Atmospheric CO2 Comprehensive Analysis

U.C. Berkeley, MIDS, Summer 2022, W271 Lab2

Aastha Khana, Eda Kavlakoglu, Don Irwin, Ben Mok

07/13/2022

Contents

1 Report from the Point of View of 1997

1.1 Introduction

In the 1950s, a geochemist, Charles David Keeling, began to collect air samples to measure the amount of carbon dioxide in the air, and these measurements would become his life's work. His initial samples indicated that the air contained more CO2 at night compared to the daytime. This finding suggested that plants absorbed CO2 during the day and released CO2 during the evening. Over time, his measurements showed a seasonal cycle, which was consistent with the seasonal patterns of terrestrial vegetation, and the CO2 data also would persistently increase over time. This rise in CO2 prompted questions about preconceived notions around the ocean's ability to absorb CO2 emitted by fossil fuels. If CO2 released by coal, natural gas, and petroleum remained in the air, this could account for the rise in CO2 over the years.

This finding is incredibly important, because the rising rate of CO2 has implications on the Earth's ability to release heat from the atmosphere. If heat isn't able to escape, then this would subsequently result in a warmer climate, which would become known as the "greenhouse effect". However, while the CO2 data demonstrated a clear upward trend over time, it was also marked by periods of decline which could not solely be attributed to plant growth seasonality. A deeper investigation revealed that El Nino weather patterns affected the amount of CO2 that was released in the air by vegetation and soils. Over time, the data would also show earlier seasonal starts, and when this data was layered with temperature data, it had become clear that the concerns of a warming planet years earlier were starting to manifest itself within the data. Since this connection has been established between rising CO2 and corresponding global temperatures, more attention has surrounded the topics of global warming and climate change. As a result, business and political decision makers are starting to incorporate this data into their longer-term strategies; however, their pace of action may be too slow to prevent some of dangers associated with climate change.

In the analysis below, we will confirm the trends identified above by Keeling himself and use this data to fit a model to make a forecast of CO2 levels in the future.

1.2 Exploratory Data Analysis

The data that we will analyze in this report is monthly data on the CO2 levels made at the Mauna Loa observatory from Jan 1959 to Dec 1997. According to the data documentation, the air at Mauna Loa is thought to be representative of much of the Northern Hemisphere and potentially the globe as well, as the observatory is at an altitude of 3400 meters and surrounded by bare lava, which allows for measurement of "background" air that is resistent to day-to-day fluctuations in CO2 levels.

The data is in units of "mole fraction", which according to the data source is "defined as the number of carbon dioxide molecules in a given number of molecules of air, after removal of water vapor. For example, 413 parts per million of CO2 (abbreviated as ppm) means that in every million molecules of (dry) air there are on average 413 CO2 molecules." The data that makes up our dataset was measured daily from Jan 1959 to Dec 1997, but in our dataset appears as an averaged mean per month in ppm units.

In raw form, the data appears as a matrix of doubles that represent the ppm measurements per month and year combination. Let's create some initial EDA plots that will allow us to better understand the data, starting by analyzing the time series, histogram, auto-correlation function (ACF), and partial auto-correlation function (PACF) plots.

1.2.1 Standard time series graph analysis
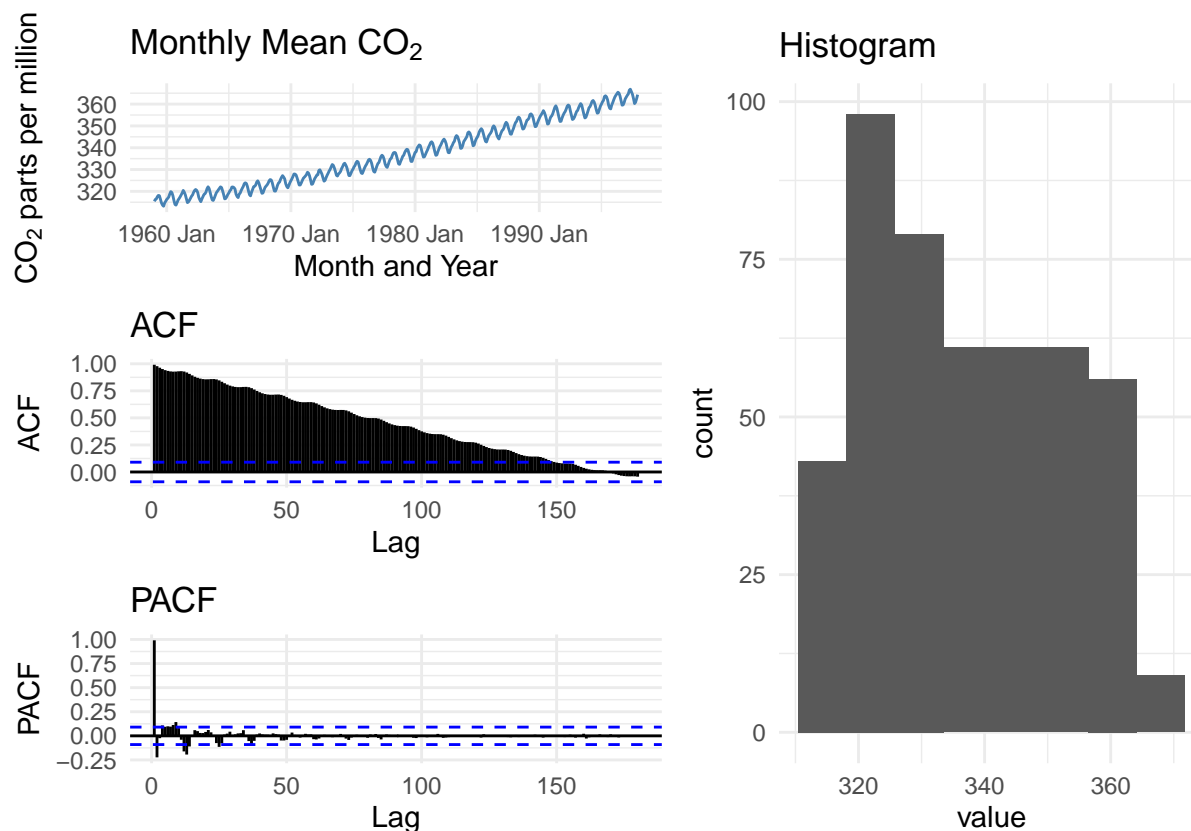
As we can see above, the data seems to follow a clear and increasing trend, with a distinct seasonal pattern that appears as "waves" that we would like to further analyze. The magnitude of the fluctuations do not appear to vary with the time series level, so in terms of decomposition, an additive model would likely fit this series best. The time series does not appear stationary from this plot, as the mean does not appear constant and in fact appears to increase over time, but the variance appears to be constant.

The ACF starts out close to 1 and declines slowly over time, losing significance but staying mostly positive and above the significance line until around lag 140. This slow decline in ACF is what we should see in a time series with a pronounced trend effect, and tracks with what we noticed in the previous time series plot. There appear to be "waves" in the ACF plot similar to the "waves" we also noticed in the time series plot, indicating that the there is a seasonal or cyclic component to our data. Thinking ahead to our modeling, the ACF seems to indicate an autoregressive component in our data generating process, as it declines slowly over time.

The PACF starts out with a single significant positive spike at lag 1, followed by a (relatively smaller) significant negative spike at lag 2, with oscillating clusters of positive and negative lags with much lower levels of significance as the lag number increases. Thinking ahead to our modeling, the PACF seems to indicate an moving average component in our data generating process, as it oscillates between positive and negative over time.

The histogram shows that the CO2 values seem to range between 300 and 380 ppm and do not appear normally distributed, with most values in between these two ranges.
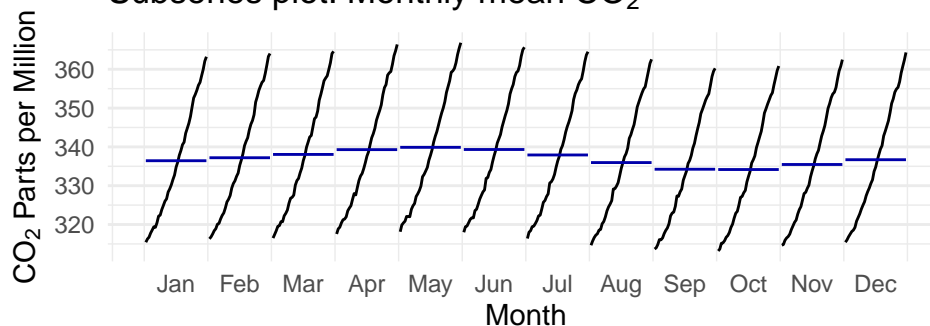
1.2.2   Analyzing seasonality

Let's take a closer look at the seasonality in the data. We'll start by creating a season plot and a subseries plot of the data.
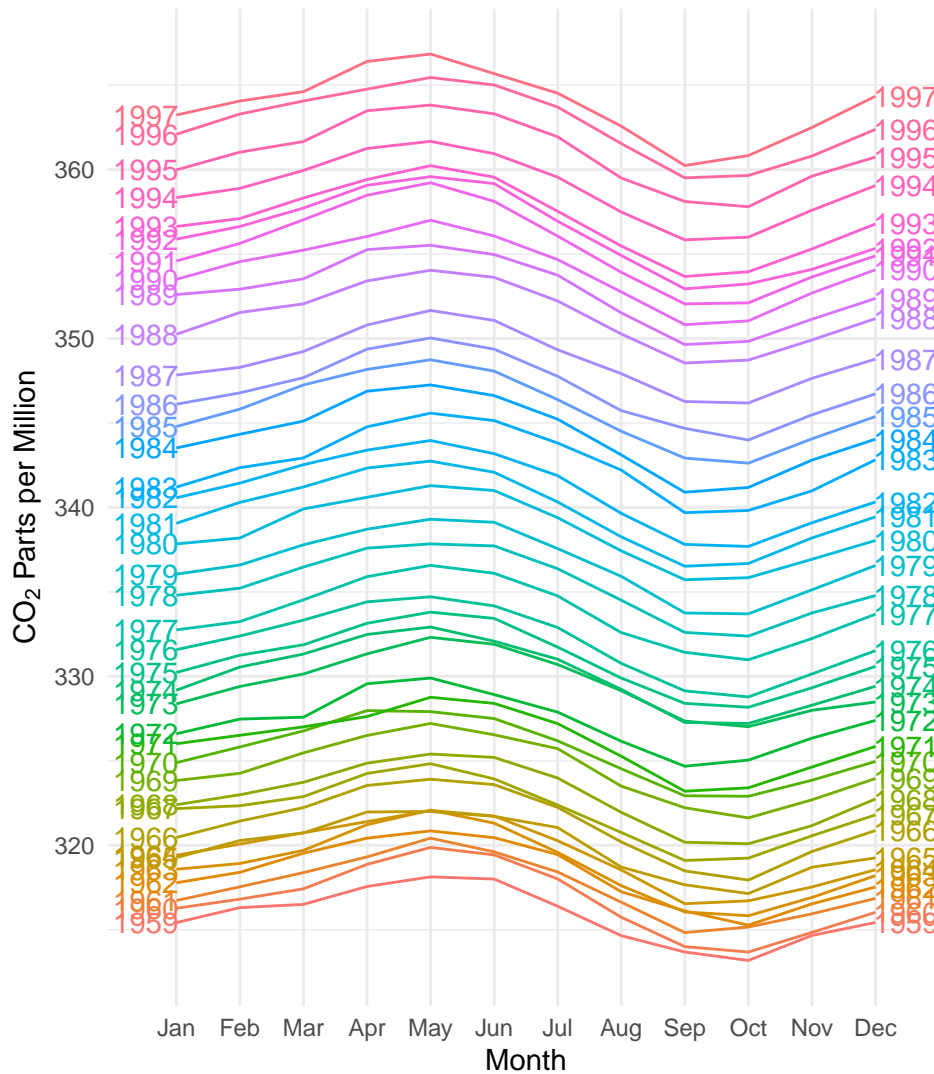
From these plots, we can see that the CO2 levels appear to increase from January through May, then decrease from June through October, hitting a low in October, and then increase again from November through end of the year. The source report does mention that plants and soil absorbing and emitting CO2 could influence these measurements. A possible explanation could be that, in the colder months, we can expect higher CO2 levels as plants die off, and in the warmer months, we can expect lower

CO2 levels as plants thrive. Of course, there is variability across the Northern Hemisphere in what counts as "colder" months and when plants thrive - for example, Hawaiian "winter" is temperate and plants can grow year round, so this could explain why the seasonal variability is slight across the year. We can again see the trend of CO2 levels increasing year by year, but the seasonality effect seems constant year after year without a noticeable increase in the magnitude of the fluctuations across years, again supporting an additive model.

We can confirm that our time series trend is increasing by removing our seasonality components and aggregating our data by year instead of month, as seen below.

Annual Mean CO$_2$ Levels
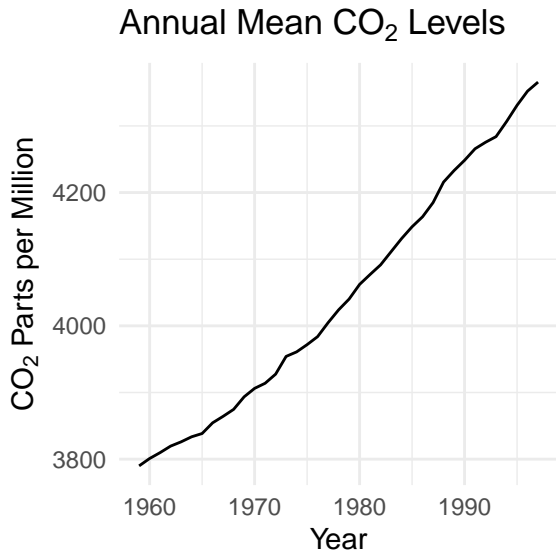
1.2.3 Removing trend and seasonal movements

We can also use both additive and multiplicative decomposition to remove both the trend and seasonal movements from our dataset and confirm that the variance is stationary. The results from these decompositions are plotted below.
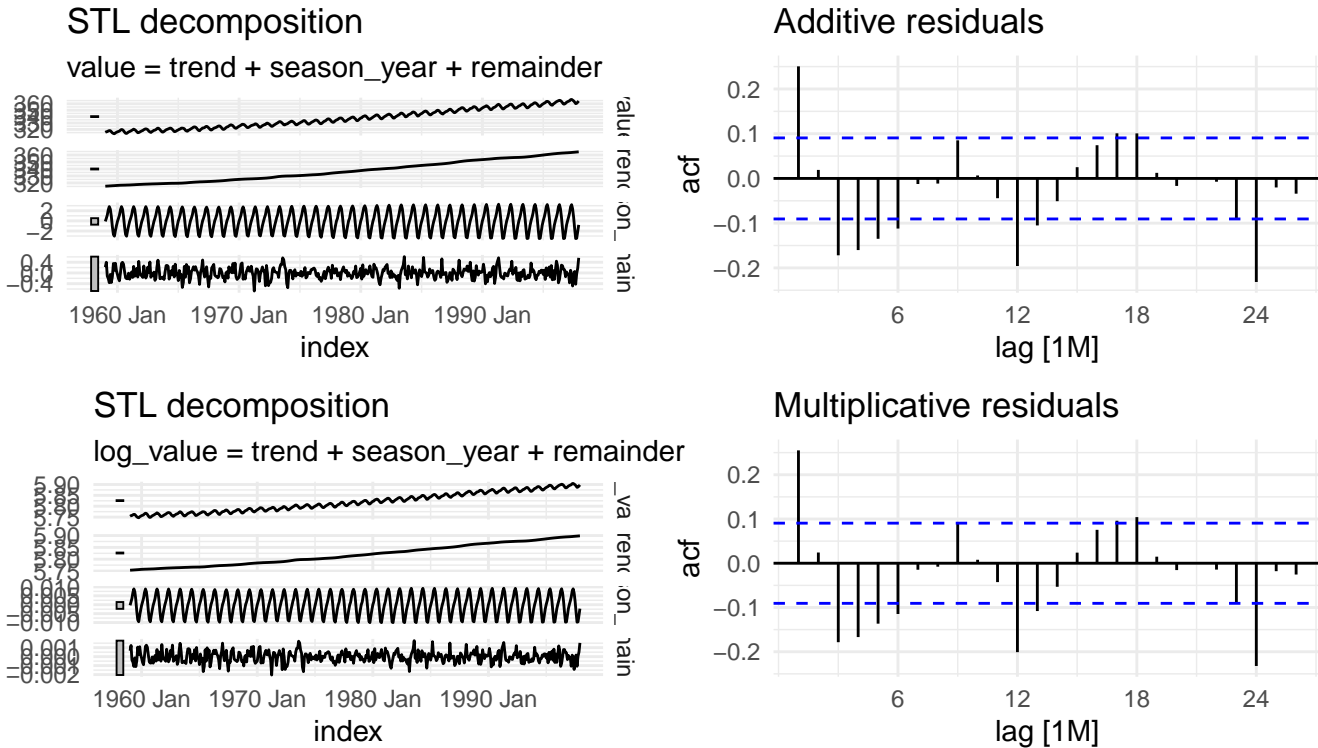
From these plots, we can confirm again that the time series is trending upwards, as seen in the "trend" sections of the decomposition plots. However, upon closer look, the fluctuations within the seasonality of the time series seem to grow slightly larger over time, which we can see in the "season_year" section of the top left plot. In the multiplicative plot on the bottom left, the "season_year" plot appears just slightly more stable, tentatively supporting the idea that this might actually be a multiplicative time series, contrary to our earlier findings. In the next section we should look at applying a log transform on our series prior to modeling.

Looking at the residual plots, the residuals on both appear stationary, meaning that the decomposition methods we are using was able to eliminate deterministic components from the time series. However, they do not appear to be white noise, meaning that there is still correlation in the data.

1.2.4 Testing whether time series is stationary

Let's complete our EDA by running statistical tests to determine whether our model is stationary or non-stationary. We will run both the Augmented Dickey-Fuller (ADF) test and the Phillips Perron (PP) test to do this, under the following hypotheses:

H0: Time series is non-stationary

H1: Time series is stationary

```
##
##   Augmented Dickey-Fuller Test
##
## data:  co2
## Dickey-Fuller = -6.842, Lag order = 5, p-value = 0.01
## alternative hypothesis: stationary
```

```
##
##   Phillips-Perron Unit Root Test
##
## data:  co2
## Dickey-Fuller Z(alpha) = -92.68, Truncation lag parameter = 5, p-value
## = 0.01
## alternative hypothesis: stationary
```

Based on the ADF and PP tests, we can reject the null hypothesis that the time series is non-stationary. This is surprising as based on our visual analysis of the time series plots, the time series does not appear to be stationary with a mean that trends upwards. Because we know that both the ADF and PP tests have low power, we will move forward with the assumption that this time series is non-stationary based on our visual EDA.

## 1.3 Modeling

1.3.1 Linear and log-linear models

Let's start by creating two linear models, one fit on our CO_2 data and one fit on the log transform of our CO_2 data. Since in our EDA we did notice that the fluctuations within the seasonality of the time series seem to grow slightly larger over time, potentially indicating a multiplicative series, we want to try out this log transform to see if it reduces variance in our model and leads to smaller residuals.

```
## Series: value
## Model: TSLM
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -6.039885  -1.947575  -0.001671   1.911271   6.514852
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.115e+02   2.424e-01  1284.9   <2e-16 ***
## trend()     1.090e-01   8.958e-04   121.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.618 on 466 degrees of freedom
## Multiple R-squared: 0.9695,  Adjusted R-squared: 0.9694
## F-statistic: 1.479e+04 on 1 and 466 DF, p-value: < 2.22e-16
```

```
## Series: log_value
## Model: TSLM
##
## Residuals:
##         Min          1Q       Median          3Q          Max
## -0.0172650  -0.0056145   0.0002764   0.0053760   0.0187770
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.744e+00   6.829e-04  8410.5   <2e-16 ***
## trend()     3.224e-04   2.523e-06   127.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.007375 on 466 degrees of freedom
## Multiple R-squared: 0.9722,  Adjusted R-squared: 0.9722
## F-statistic: 1.633e+04 on 1 and 466 DF, p-value: < 2.22e-16
```

We can see from the output here that the R-squared values on both models are high and the input coefficients are highly significant on both. We will also plot the models on top of our data.
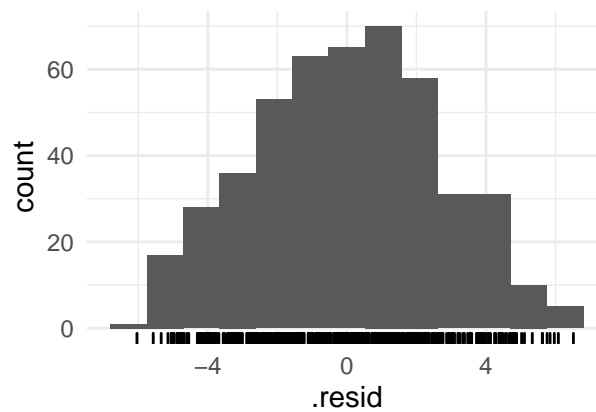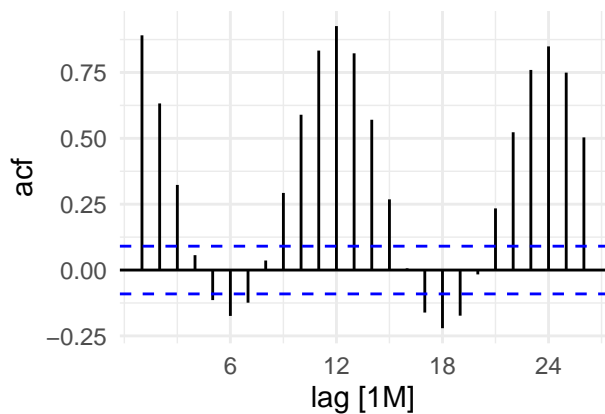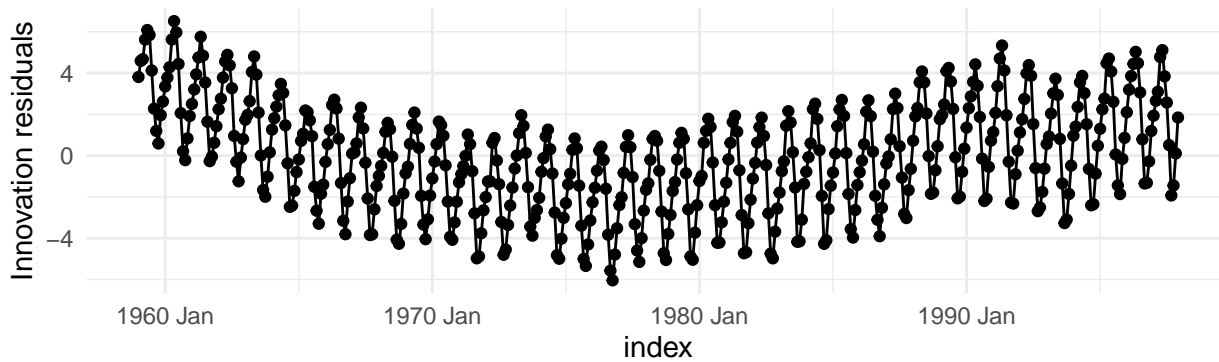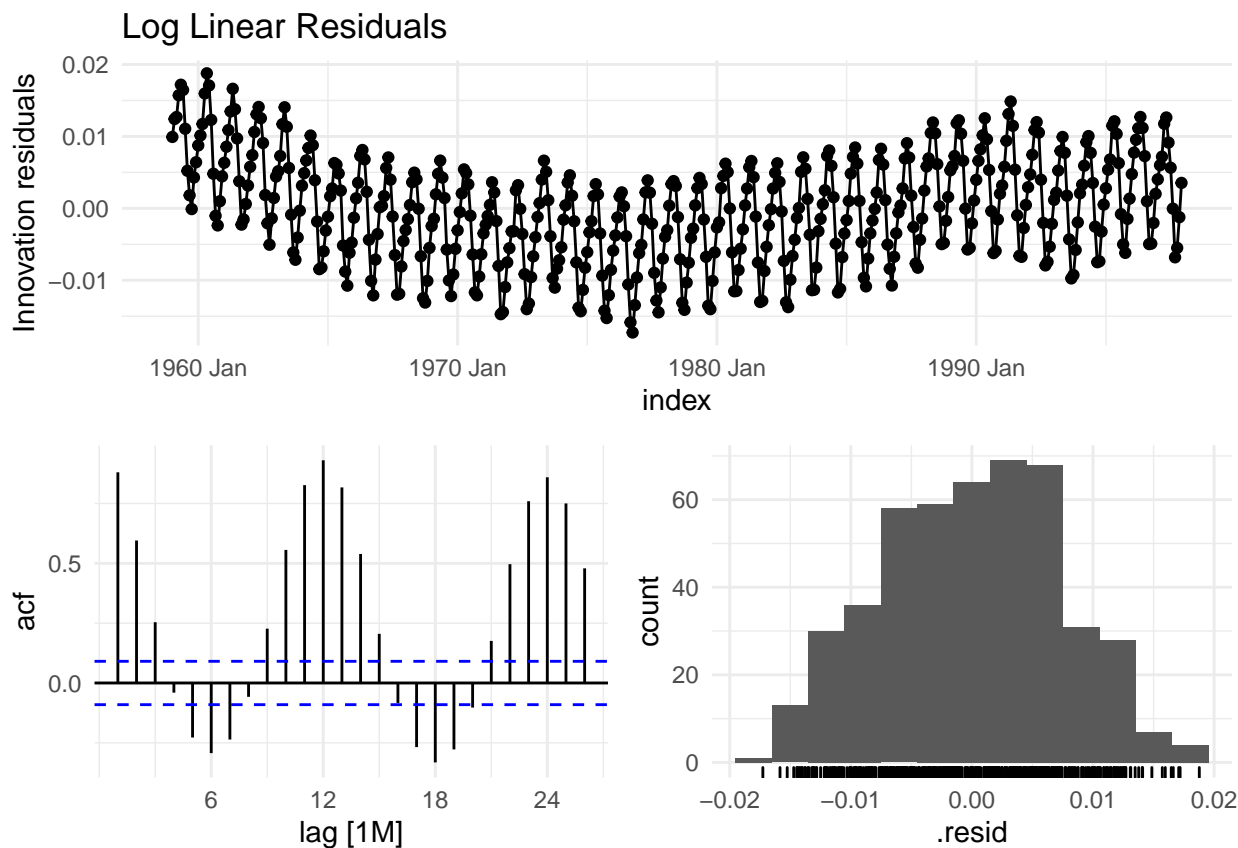
Linear Model / Mean CO$_2$ Levels
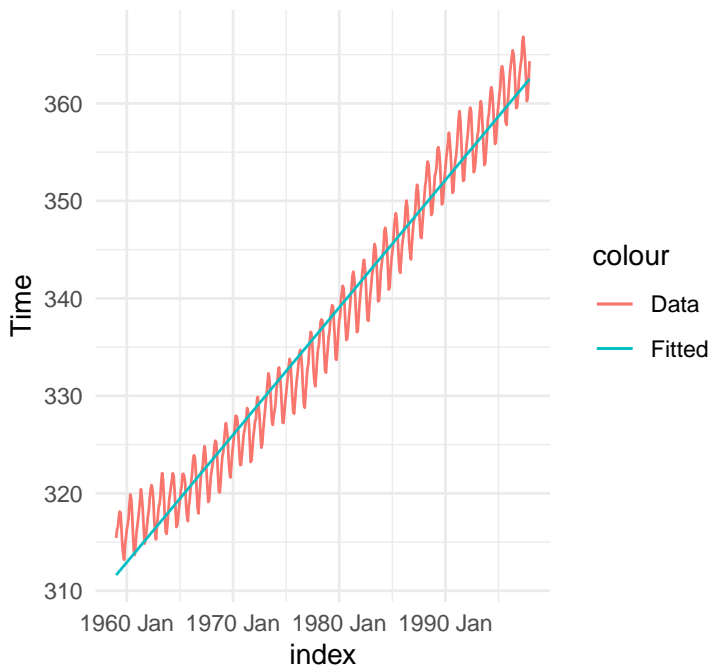
Log Linear Model / Log Mean CO$_2$ Levels
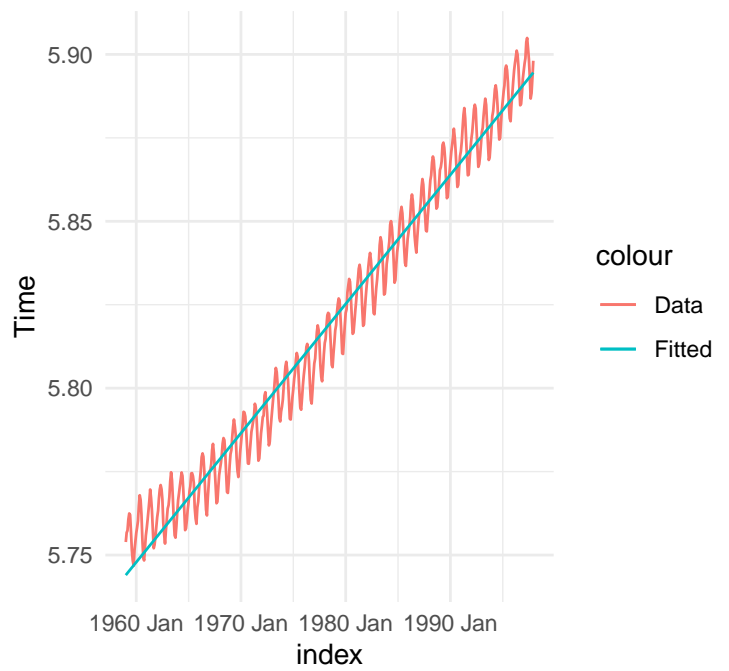
From these plots, both models seem to do a similar job fitting the data but fail to account for any of the seasonal "waves". Next, let's examine the model residuals.

