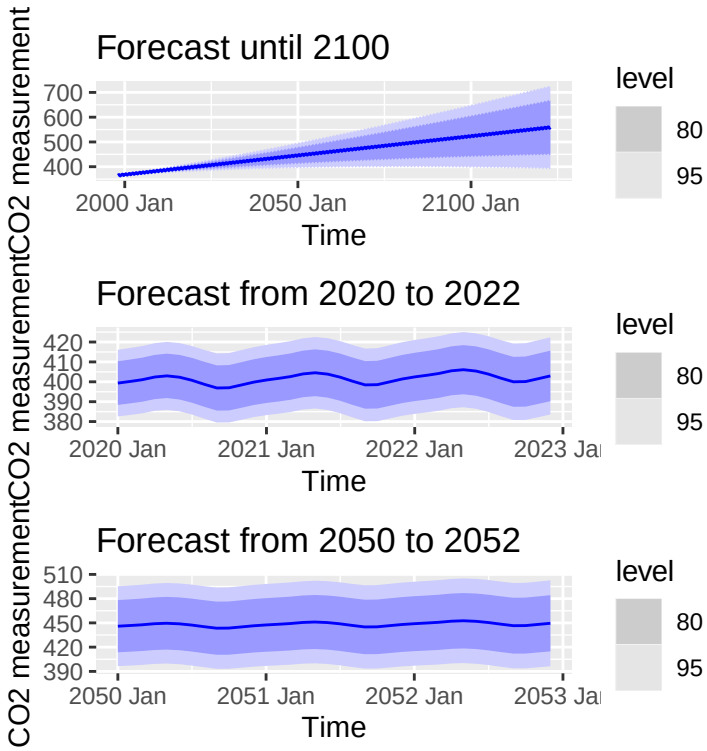
We can see that the forecast for year 2022 using the ARIMA model produces a point estimate of 402 for January 2022 and 403 for December 2022, and estimates as low as 400 and as high as 406 throughout the year. In comparison with the linear model estimated in a previous section, which shows a stagnant estimate around 2020, the ARIMA model seems to produce a prediction that follows the consistent upward trend of CO2 levels seen in previous years.

Forecasting Atmospheric CO2 Growth

We will use our ARIMA model to forecast different levels of CO2 concentration.

Our model predicts that CO2 levels might reach 420 ppm for the first time , with a 95% prediction interval, in May 2020. The upper level of the 95% prediction interval drops below 420 ppm for the last time in October 2022. The model predicts that the estimate will reach 420 for the first time in May 2031, and go below 420 for the last time in October 2034.

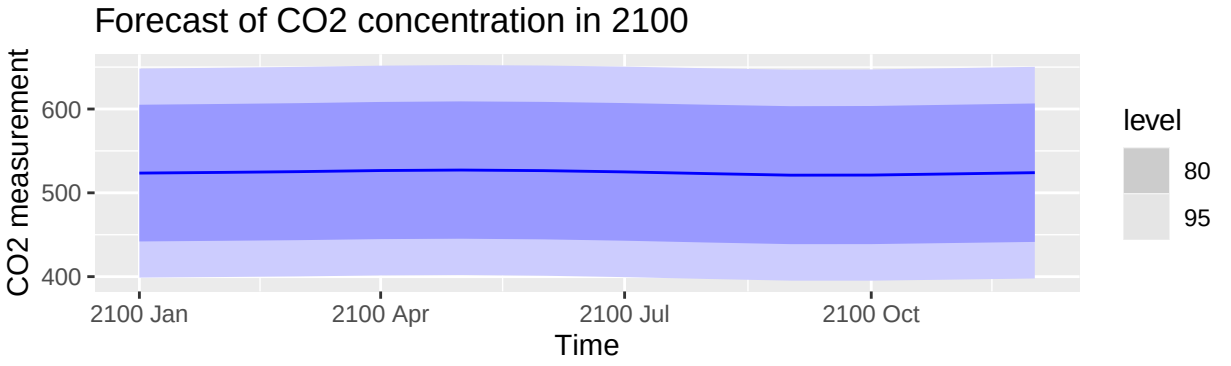
Below is an exploration of when CO2 levels might reach 420 ppm and 500 ppm. We have zoomed into the forecast of 2020-2022 and 2050-2052, which is where the 95% prediction intervals include and exclude the 420 and 500 ppm levels, respectively. We can see the estimates reaching these levels in the forecast until 2100 plot.



Our model predicts that CO2 levels might reach 420 ppm for the first time , with a 95% prediction interval, in May 2020. The upper level of the 95% prediction interval drops below 420 ppm for the last time in October 2022. The model predicts that the estimate will reach 420 for the first time in May 2031, and go below 420 for the last time in October 2034.

Our model predicts that CO2 levels might reach 500 ppm for the first time , with a 95% prediction interval, in March 2051. The upper level of the 95% prediction interval drops below 500 ppm for the last time in August 2052. The model predicts that the estimate will reach 500 ppm for the first time in April 2083, and go below 500 ppm for the last time in October 2086.

We now produce forecasts of concentration levels far into the future, for the year 2100.



