Present and Future Global CO_2 Emissions

true true true true

Abstract

Year 1998 is upon us and global attention is turning toward the consequences of human-actions in our environmental system. While industrial pollution, water contamination, and others have their own impacts on health, none affects the entire globe as much as the global rise in temperature. The Intergovernmental Panel on Climate Change (IPCC) has been examining these trends for more than ten years. In the second report released in 1995 the IPCC notes that the balance of the evidence suggests that climate changes, in particular global warming, is attributable to human activities. Skepticism on the findings, that the global temperature rise due to human activities, exists but hasn't yet degenerated in a polarized pro- and anti- global warming camps. That said, there is also little political will to mitigate the global warming effects through tangible actions and/or policies. The effect of CO_2 causing "Green House" effect is no more under debate; nor the effect of Green House to raise the temprature. This behooves us to analyze the CO_2 levels in the atmosphere. Data from the Mona Loa Observatory (MLO) is analyzed in this report to describe and predict global CO_2 concentrations under several possible scenarios. What we find, when we run the analysis, may paint a grim picture.

For the last several hundreds of years we earthlings have been living under a "Golden Period" of moderate climactic conditions, barring few locations that have extremes. The vast majority of the earth is inhabitable. We enjoy regular seasons which allows for a steady stream of food, fresh water, flora, and fauna. Overall, we have had a balanced environment around us. Increased human activities and the desire to exploit nature at an ever increasing rate for our needs have impacted this balance and, consequently, the climate that we otherwise take for granted. Understanding changing climate, and what it means for the earth's inhabitants is of growing interest to the scientific and policy community. We do not as yet reckoned all the effects of the human actions on climate. One of the key stabilizing parameters is the temperature of the earth itself. Huge quantities of fresh water are frozen in the form of glaciers. An increase in temperature, for instance, can melt these glaciers and cause water levels to raise. The impact of this action on coastal community is a disaster, to put it mildly. There are several other disastrous outcomes that temperature increase can cause. One of the main causes for temperature increase is the green house effect of gases in our atmosphere, mainly in the form of CO_2. This report analyses what we've seen so far in the carbon footprint in our atmosphere, and what possible predictions we can make. We hope that this analysis provides an understanding of how CO_2 level has changed and what the forecast for CO_2 level is if the current dynamics continue. We hope that this motivates us to take actions to reduce CO_2 emissions.

Background

Carbon Emissions

What do we mean by carbon emission, and why do we care about this? The term carbon emission is used in the context of how carbon, in the form of Carbon Dioxide (CO_2), in our atmosphere impacts our lives. A lot of carbon is sequestered in the form of plants. Plants consume CO_2, along with water and sunlight, to produce glucose in a process called Photosynthesis. Thus, not only carbon stays in plants but plants remove CO_2 from the atmosphere. When we lose plant life on earth we're denied of the CO_2 cleaning mechanism. Carbon is also present in fossil fuel and coal. When we burn fossil fuel, coal, and wood for our energy needs we release CO_2. Nature maintains a delicate balance where CO_2 emissions in moderation can be consumed

by the plants. When we destroy forests to expand our habitable land and burn fossil fuel, coal, and wood we're creating an avalanche effect.

There is little dispute that burning fossil fuel releases CO_2. What we have understood is that CO_2 is a Green House gas. CO_2 acts like a glass which allows light to get through while trapping heat. A simple analogy is that when we sit inside a car, windows drawn up, on a bright cold day, we can feel the temperature inside the car rise. The light energy coming through the glass window heats up the air in the car and the heat is trapped by the glass. This phenomenon is called Green House effect, named after the techniques used to grow vegetables inside glass enclosures in colder climates. CO_2 plays the role of the glass around the earth. Thus, more CO_2 more heat is trapped.

While there are other green house gases such as Methane (CH_4) which are far more damaging than CO_2 our focus is on CO_2. Methane has relatively lower concentration and its emission is not as prevalent as that of CO_2. Hence, we focus the rest of the report to understand the trend we have seen thus far of CO_2 levels in the atmosphere, and what it forebodes.

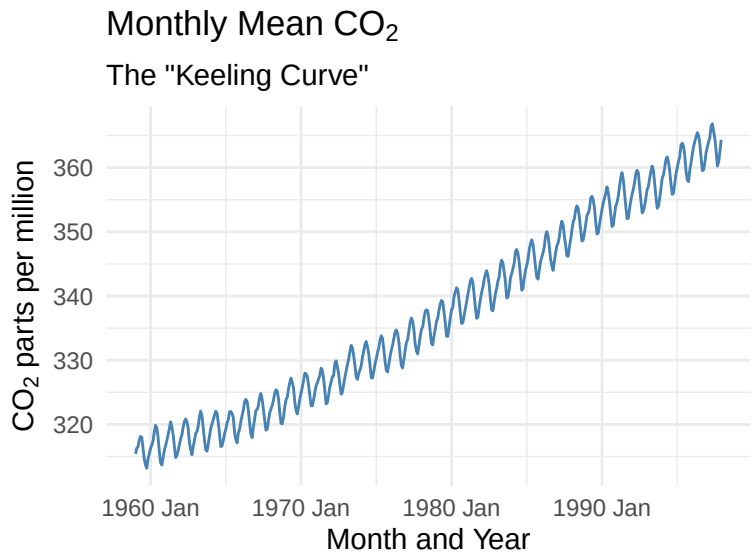
Measurement and Data

Measuring Atmospheric Carbon

Crucial to studying trends and forecasting levels of atmospheric carbon is reliable measurement of this concept. The importance of measuring CO_2 levels at relatively isolated places is important to understand the CO_2 levels in our atmosphere. For instance, measuring CO_2 in the vicinity a refinery or some other factories will provide very skewed readings. We may also see huge variance depending on the load the factory is handling. Similarly, measuring CO_2 levels in a busy downtown thoroughfare is also avoidable. Clearly, we will see elevated levels, and the measurements depend on the commute hours, holidays etc.

Thus, we need a neutral ground to measure the CO_2 levels. The data we are analyzing comes from Mouna Loa Observatory (MLO) in Hawaii's Big Island. At a height of over $13,500$ feet on a land that is surrounded by the Pacific Ocean on all sides, with no major industries, no major automobile movement in the area, and surrounded by nature, we believe that the data from MLO provides reasonable measurements. While there are active volcanoes in Hawaii in the Big Island, the altitude of the MLO allows for a neutral ground to measure CO_2 levels.

Historical Trends in Atmospheric Carbon



Atmospheric carbon is plotted in **??**, and shows some worrying trends. As is evident in the plot titled *Monthly Mean* CO_2 we see an increasing trend. This plot is referred to as the **The Keeling Curve**, named after the scientist Dr. Charles David Keeling, shows seasonal variations but the trend is definitely upward.