Linear Residuals

Log Linear Residuals

The residuals don't look much different here between the linear and log linear models, signaling that a log transformation is unnecessary for the linear model. In both models, the ACFs show a periodic oscillating pattern which signals that something is missing from our model. Based on our EDA, this is most likely the seasonality that we discussed but did not account for in this model. Additionally, the residuals appear somewhat normally distributed for both models. Let's run a Ljung Box test to understand whether we have white noise residuals, under the following hypotheses:

H0: Data are independently distributed.

H1: Data are not independently distributed.

```
##
##  Box-Ljung test
##
## data:  linear_residuals
## X-squared = 373.94, df = 1, p-value < 2.2e-16
```

```
##
##  Box-Ljung test
##
## data:  linear_residuals
## X-squared = 850.26, df = 10, p-value < 2.2e-16
```

```
##
##  Box-Ljung test
##
## data:  linear_log_residuals
## X-squared = 365.1, df = 1, p-value < 2.2e-16
```

```
##
##  Box-Ljung test
##
## data:  linear_log_residuals
## X-squared = 830.65, df = 10, p-value < 2.2e-16
```

Based on the Ljung Box test, we can reject the null that the data is independently distributed up to 10 lags, which means that we likely do not have white noise residuals and both our models are failing to account for some variance in our data (likely the seasonality).

8

### 1.3.2 Explaining variance with quadradic trend models

Let's repeat this process with a quadratic trend model to see if this model accounts for additional variance in our process.

```
## Series: value
## Model: TSLM
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0195 -1.7120  0.2144  1.7957  4.8345
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.148e+02  3.039e-01 1035.65   <2e-16 ***
## trend()      6.739e-02  2.993e-03   22.52   <2e-16 ***
## I(trend()^2) 8.862e-05  6.179e-06   14.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.182 on 465 degrees of freedom
## Multiple R-squared: 0.9788,  Adjusted R-squared: 0.9787
## F-statistic: 1.075e+04 on 2 and 465 DF, p-value: < 2.22e-16
```

```
## Series: log_value
## Model: TSLM
##
## Residuals:
##         Min         1Q     Median         3Q        Max
## -0.0143052 -0.0050832  0.0005277  0.0052757  0.0136508
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.752e+00  9.039e-04 6363.18   <2e-16 ***
## trend()      2.216e-04  8.900e-06   24.90   <2e-16 ***
## I(trend()^2) 2.149e-07  1.838e-08   11.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.00649 on 465 degrees of freedom
## Multiple R-squared: 0.9786,  Adjusted R-squared: 0.9785
## F-statistic: 1.061e+04 on 2 and 465 DF, p-value: < 2.22e-16
```

We can see from the output here that the R-squared values on both models are high and the input coefficients are highly significant on both. We will also plot the models on top of our data.

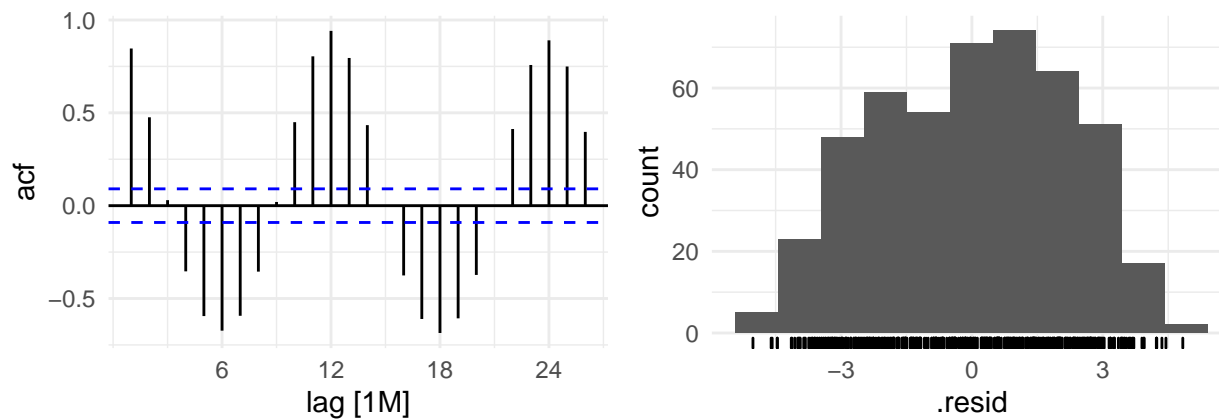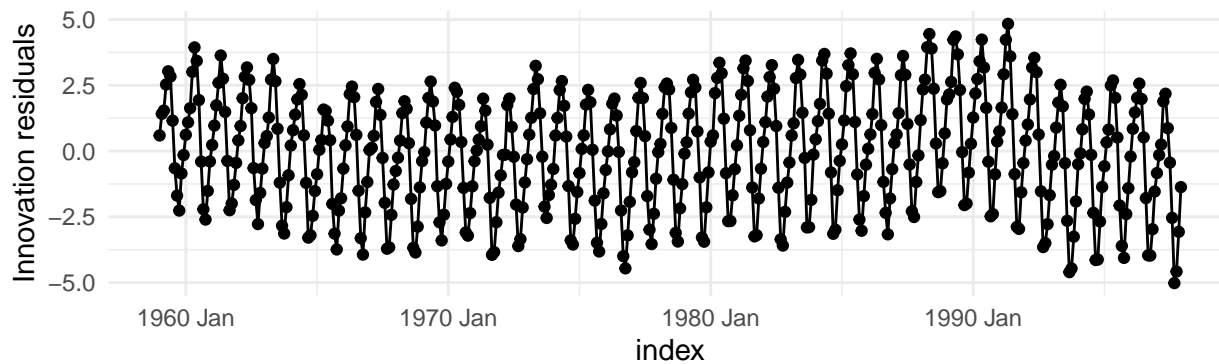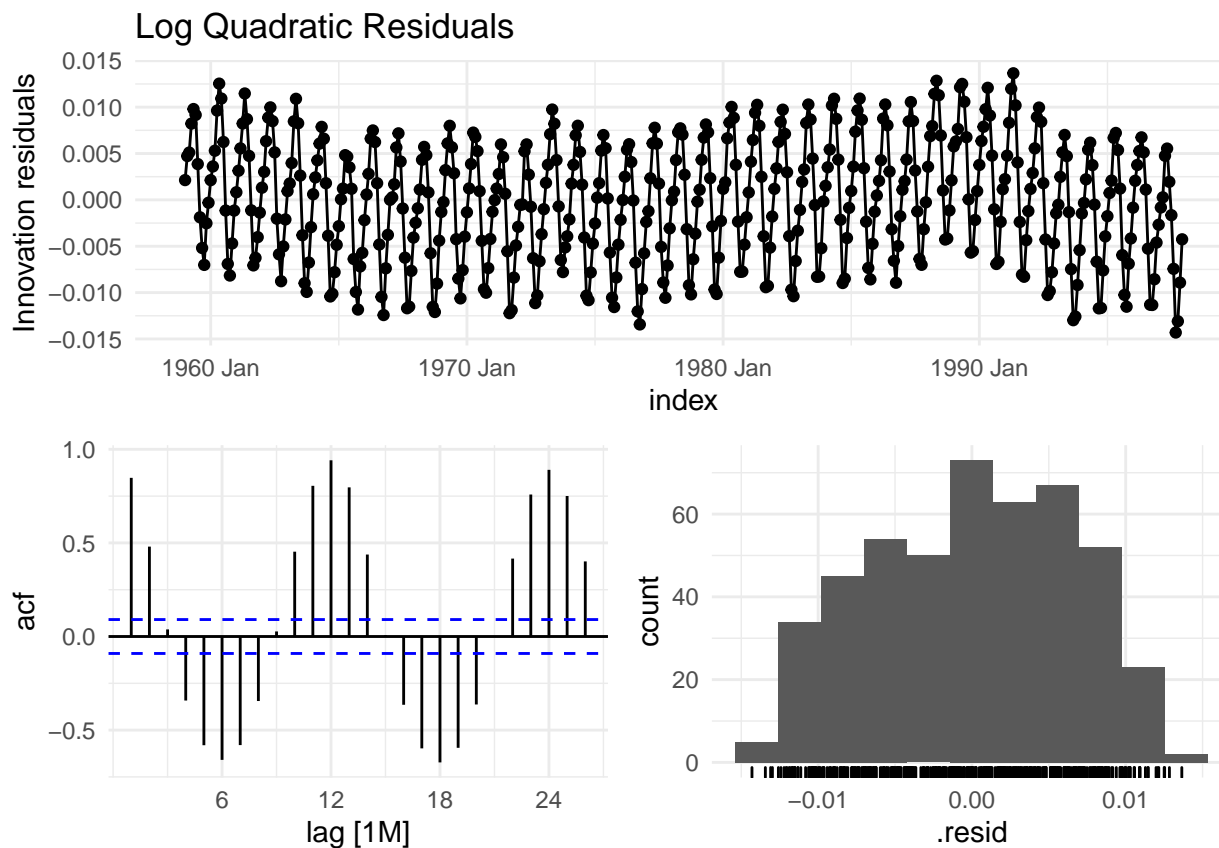## Quadratic Model / Mean CO$_2$ Levels
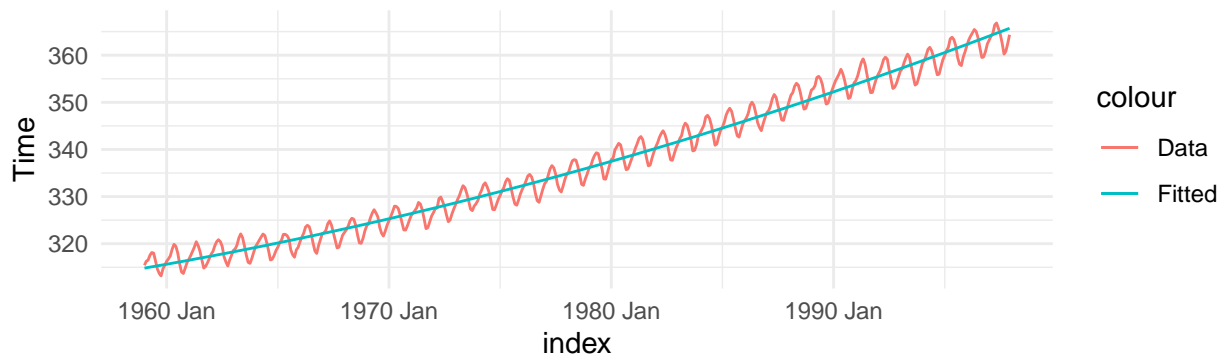
Log Quadratic Model / Log Mean CO$_2$ Levels
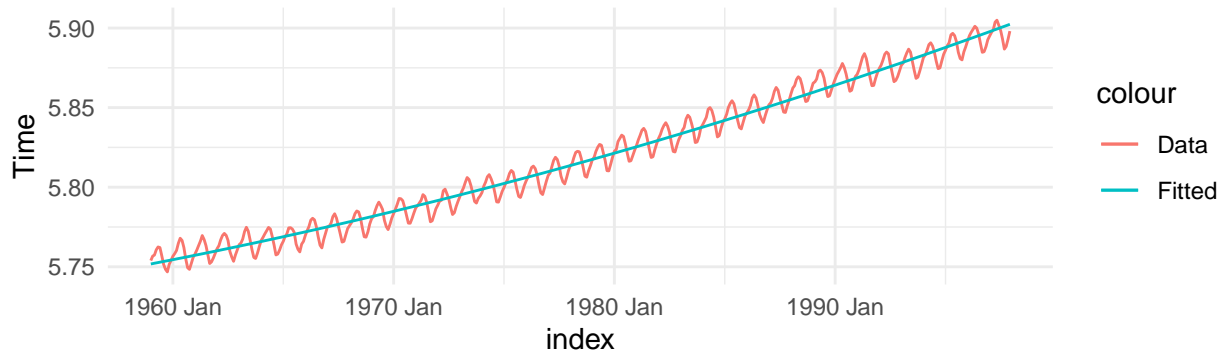
From these plots, both models seem to do a similar job fitting the data. Next, let's examine the model residuals.

Quadratic Residuals

The residuals don't look much different here between both models, signaling that a log transformation is unnecessary for the quadratic model as well. In both models, the ACFs show the same periodic oscillating pattern which signals that the seasonality is likely missing from our model. The residuals also appear somewhat normally distributed for both models. Let's run a Ljung Box test again to understand whether we have white noise residuals under the same hypotheses as before.

```
##
##  Box-Ljung test
##
## data:  quadratic_residuals
## X-squared = 337.42, df = 1, p-value < 2.2e-16
```

```
##
##  Box-Ljung test
##
## data:  quadratic_residuals
## X-squared = 1213.2, df = 10, p-value < 2.2e-16
```

```
##
##  Box-Ljung test
##
## data:  quadratic_log_residuals
## X-squared = 338.34, df = 1, p-value < 2.2e-16
```

```
##
##  Box-Ljung test
##
## data:  quadratic_log_residuals
## X-squared = 1186.5, df = 10, p-value < 2.2e-16
```

As before, based on the Ljung Box test, we can reject the null that the data is independently distributed up to 10 lags, which means that we likely do not have white noise residuals and both our models are failing to account for some variance in our data (likely again the seasonality).

Since the log transforms don't appear to reduce any of the variance in our models, I am going to compare the linear model directly with the quadratic model on non-transformed data. Our previous analysis shows that both models are behaving similarly

but the quadratic model has smaller residuals compared to the linear model, making it our more favorable model. In addition to the residual analysis, I going to compare the two using the Bayesian Information Criteria (BIC) as my metric.

```
## # A tibble: 1 x 1
##     BIC
##   <dbl>
## 1  917.
```

```
## # A tibble: 1 x 1
##     BIC
##   <dbl>
## 1  752.
```

The quadratic model has a lower BIC, so I will move forward with our modeling process using this quadratic model. Our next step is to fit a polynomial model that incorporates seasonal dummy variables, which will hopefully account for some of the variance that we failed to capture in our previous models. We'll create models using both the log transform and the normal CO2 values, although it has appeared from our past models that the log transform does not seem to account for much, if any, variance in our data.

```
## Series: value
## Model: TSLM
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.99478 -0.54468 -0.06017  0.47265  1.95480
##
## Coefficients:
##                   Estimate Std. Error  t value Pr(>|t|)
## (Intercept)      3.147e+02  1.494e-01 2105.894  < 2e-16 ***
## trend()          6.763e-02  9.929e-04   68.114  < 2e-16 ***
## I(trend()^2)     8.865e-05  2.050e-06   43.242  < 2e-16 ***
## season()year2    6.642e-01  1.640e-01    4.051 5.99e-05 ***
## season()year3    1.407e+00  1.640e-01    8.582  < 2e-16 ***
## season()year4    2.538e+00  1.640e-01   15.480  < 2e-16 ***
## season()year5    3.017e+00  1.640e-01   18.400  < 2e-16 ***
## season()year6    2.354e+00  1.640e-01   14.357  < 2e-16 ***
## season()year7    8.331e-01  1.640e-01    5.081 5.50e-07 ***
## season()year8   -1.235e+00  1.640e-01   -7.531 2.75e-13 ***
## season()year9   -3.059e+00  1.640e-01  -18.659  < 2e-16 ***
## season()year10  -3.243e+00  1.640e-01  -19.777  < 2e-16 ***
## season()year11  -2.054e+00  1.640e-01  -12.526  < 2e-16 ***
## season()year12  -9.374e-01  1.640e-01   -5.717 1.97e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.724 on 454 degrees of freedom
## Multiple R-squared: 0.9977,  Adjusted R-squared: 0.9977
## F-statistic: 1.531e+04 on 13 and 454 DF, p-value: < 2.22e-16
```

```
## Series: log_value
## Model: TSLM
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0053270 -0.0017362 -0.0001774  0.0015139  0.0057292
##
## Coefficients:
##                   Estimate Std. Error   t value Pr(>|t|)
## (Intercept)      5.751e+00  4.533e-04 12687.967  < 2e-16 ***
## trend()          2.223e-04  3.012e-06    73.817  < 2e-16 ***
## I(trend()^2)     2.150e-07  6.219e-09    34.566  < 2e-16 ***
## season()year2    1.969e-03  4.974e-04     3.959 8.73e-05 ***
## season()year3    4.163e-03  4.974e-04     8.371 7.16e-16 ***
## season()year4    7.498e-03  4.974e-04    15.075  < 2e-16 ***
## season()year5    8.911e-03  4.974e-04    17.916  < 2e-16 ***
## season()year6    6.965e-03  4.974e-04    14.004  < 2e-16 ***
```

```
## season()year7    2.480e-03  4.974e-04     4.986 8.78e-07 ***
## season()year8   -3.662e-03  4.974e-04    -7.362 8.61e-13 ***
## season()year9   -9.098e-03  4.974e-04   -18.290  < 2e-16 ***
## season()year10  -9.661e-03  4.974e-04   -19.423  < 2e-16 ***
## season()year11  -6.113e-03  4.974e-04   -12.290  < 2e-16 ***
## season()year12  -2.799e-03  4.974e-04    -5.627 3.21e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002196 on 454 degrees of freedom
## Multiple R-squared: 0.9976,  Adjusted R-squared: 0.9975
## F-statistic: 1.453e+04 on 13 and 454 DF, p-value: < 2.22e-16
```

Just from this summary, we can see that the R-squared values are slightly higher on both of these seasonal models than either the linear or quadratic models. We can also see that all the input coefficients are highly significant. Let's plot our models next.