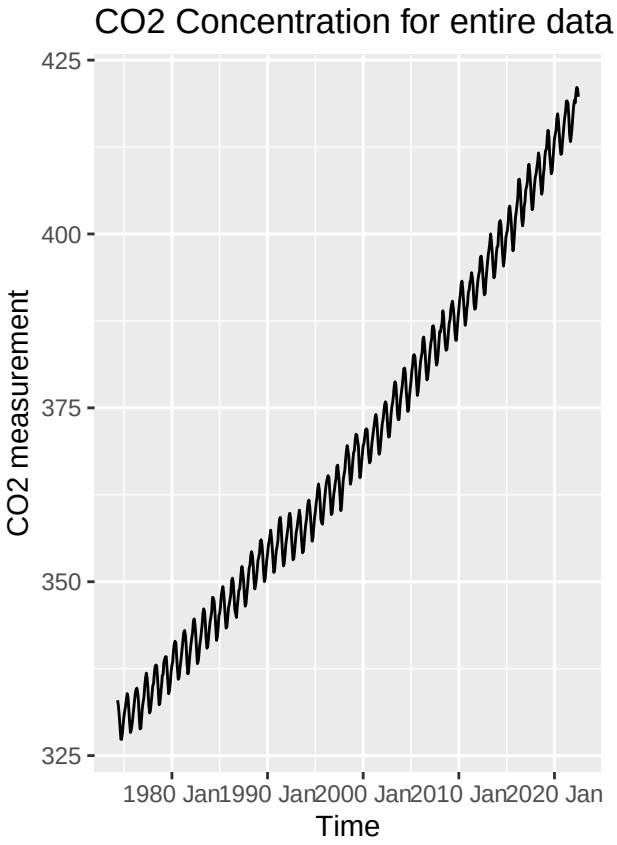
The CO2 levels predicted by the ARIMA model in 2100 range from 521 to 527, January and December being both 524 ppm. It's important to note that that these predictions are so far into the future, and our model has lags of order less or equal to two months, making it unlikely that the predictions of CO2 levels one hundred years away will be highly accurate. We could be more confident in our model either by forecasting periods closer to 1997 or evaluating the model as more data becomes available throughout the years.

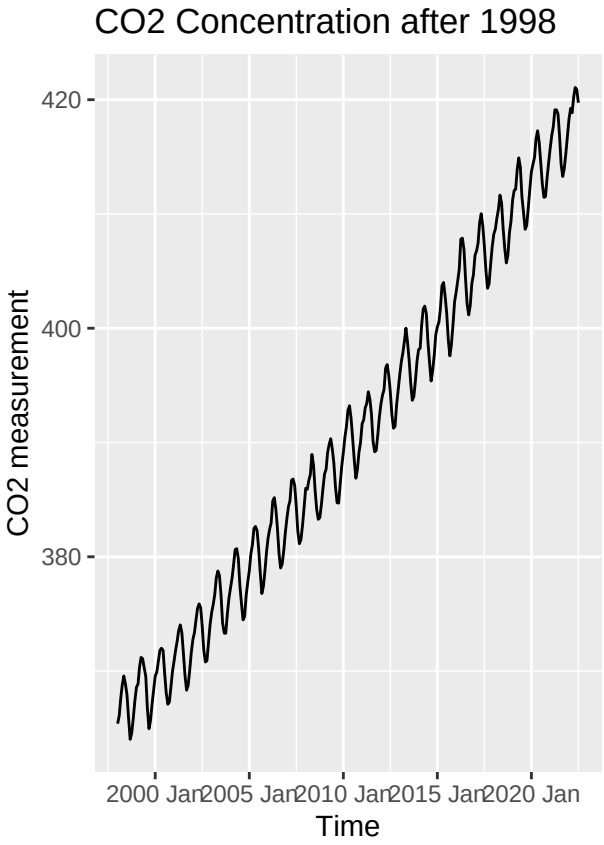
Report from the Point of View of the Present

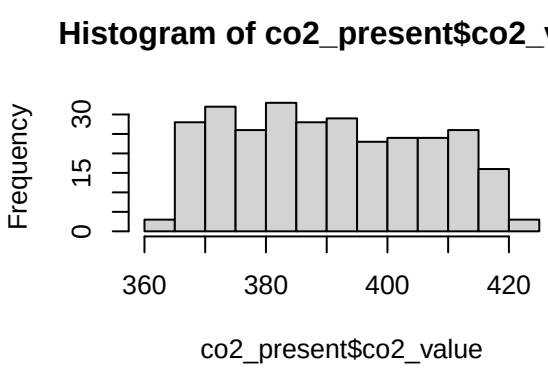
In this part of the project, we will be using CO2 data that is readily available in NOAA/GML to answer the question of whether there are any trends or seasonality factors that we can use to forecast global CO2 levels accurately. NOAA/GLM uses state of the art technology to analyze CO2 and continuously improving. In April of 2019, a new CO2 analyzer was installed at Mauna Loa that uses a technique called Cavity Ring-Down Spectroscopy (CRDS).

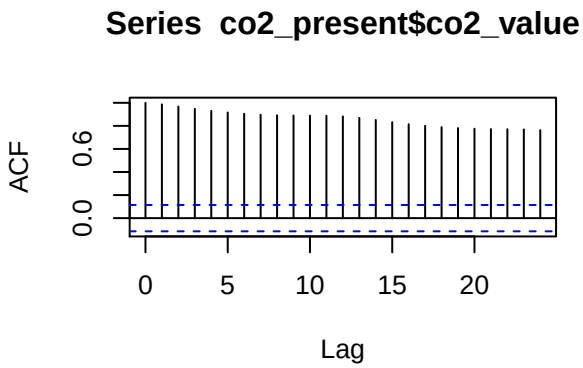
We have acquired the full data all the way from 1974 to current day. We will, however, only use the data after 1998 to current day. First, we will perform a full EDA on the data set after 1998 to current date. Then we will be compare the realized CO2 data to the forecasted CO2 level.