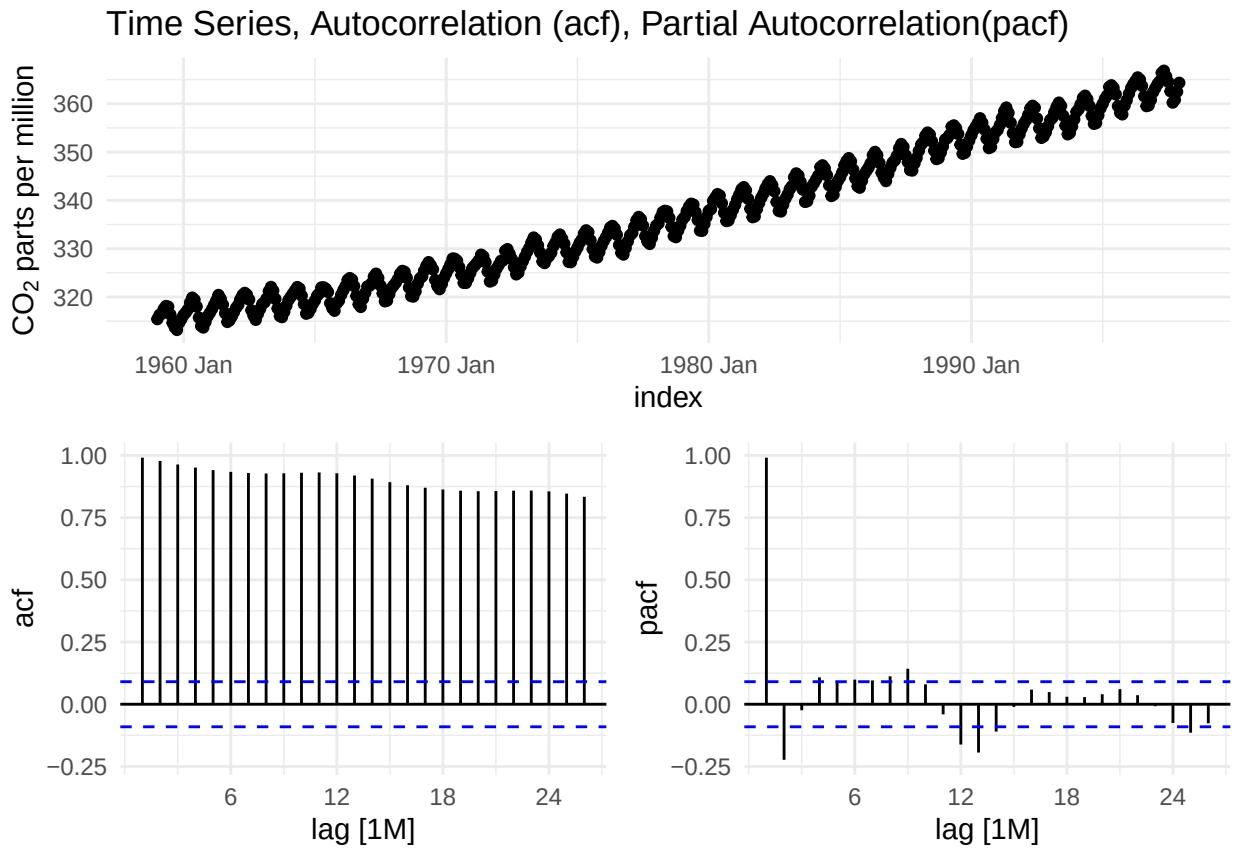
We will further examine the MLO data to understand better what we have observed till date, and what the forecast may look like. Let us look at the first few observations of the data.

```
##      Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
## 1959 315 316 316 318 318 318 316 315 314 313 315 315
## 1960 316 317 317 319 320 319 318 316 314 314 315 316
```

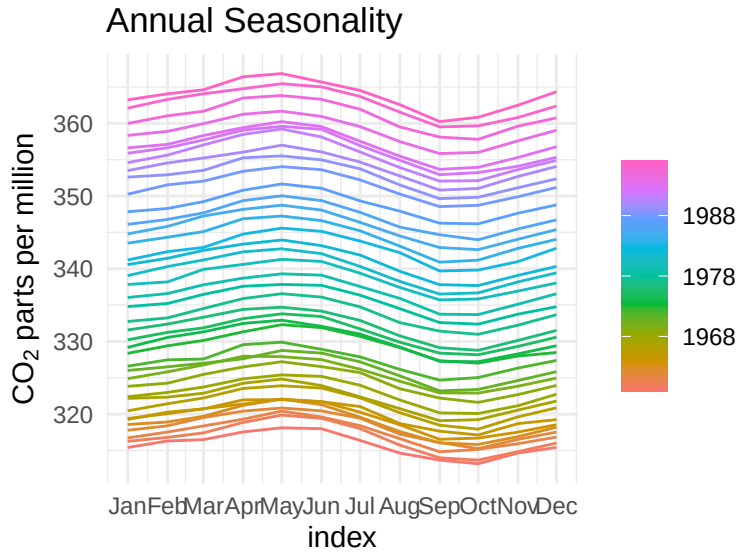
```
# check for missing values
num_na <- sum(is.na(co2_tsb$co2_ppm))
if (num_na > 0) {
  print(paste("There are", num_na, "missing values"))
}
```

The data, when organized as a table indexed by month, has 2 columns, named **index** and **co2_ppm**. The column names are somewhat self explanatory - the index of this table is the month of each year, and for each month we've the mean CO_2 level in Parts Per Million (ppm). There are 468 observations, starting from 1959 Jan and ending at 1997 Dec.

Let us examine the data further. Let's look at the time series data, the correlation among observations, referred to as Serial Correlation of Autocorrelation (ACF), and also the Partial Correlation (PACF), which is the correlation between two readings separated by lag 'K' after removing the observations that are between the two observations being correlated.