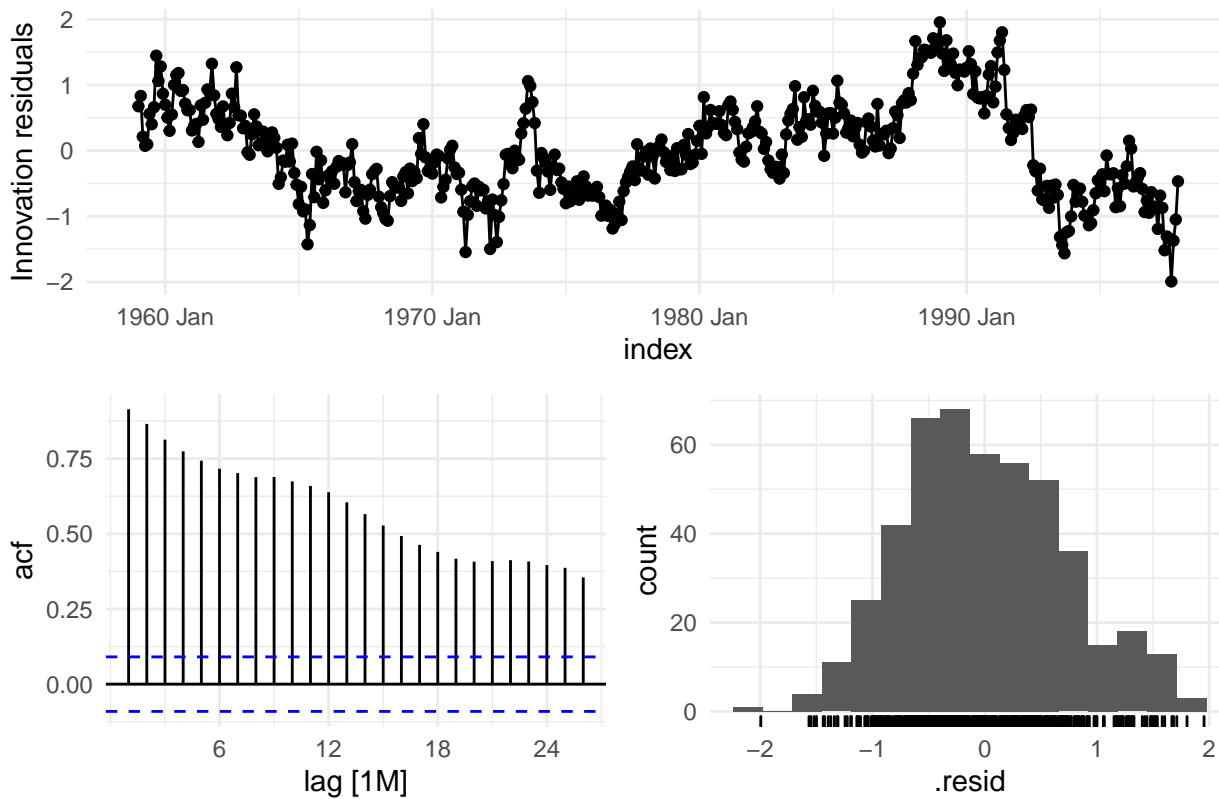Quadratic Seasonal Model / Mean CO_2 Levels
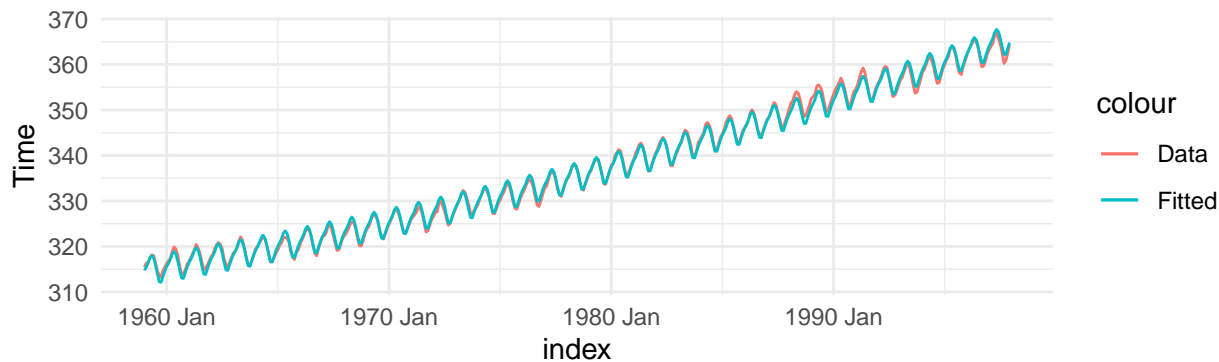
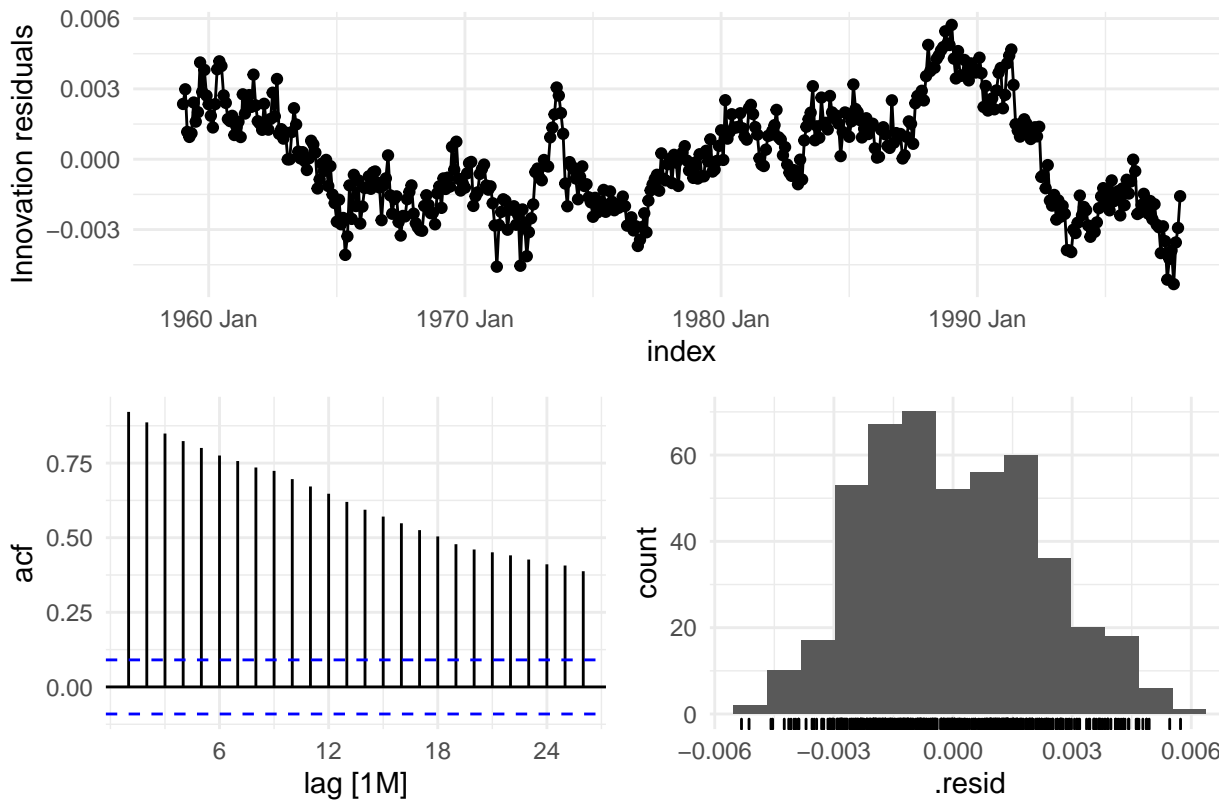Log Quadratic Seasonal Model / Log Mean CO_2 Levels
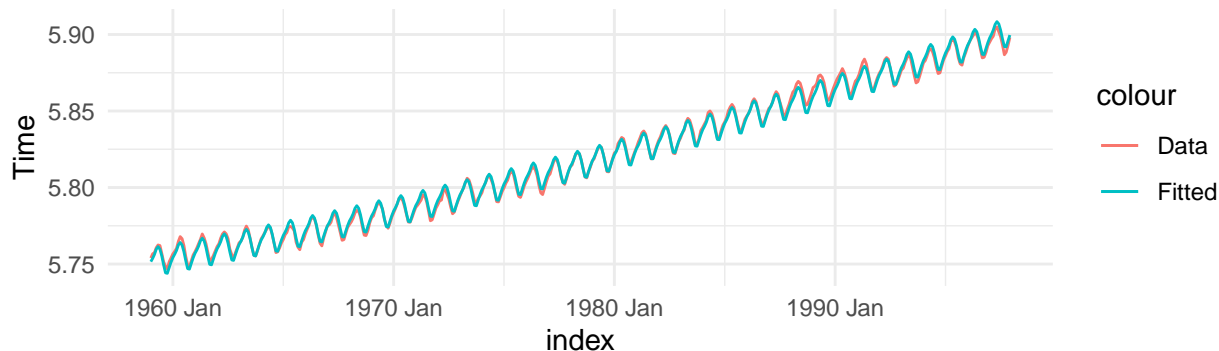
From these plots, both models seem to do a similar job fitting the data and, unlike the previous models we've seen, actually account for the seasonal "waves" in our plot and track them pretty closely. Next, let's examine the model residuals.

Quadratic Season Residuals

Log Quadratic Season Residuals

The residuals again don't look much different here between both models, signaling that a log transformation is unnecessary for this seasonal quadratic model as well. In both models, the ACFs now shows highly significant and slowly declining lags, unlike the ACF plots on the previous models which showed an oscillating trend that was likely related to seasonality variance that our latest models are now capturing. The residuals appear somewhat normally distributed for both models. In comparison to both the non-seasonal linear and quadratic models, the residuals are small in magnitude. Let's run a Ljung Box test again to

understand whether we have white noise residuals under the same hypotheses as before:

```
##
##   Box-Ljung test
##
## data:   quadratic_season_residuals
## X-squared = 393.48, df = 1, p-value < 2.2e-16
```

```
##
##   Box-Ljung test
##
## data:   quadratic_season_residuals
## X-squared = 2758, df = 10, p-value < 2.2e-16
```

```
##
##   Box-Ljung test
##
## data:   quadratic_log_season_residuals
## X-squared = 399.97, df = 1, p-value < 2.2e-16
```

```
##
##   Box-Ljung test
##
## data:   quadratic_log_season_residuals
## X-squared = 3040.2, df = 10, p-value < 2.2e-16
```

As before, based on the Ljung Box test, we can reject the null that the data is independently distributed up to 10 lags, which means that we likely do not have white noise residuals and both our models are failing to account for some variance in our data. This is somewhat surprising as our residuals are quite small, especially in comparison to our previous models. Let's compare the BIC of the quadratic seasonal model without the log transform to our previous BICs.

1.3.3 Model selection

Let's compare the BIC of this model against the pure linear and quadratic models without seasonal components.

```
## # A tibble: 1 x 1
##      BIC
##    <dbl>
## 1 -224.
```

The BIC for our quadratic seasonal model is much smaller than the BICs for either our linear or quadratic model, and by this criteria is our best model that we have fit so far. Let's use this model to generate forecasts to the year 2020.

1.3.4 Forecasts based on models
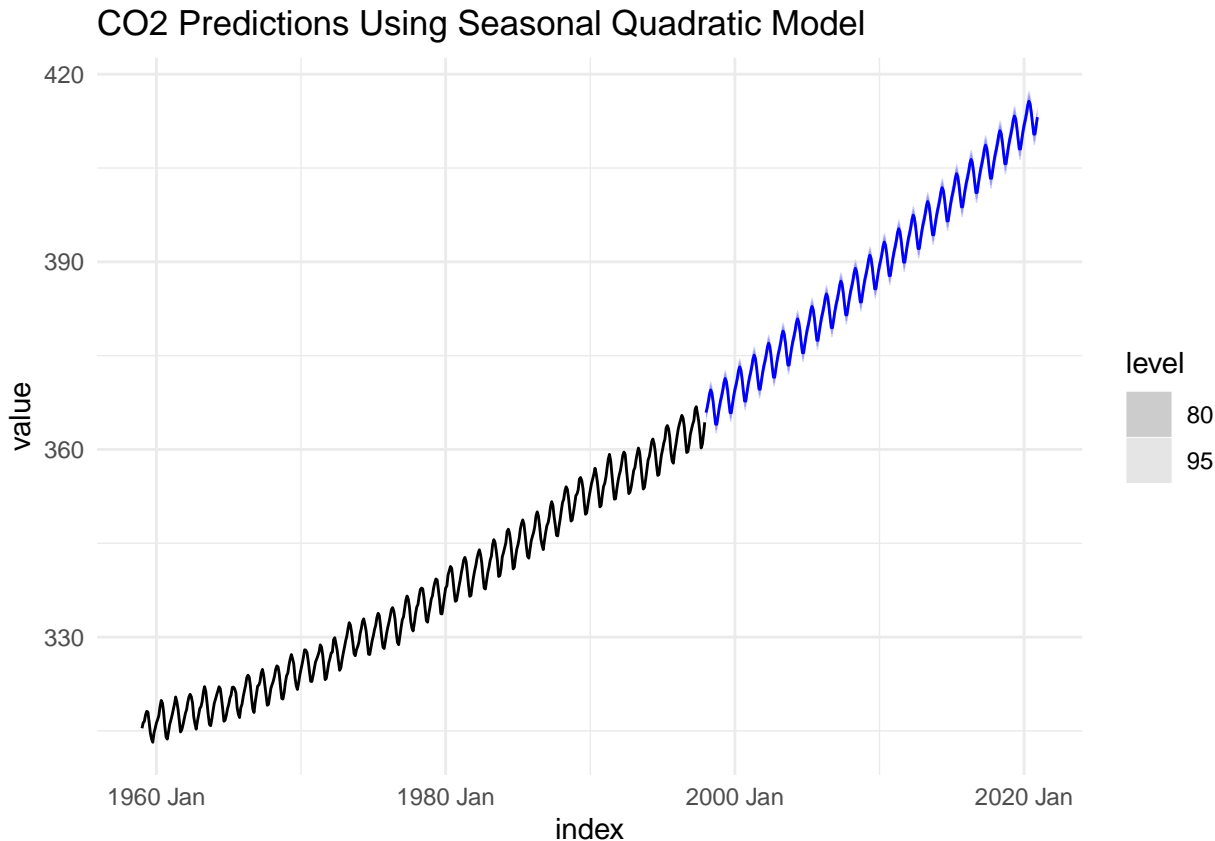
CO2 Predictions Using Seasonal Quadratic Model

We can see that these predictions continue to follow the trend and seasonal fluctuations that we noticed in our data. Let's complete our linear modeling by examining the CLM assumptions for our best model, the seasonal quadratic model.

The first CLM assumption is that the underlying data generating process follows a linear model. We cannot check this assumption and will just assume it is true, and, if untrue, assume that we are just fitting the best linear model to the data.

The second assumption is ergodatic stationarity. Since this model is including trend and seasonal variables, we are eliminating the deterministic parts in the process and are satisfying this assumption. From the ACF plot, the process does look stationary but the residuals do appear serially correlated.

The third assumption is no perfect multicollinearity. When we fit our model, R did not report any warning or missing coefficients, which means that this assumption is violated.

The fourth assumption is the zero conditional mean assumption. We can evaluate this by looking at these residuals versus the fitted value plot, plotted below. If this assumption is satisfied, there should be no relationship between the fitted values and the residuals, a.k.a. there should be a straight line in the plot.

Seasonal Quadratic Model Residuals vs Fitted
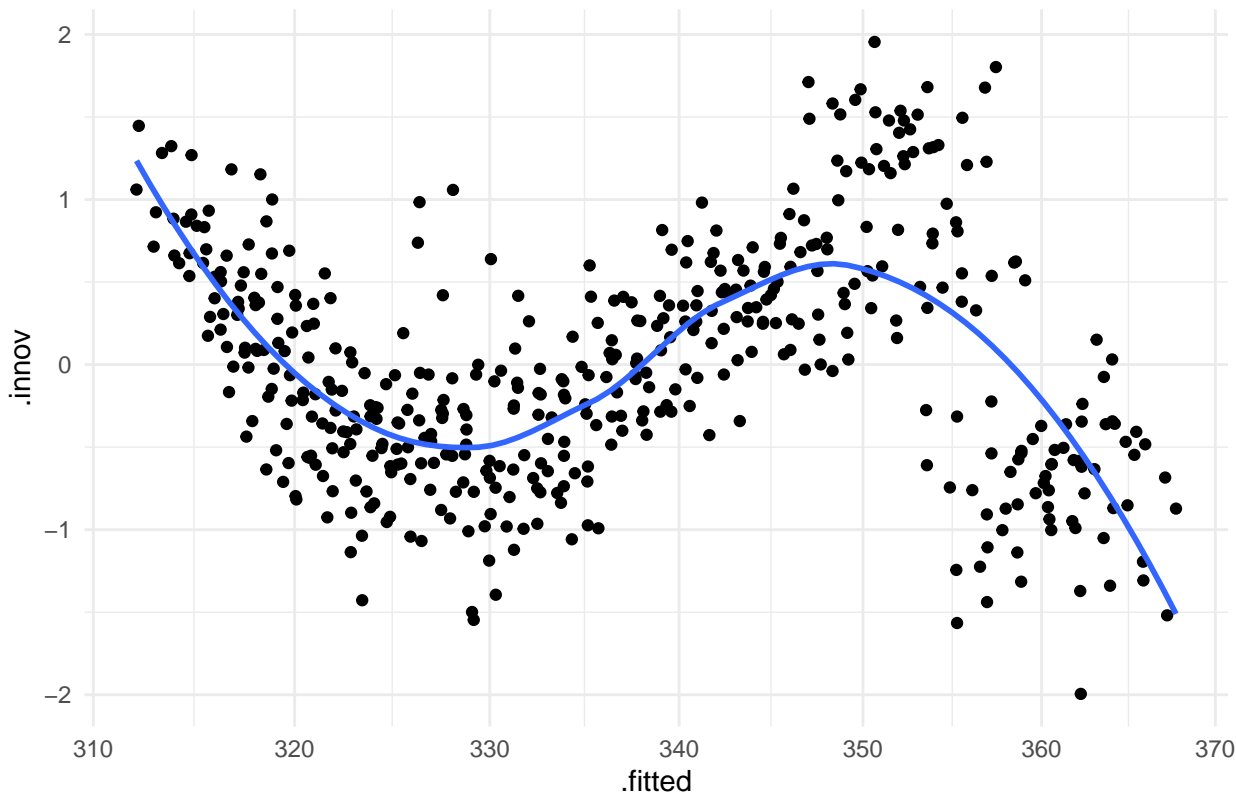
The line is curved, so we cannot justify that the zero conditional mean assumption is satisfied, meaning that there are likely additional variables that should be included in our model.

The fifth CLM assumption is homoskedasticity. The residuals are quite spread out with a large range of variance, so it seems as though this assumption is not satisfied.

The last assumption is no serial correlation. From our ACF plot, the residuals do appear to be serially correlated. Additionally, the Ljung Box test returns a high statistic with a small p-value. We can reject the null hypothesis that there is no autocorrelation and conclude that our residuals here are autocorrelated, thereby not satisfying this assumption.

The residuals do look close to a normal distribution, which gives us more confidence in our generated prediction intervals.

1.4 ARIMA Modeling

Now, let's choose an ARIMA model to fit to the CO2 data. Recall that our EDA indicated that the CO2 data had the following characteristics:

- non-stationary data series
- slowly decaying ACF plot, indicative of an AR process
- oscillating PACF plot, indicative of an MA process
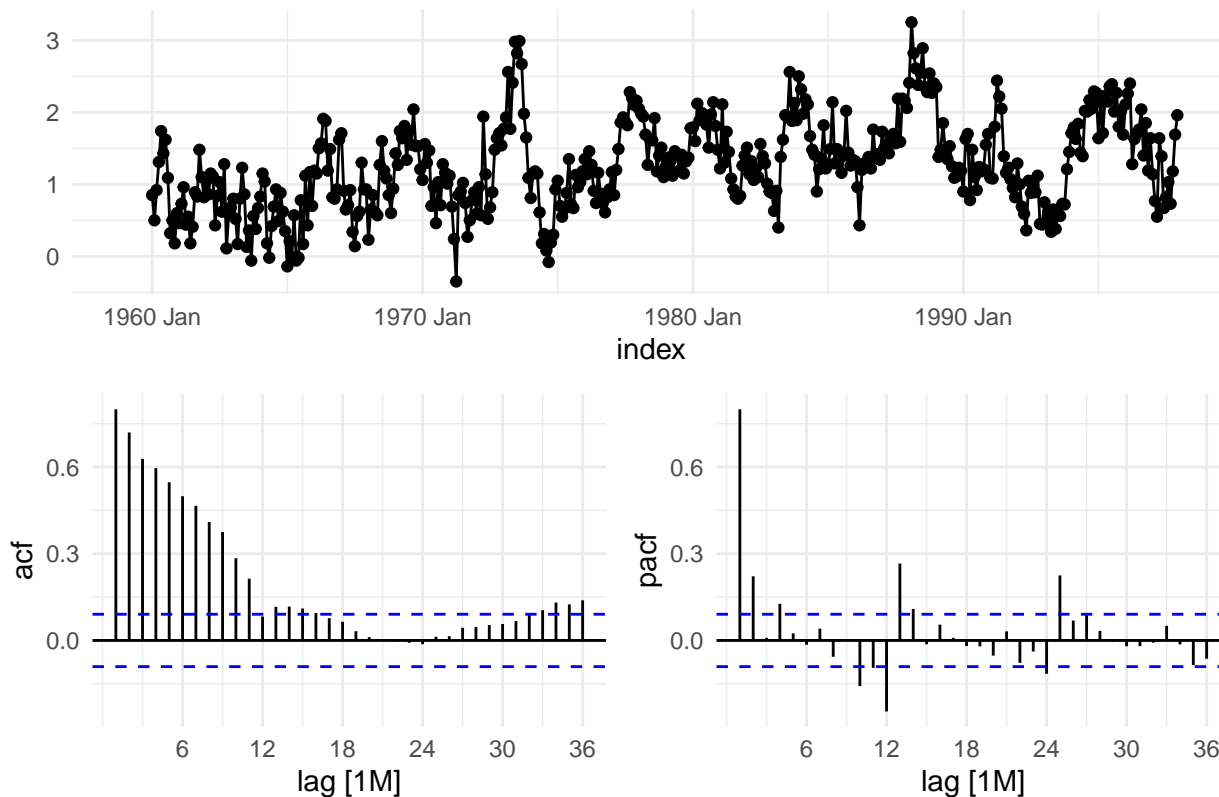- seasonal data

In order to transform the data into a stationary series, we can start by applying a first-order difference to the series and plotting the characteristics.

First–order differenced
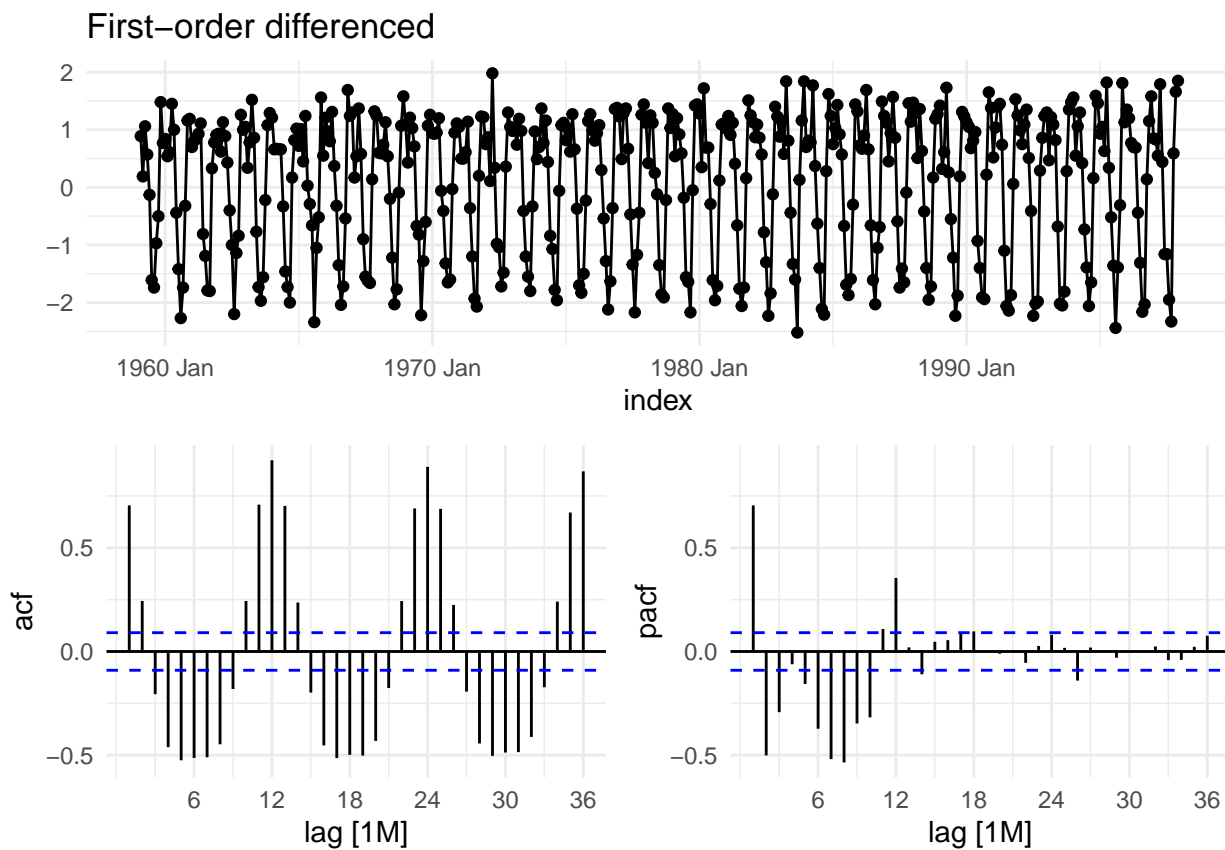
Based on the trend plot, the data seems to stabilize around a mean of -0.5, and it appears to have a constant variance, suggesting that the data only needs a first-order difference. The ACF plot continues to decay gradually but the oscillating behavior persists here. This indicates that the seasonal pattern is strong and stable after a first-order difference. As a result, we will want to use a seasonal difference in addition to this first difference. Since the seasonal behavior repeats every year, we will set m = 12, meaning the seasonal pattern repeats once every twelve months. The output of this seasonal model is below.

```
## # A tsibble: 468 x 4 [1M]
##       index value log_value sdiff_co2
##       <mth> <dbl>     <dbl>     <dbl>
##  1 1959 Jan  315.      5.75        NA
##  2 1959 Feb  316.      5.76        NA
##  3 1959 Mar  316.      5.76        NA
##  4 1959 Apr  318.      5.76        NA
##  5 1959 May  318.      5.76        NA
##  6 1959 Jun  318       5.76        NA
##  7 1959 Jul  316.      5.76        NA
##  8 1959 Aug  315.      5.75        NA
##  9 1959 Sep  314.      5.75        NA
## 10 1959 Oct  313.      5.75        NA
## # ... with 458 more rows
```

Seasonally differenced

On its own, seasonal differencing does not appear to make the data stationary. However, based on the ACF plot, we can confirm that the seasonal differencing either removes or significantly reduces the oscillation as it does not seem apparent anymore. As a result, we can conclude that we need both a first difference and a seasonal difference on the series. We can confirm these observations using the following unit root tests. We will first use the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test under the following hypothesis:

H0: The data are stationary

HA: The data are not stationary

```
## # A tibble: 1 x 2
##   kpss_stat kpss_pvalue
##       <dbl>       <dbl>
## 1      7.82        0.01
```

```
## # A tibble: 1 x 2
##   kpss_stat kpss_pvalue
##       <dbl>       <dbl>
## 1    0.0124         0.1
```

```
## # A tibble: 1 x 2
##   kpss_stat kpss_pvalue
##       <dbl>       <dbl>
## 1      1.94        0.01
```

We can interpret the tests in the following manner:

- For the undifferenced data, the test statistic (7.81) is bigger than the 1% critical value, so the p-value is less than 0.01, indicating that the null hypothesis is rejected. That is, the data are not stationary.

- For the first differenced data, the p-value is greater than 0.1, indicating the test fails to reject the null hypothesis. That is, the data are stationary.

- For the seasonally differenced data, the test statistic (1.94) is bigger than the 1% critical value, so the p-value is less than 0.01, indicating that the null hypothesis is rejected. That is, the data are not stationary.