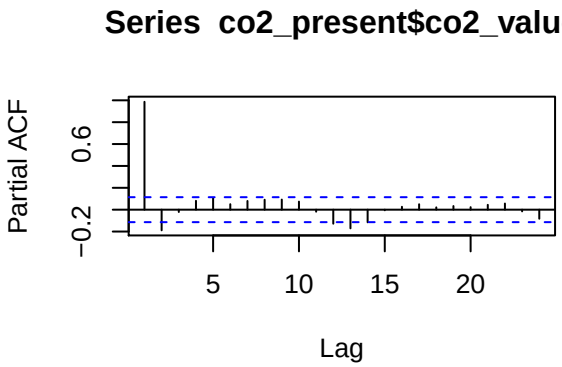
The weekly data that we pulled from the website included 2511 observation with 9 variables. We got rid of the columns that we did not need for this analysis and extracted a subset of the data after 1998 that left us with 1279 observations. We noticed that about 4 weekly values had a nonsensical value of "-1000," so we decided to get rid of those and aggregate the data by month. We then casted the dataframe to a tsibble.









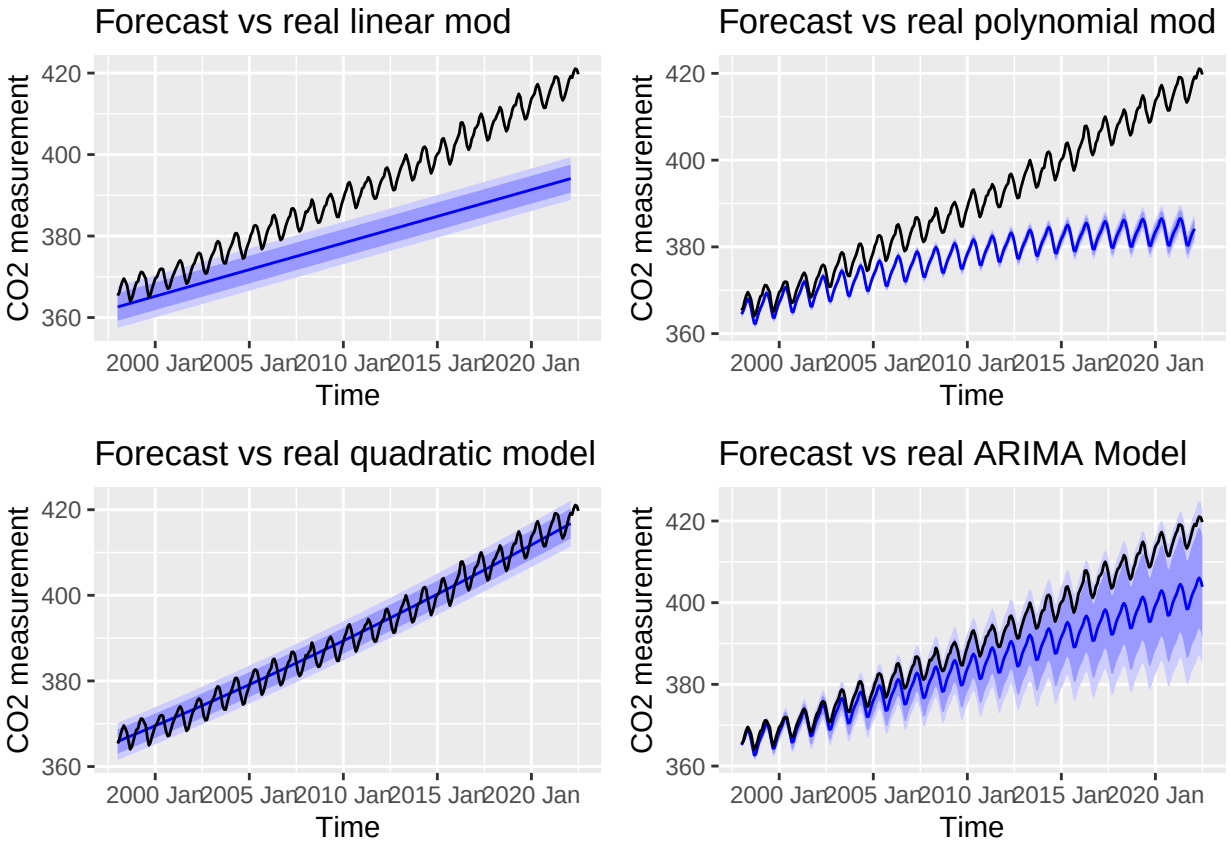


As we can see in the graphs, the CO2 measurements are still showing a upward trend with a seasonal pattern for the years after 1998. For the data after 1998,there are no missing values. The histogram does not really tell us much. The ACF plot shows a slow decay with some oscillations. The PACF graphs shows a spike in lag 1 as expected but then a sinusoidal pattern with statistical significance at different lags.

We will, now, use the 4 generated models to compare their forecast to the realized CO2 data.

First let's look at the linear forecast using the model that was generated using the data until 1997. As we can see in graph, the model is underestimating the realized CO2. The realized CO2 in year 2022 is around 420, where the linear model forecasts it to be around in a range between 393-397.

The polynomial model appears to flatten after 2015, and it is underestimating the realized CO2. The quadratic model appears to be performing very well. The forecast seems to be following the available realized CO2 quite closely. Looking at the forecast made by the ARIMA model generated, we can see that ARIMA model forecast is much better than both the linear model and the polynomial model forecasts. The ARIMA model seems to be more confident in its forecast closer to 1997, but as the years pass by, the forecast's confidence interval widens. Comparing the model forecast to the realized CO2 values, we can see the realized CO2 rises much sharper than the model forecasts. The wide confidence interval towards the later years of the forecast tries to capture this, however, the realized CO2 towards the later years lie slightly outside the confidence interval. The quadratic model appears to be outperforming the ARIMA model. The ARIMA model had predicted that CO2 levels might reach 420 ppm for the first time in May 2020 with a confidence interval of 95%. The realized CO2 first hit 420 ppm in April 2022. The model seems to have performed relatively well in predicting when 420 ppm will be reached.