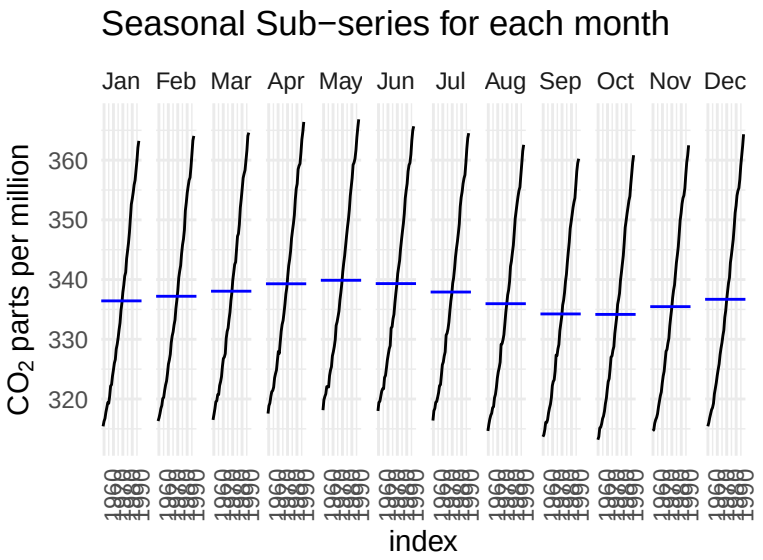
The time series graph is same as the Keeling Curve seen above. A visual analysis of the ACF indicates possible seasonality. The ACF values are not monotonically decreasing. We see a wavy pattern, indicative of possible seasonality. We'll drill down further to examine the seasonality.



In the plot above, titled **Annual Seasonality** we see an almost same degree of seasonality year after year. Two pieces of information are evident in the plot - there is a clear seasonality in the amount of CO_2 that hits a peak in the mid-May to mid-June time frame, and hits a minimum in the late-September to early October time frame, and the $CO-2$ levels have been going up without exception year over year. As figured out by

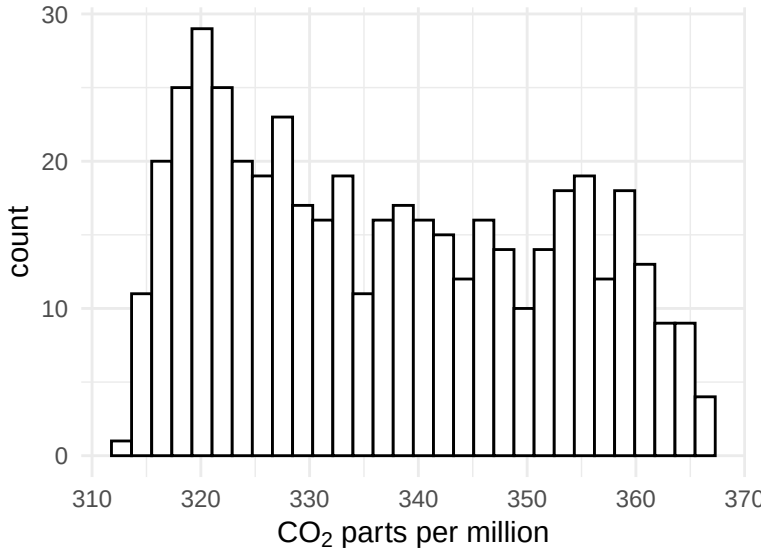
Dr. Keeling the plant life in the Northern Hemisphere starts growing in late spring onward which causes CO_2 to be absorbed by them. When the fall season kicks in the leaves fall and the plants stop growing leading to higher concentration both because of lack of vegetation to absorb CO_2 and also in small part by the fallen leaves and vegetation.

Additionally, let us examine the sub-series. This plot shows us how each month fared across years of observation,



In the plot titled **Seasonal Sub-series for each month** we see that the trend we examined earlier is clearly visible. The blue lines, that represent the mean for the month across all years of observations show a peak in the May-June time frame and a low in September-October time frame. It is also quite evident that for a given month the amount of CO_2 has monotonically increase from 1959 to 1997.

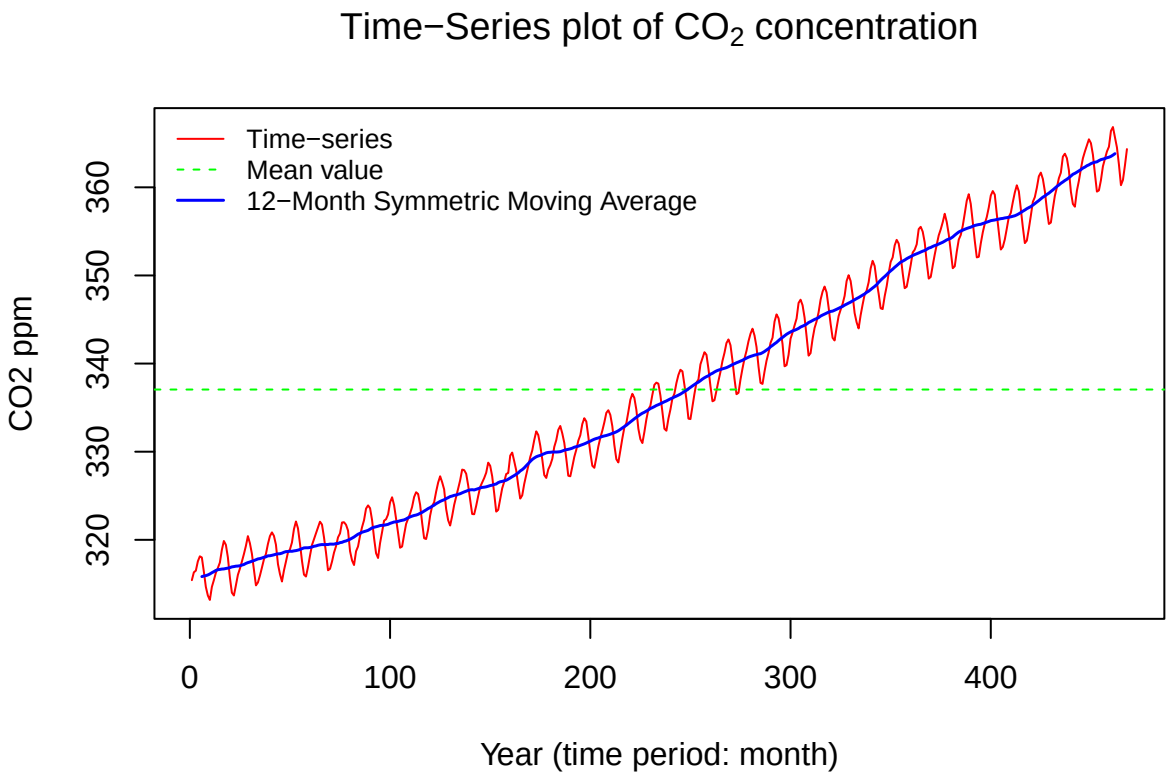
We can also take a quick look at the histogram to see if any obvious pattern emerges.



The histogram is not particularly interesting. Given the seasonality the value of around 320ppm may appear more often.