Thus, we can conclude that we need both a first difference and a seasonal difference. But, how many differences do we need? To confirm that we are applying the appropriate amount of differences, we can use the following:

```
## # A tibble: 1 x 1
##    ndiffs
##     <int>
## 1       1

## # A tibble: 1 x 1
##    ndiffs
##     <int>
## 1       0

## # A tibble: 1 x 1
##    ndiffs
##     <int>
## 1       1

## # A tibble: 1 x 1
##    nsdiffs
##      <int>
## 1        1

## # A tibble: 1 x 1
##    nsdiffs
##      <int>
## 1        1

## # A tibble: 1 x 1
##    nsdiffs
##      <int>
## 1        0
```
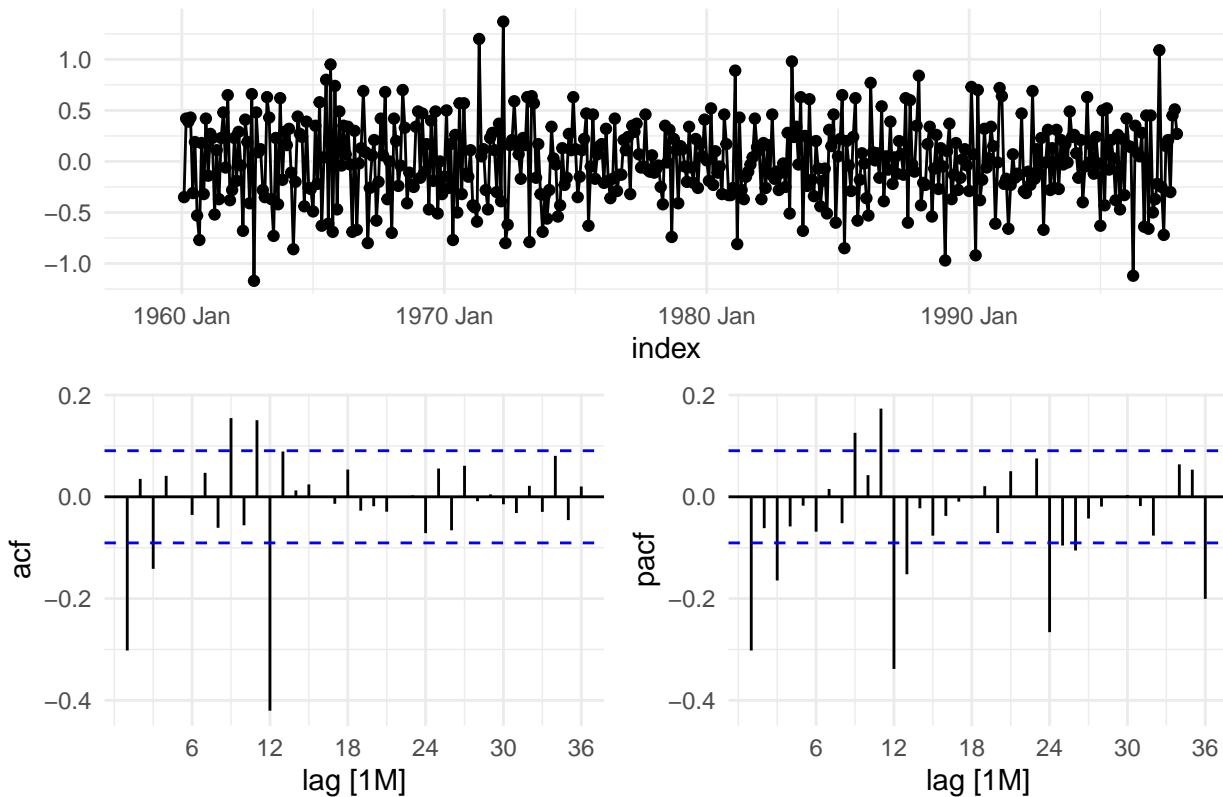
Here we can see that once we apply a first difference and a seasonal difference, the test returns 0, indicating no additional differences are required. It is important that we correctly identify the correct number of differences as we could potentially introduce false dynamics or autocorrelations into the time series.


We will now use the ACF and PACF plots of our double differenced data to determine an appropriate ARIMA model via brute force. The significant spike at lag 3 in the ACF suggests a non-seasonal MA(3) component. The significant spike at lag 12 in

the ACF suggests a seasonal MA(1) component. So, we might begin with an ARIMA(0,1,3)(0,1,1)12 model, indicating a first difference, a seasonal difference, and non-seasonal MA(3) and seasonal MA(1) component. The output of fitting this model appears below.

```
##
## Call:
## arima(x = co2_df$value, order = c(0, 1, 3), seasonal = list(order = c(0, 1,
##     1), period = 12))
##
## Coefficients:
##           ma1      ma2      ma3     sma1
##       -0.3394  -0.0180  -0.0973  -0.8538
## s.e.   0.0475   0.0497   0.0467   0.0256
##
## sigma^2 estimated as 0.0816:  log likelihood = -83.43,  aic = 176.86
```


Time Series of Residuals
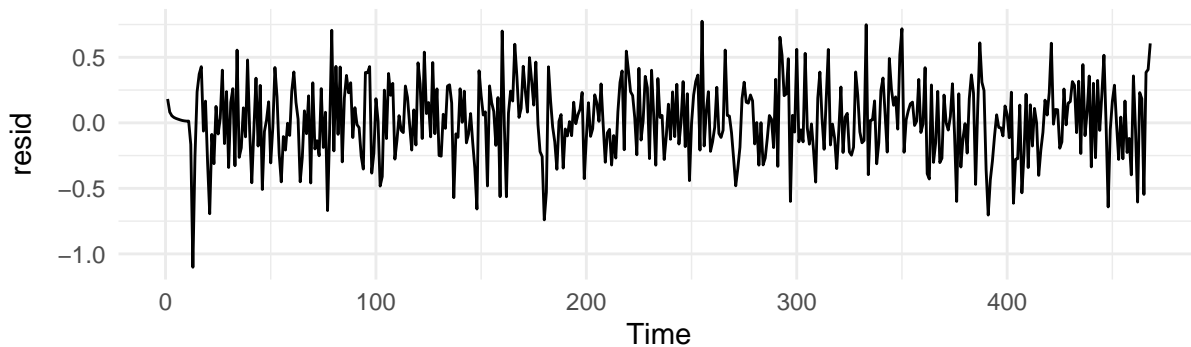
ACF of Residuals
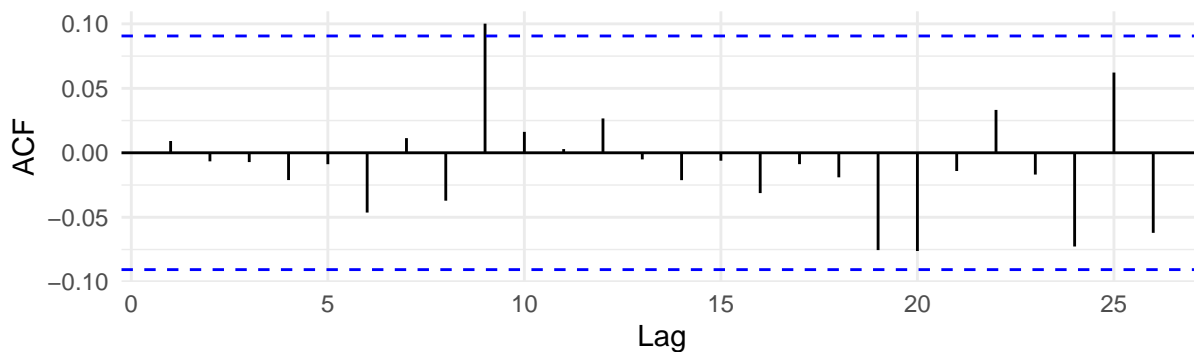
```
Box.test(resid,lag=10, fitdf=0, type="Lj")
```

```
##
##  Box-Ljung test
##
## data:  resid
## X-squared = 7.0116, df = 10, p-value = 0.7243
```

The ACF plot indicates that the residuals are following a white noise process as each autocorrelation is close to zero and 95% of spikes lie within significance thresholds. There is one small but significant spike at lag 8. Additionally, since the p-value from the Box-Ljung test is greater than 0.05, we fail to reject the null hypothesis and conclude that there is no serial correlation in the data.

If we had evaluated the PACF as well, we may have chosen ARIMA(3,1,0)(1,1,1)12 model, indicating a first difference, a seasonal difference, non-seasonal AR(3) and MA(1) components, and seasonal AR(1) and AR(1) component. The significant spike at lag 3 in the PACF suggests a non-seasonal AR(3) component. The significant spike at lag 12 in the PACF suggests a seasonal AR(1) component. However, by comparing the AIC values of the first and second model, the first selection was a better fit based on the lower AIC value.

```
## 
## Call:
## arima(x = co2_df$value, order = c(3, 1, 1), seasonal = list(order = c(1, 1,
##     1), period = 12))
## 
## Coefficients:
##          ar1     ar2      ar3      ma1     sar1     sma1
##       0.2972  0.0842  -0.0635  -0.6335   0.0394  -0.8637
## s.e.  0.1822  0.0785   0.0578   0.1765   0.0545   0.0276
## 
## sigma^2 estimated as 0.0815:  log likelihood = -83.15,  aic = 180.29
```

We can also try automatically fitting the model:

```
## Series: value
## Model: ARIMA(1,1,1)(1,1,2)[12]
## 
## Coefficients:
##          ar1      ma1     sar1     sma1     sma2
##       0.2569  -0.5847  -0.5489  -0.2620  -0.5123
## s.e.  0.1406   0.1204   0.5881   0.5703   0.4820
## 
## sigma^2 estimated as 0.08576:  log likelihood=-84.39
## AIC=180.78   AICc=180.97   BIC=205.5
```
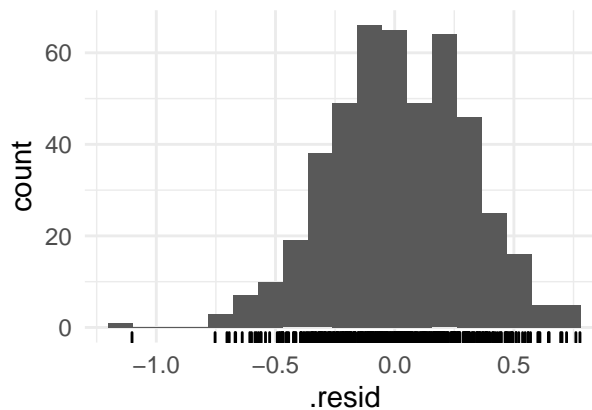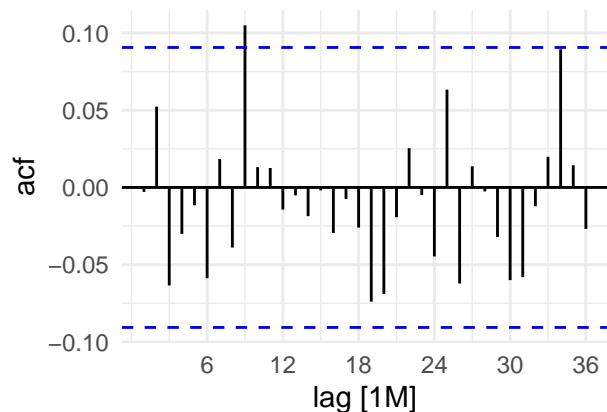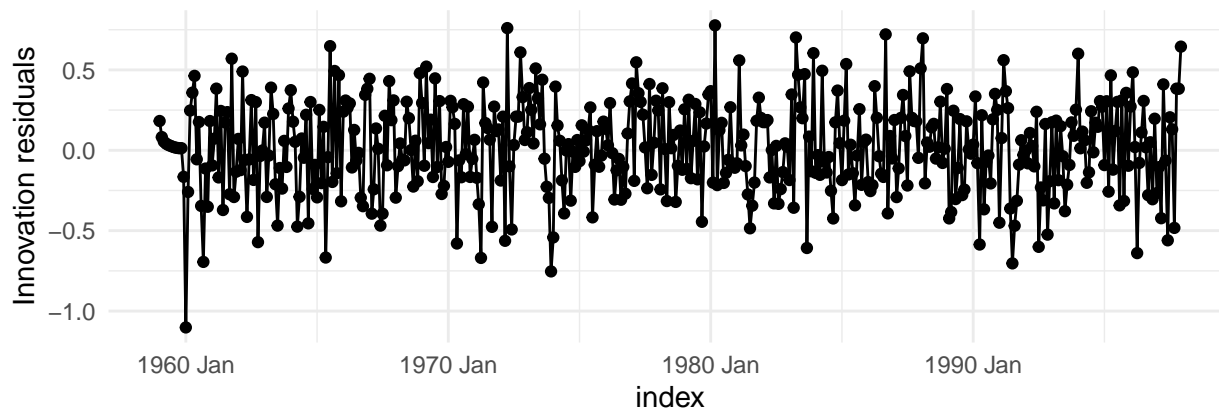


```
## # A tsibble: 468 x 3 [1M]
## # Key:       .model [1]
##    .model          index .resid
##    <chr>           <mth>  <dbl>
## 1 ARIMA(value) 1959 Jan 0.182
## 2 ARIMA(value) 1959 Feb 0.0821
## 3 ARIMA(value) 1959 Mar 0.0539
## 4 ARIMA(value) 1959 Apr 0.0411
## 5 ARIMA(value) 1959 May 0.0333
## 6 ARIMA(value) 1959 Jun 0.0276
```

```
##  7 ARIMA(value) 1959 Jul 0.0221
##  8 ARIMA(value) 1959 Aug 0.0177
##  9 ARIMA(value) 1959 Sep 0.0148
## 10 ARIMA(value) 1959 Oct 0.0129
## # ... with 458 more rows
```

```
##
##  Box-Ljung test
##
## data:  auto_resid[3]
## X-squared = 11.577, df = 10, p-value = 0.3143
```

The trend and ACF plots indicate that the residuals of the automatically fitted ARIMA model are following a white noise process. Autocorrelations are close to zero and 95% of spikes lie within significance thresholds. There is one small but significant spike at lag 9. Also, since the p-value from the Box-Ljung test is greater than 0.05, we fail to reject the null hypothesis and conclude that there is no serial correlation in the data. However, the automatically fitted ARIMA model does not outperform our first guess as the AIC in the autofit model is higher. We will continue to proceed with the ARIMA(0,1,3)(0,1,1)12.
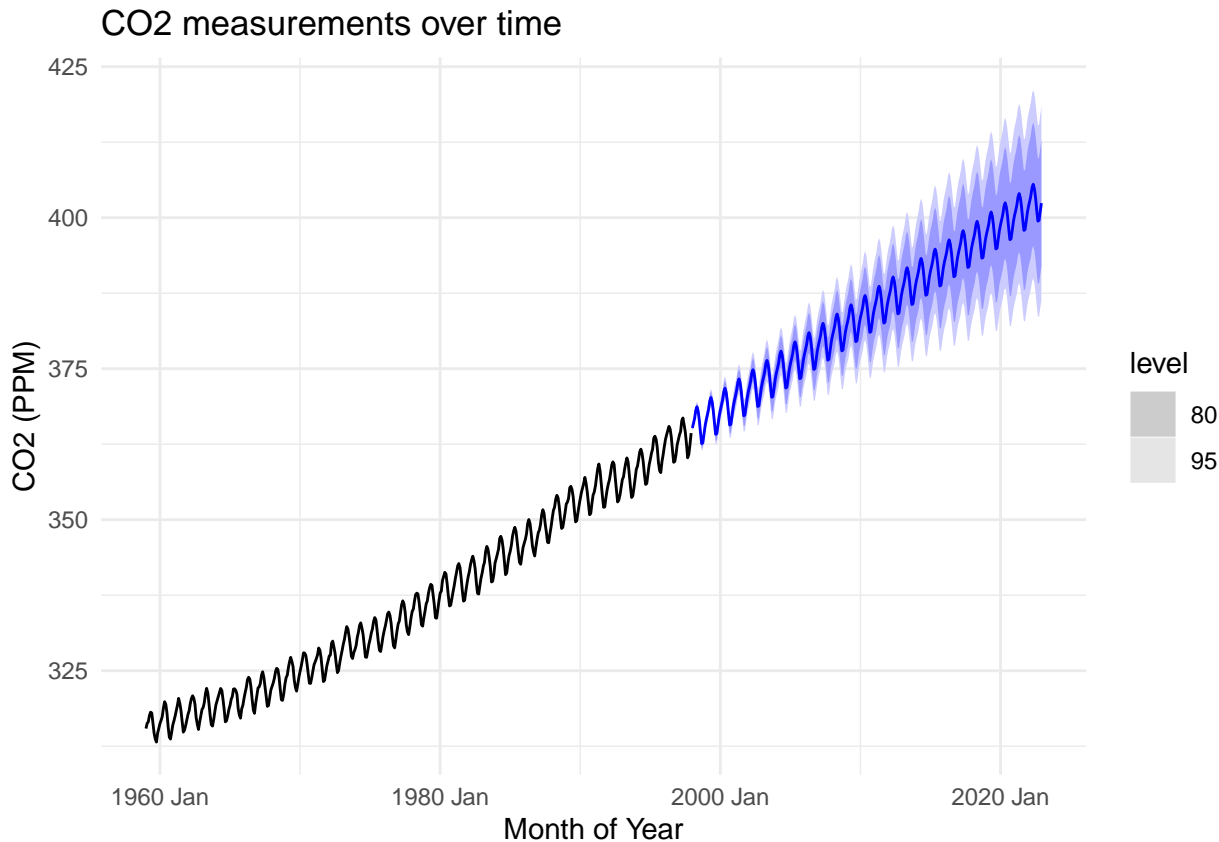
Again, we can see that these predictions continue to follow the trend and seasonal fluctuations that we noticed earlier in the data. However, our confidence intervals continue to widen with each year that we forecast.

1.5   Forecast atmospheric CO2 growth

If we generate predictions for when atmospheric CO2 is expected to be at 420 ppm and 500 ppm levels at the 95% confidence level, we can observe the wide confidence intervals noted in the previous graph. Predictions for 420ppm were as early as April 2022 and as late as December 2100. Surprisingly, when we make predictions on for 500ppm the confidence intervals narrow. According to this model's predictions, 500ppm CO2 levels will occur as early as March 2055 and as late as August 2086, as seen below.

```
## [1] "Earliest 420ppm Point Estimate: Apr 2022"
```

```
## [1] "Latest 420ppm Point Estimate: Dec 2100"
```

```
## [1] "Range of 420ppm Interval: 78.6666666666665"
```

```
## [1] "Earliest 500ppm Point Estimate: Mar 2055"
```

```
## [1] "Latest 500ppm Point Estimate: Aug 2086"
```

```
## [1] "Range of 500ppm Interval: 31.416666666667"
```

When we look specifically at CO2 levels in 2100, we can expect the ppm values to be around 525 throughout the course of the year with fairly wide confidence intervals about 100ppms lower or higher.

```
##       time_index  point_forecast              lo_95           high_95
##        "yearmon"       "numeric"        "numeric"        "numeric"
```

```
##     time_index point_forecast     lo_95  high_95
## 1     Jan 2100        521.6461 420.7014 622.5907
## 2     Feb 2100        522.4843 421.4223 623.5463
## 3     Mar 2100        523.3405 422.1614 624.5197
## 4     Apr 2100        524.6894 423.3936 625.9852
## 5     May 2100        525.2589 423.8467 626.6712
## 6     Jun 2100        524.5735 423.0449 626.1021
## 7     Jul 2100        523.0702 421.4254 624.7151
## 8     Aug 2100        521.0166 419.2556 622.7775
## 9     Sep 2100        519.1595 417.2825 621.0364
## 10    Oct 2100        519.2845 417.2918 621.2773
## 11    Nov 2100        520.7157 418.6073 622.8241
## 12    Dec 2100        522.1267 419.9027 624.3507
```

2100 CO2 predictions
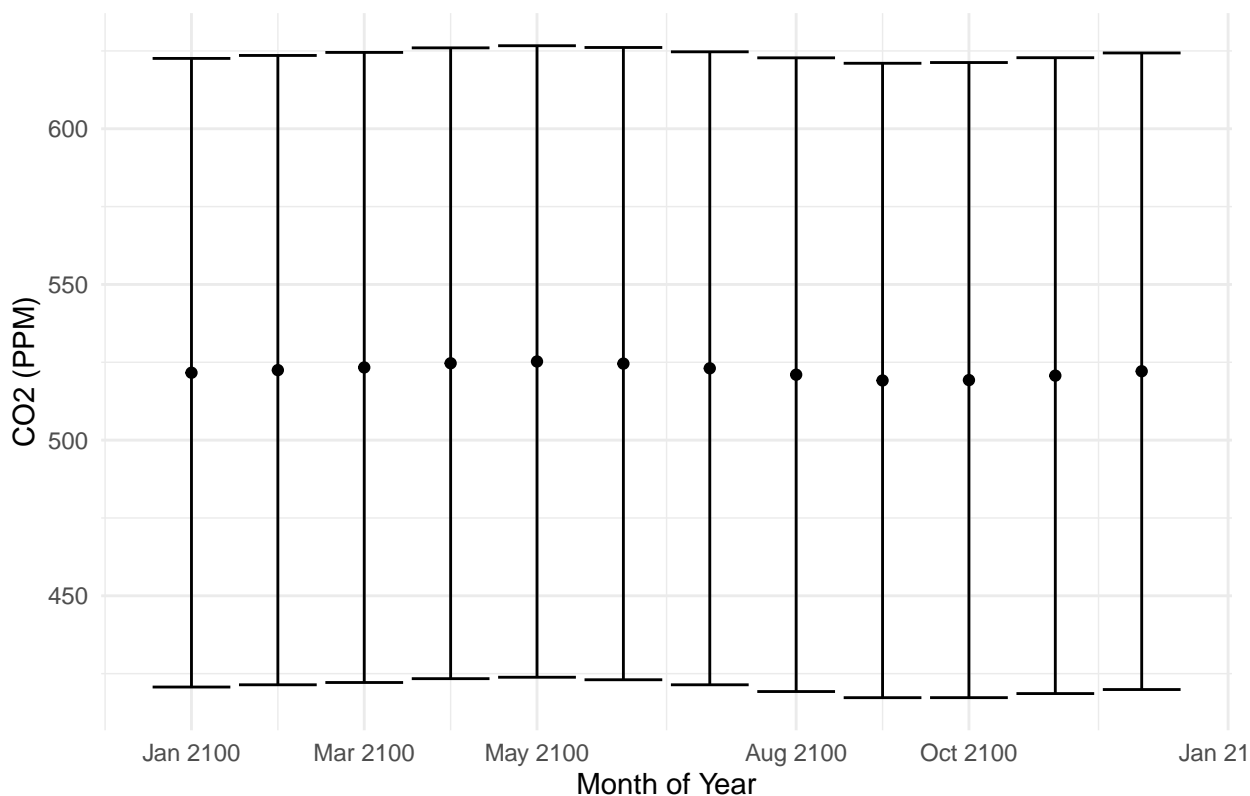
Based on the very wide confidence intervals, we're not very confident that these are accurate predictions. To assess the accuracy of our predictions more methodically, we can create a training set and a test set. From the plot, we can see that our predictions widen from the actuals with each consecutive year in just the last 8 years. While our predictions are probably a best guess given the information that we have, the ARIMA model assumes that seasonality is fixed over time, and this may be a strong assumption to make given that there are other factors that could likely impact CO2 over time.