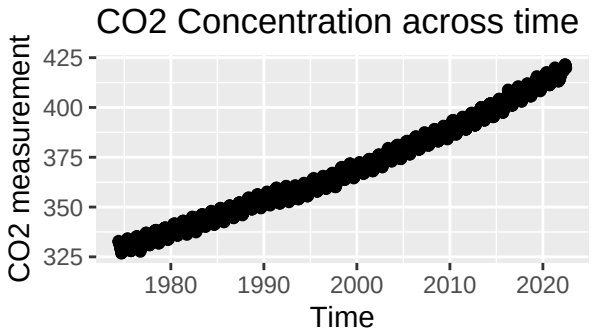


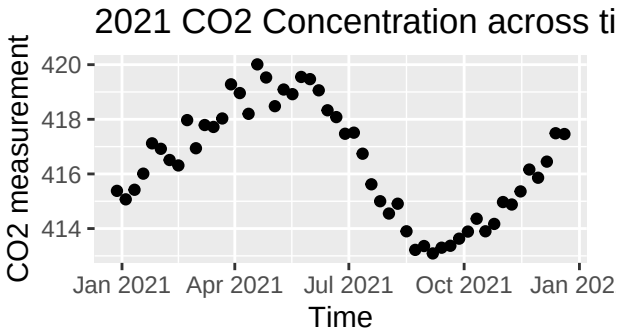
Next let's look at the accuracy of the four models. As expected, the root mean squared error for the the ARIMA model is lower than both the linear model as well as the polynomial model, indicating that ARIMA model did a better job at predicting CO2 levels. As for the linear model, it appears to be performing better than the polynomial model. Interestingly, comparing the quadratic model to the ARIMA model, we get a better fit with a RMSE of 2.28 for quadratic model and 7.91 for ARIMA model. This indicates that our ARIMA model has more room to improve. In the next section we will try to improve ARIMA model.

```
##   model_type      ME  RMSE   MAE     MPE  MAPE MASE RMSSE  ACF1
## 1     linear 11.7821 13.45 11.78 2.96138 2.961  NaN   NaN 0.969
## 2 Polynomial 13.1375 16.38 13.14 3.27732 3.277  NaN   NaN 0.988
## 3      ARIMA  6.4428  7.91  6.44 1.60618 1.606  NaN   NaN 0.986
## 4  quadratic  0.0484  2.28  1.93 0.00513 0.494  NaN   NaN 0.836
```

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

Repull the data to get weekly data

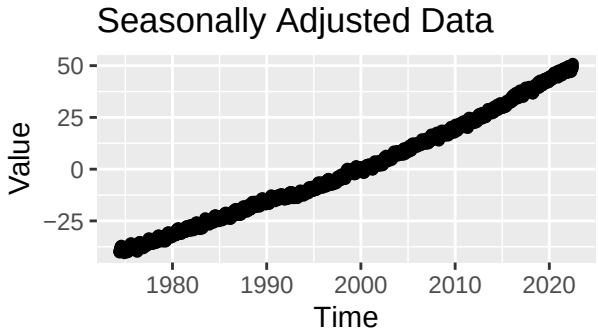


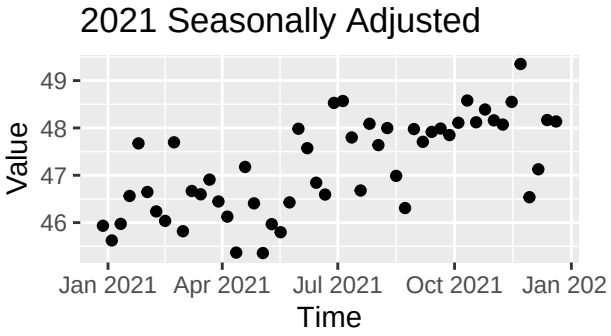


- Graphically, the weekly CO2 data follows a monthly seasonal trend.

seasonally adjust the data

```
co2_sa = co2_present_full
sea_mod = lm(co2_value ~ feb + mar + apr + may + jun + jul + aug + sep + oct + nov + dec, co2_present_fu
co2_sa$co2_value_sa = sea_mod$residuals
```