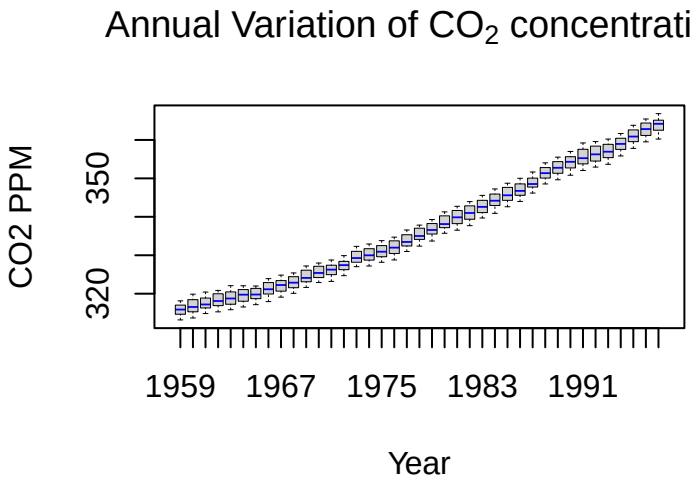
As we examine data we want to see the trend component, seasonal component, and the reminder or ran-

dom component. We use two different methods - classical decomposition and STL (Seasonal and Trend decomposition using Loess).



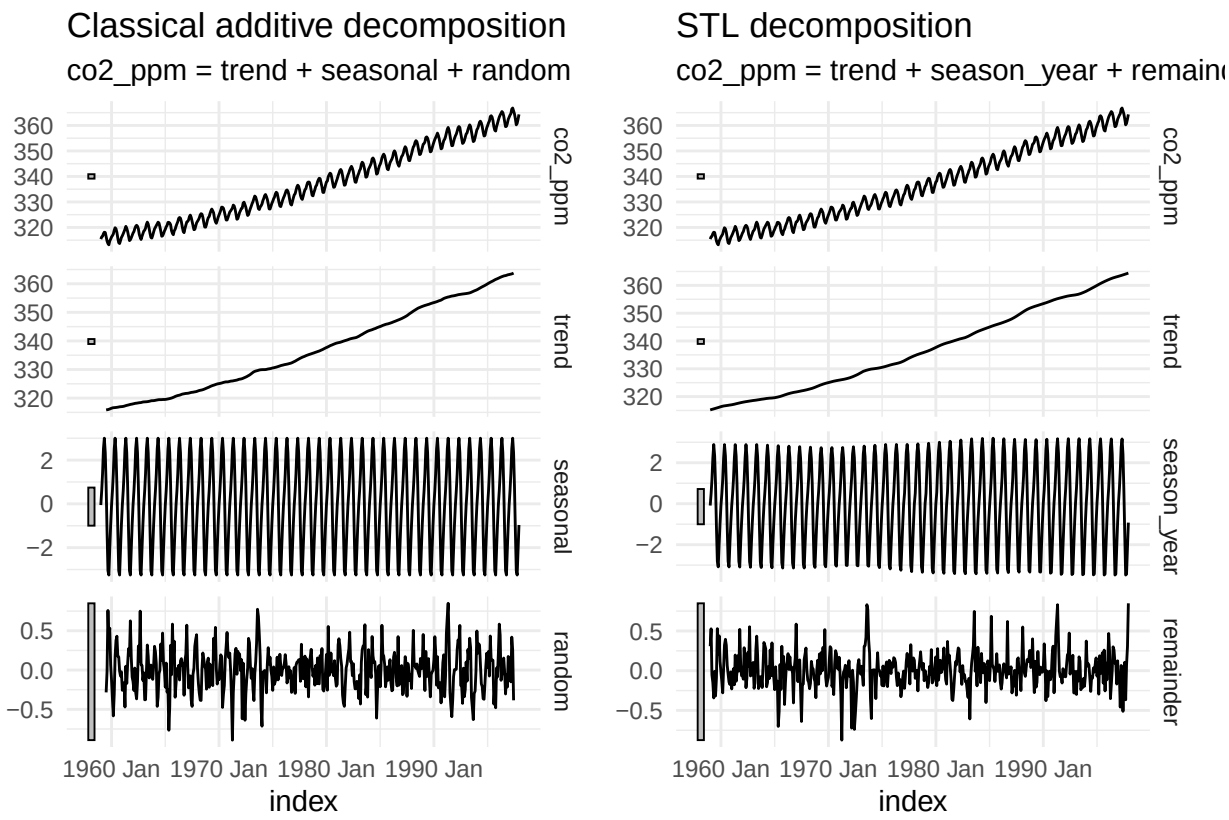
The plot titled **Time-series plot of the CO_2 concentration** shows the time series and the trend together. The visual clearly shows the increasing trend in the CO_2 level as a function of time.

We can get a visual of how the CO_2 level varied across years.

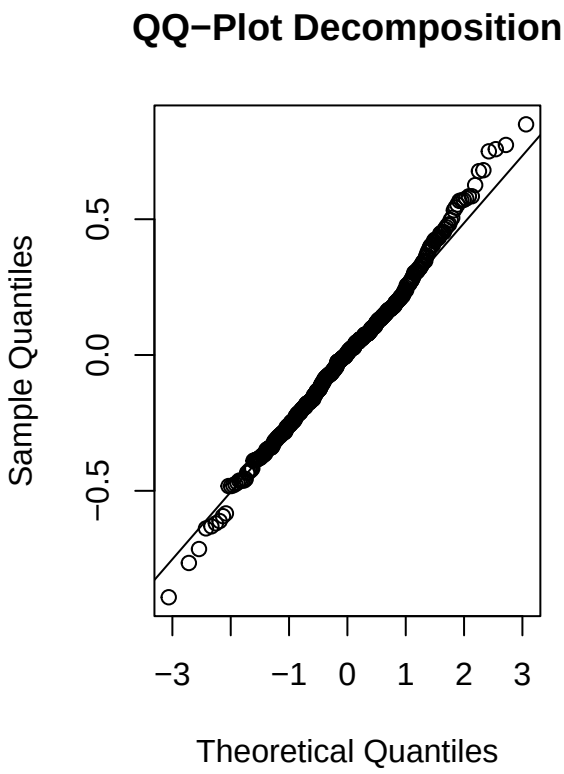


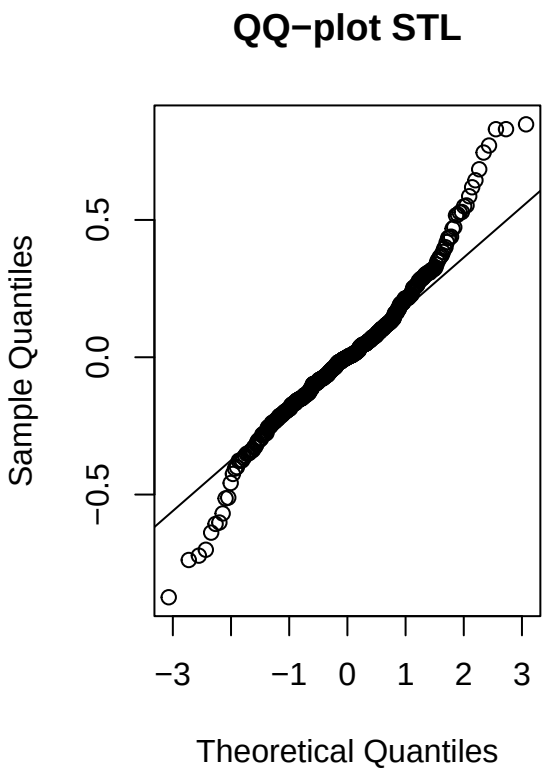
As we examine data we want to see the trend component, seasonal component, and the reminder or ran-

dom component. We use two different methods - classical decomposition and STL (Seasonal and Trend decomposition using Loess).