When we observe the residuals, we also notice that there are two significant spikes in the ACF plot at lag 3 and lag 9. Our p-value from the Box-Ljung test is also greater than 0.05, meaning that we fail to reject the null hypothesis of no serial correlation. However, since both of these observations are close to the significance threshold, this may be cause for concern around the predictive power of our model, particularly for such a far projection in time, as seen below.

```
## Series: value
## Model: ARIMA(0,1,1)(2,1,2)[12]
##
## Coefficients:
##            ma1      sar1      sar2      sma1      sma2
##        -0.3595   -0.1484   -0.0750   -0.6937   -0.1217
## s.e.    0.0575    0.5648    0.0658    0.5646    0.4868
##
## sigma^2 estimated as 0.08563:  log likelihood=-63.83
## AIC=139.66    AICc=139.9    BIC=162.77
```

Forecasts for co2 levels
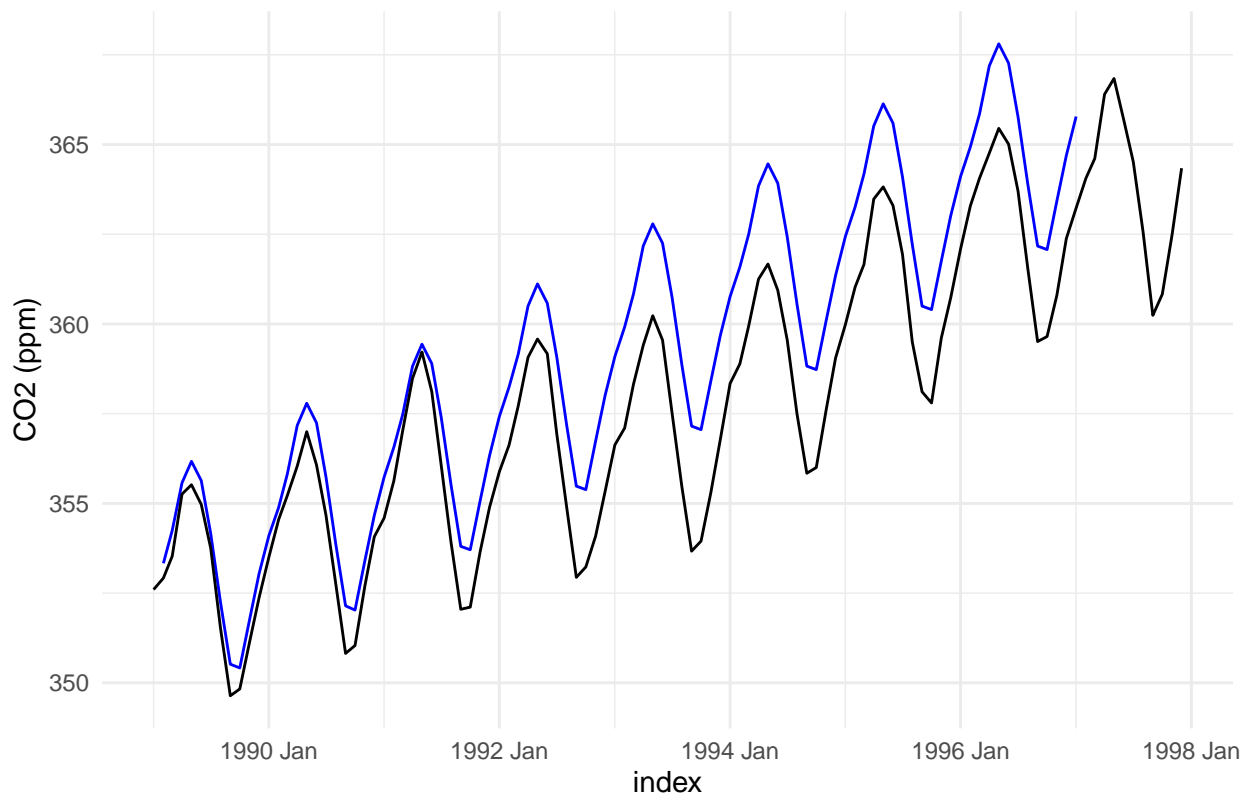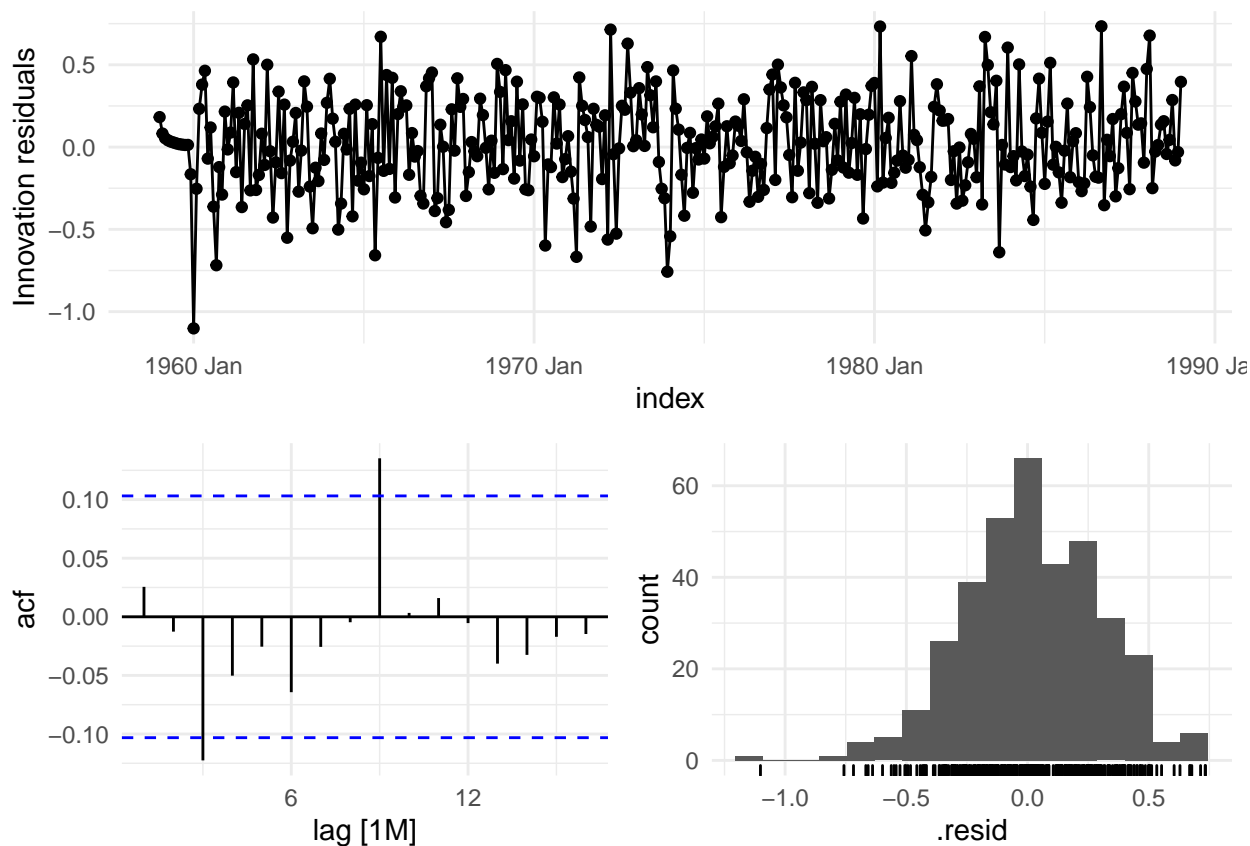
```
## # A tibble: 1 x 3
##   .model      lb_stat lb_pvalue
##   <chr>         <dbl>     <dbl>
## 1 ARIMA(value)   16.8    0.0778
```

2 Report From the Perspective of the Present

2.1 Introduction

We now ask the question of whether the data collected from January 1998 to the present has different characteristics which could lead us to additional or different conclusions than those which the dataset from 1958 until 1997 yielded.

Based on Keeling's autobiography, we know that there were differences between Keeling's data and the National Oceanic and Atmospheric Administration (NOAA) data until 1983, where they reconciled their measurements. However, despite this, we assume that the both datasets behave similarly.

It is worth noting that the "present" dataset contains only about 24 years, or 293 months, compared to the 39 years, or 468 months contained in the "1997" dataset.

2.2 Create a modern data pipeline for Mona Loa CO2 data.

As instructed we pulled created a pipeline to pull down the weekly Mona Loa CO2 data.

Within the dataset the weeks ending on dates; 2008-06-29, 2008-07-06 , 2008-07-13 , 2005-10-16, contained the measurement value -999.99, which cannot be accurate. We "stubbed" those four dates with values similar to the weekly values for the weeks directly before or after. We did not remove those data outright as doing so could cause problems within the utilization of the time series downstream.

Weekly Mean CO$_2$

1998 through Present

ACF 1998 through Present

PACF 1998 through Present
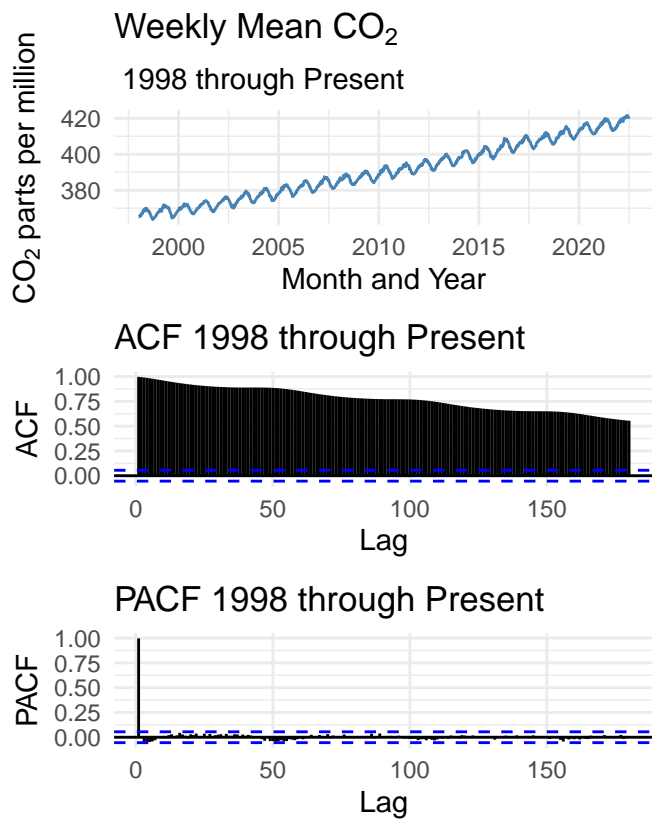
Histogram 1998 through Present
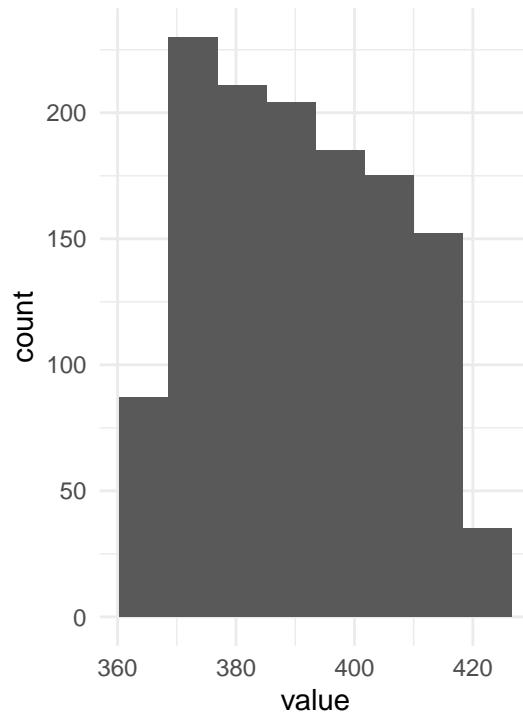
As with the "1997" dataset, we see a clear increasing trend with the same seasonal "wave" pattern. Seasonal fluctuation does not appear to increase with time, suggesting, again, that an additive model explains the decomposition in this series. The time series is not stationary, as its mean increases. Variance presents as constant.

The ACF displays a slow, over-time, decline consistent with a trending time series. While the ACF plot of the "1997" plot remains above the significance line until lag 140, this ACF plot never crosses the significance line.

The PACF of this time series has a strong postive spike a lag 1 followed by a pattern consistant with partial autocorrelation.

Our histogram of has no values lower than 365, or higher than 421, and that these values are not normally distributed.
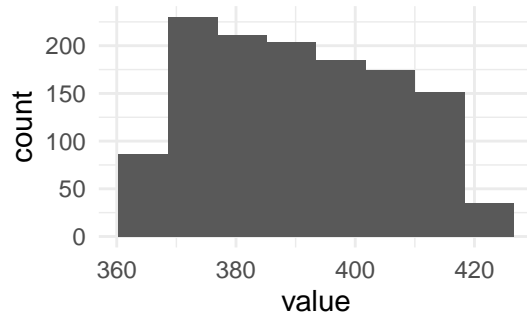
In order to better understand the differences between these two datasets we examine them side by side.
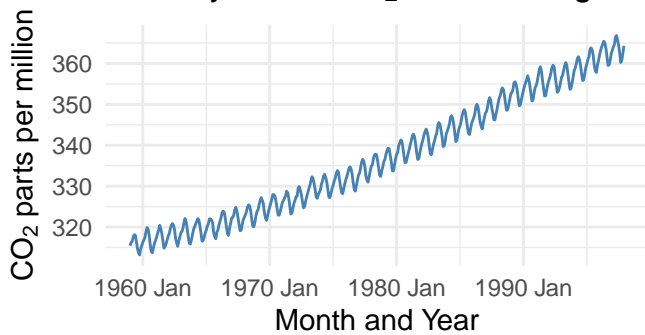
Weekly Mean CO$_2$ 1998 through Present
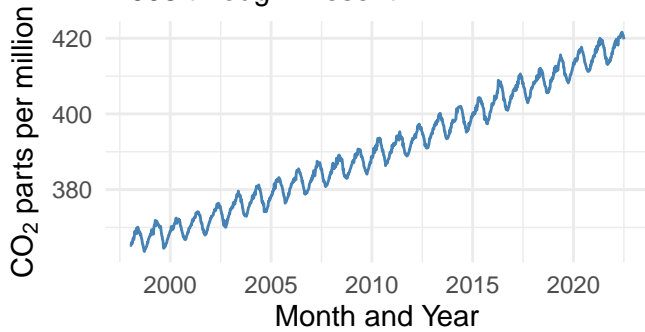
Histogram 1998 through Present

Monthly Mean CO$_2$ 1958 through 1997

Histogram 1958 through 1997
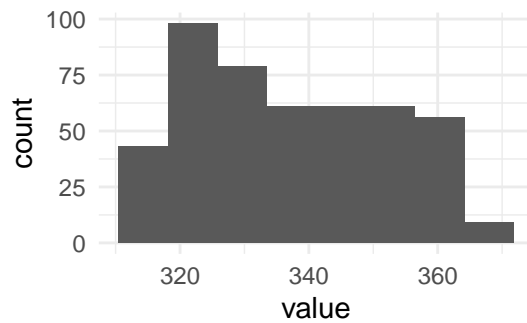
When we start looking graphing the two datasets side by side, the difference of one containing 39 years, and the other 24, begins to hamper our understanding.

We take only the last 293 months of the "1997" dataset and graph it on the same space as the "present" dataset.

293 month comparison: 1973 through 1997, and 1998 through present.

Periods
— 1973 through 1997
— 1997 through present

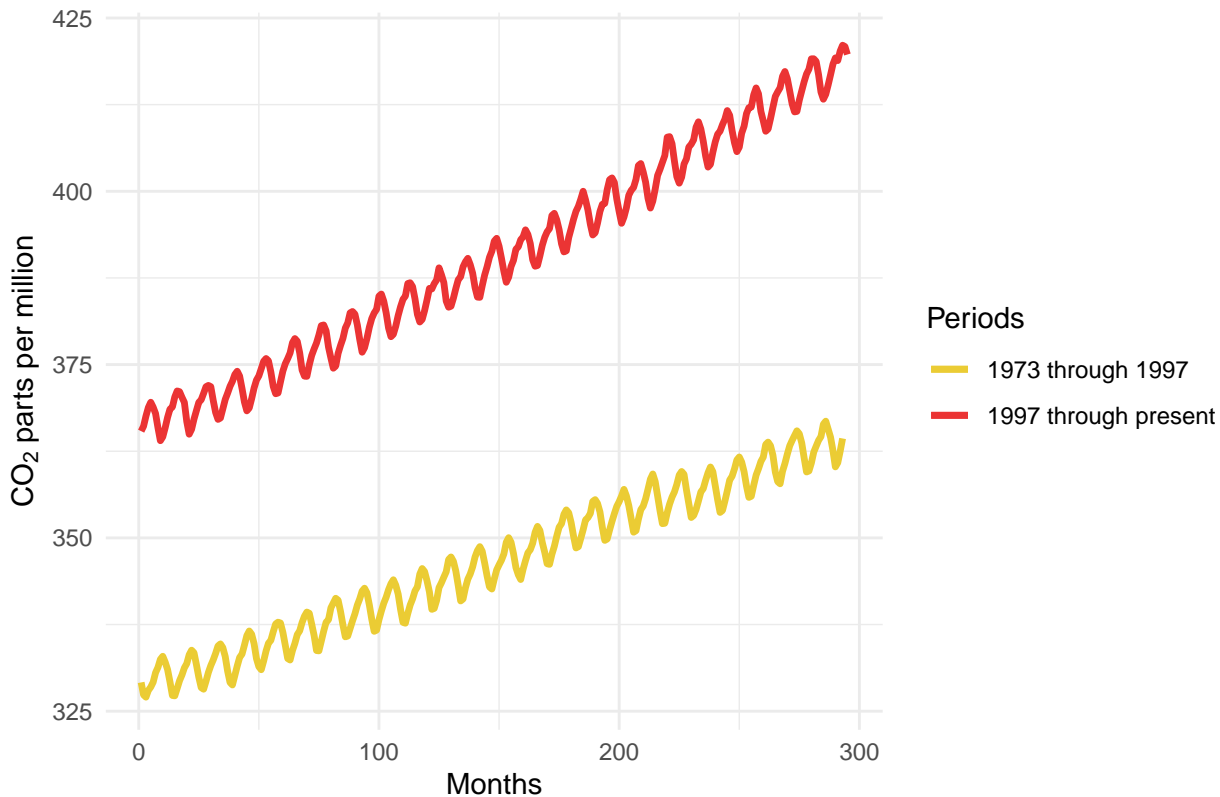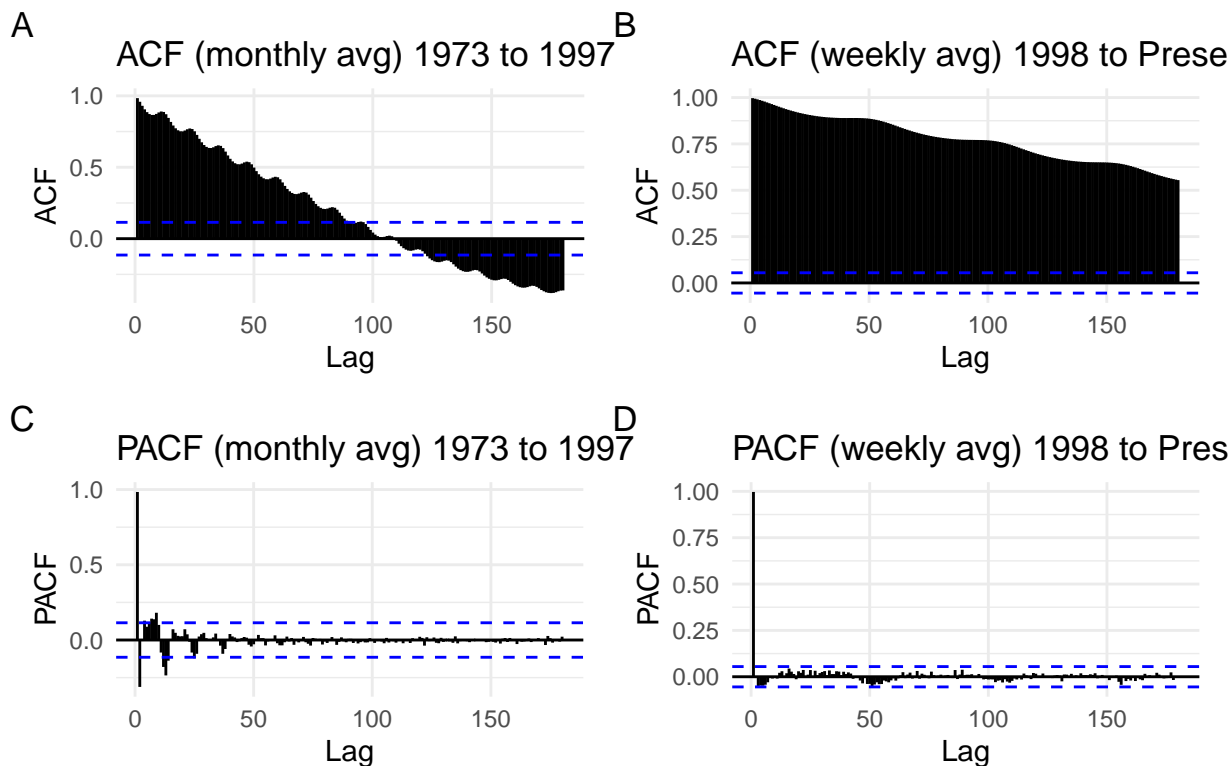In the 293 months prior to Dec 31, 1997, the atmospheric CO2 rose by 39.82 PPM. By comparison, in the 293 months prior to May 31, 2022 (the most recent month we have a measurement for), the atmospheric CO2 rose by 57.054 PPM. This is a difference of 17.234 PPM during the same time interval. This could indicate an acceleration in the increase in PPM.

We apply the same "parsing" of the original "1997" dataset and compare the parsed dataset's ACF and PACF graphs to those of the "present" dataset.

ACF and PACF comparison

A

ACF (monthly avg) 1973 to 1997

B

ACF (weekly avg) 1998 to Prese

C

PACF (monthly avg) 1973 to 1997

D

PACF (weekly avg) 1998 to Pres

Parsing the "1997" dataset, to include 293 months only, yields a ACF which crosses the significance line 100, while the weekly data does not cross the significance line. We notice that the amplitude of oscillation on the "1997" datasets' ACF, is greater than the amplitude of oscillation in the "present" dataset's ACF.

Next we examine the seasonality of the "present" comparing it to the "1997" seasonality.

In order to the present dataset's seasonality 1997 seasonality we aggregate values within the weekly dataset to monthly values, for the purpose of this comparison.

Seasonal plot: Monthly mean CO$_2$
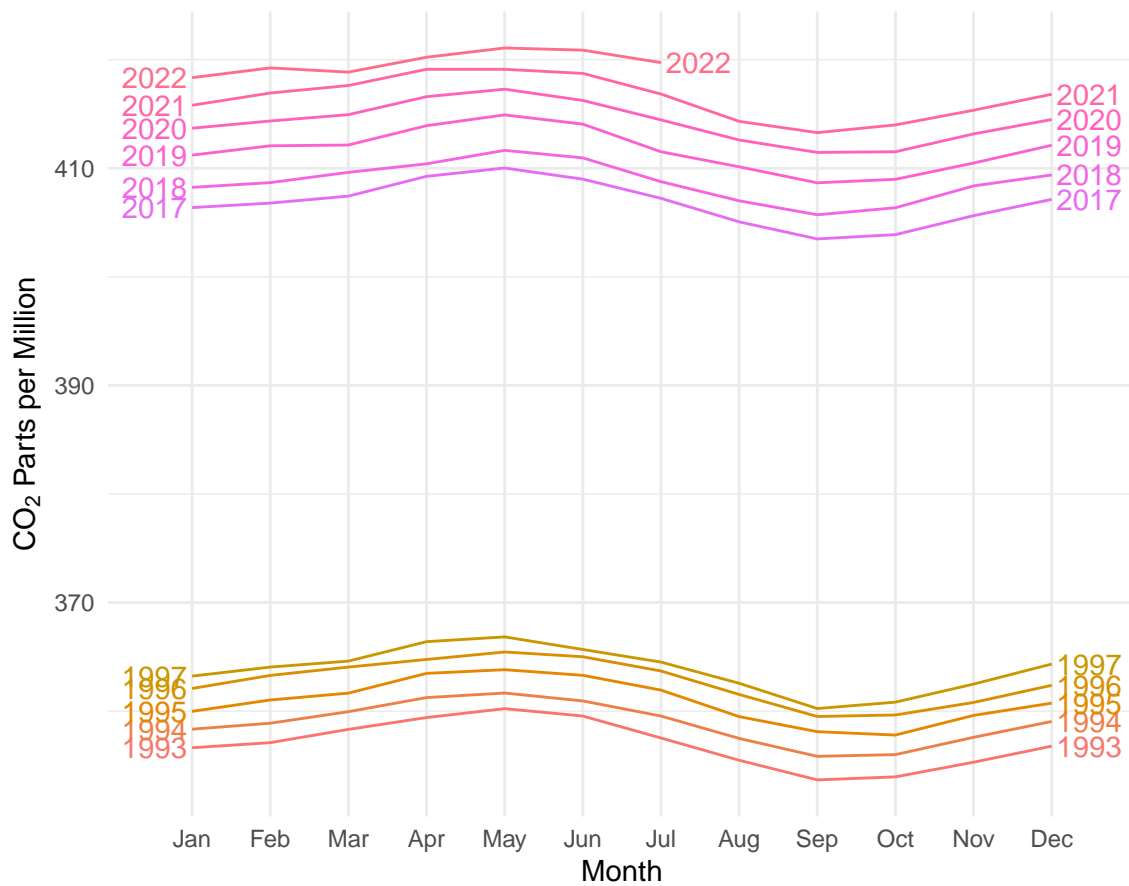
In the plot above we look at the last 5 years in the "1997" dataset and the last 5 years of the "present" dataset.
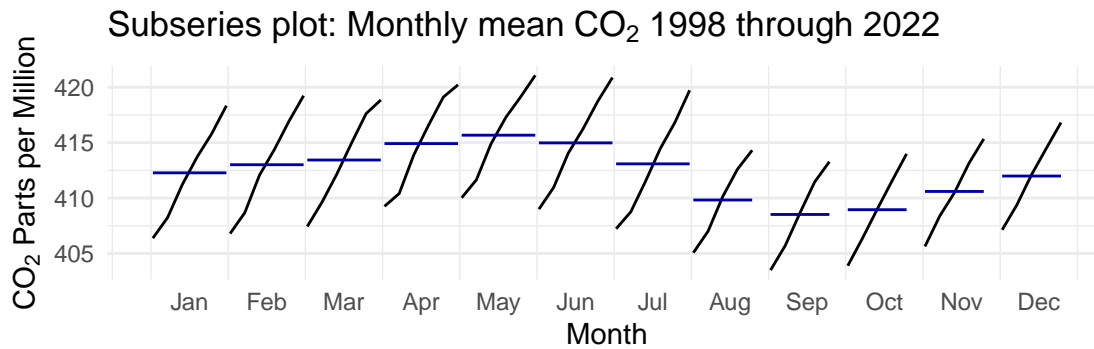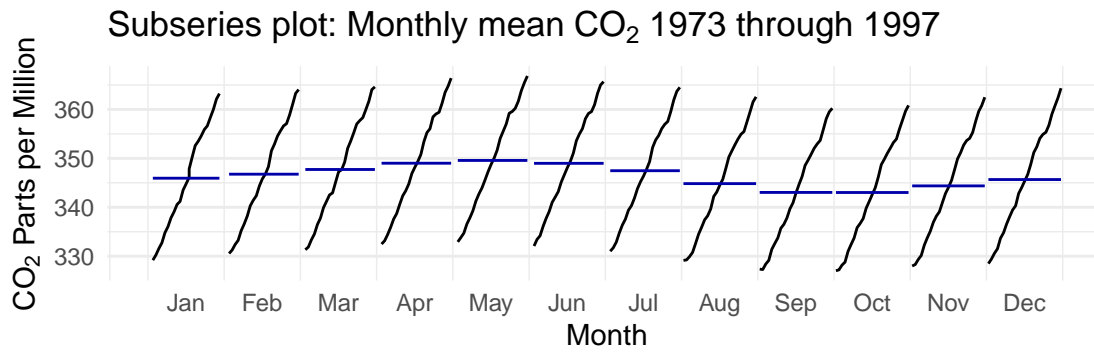
What we are trying to ascertain is if the seasonality has become more extreme. It is difficult to ascertain if the seasonality has become more extreme based on the graph above.