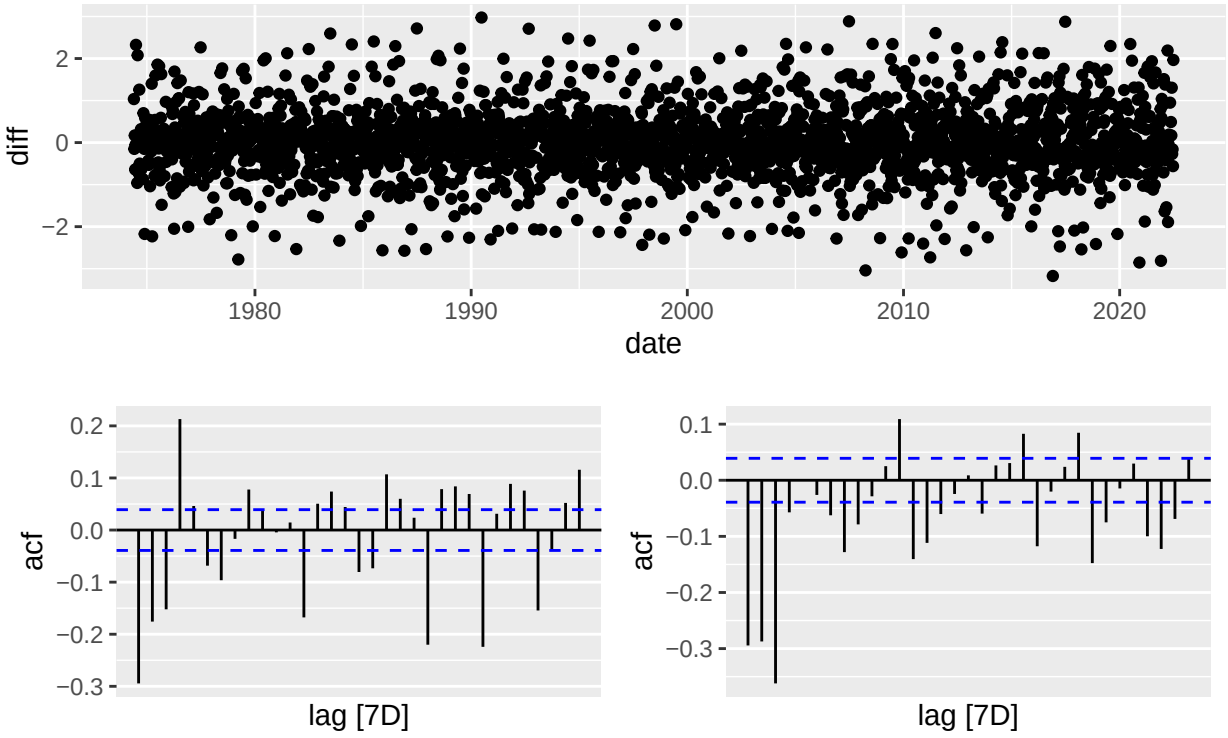




- Once we control for monthly seasonality using month dummy variables, we have only time trend left.

manually select best arima for seasonally adjusted data

```
data <- co2_sa %>%
  mutate(diff = difference(co2_value_sa))
data <- na.omit(data)
data <- as_tsibble(data, index = date, regular = TRUE)
```



- After differencing the data, it looks like there is still significant lag at 30. Therefore we will include a lag of 26 for the ARIMA model, because 26 weeks are a half of year.

manually select the best polynomial time trend model for seasonally adjusted data

```
linear_models = co2_sa %>%
  model(model_poly1 = TSLM(co2_value_sa ~ trend()),
        model_poly2 = TSLM(co2_value_sa ~ I(trend()) + I(trend()^2)),
        model_poly3 = TSLM(co2_value_sa ~ I(trend()^1) + I(trend()^2) + I(trend()^3))
        )

linear_models %>% report
```

```
## # A tibble: 3 x 15
##   .model    r_squared adj_r_squared sigma2 statistic p_value    df log_lik   AIC
##   <chr>         <dbl>         <dbl>  <dbl>     <dbl>   <dbl> <int>   <dbl> <dbl>
## 1 model_po~     0.989         0.989  6.85    232834.       0     2  -5978. 4836.
## 2 model_po~     0.998         0.998  1.15    697736.       0     3  -3740.  363.
## 3 model_po~     0.999         0.999  0.864   621002.       0     4  -3378. -361.
## # ... with 6 more variables: AICc <dbl>, BIC <dbl>, CV <dbl>, deviance <dbl>,
## #   df.residual <int>, rank <int>
```

- Polynomial with order of 2 has the lowest **non-negative** BIC score, therefore we will use it as our polynomial time-trend model to seasonal adjusted series.

splitting seasonal data into test and training set

```
# split seasonal adjusted data
dat <- co2_sa %>% select(c('date','co2_value_sa'))

test.size<-104 # last two years 2*52 = 104 weeks

#split training and test
dat.train<-dat %>%
  slice(1:(n()-test.size))

dat.test<-dat %>%
  slice((n()-test.size+1):n())
```

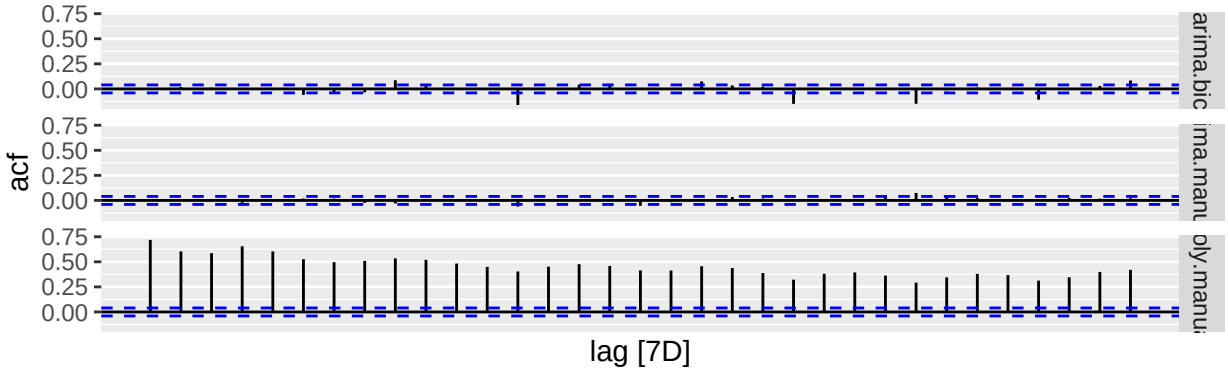
in sample bic comparisons

```
model.comp<-dat.train %>%
  model(arima.bic=ARIMA(co2_value_sa ~ 1 + pdq(0:30,0:2,0:30) + PDQ(0,0,0), ic="bic", stepwise=F, greedy
        arima.manual=ARIMA(co2_value_sa ~ 1 + pdq(26,1,1) + PDQ(0,0,0), ic="bic", stepwise=F, greedy=F)
        poly.manual = TSLM(co2_value_sa ~ I(trend()) + I(trend()^2))
        )

model.comp %>% report
```

```
## # A tibble: 3 x 17
##   .model       sigma2 log_lik   AIC  AICc   BIC ar_roots   ma_roots  r_squared
##   <chr>         <dbl>   <dbl> <dbl> <dbl> <dbl> <list>     <list>        <dbl>
## 1 arima.bic     0.451  -2453. 4920. 4921. 4961. <cpl [4]>  <cpl [1]>    NA
## 2 arima.manual  0.403  -2309. 4675. 4676. 4843. <cpl [26]> <cpl [1]>    NA
## 3 poly.manual   1.14   -3575.  327.  327.  351. <NULL>     <NULL>        0.998
## # ... with 8 more variables: adj_r_squared <dbl>, statistic <dbl>,
## #   p_value <dbl>, df <int>, CV <dbl>, deviance <dbl>, df.residual <int>,
## #   rank <int>
```