Clearly, both the Classical Decomposition and the STL Decomposition call out the trend line and the seasonality. There is an upward, monotonically increasing, trend. The seasonality is also clearly visible across years. The earlier graph titled **Annual Seasonality** showed a pattern of seasonality that repeats every year. The plots above is a different way to look at the seasonality. Suffice to say that we see a clear seasonal pattern. The random/remainder component look like white noise. Let's conduct a test to see if indeed these are normally distributed.





The QQ plots above show that in the case of Classical Decomposition we do see a the random component being normally distributed. The same isn't quite true with the STL Decomposition. While majority of the remainder values align on the straight line we do see outliers and deviations. We can do "Shapiro Test" to get a quantitative feel for the normal distribution.

```
# check if remainder (STL) & random (decomposition) are normally distributed
shapiro.test(decomp_rand$.val)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  decomp_rand$.val
## W = 1, p-value = 0.2
```

```
shapiro.test(stl_reminder$.val)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  stl_reminder$.val
## W = 1, p-value = 2e-08
```

The Shapiro test for the case of Classical Decomposition shows normality. The P-value is well above the 0.05 for a 95% confidence level. Again, the same can't be concluded of the STL test. According to the test the remainder isn't normally distributed. The mean values for both Classical Decomposition (0.002) and STL

Decomposition (0.008) are both close to zero. We can further test if the random/remainder component has seasonality.

```
# augmented Dickey-Fuller test to check for stationarity
x <- decomp_rand$.val
x <- x[!is.na(x)]
adf.test(x)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  x
## Dickey-Fuller = -14, Lag order = 7, p-value = 0.01
## alternative hypothesis: stationary
```

```
# phillips-peron test
x <- stl_reminder$.val
x <- x[!is.na(x)]
pp.test(x)
```

```
##
##  Phillips-Perron Unit Root Test
##
## data:  x
## Dickey-Fuller Z(alpha) = -292, Truncation lag parameter = 5, p-value =
## 0.01
## alternative hypothesis: stationary
```

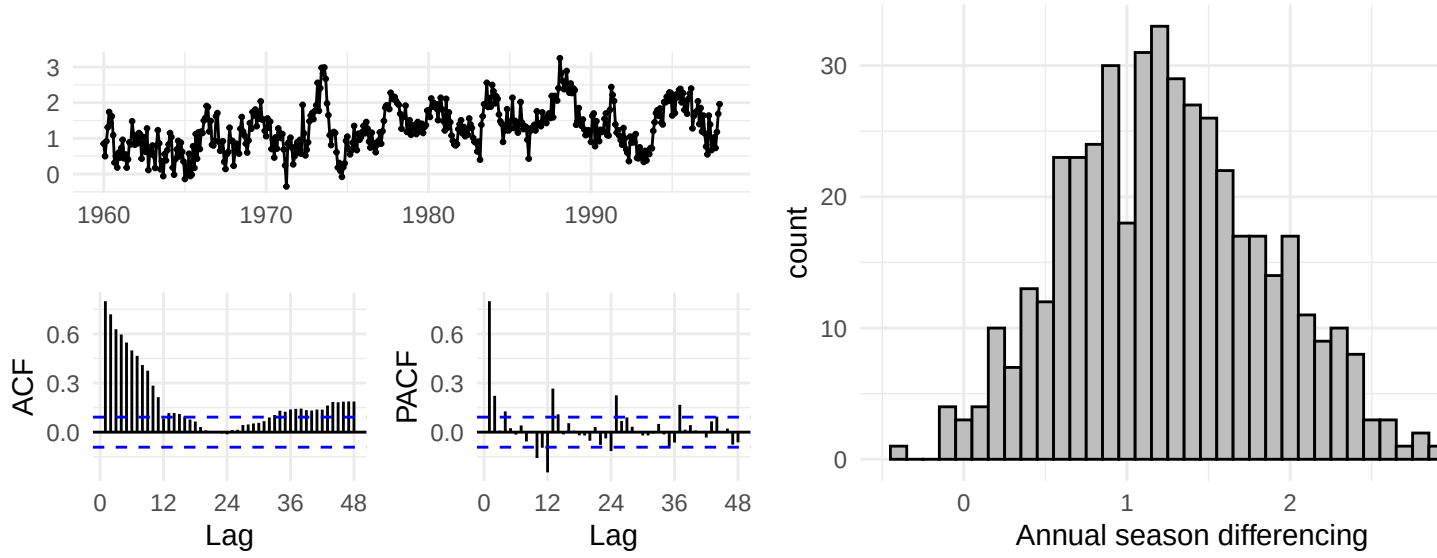
The results of the Augmented Dickey-Fuller test and the Phillips-Peron test confirm that the random/remainder components are stationary - p-values in both cases are 0.01, which less that 0.05, for a 95% confidence level.

We now turn our attention to the stationarity of the observations itself. We test this with KPSS (Kwiatkowski) and PP (Phillips-Peron) methods which check for unit roots for a time-series process. The presence of unit root renders the process non-stationary.

```
##   test_type test_stat p_value
## 1      KPSS      7.82    0.01
## 2        PP     -0.91    0.10
```

The test results show that we've a process that merits considerations as being stationary. The p-value for the case KPSS gives us a 99% confidence level, where as the PP test gives a 90% confidence.