We next compare the two subseries graphs looking at each dataset.

Subseries plot: Monthly mean CO_2 1973 through 1997

Subseries plot: Monthly mean CO_2 1998 through 2022

With this new plot it becomes clearer that the yearly oscillations in PPM are more extreme in the present dataset than in the one ending in 1997.

That is to say the seasonality effect shows a noticeable increase in the magnitude of the fluctuations in the second dataset.

We now remove the seasonal fluctuations in both datasets and graph the increase in atmospheric CO2 PPM.

In order to do this we aggregate our weekly data to monthly data in order to compare on the same scales.

293 month comparison: 1973 through 1997, and 1998 through present
Smoothed
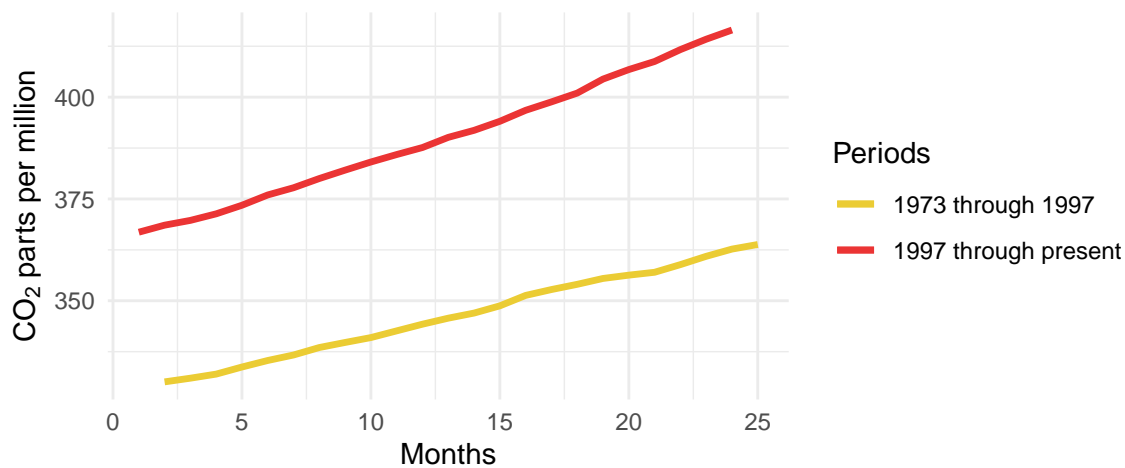
We now perform additive and multiplicative decomposition to remove the trend and seasonal movements from our dataset to confirm that the

STL decomposition

value = trend + season_year + remainder

Additive residuals

STL decomposition

log_value = trend + season_year + remainder
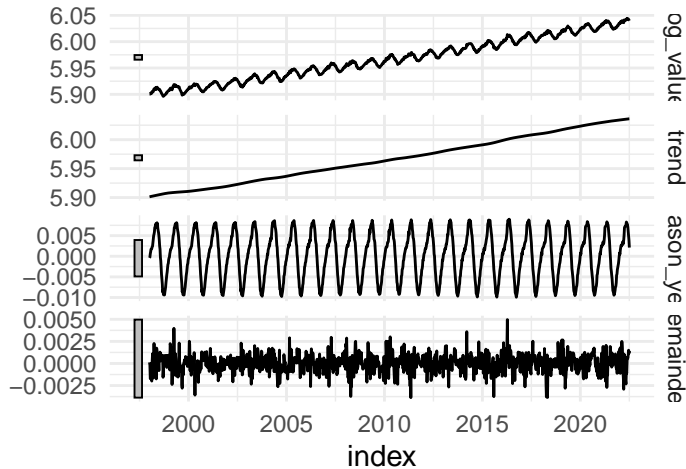
Multiplicative residuals

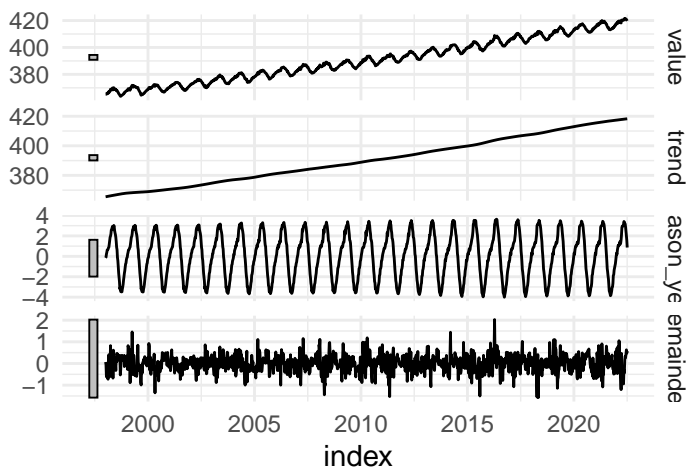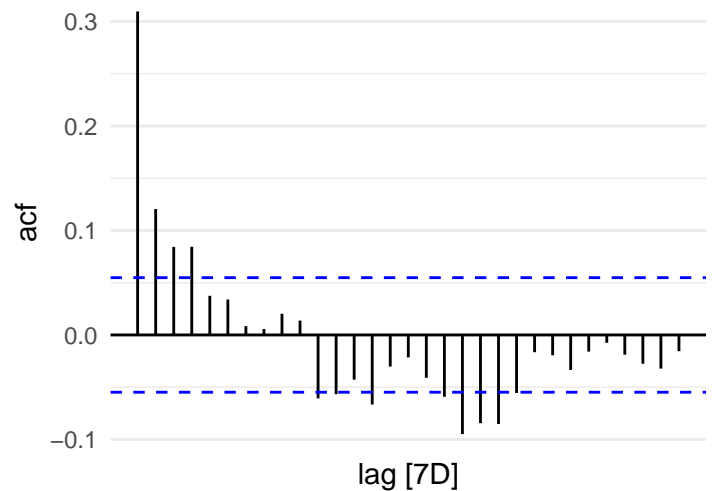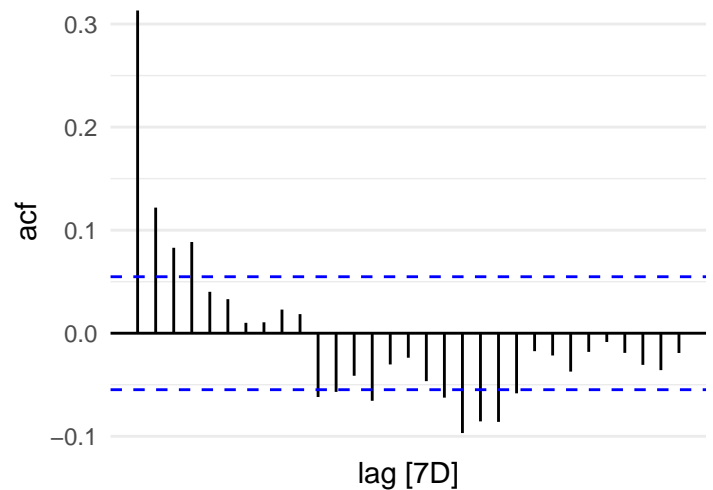These graphs confirm that the time series is trending upwards. In the additive decomposition plots (top left), we see that seasonal fluctuations increase over time. This confirms our side-by-side graph above showing the increase in seasonal fluctuations between the two time series.

The multiplicative "season_year" decomposition graph is stable, and not visibly increasing. This means like the "1997" time series, the "present" time series may be multiplicative.

Both residuals plots, again, show our time series as stationary, but not white noise. This means correlation in the data still exists.

Next we run the ADF test, as well as the PP test to see if the time series is stationary.

H0: The data are not stationary HA: The data are stationary

```
##
##   Augmented Dickey-Fuller Test
##
## data:  co2_present_ts
## Dickey-Fuller = -9.6925, Lag order = 5, p-value = 0.01
## alternative hypothesis: stationary
```

```
##
##   Phillips-Perron Unit Root Test
##
## data:  co2_present_ts
## Dickey-Fuller Z(alpha) = -71.682, Truncation lag parameter = 7, p-value
## = 0.01
## alternative hypothesis: stationary
```

The tests above indicate that we can reject the null hypothesis that the time series is non-stationary.

The Unit Root tests above indicate that this time series is difference stationary.
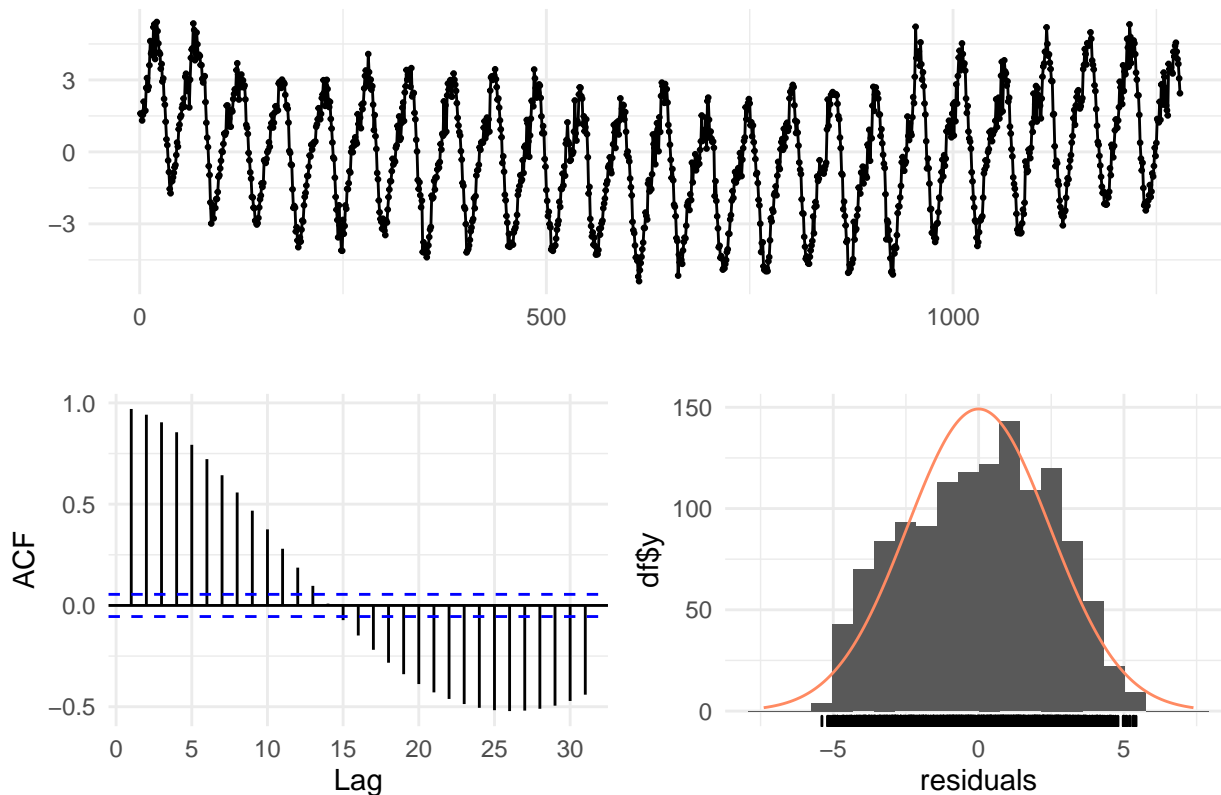
This contradicts our visual EDA, which shows a changing mean, given the strong upward trend.

We should note that the tests above are testing whether the time series are difference stationary. We apply a linear model and then graph its residuals.

Residuals


```
##
##  Breusch-Godfrey test for serial correlation of order up to 10
##
## data:  Residuals
## LM test = 1220.4, df = 10, p-value < 2.2e-16
```

Graphing the residuals appears to show a mean reverting process. That is to say the residuals are stationary and not trending, and appear to be normally distributed.

The oscillating ADF appears to show seasonality.

Finally, we run an ADF and PP test on the residuals, and we again, see that the null hypothesis of non-trend-stationary can be ignored.

H0: The data are not stationary HA: The data are stationary

```
## Warning in adf.test(trModel$residuals, alternative = "stationary", k = 5): p-
## value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  trModel$residuals
## Dickey-Fuller = -9.6925, Lag order = 5, p-value = 0.01
## alternative hypothesis: stationary
```

```
## Warning in pp.test(trModel$residuals, alternative = "stationary"): p-value
## smaller than printed p-value
```

```
##
##   Phillips-Perron Unit Root Test
##
## data:  trModel$residuals
## Dickey-Fuller Z(alpha) = -71.682, Truncation lag parameter = 7, p-value
## = 0.01
## alternative hypothesis: stationary
```

Despite the fact that the ADF and PP tests indicate the data series is stationary, our visual analysis of this data set suggests that it is non-stationary as its mean is increasing over time. We will ignore the ADF and PP test results as they can sometimes be of low power.

2.3   Compare linear model forecasts against realized CO2

In section 1 we fit two models, one linear model and one quadratic model. In this section we're going to compare both models to the "present" data set and see how well they predict future data.

2.3.1   Simple linear model with no polynomial terms or seasonality
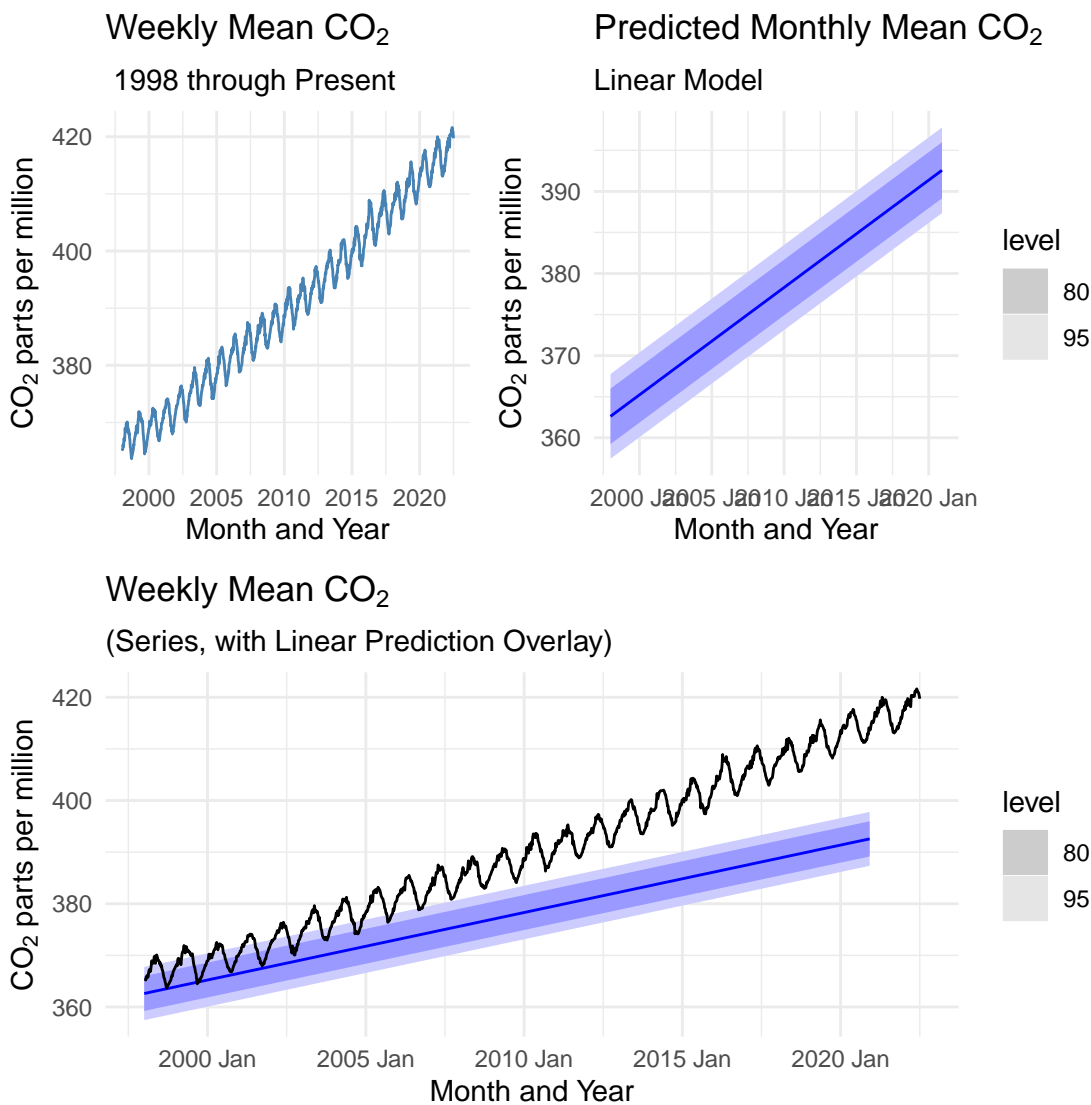
The simple linear model above which uses only trend as a coefficient has a slope that is too shallow to be useful in prediction.

2.3.2   Quadradic model with polynomial terms and seasonality

Below, we utilize a model with polynomial terms and seasonality.

Weekly Mean CO$_2$

1998 through Present

Predicted Monthly Mean CO$_2$

Quadratic Model w/ Trend & Seasonal Eff

Weekly Mean CO$_2$

(Series, with Quadratic Prediction Overlay)
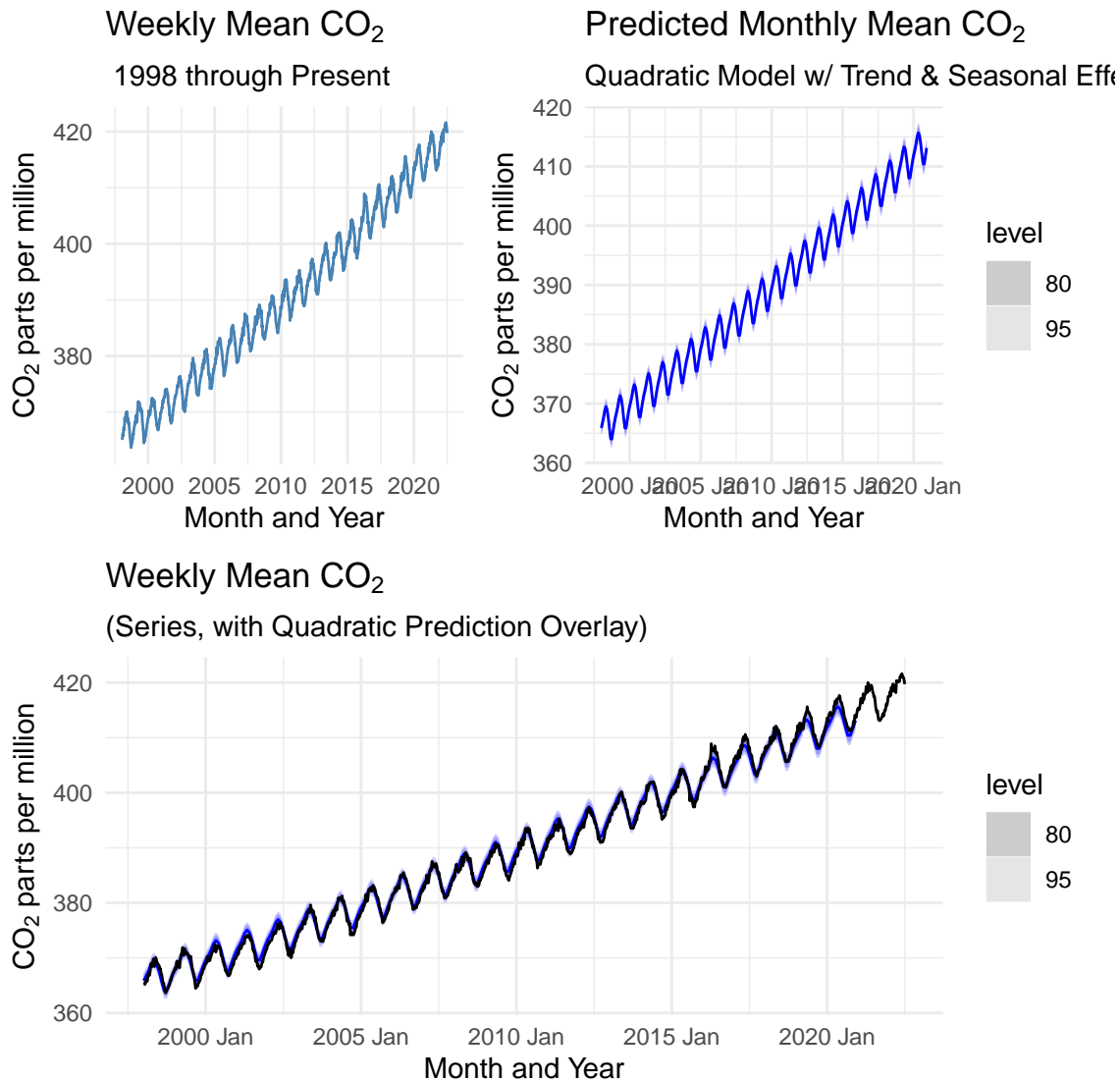
In the top-left, we plot the 1998-present time series, next to it, we plot the prediction yielded by the best-fitting model with polynomial terms and seasonality, model we developed against the "1997" data set. Beneath the two plots we overlay the original time series with the prediction. Towards the middle of time series, the fit is nearly perfect. At both the beginning and end of the time series, there are barely visible differences the prediction and the "real" underlying time series.

We predict 23 years from Dec 31, 1997, which means the predict values are through December 31 2021. Our "present" time series contains data "through" May 2022, which means we have an addition 5 months of data, indicated by the solid black line extending upwards out of the prediction.

The linear model is very accurate.

2.4 Compare ARIMA models forecasts against realized CO2

Weekly Mean CO$_2$
1998 through Present

Predicted Monthly Mean CO$_2$
(ARIMA Model w/ Trend & Seasonal Effec

Weekly Mean CO$_2$
(Series, with ARIMA Prediction Overlay)
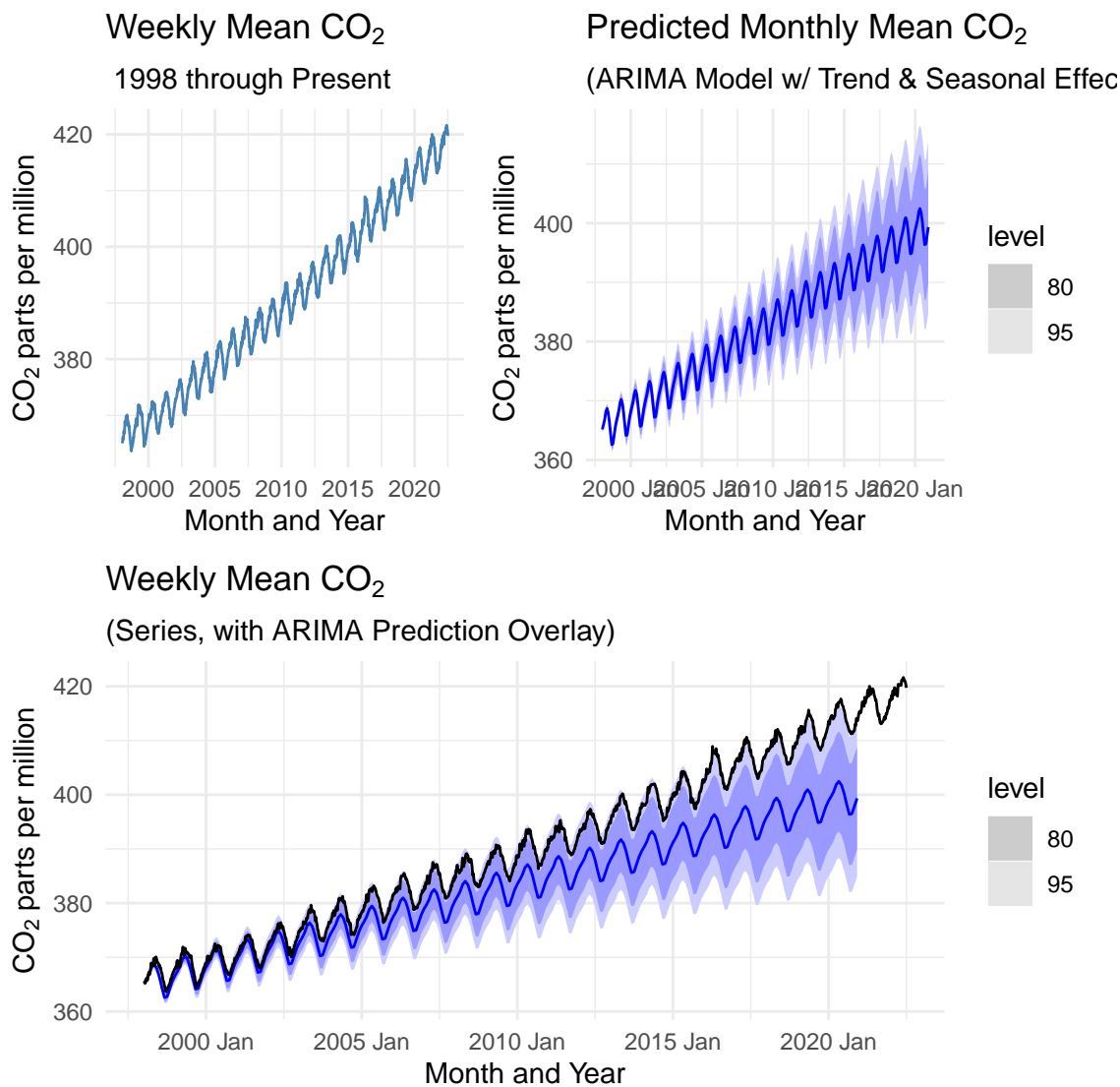
In the top-left graph, we plot the 1998-present time series, next to it, we plot the prediction yielded by the best-fitting ARIMA model we developed against the "1997" data set. Beneath the two plots we overlay the original time series with the prediction. The overlay shows a divergence between predicted ARIMA values and the actual trend.