- In terms of BIC score, polynomial time trend data actually performs significantly better on in sample data compare to both automatically selected, and manually selected ARIMA data.

in sample residual comparisons




```
## [1] "arima.bic Ljung-Box test"
```

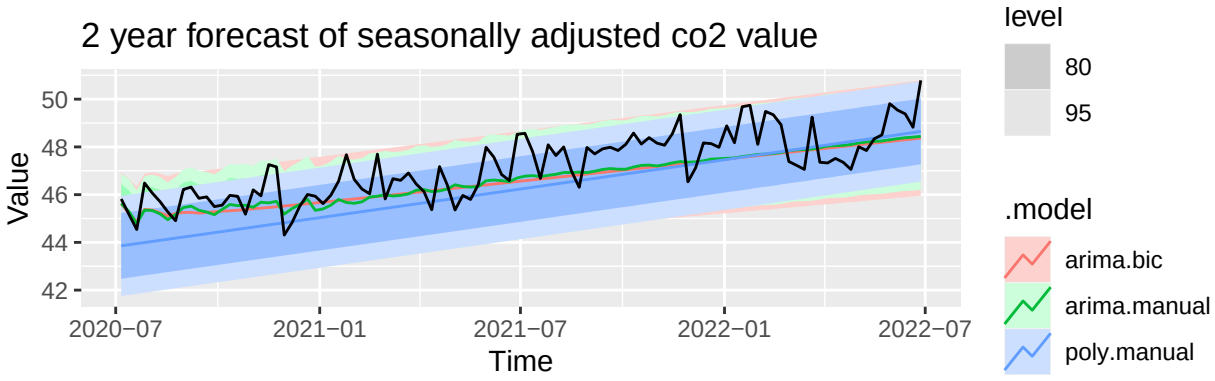
```
##
```

```
##  Box-Ljung test
##
## data:  .
## X-squared = 39, df = 10, p-value = 2e-05

## [1] "arima.manual Ljung-Box test"

##
##  Box-Ljung test
##
## data:  .
## X-squared = 7, df = 10, p-value = 0.7

## [1] "poly.manual Ljung-Box test"

##
##  Box-Ljung test
##
## data:  .
## X-squared = 8099, df = 10, p-value <2e-16
```

- In contrast when looking at in-sample residuals and each model's Ljung-Box test, only the residual from manually selected ARIMA appears to be stationary.

test sample forecast comparisons

```
## Plot variable not specified, automatically selected `.vars = co2_value_sa`
```

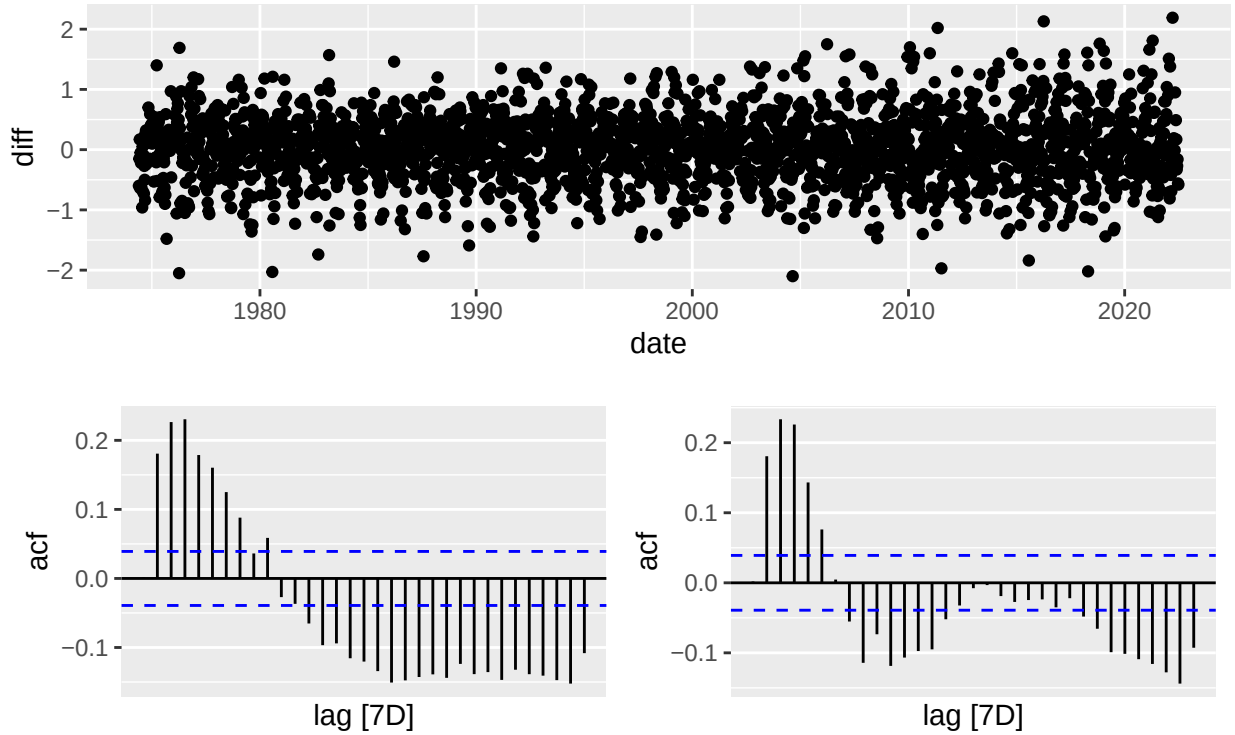



```
## # A tibble: 3 x 10
##   .model       .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
##   <chr>        <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima.bic    Test  0.569 0.990 0.812  1.17  1.70   NaN   NaN 0.338
## 2 arima.manual Test  0.524 0.945 0.781  1.08  1.64   NaN   NaN 0.333
## 3 poly.manual  Test  0.933 1.28  1.12   1.97  2.36   NaN   NaN 0.426
```