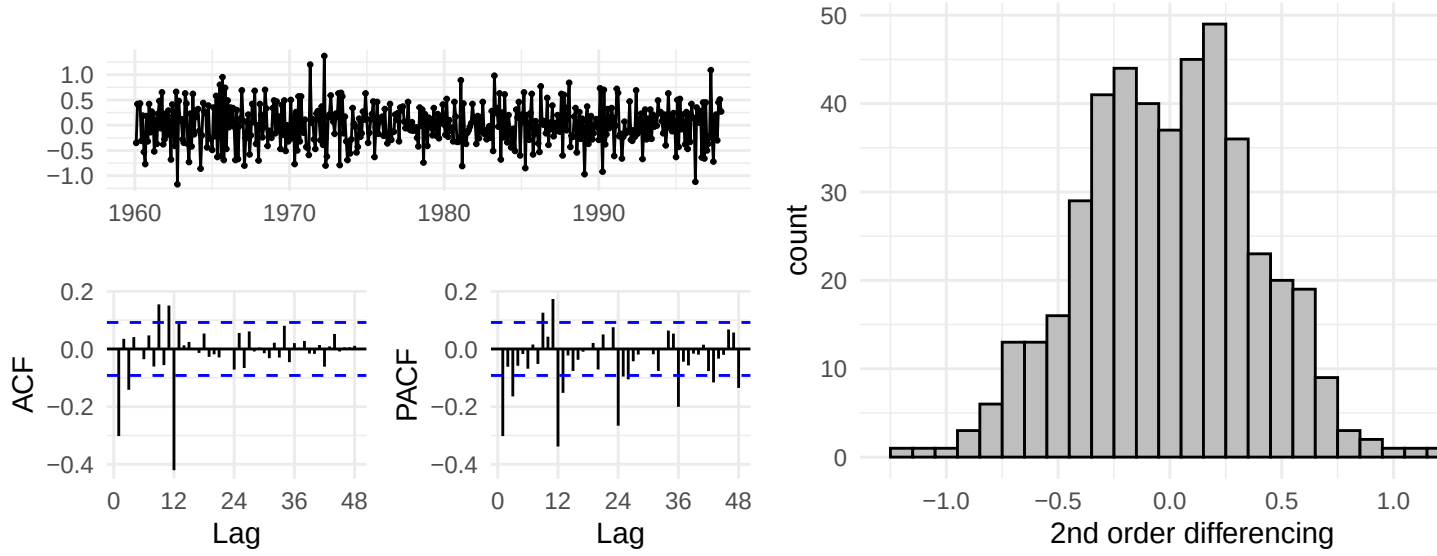
We can check to see if difference series provides additional insight and/or makes the date more amenable to a suitable model. We apply difference at 12th lag to remove seasonality and retest updated time-series using unit root test.



While the histogram above looks approximately bell-shaped the ACF and PACF plots still show seasonality. We can do an additional check with KPSS and PP tests as above on the differenced data.

```
##   test_type test_stat p_value
## 1      KPSS      1.94    0.01
## 2        PP     -6.67    0.01
```

When we difference the data we do see some improvement. The p-values reported by the checks above are consistent and provide a 99% confidence level. We stretch ourselves a bit more to see if the second-order difference provides any additiona improvements. We repeat the tests above.




```
##   test_type test_stat p_value
## 1      KPSS    0.0115    0.10
## 2        PP  -31.1688    0.01
```

Difference of difference time-series looks reasonably stationary around mean of 0. Histogram of second order difference is nearly normal in distribution and gives confidence that data is centered around mean or in other

words, stationary. Unit root test results confirms that series is now stationary at 90% or 99% confidence level. ACF plots still shows out of significance bound correlations up to 12th lag but follow up lags are well within bounds. PACF also reflects sinusoidal pattern indicating residual seasonality.

Overall, we can proceed with model building, as modified time series seems reasonably stationary based upon p-value and we would prefer not to over fit our model.

Models and Forecasts

While these plots might be compelling, it is often challenging to learn the exact nature of a time series process from only these overview, "time vs. outcome" style of plots. In this section, we present and evaluate two classes of models to assess which time series model is most appropriate to use.

Linear Models

To begin, we can consider a naive model of the form:

$$\text{CO}_2 = \phi_0 + \phi_1 t + \epsilon_t \tag{1}$$

The above model is essentially a linear function of time and the mean of this model is $\phi_0 + \phi_1 t$. Clearly, we know that this is not the case. This model doesn't capture the seasonality that we see in the CO_2 observations. Recall that CO_2 level peaks in mid-May to mid-June time frame, and hits a low in mid-September to mid-October time frame. We can consider the yearly average value and regress that as a function of time. The new model will look like

$$\text{yearlyCO}_2 = \phi_0 + \phi_1 t + \epsilon_t \tag{2}$$

```
##
## Call:
## lm(formula = co2_ppm ~ I(year_ind - min_year), data = co2_yearly)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.447 -1.192 -0.501  1.271  3.672
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             312.1540     0.5120   609.6   <2e-16 ***
## I(year_ind - min_year)   1.3105     0.0232    56.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.63 on 37 degrees of freedom
## Multiple R-squared:  0.989,  Adjusted R-squared:  0.988
## F-statistic: 3.19e+03 on 1 and 37 DF,  p-value: <2e-16
```

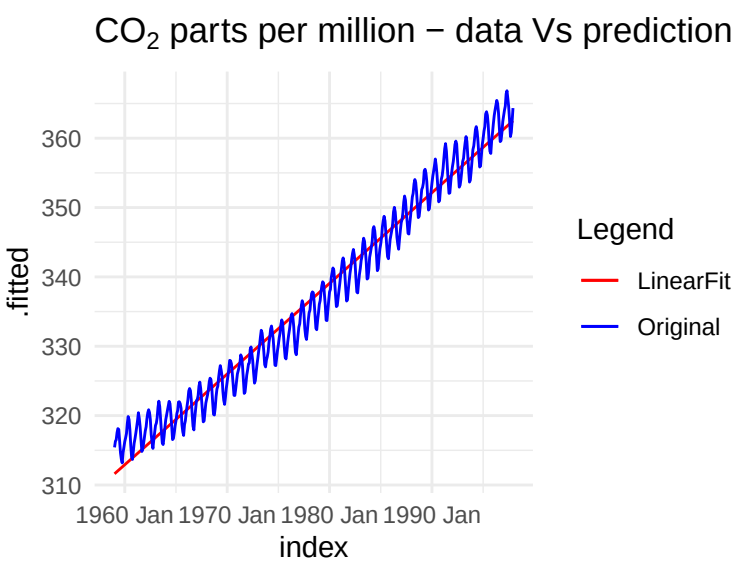
The model above looks reasonable. What we see is that the base is approximately 312 ppm and an increase of 1 every year. The p-values are low enough for us to consider this a reasonable model. The model, however, doesn't capture the seasonality and other nuances that may be better modeled with a time series. We will use time-series aware linear models to see if we get additional insights. The model we're exploring is