We predict 23 years from Dec 31, 1997, which means the predict values are through December 31 2021. Our "present" time series contains data "through" May 2022, which means we have an addition 5 months of data, indicated by the solid black line extending upwards out of the prediction.

The ARIMA model does not predict as accurately as the quadradic model.

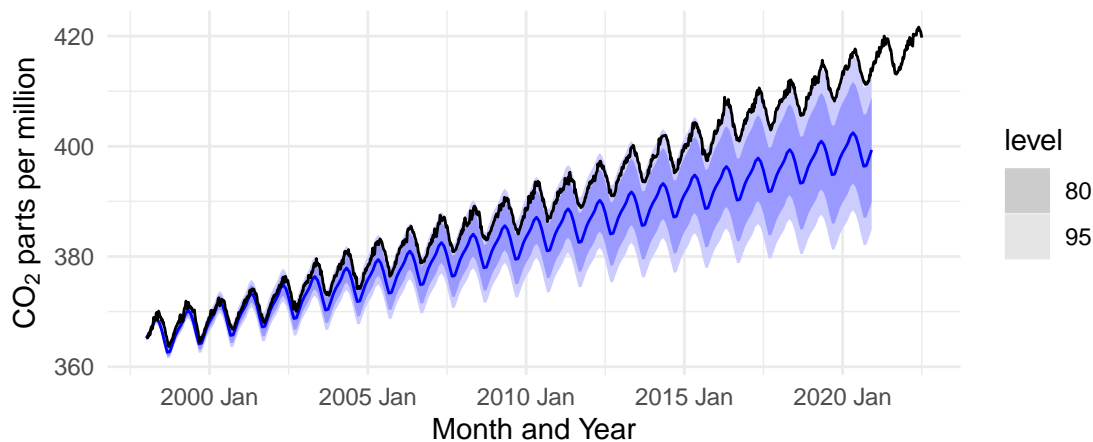To demonstrate this we graph both on top of one another below.

Weekly Mean CO_2

(Series, with Quadratic Prediction Overlay)
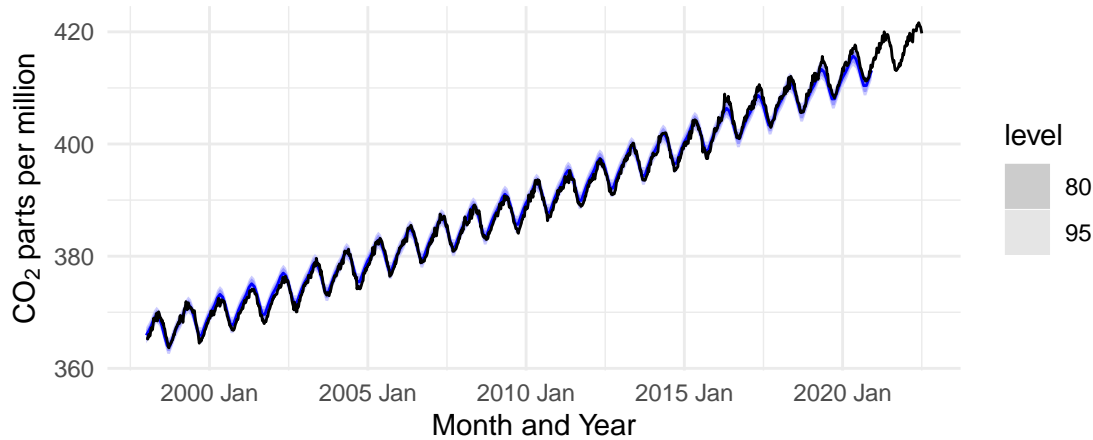
Weekly Mean CO_2

(Series, with ARIMA Prediction Overlay)

2.5 Evaluate the performance of 1997 linear and ARIMA models

From our present CO_2 data, we observe that the date that Mona Lua crossed 420 ppm was on April 25, 2021. When comparing these across the models in the 1997 report: in the `trend model`, it is forecasted that 420ppm will not be reached until December 2041, whereas the point forecast for the ARIMA model is April 2032. The `arima fit` also had a 95% confidence interval for 420ppm where lower bound is a very wide range of April 2022 and upper bound is December 2100. Concluding, our observed actual does not match what were forecasted in either of the models. Further, our observed date is also outside of the range of the confidence interval for their estimated ARIMA model. Let's next compute and observe the accuracies of both the linear model and ARIMA model that were estimated in 1997.

```
## # A tsibble: 6 x 3 [7D]
##    index      value log_value
##    <date>     <dbl>     <dbl>
## 1 2021-04-25  420.      6.04
## 2 2022-03-27  420.      6.04
## 3 2022-04-03  420.      6.04
## 4 2022-04-10  420.      6.04
## 5 2022-04-17  420.      6.04
## 6 2022-04-24  420.      6.04
```

```
## [1] "Forecasts from 1997 report"
```

```
## [1] "Point Forecast LM: 2041 Dec"
```

```
## [1] "Point Forecast ARIMA: Apr 2032"
```

```
## [1] "Lower bound forecast ARIMA: Apr 2022"

## [1] "Upper bound forecast ARIMA: Dec 2100"

## # A tibble: 1 x 10
##   .model       .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
##   <chr>        <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 trend_model  Test   12.0  13.7  12.0  3.02  3.02   NaN   NaN 0.971

## # A tibble: 1 x 10
##   .model    .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
##   <chr>     <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_fit Test   6.78  8.26  6.78  1.69  1.69   NaN   NaN 0.986
```

Within these accuracy measurements displayed we will choose to talk about RMSE for the remainder of this analysis. The RMSE, or Root Mean Squared Error is the standard deviation of the observed residauls and is used to measure the overall spread of residuals, and in our context will be used to assess the accuracy of our model fits. The RMSE of the `trend model` is `13.7` whereas the RMSE of the `arima fit` is `8.26`, concluding that the `arima fit` is the more accurate model of the two. Which is loosely true, though `arima fit` did not have confidence bounds for forecasted date of observed 420ppm, it still has a point estimate that is closer than that of the `trend model`.

2.6 ARIMA modeling

To better understand our data and predictive modeling for atmospheric CO_2 and to further emulate the analysis done in the 1997 report, we will move forward with using monthly data for this time series analysis. Another reason for why we have chose this method of using monthly data, is because the predictability using monthly data is theoretically and historically more accurate, due to monthly data;s ability flatten variations and pick up seasonality more reliably, thus leading to forecasts that are more stable.

Let's choose an ARIMA model to fit the present CO2 data from NOAA. First we will assess two datasets and two models, one dataset for seasonally adjusted data (SA) and one for non-sesonally adjusted data (NSA). Then, within both datasets, we will split the data into test and train sets with the last two years of data being the test set. Following the report from 1997 and observing the time series we have for our 1998-present data of CO_2 we will conclude with also applying linear differencing methods to our time series.

2.6.1 Seasonally adjusted data assessment

Now that we have all of our necessary data, let's get to creating models for both. First let's assess the number of differences needed for the seasonally adjusted data to be stationary, since this data already has the season component stripped out, we know there are no seasonal differences needed. Let's also assess the acf and pacf.

Assessing the ACF and PACF plots below, the ACF has a slow decay while the PACF has a large spike at lag 1, indicating that this series has a unit root and that the process is of AR(1).

Seasonally adjusted ACF

Seasonally adjusted PACF
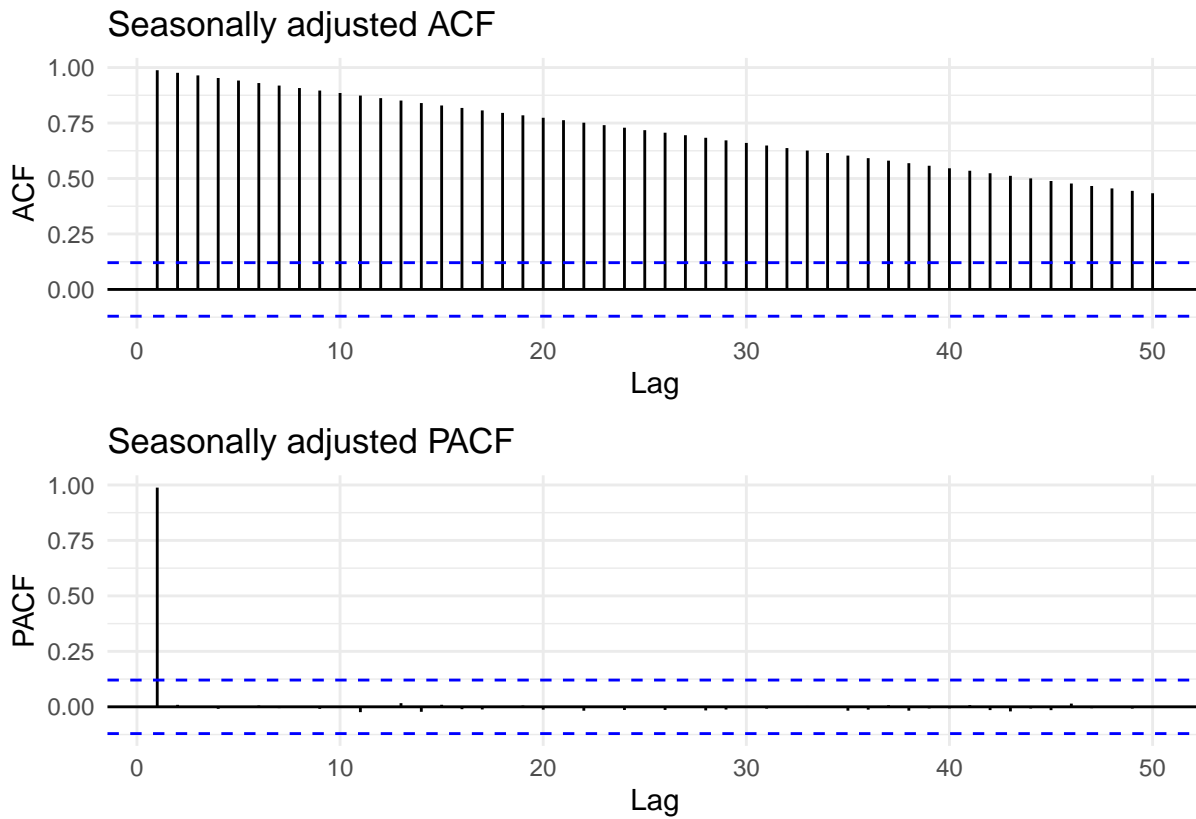
Next, we run a KPSS test to test for stationarity, and since the pvalue is smaller than 0.05 critical value, we know the data is not stationary. We then run ndiffs to check how many differences are needed to make the data stationary, which is 1. Let's next difference the data once and assess the ACF and PACFs again.

```
## # A tibble: 1 x 3
##   .model kpss_stat kpss_pvalue
##   <chr>      <dbl>       <dbl>
## 1 stl         4.48        0.01
```

```
## # A tibble: 1 x 2
##   .model ndiffs
##   <chr>   <int>
## 1 stl         1
```

Seasonally adjusted ACF 1 diff

Seasonally adjusted PACF 1 diff
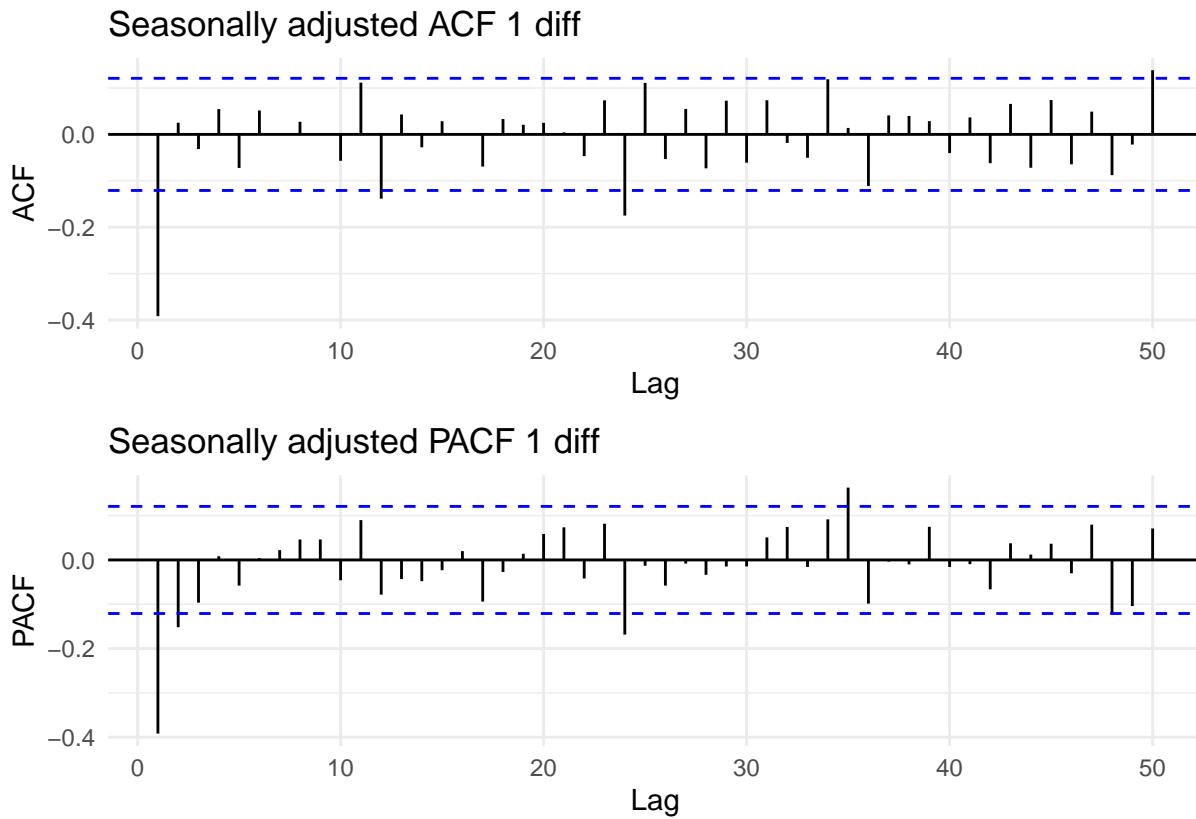
Looking at the SA differenced ACF and PACF, the significant spike at lag=1 in the ACF suggests a non-seasonal MA(1) component while the PACF yields no new information. Let's fit a couple of models to compare, one being an $ARIMA(1, 1, 1)$ and one that is automatically computed through the `ARIMA` func call while adjusting to assure the estimated model retains no seasonal components.

```
## Series: sa_co2_present_train$season_adjust
## ARIMA(1,1,1) with drift
##
## Coefficients:
##           ar1      ma1    drift
##       -0.0341  -0.4334   0.1801
## s.e.   0.1339   0.1212   0.0101
##
## sigma^2 = 0.09048:  log likelihood = -55.1
## AIC=118.2    AICc=118.36    BIC=132.49
```

```
## Series: season_adjust
## Model: ARIMA(0,1,1) w/ drift
##
## Coefficients:
##           ma1  constant
##       -0.4604    0.1801
## s.e.   0.0541    0.0100
##
## sigma^2 estimated as 0.09016:  log likelihood=-55.13
## AIC=116.27    AICc=116.36    BIC=126.98
```

Assessing the summaries from the model we chose to fit $ARIMA(1, 1, 1)$ with drift and without a seasonal-ARIMA process, we get an $AIC = 118$ and $BIC = 132$. Furthermore, the algorithmically chosen model gave an $ARIMA(0, 1, 1)$ model with $AIC = 116$ and $BIC = 126$. In either models, we observe that the automatic ARIMA call yields lower AIC and BIC values, which leads us to choose $ARIMA(0, 1, 1)$ for the remainder of our analysis.

Now that we have our model, let's check the residuals for stationarity.

Residual Plot SA Model
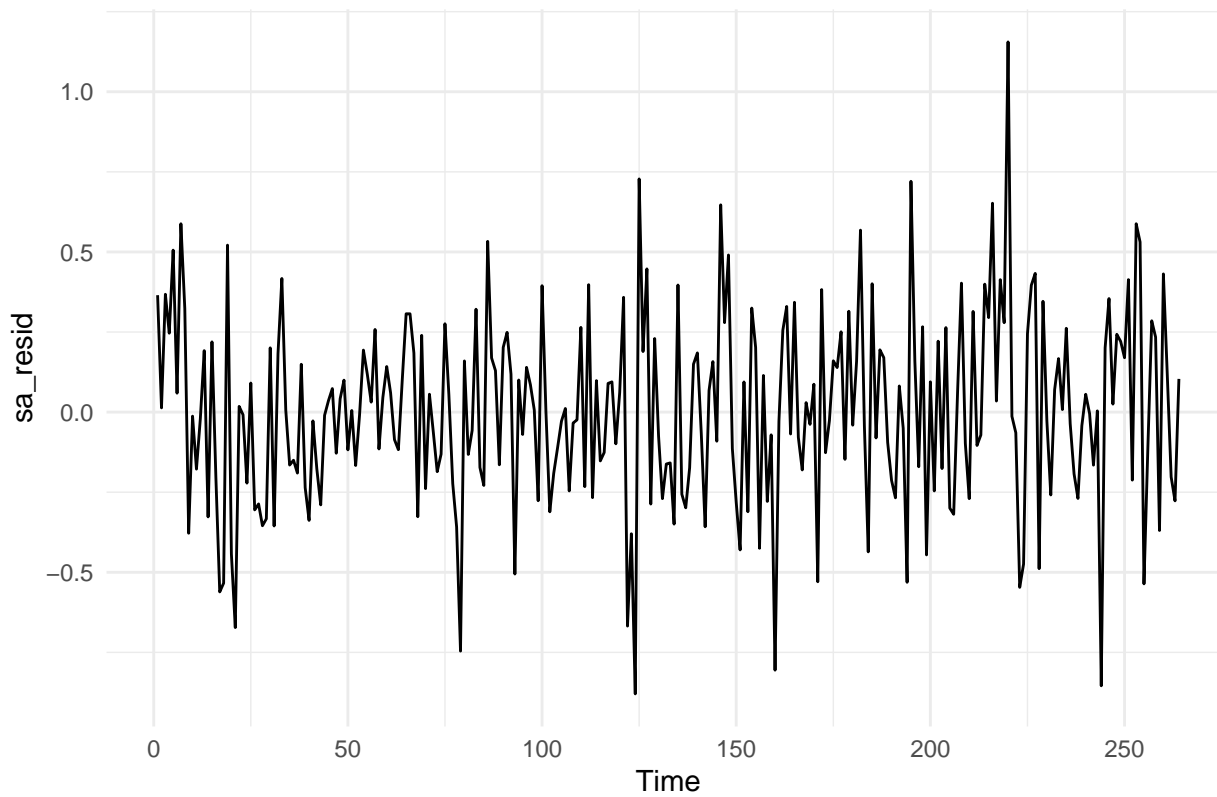
ACF Plot SA Model
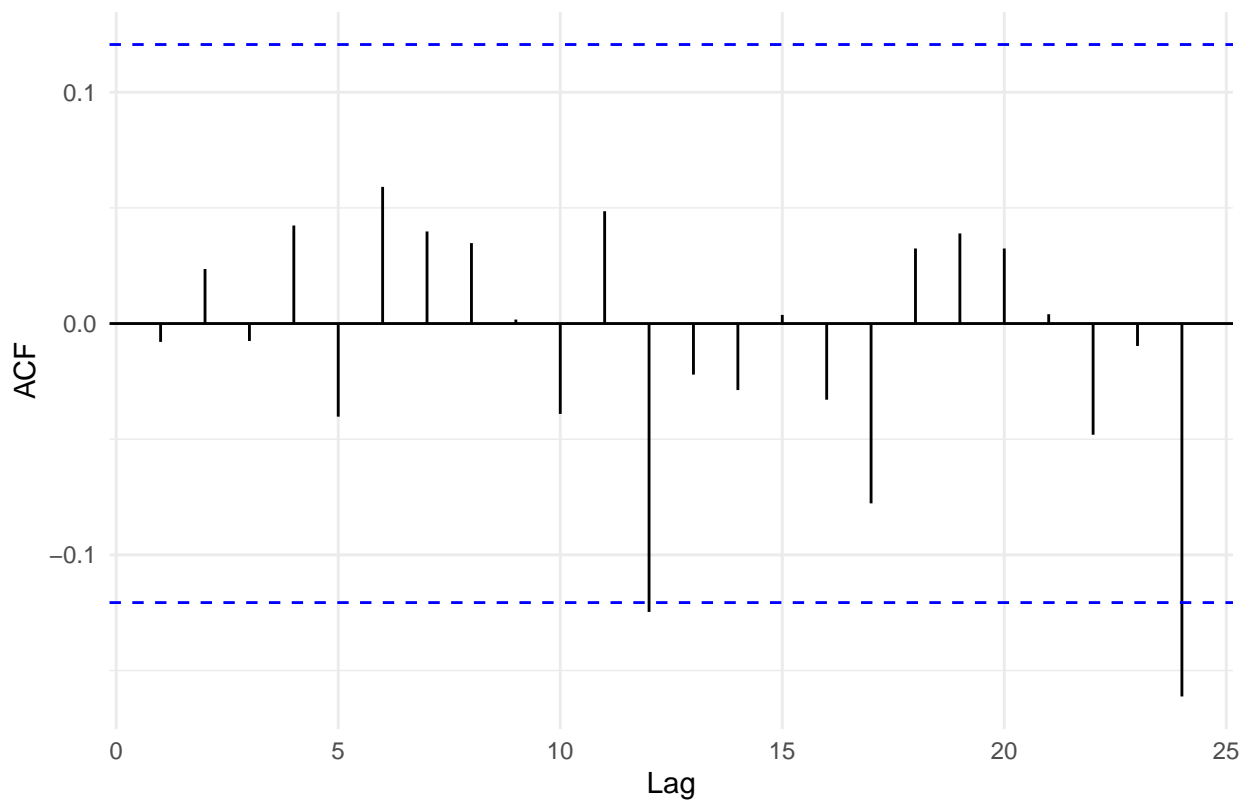

```
##
## 	Box-Ljung test
##
## data:  sa_resid
## X-squared = 8.6007, df = 15, p-value = 0.8975
```

41

From observing the residual plot, the data loosely resembles that of white noise, while the ACF of the residuals suggests that the data is stationary, with a small lag spike at lag=24 and the rest of the lags within the 95% bounds. Assessing the results from the Ljung-Box test, we have a p-value that is much larger than the 0.05 critical value leading us to fail to reject the null hypothesis and conclude that there is no serial correlation within our time series. As a note, it is interesting to observe the ACF in this case with lag=24, which indicates a seasonal MA(1) process remains, however, since we are to fit an ARIMA model to our seasonally adjusted time series, we will stay true to our subjugated methods.

2.6.2   Non-seasonally adjusted data assessment

Let's follow the same procedure above for the NSA data. Furthermore, we know there is a seasonal component to this data so let's see how many differences it will take for the data to become stationary. We will also observe the PACF and ACF plots of the NSA data.
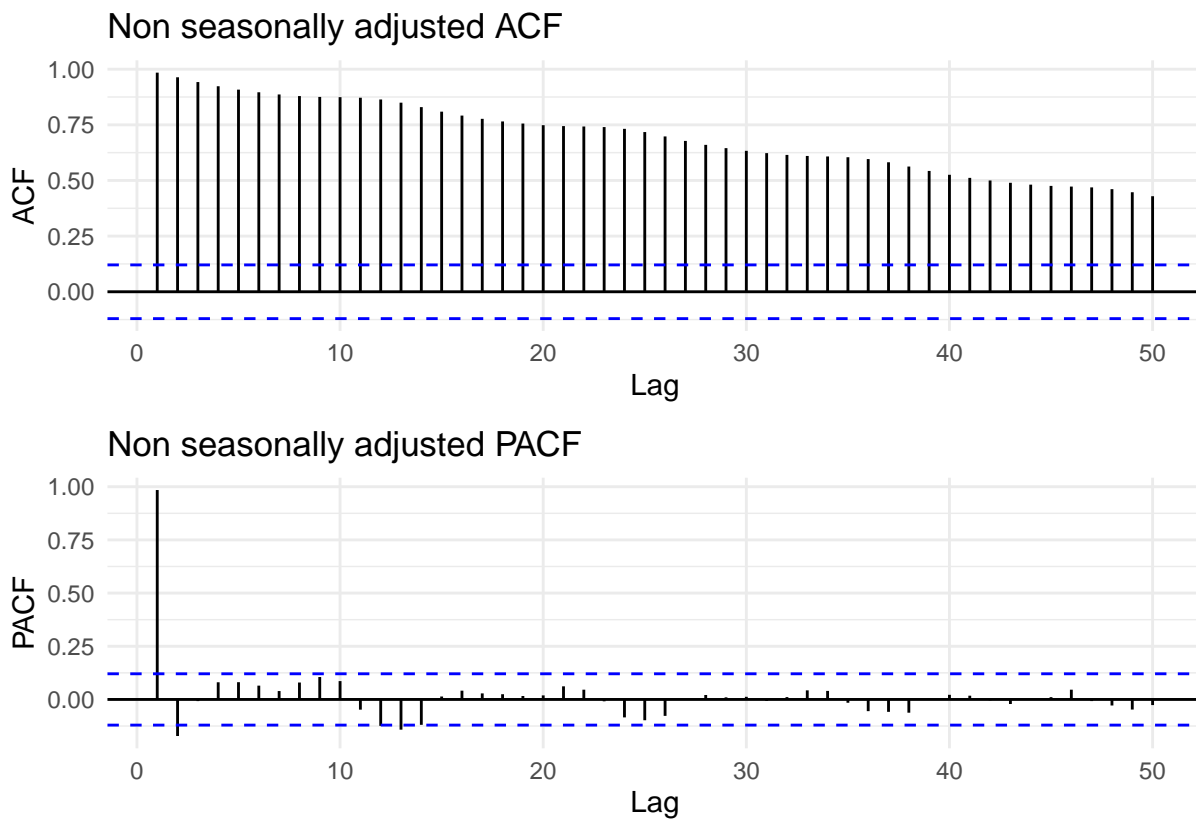
According to the PACF plot of NSA data, we observe a spike at lag=1 with a slight spike at lag=2, this indicates an AR(1) process is needed for this data. Furthermore, the ACF of this data indicates a slow decay of spikes. Next let's test for the number of differences needed for NSA data.