- Now looking at forecast on the test samples,
  - Polynomial tend to underestimate the co2 value in the earlier weeks, but it gets better at later weeks. This suggest as time go further out, it may start overestimating values.
  - While BIC-selected and manually-selected Arima shows similar trend, the manually-selected model better captures the fluctuation within the data due to using more lagging terms. However, the effect of lagging term will diminish overtime. Therefore for longer term forecast manually-selected ARIMA may not do a better job than BIC-selected ARIMA.

manually select best arima for non-seasonally adjusted data

```
data <- co2_present_full %>%
  mutate(diff = difference(co2_value))
data <- na.omit(data)
data <- as_tsibble(data, index = date, regular = TRUE)
```



* Even after differencing we notice a strong seasonality in the plots. It's not clear how many lag terms we should use to account for seasonality, but we can potentially use 26 lag terms for AR because 26 weeks is 6 months of data, and let ARIMA function automatically to select for seasonality adjustment.

splitting non-seasonal data into test and training set

```
# split seasonal adjusted data
dat <- co2_present_full %>% select(c('date','co2_value'))

test.size<-104 # last two years 2*52 = 104 weeks

#split training and test
dat.train<-dat %>%
  slice(1:(n()-test.size))

dat.test<-dat %>%
  slice((n()-test.size+1):n())
```

in sample bic comparisons

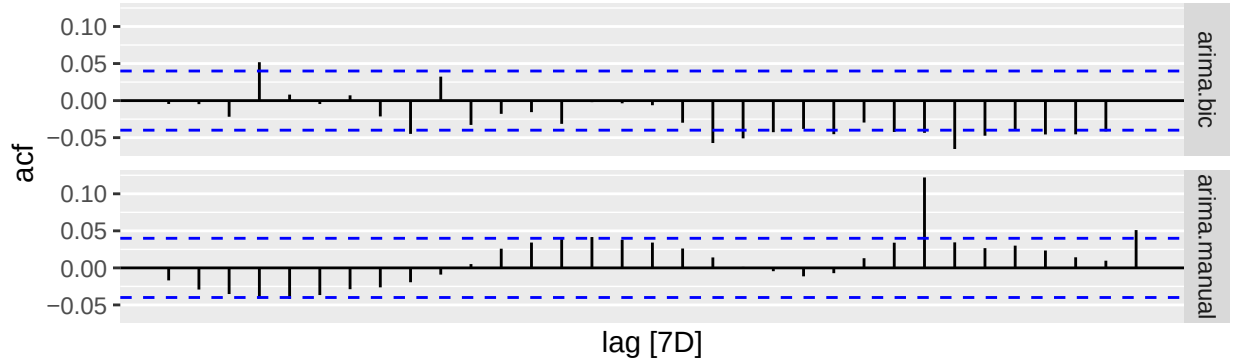
```
model.comp<-dat.train %>%
  model(arima.bic=ARIMA(co2_value ~ 1 + pdq(0:30,0:2,0:30) + PDQ(0:30,0:2,0:30), ic="bic", stepwise=F, g
```

```
        arima.manual=ARIMA(co2_value ~ 1 + pdq(26,1,1) + PDQ(0:30,0:2,0:30), ic="bic", stepwise=F, greed
        )

model.comp %>% report
```

```
## # A tibble: 2 x 8
##   .model       sigma2 log_lik   AIC  AICc   BIC ar_roots   ma_roots
##   <chr>         <dbl>   <dbl> <dbl> <dbl> <dbl> <list>     <list>
## 1 arima.bic     0.246  -1722. 3459. 3460. 3506. <cpl [4]>  <cpl [2]>
## 2 arima.manual  0.211  -1531. 3120. 3121. 3288. <cpl [26]> <cpl [1]>
```

- In terms of BIC score, manually-selected ARIMA data actually performs a lot better on in sample data compare to BIC-selected ARIMA, but this could also be due to overfitting.

in sample residual comparisons




```
## [1] "arima.bic Ljung-Box test"
```

```
##
##  Box-Ljung test
##
## data:  .
## X-squared = 17, df = 10, p-value = 0.08
```