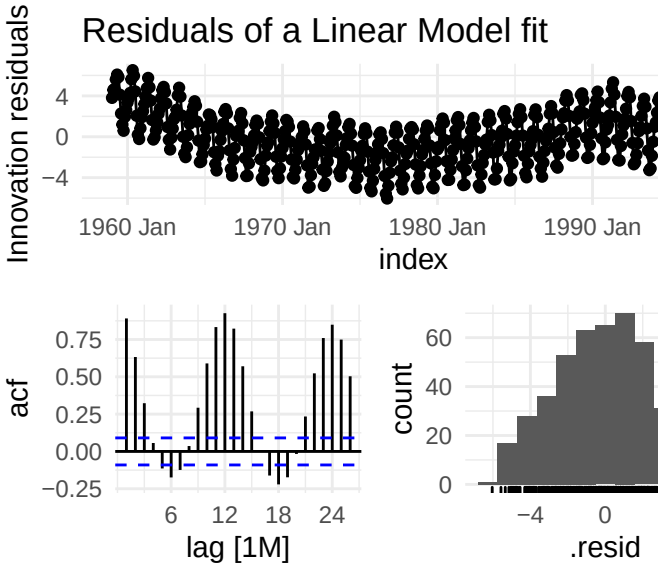
$$\mathrm{CO}_2 = \phi_0 + \phi_1 * trend + \epsilon_t \tag{3}$$

```
## Series: co2_ppm
## 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
```

It shouldn't come as a surprise that $TLSM()$ fitted a model with identical coefficients as that of the $lm()$ method earlier. The trend is essentially yearly. The way we invoked $lm()$, based on yearly average, gives us the same results as $TSLNM()$.

Now, let's compare the fit with the original.





As expected the linear prediction model is a good fit. The plots above corroborates the finding.

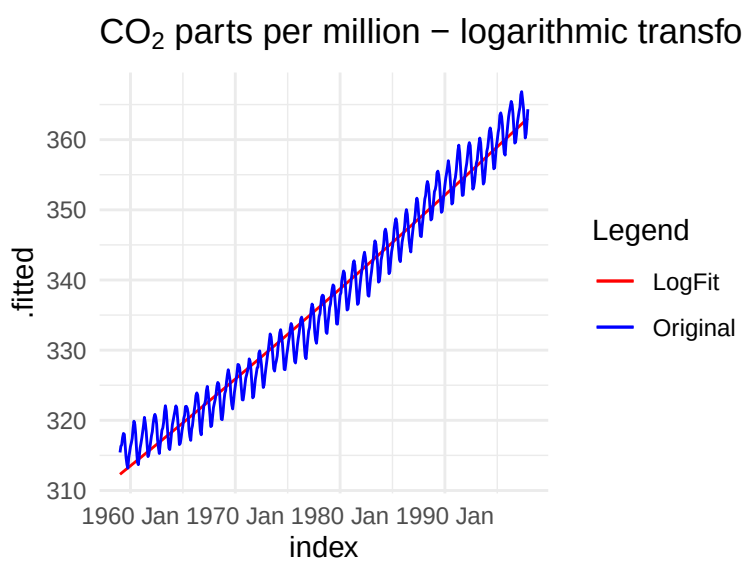
We can examine if data transformation, to logarithmic values, improves the model. As we did for the linear case we'll examine the output of the log-transformed model, and the residuals. The analytical model is

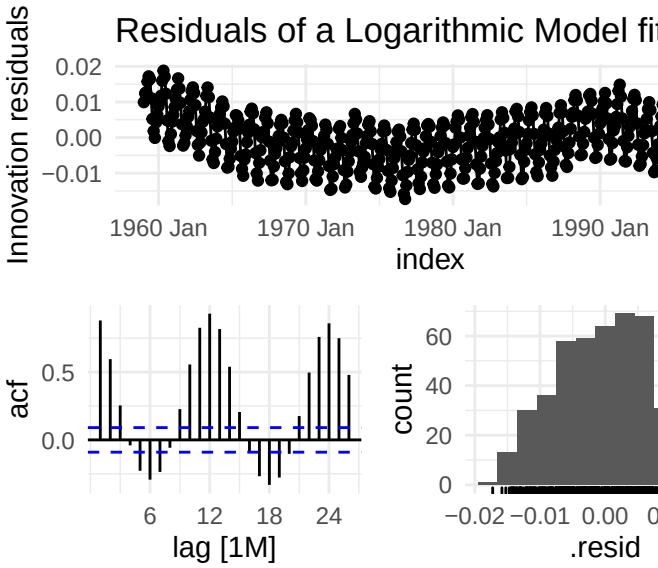
$$\log(CO_2) = \phi_0 + \phi_1 t + \epsilon_t \tag{4}$$

```
## Series: co2_ppm
## Model: TSLM
```

```
## Transformation: log(co2_ppm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.01727 -0.00561  0.00028  0.00538  0.01878
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.74e+00   6.83e-04    8410   <2e-16 ***
## trend()     3.22e-04   2.52e-06     128   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.00738 on 466 degrees of freedom
## Multiple R-squared: 0.972,   Adjusted R-squared: 0.972
## F-statistic: 1.63e+04 on 1 and 466 DF, p-value: <2e-16
```

Logarithmic transformation has reduced the magnitude of the coefficients. This is to be expected as log transformations tends to grow slower than linear transformation. Consequently, the intercept and the slope show smaller values.





Residual plot convey information similar to what we saw earlier.

As for linear model, we will extend to see the impact of quadratic terms.

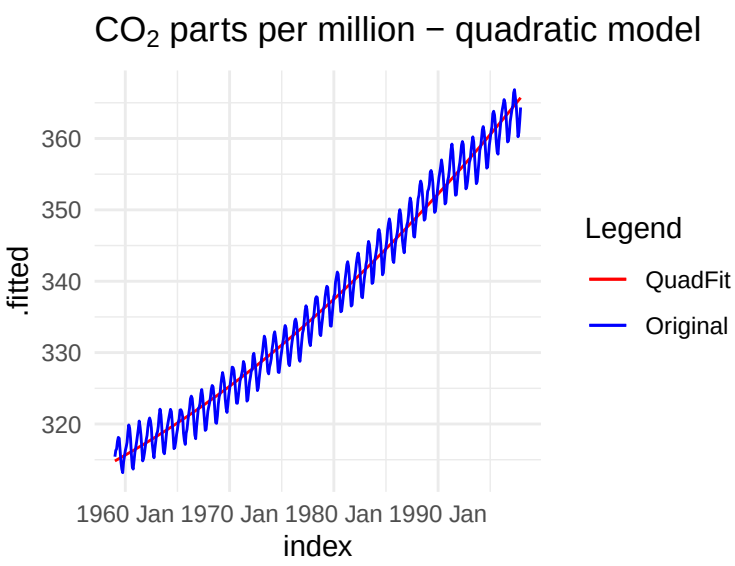
$$\text{CO}_2 = \phi_0 + \phi_1 t + \phi_2 t^2 + \epsilon_t \tag{5}$$