From the below metrics we've found that 1 non-seasonal difference and 1 seasonal difference is needed.

```
## # A tibble: 1 x 1
##   ndiffs
##    <int>
## 1      1
```

```
## # A tibble: 1 x 1
##   nsdiffs
##     <int>
## 1       1
```

Non seasonally adjusted ACF double difference

Non seasonally adjusted PACF double difference
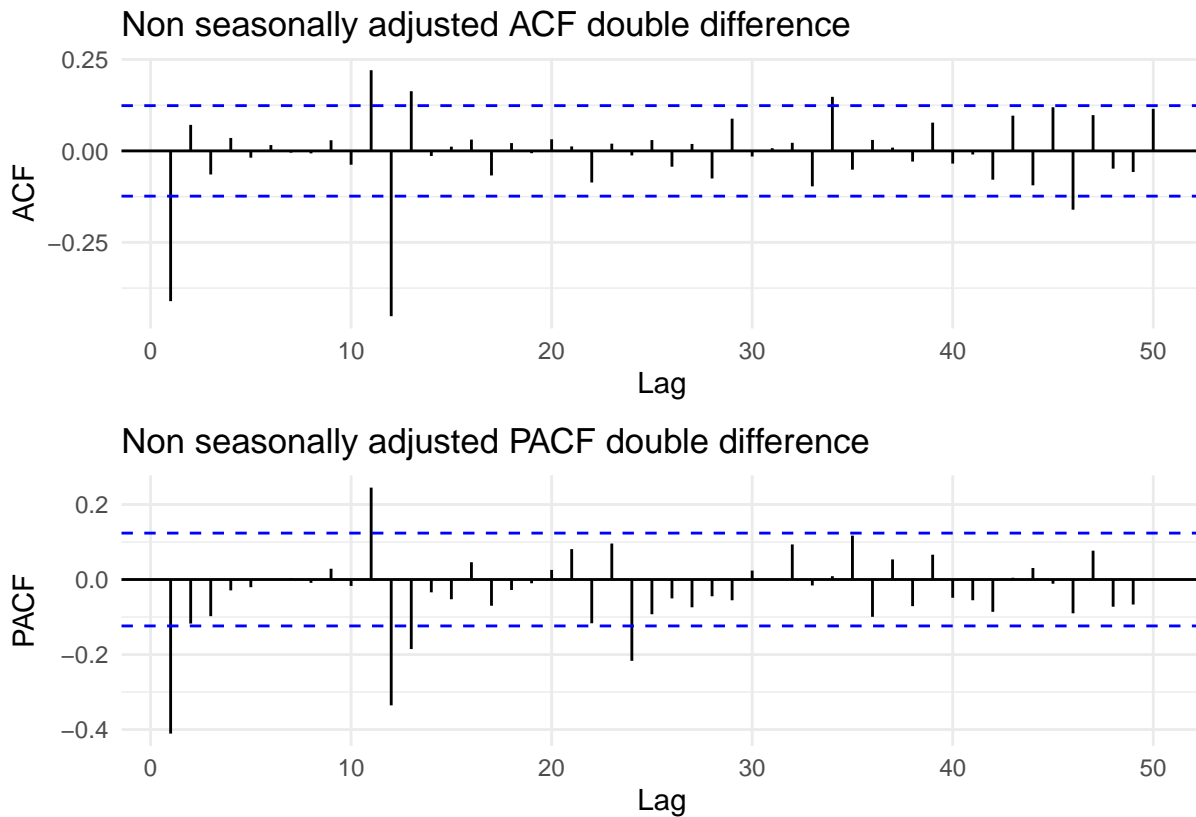
The ACF of the doubly differenced NSA time series shows a significant spike at lag=1 and lag=12 indicating that we need non-seasonal MA(1) and seasonal MA(1) components respectively. In the PACF there is 1 significant spike at lag=1 suggesting a non-seasonal AR(1) while there is also a signifncant spike at lag=12 and lag=24 indicating a non-seasonal AR(2) component.

Now let's fit two models, same as before. One that is automatically chosen for us and another one by our assessment, which will be an $ARIMA(1,1,1)(2,1,1)[12]$ model.

```
## Series: nsa_co2_present_train$value
## ARIMA(1,1,1)(2,1,1)[12]
##
## Coefficients:
##           ar1       ma1     sar1     sar2      sma1
##        0.0742  -0.5347   0.0629   0.0084   -0.8488
## s.e.   0.1371   0.1230   0.0862   0.0869    0.0660
##
## sigma^2 = 0.1356:  log likelihood = -103.27
## AIC=218.54   AICc=218.89   BIC=239.7
```

```
## Series: value
## Model: ARIMA(1,1,2)(0,1,1)[12]
##
## Coefficients:
##           ar1       ma1      ma2     sma1
##        0.8957  -1.3679   0.4042   -0.806
## s.e.   0.1072   0.1321   0.0960    0.053
##
## sigma^2 estimated as 0.1353:  log likelihood=-103.15
## AIC=216.31   AICc=216.55   BIC=233.93
```

Observing the above summaries of both estimated models: the BIC is lower in the automatically generated model $ARIMA(1,1,2)(0,1,1)[12]$. We also observe that our assessments were correct and not too far off from what the auto ARIMA call concluded with. Because we are using BIC as the main criteria of model selection, we will select the algorithmically computed model as our model ($ARIMA(1,1,2)(0,1,1)[12]$) of interest.

Next, we should check the residuals and run tests on the model we have estimated.

Residual Plot NSA Model

ACF Plot NSA Model
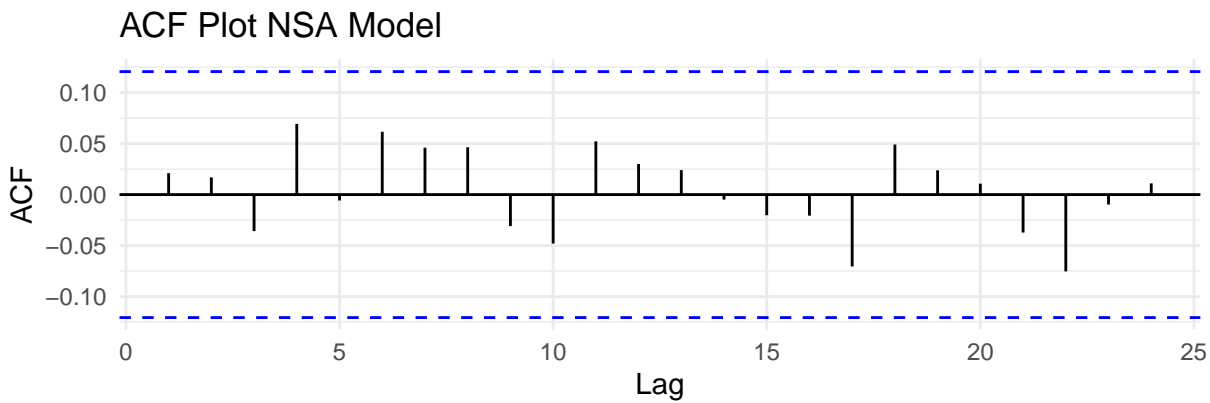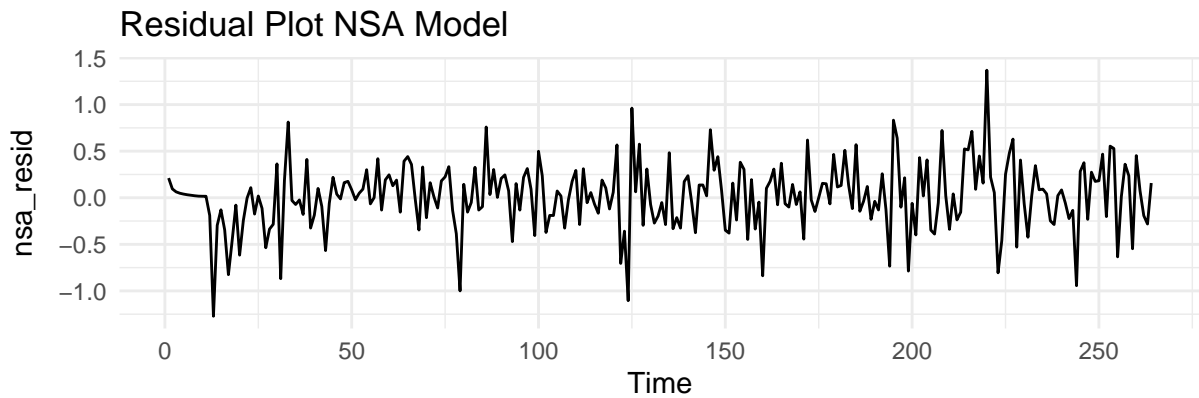

```
## 
##  Box-Ljung test
## 
## data:  nsa_resid
## X-squared = 6.2455, df = 15, p-value = 0.9753
```

From the results of the ACF of residuals, we observe that all lags are within the significant 95% threshold which means it closely resembles that of a white noise. The time series plot of the residuals loosely resemble that of a white noise, with fluctuations floating around 0 and very rarely crossing 1 and -1. From the Ljung-Box test, we observe a p-value that is higher than 0.05 critical value, leading us to reject the null hypothesis and conclude that there is no correlation within our time series.

From these results we will continue with using $ARIMA(0,1,1)$ for our seasonally adjusted time series, and $ARIMA(1,1,2)(0,1,1)[12]$ for our non-seasonally adjusted time series.

2.6.3 Assessing model accuracy

With creating and assessing the goodness of fit of so many models, we'd also like to observe how well our models can do against our test data. In order to do so, we will measure the accuracy of our models against a test set while also observing their forecasts into the future.

- First, we create forecasts two years into the future with both models.
- Next, we will compute the accuracy measurements of both models.
- And finally, we will also train a polynomial time-trend model to the seasonally adjusted series and run the same accuracy computations as previous models.

Here are the cross validation results of the seasonally adjusted time series:

```
##                         ME       RMSE        MAE          MPE       MAPE       MASE
## Training set 0.002074176 0.2985506 0.2349384 0.0002336546 0.06052062 0.8082674
## Test set     0.470844616 0.5381627 0.4780406 0.1131382109 0.11486154 1.6446210
##                     ACF1
## Training set -0.007911261
## Test set              NA
```

And below are the cross validation results of the non-seasonally adjusted time series:

```
##                          ME       RMSE        MAE         MPE        MAPE        MASE
## Training set 0.01419363 0.3557887 0.2649963 0.003252174 0.06808968 0.2269351
## Test set     0.13210426 0.3983645 0.3229973 0.031926712 0.07744393 0.2766054
##                        ACF1
## Training set 0.02102272
## Test set             NA
```

Looking at the RMSE of both models against our test set, we observe that the model fitted to the non-seasonally adjusted data yields a lower RMSE (0.398) compared to the seasonally adjusted model with an RMSE of 0.54. Let's take a look at the forecasting abilities descriptively below.
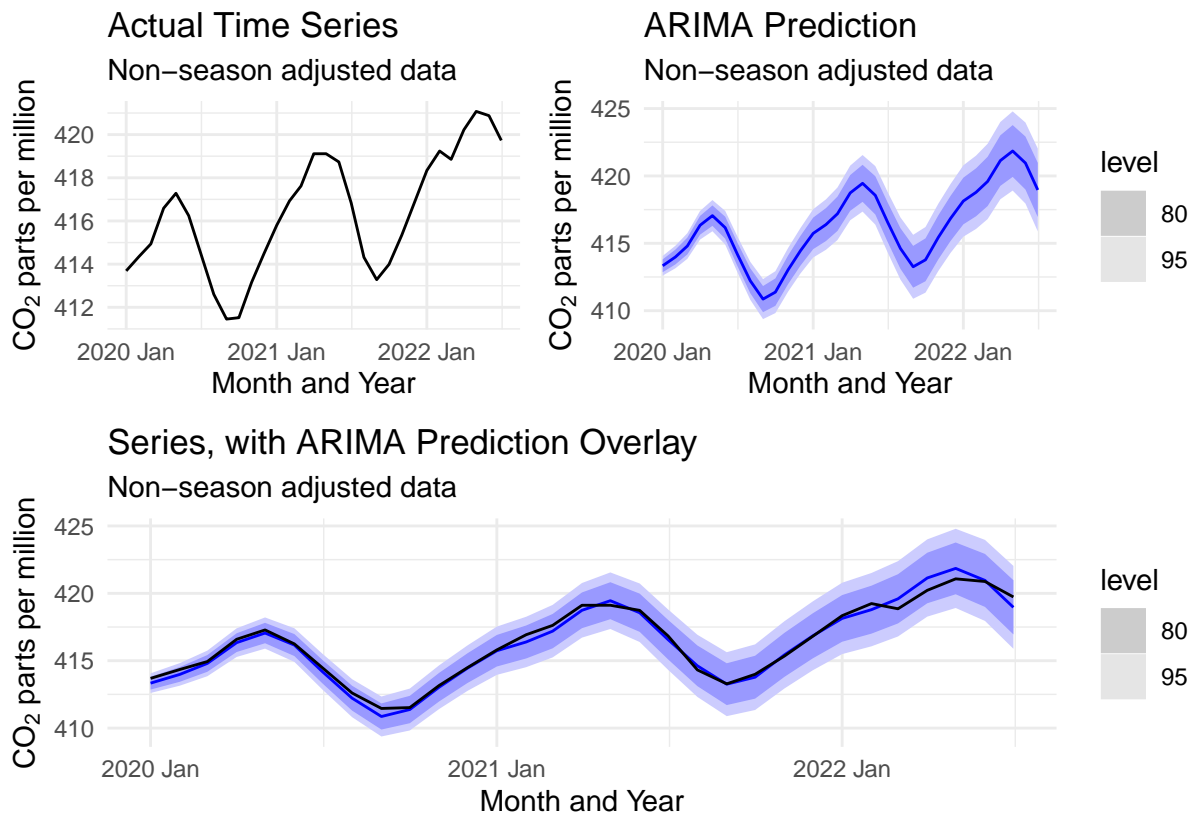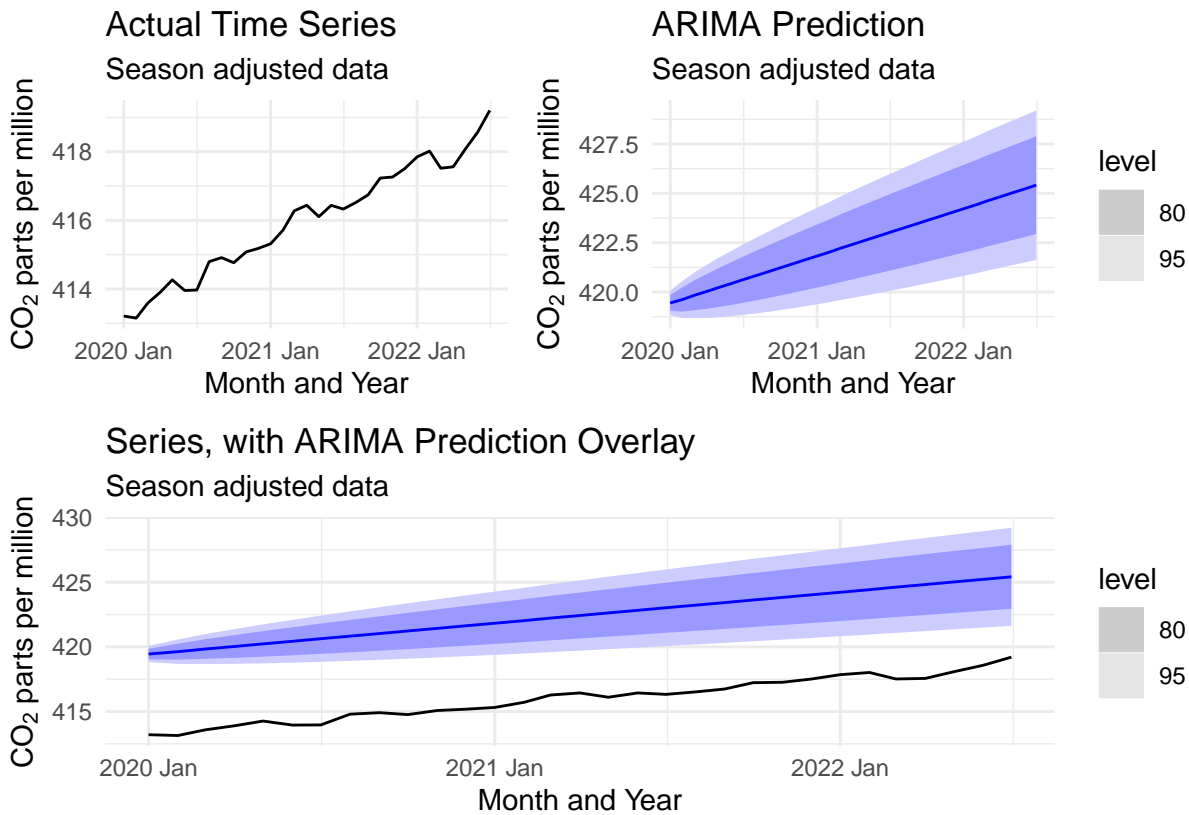
With the NSA data, we observe that the predicted values are very similar to the actual values, nearly 100% following the pattern that is observed in the actual data from Jan 2020 to May 2022. Conversely, with the predicted values of the model fitted to the SA data, we observe that the model does not perform as well and is not as accurate. It seems as though the predicted values are about a constant amount higher the actual values of CO_2

And now, we will fit a polynomial time trend model to the seasonally adjusted series. Here are the cross validation metrics of the polynomial time trend model:

```
## Series: season_adjust
## Model: TSLM
##
## Residuals:
##      Min        1Q    Median        3Q       Max
## -1.05399  -0.33613  -0.01203   0.36819   1.38438
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.657e+02  8.431e-02 4338.11   <2e-16 ***
## trend()       1.400e-01  1.469e-03   95.30   <2e-16 ***
## I(trend()^2)  1.427e-04  5.369e-06   26.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4532 on 261 degrees of freedom
## Multiple R-squared: 0.9989,  Adjusted R-squared: 0.9989
## F-statistic: 1.184e+05 on 2 and 261 DF, p-value: < 2.22e-16
```

```
## # A tibble: 1 x 10
##   .model .type     ME  RMSE   MAE     MPE   MAPE  MASE RMSSE  ACF1
##   <chr>  <chr>  <dbl> <dbl> <dbl>   <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 lm     Test  -0.157 0.459 0.360 -0.0375 0.0865   NaN   NaN 0.750
```

Observing the RMSE displayed above, we can see that the $RMSE = 0.459$ is less than the RMSE for the ARIMA model fit on the SA data ($RMSE = 0.54$), based on this observed metric we can conclude that the polynomial time trend model is the more accurate model of the two. And below are some descriptive analyses plots of the polynomial time trend model similar to those above, analysis of which shows an even more clear picture of predictability:
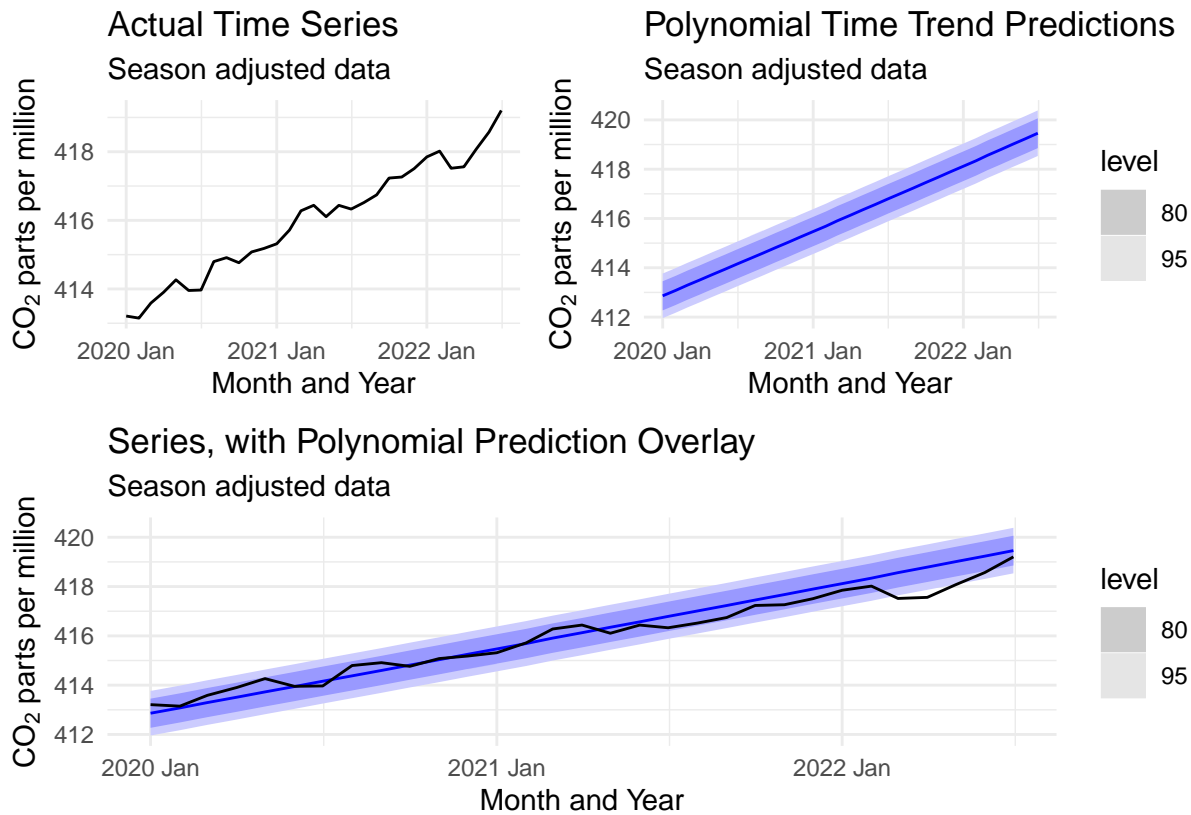
Here we observe the forecasted values as being very similar to that of the actual values within the test set, where nearly all of the actual predicted CO2 values are within the 80% threshold, and only a small subset of data in 2022 being outside of the 80% confidence range. We can conclude with these metrics and analyses that the polynomial time trend model is the better performant model and is more accurate model than the fitted ARIMA model.

2.7 How bad could it get?

Next, the question arose from a colleague of whether our model could predict the first and final times that the CO_2 levels would be 420ppm and 500ppm, we figured in doing so we could forecast how high the CO_2 levels could reach with the current state of the world, according to the NSA model we've estimated.

```
##   CO2_levels first_low point_estimates last_upper
## 1     420ppm  2021 Apr        2022 Mar   2606 May
## 2     500ppm  2047 May        2055 Apr   2529 May
```

From our observed forecasts above, it looks like the point estimates for when CO2 levels reach 420ppm is Mar 2022 while the theoretical `first time` and `last time` this model forecasts to be 420ppm is April 2021 and May 2606 respectively (according to our 95% confidence interval). Whereas, the point estimates for when CO2 levels reach 500ppm is April 2055, while the theoretical `first time` and `last time` this model would forecast 500ppm is May 2047 and May 2529 respectively.

We'd also like to generate predictions of atmospheric CO2 levels for the year 2122:

```
## # A tsibble: 12 x 4 [1M]
##    time_index point_forecast lo_95 high_95
##         <mth>          <dbl> <dbl>   <dbl>
## 1    2122 Jan           657.  549.    765.
## 2    2122 Feb           658.  550.    766.
## 3    2122 Mar           659.  550.    767.
## 4    2122 Apr           660.  552.    769.
## 5    2122 May           661.  552.    770.
## 6    2122 Jun           660.  551.    769.
## 7    2122 Jul           658.  549.    767.
## 8    2122 Aug           656.  547.    765.
## 9    2122 Sep           655.  546.    764.
## 10   2122 Oct           655.  546.    765.
## 11   2122 Nov           657.  548.    766.
```

```
## 12   2122 Dec           658.   549.    768.
```

2.7.1 Closing remarks

According to the dataframe displayed above, according to the NSA model forecasts in the year 2122 we'd expect to see CO2 levels of 654ppm to around 660ppm. The 95% confidence intervals float around 550ppm for the low bound and 765ppm for the high bound. This is evidence of a very wide confidence interval. Observing these forecasted CO2 levels is quite daunting to be honest, as a team we were very surprised to find that in just 100 years, the CO2 levels could potentially rise by more than 100ppm.

It is interesting to note here that similar to the analysis done in the 1997 report, we have also estimate confidence intervals that are very wide. Based on this fact and combined with how this Keeling curve behaves, we conclude that we are not entirely confident our forecasts are accurate.