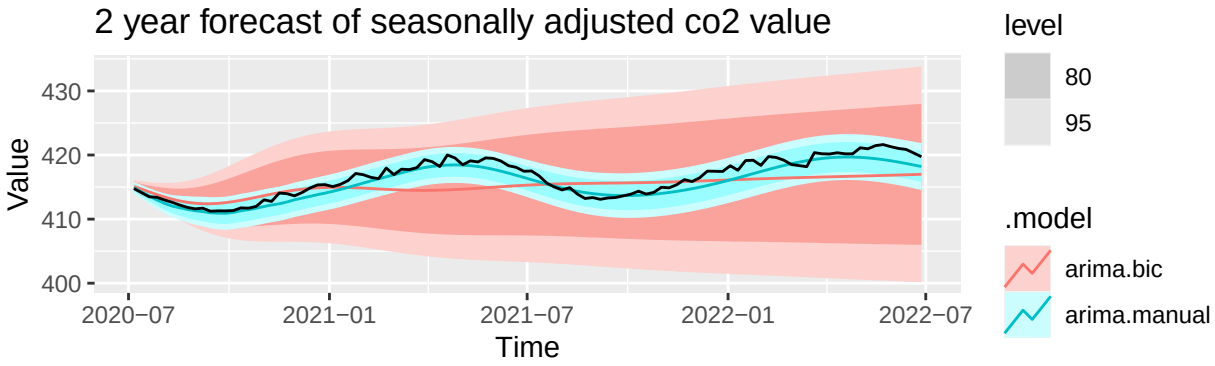
```
## [1] "arima.manual Ljung-Box test"
```

```
##
##  Box-Ljung test
##
## data:  .
## X-squared = 22, df = 10, p-value = 0.01
```

- In contrast when looking at in-sample residuals and each model's Ljung-Box test, only the residual from BIC-selected ARIMA appears to be stationary, while there is still some seasonality left for the manually-selected ARIMA.

test sample forecast comparisons

```
## Plot variable not specified, automatically selected `.vars = co2_value`
```




```
## # A tibble: 2 x 10
##   .model       .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
##   <chr>        <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima.bic    Test  1.27   2.58 2.13  0.302 0.510   NaN   NaN 0.956
## 2 arima.manual Test  0.805  1.13 0.922 0.193 0.221   NaN   NaN 0.709
```

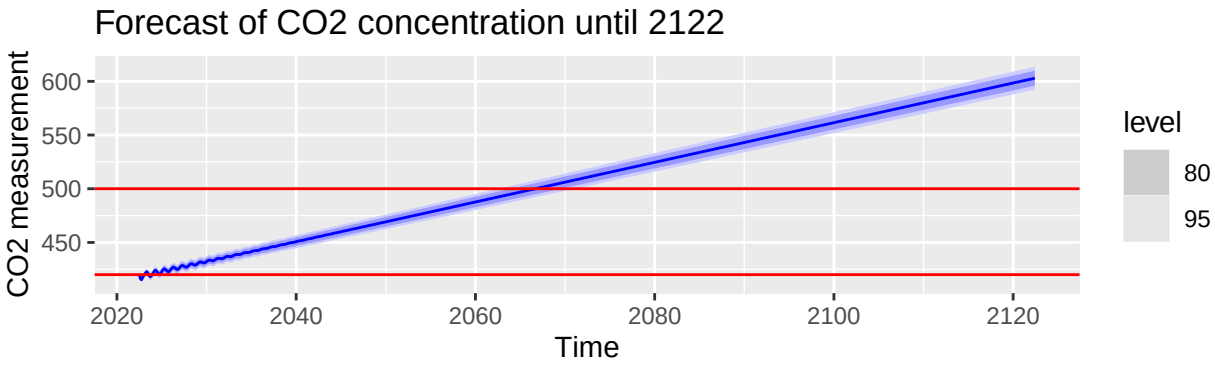
- Now looking at forecast on the test samples,
  - manually selected model better captures the fluctuation within the data due to using more lagging terms. Furthermore, it has a much narrower confidence interval, which means we can have more confidence in its result.

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

train the best ARIMA model on the whole dataset

```
dat <- co2_present_full %>% select(c('date','co2_value'))
arima.manual<-dat %>%
  model(arima.manual=ARIMA(co2_value ~ 1 + pdq(26,1,1) + PDQ(0:30,0:2,0:30), ic="bic", stepwise=F, greed
```

forecast with the best model



calculate first and last time for 420

- Using 95% confidence interval the first time we reach 420 is during the week of 2022-07-04. The last time we reach 420 is during the week of 2025-12-22
- Using mean estimation the first time we reach 420 is during the week of 2023-01-30. The last time we reach 420 is during the week of 2023-12-04

calculate first and last time for 500

- Using 95% confidence interval the first time we reach 500 ppm is during the week of 2062-07-31. The last time we reach 500 ppm is during the week of 2071-01-26
- Using mean estimation the first time we reach 500 ppm is during the week of 2066-09-13, which is also the last time we reach 500 ppm, because the trend is linearly upward.

Generate a prediction for atmospheric CO2 levels in the year 2122

```
## # A tsibble: 1 x 8 [7D]
## # Key:       .model [1]
##   .model       date        co2_value .mean `80%_lower` `80%_upper` `95%_lower`
##   <chr>        <date>         <dist> <dbl>       <dbl>       <dbl>       <dbl>
## 1 arima.manual 2122-06-01 N(603, 30)  603.        596.        610.        592.
## # ... with 1 more variable: 95%_upper <dbl>
```

- At June 2122, our best model predicts that our mean CO2 concentration is 603 with 95% upper equals 614, and 95 lower equals 592. With such a small confidence interval, we do have some confidence that CO2 levels can be significantly dangerous by 2122.
- However, the time series model assumes all important and relevant factors stays the same as they did historically when we first collect the data. We have no confidence that this assumption will stay valid in the future.

Conclusion

Based on our forecasts, the CO2 level will quickly enter a territory that makes living on earth difficult for humans. We urge everyone to take actions that will help reduce CO2 emissions.