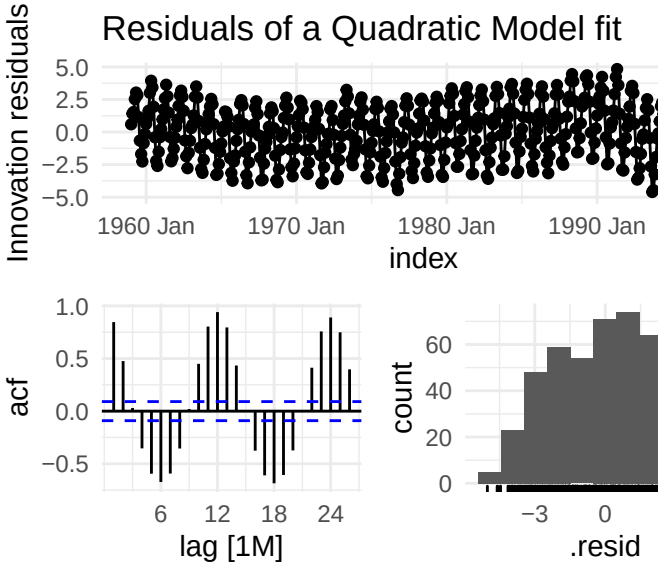
```
co2_1997_quad <- co2_tsb %>%
  model(TSLM(co2_ppm ~ trend() + I(trend()^2)))
report(co2_1997_quad)
```

```
## Series: co2_ppm
## 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 ***
## 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
```

While the quadratic term is statistically significant the coefficient associated with the quadratic term is quite small. Thus, the contribution from the quadratic term can be ignored, without significantly impacting the model. The following plots show the fit Vs actual data and the residual characteristics





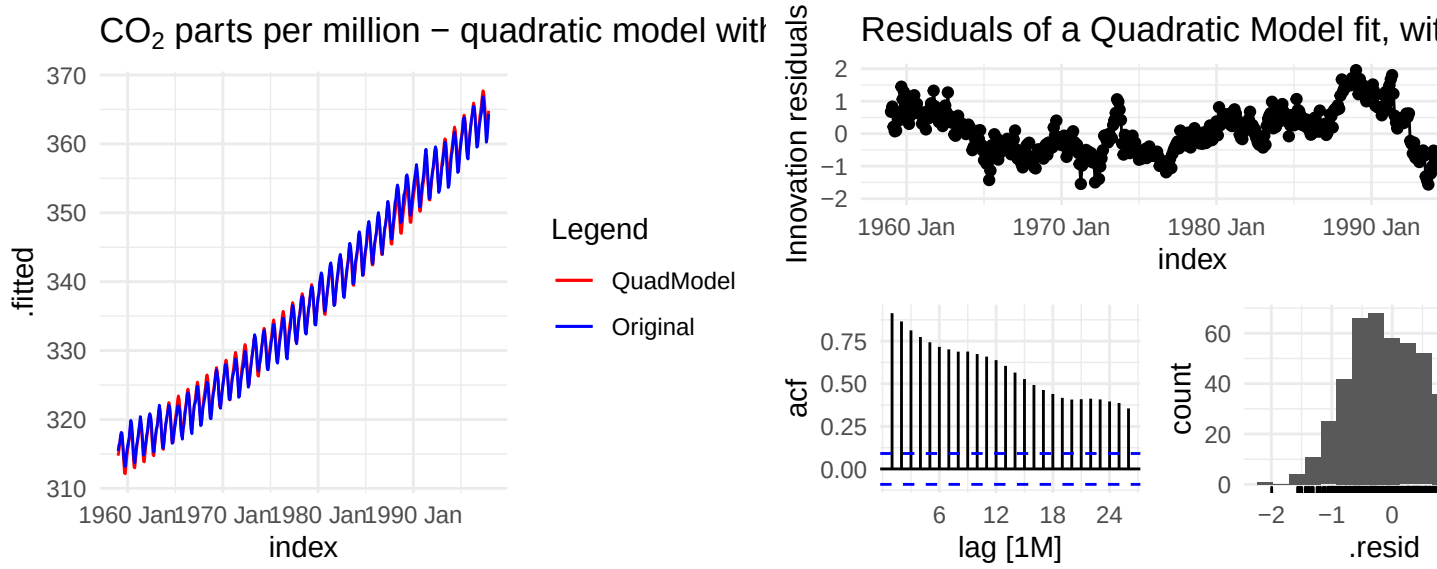
At this stage including higher order polynomials may not get us anything more, and further, may be lead to an over-fitted model. The predictions from these models could deviate from earlier model due to the lack of generality in the model.

As a next step we include seasonality to the quadratic model.

```
co2_1997_quad_Final <- co2_tsb %>%
  model(TSLM(co2_ppm ~ trend() + I(trend()^2) + season()))
report(co2_1997_quad_Final)
```

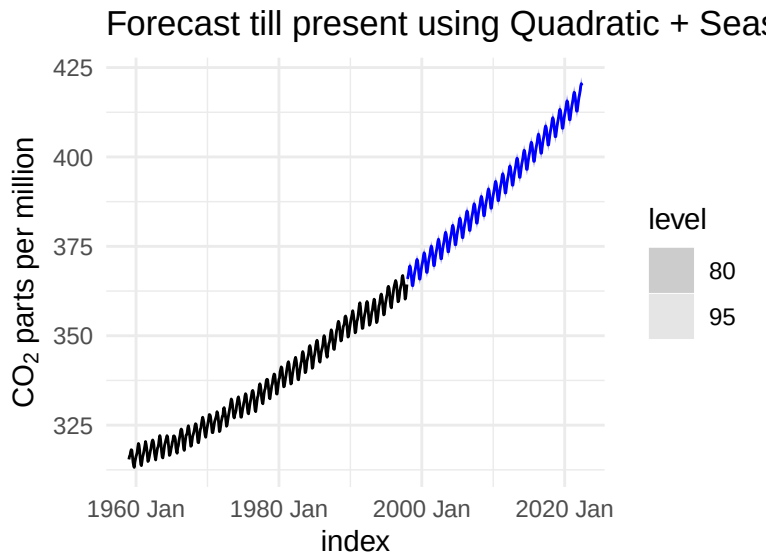
```
## Series: co2_ppm
## Model: TSLM
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.995 -0.545 -0.060  0.473  1.955
```

```
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.15e+02   1.49e-01 2105.89  < 2e-16 ***
## trend()         6.76e-02   9.93e-04   68.11  < 2e-16 ***
## I(trend()^2)    8.86e-05   2.05e-06   43.24  < 2e-16 ***
## season()year2   6.64e-01   1.64e-01    4.05  6.0e-05 ***
## season()year3   1.41e+00   1.64e-01    8.58  < 2e-16 ***
## season()year4   2.54e+00   1.64e-01   15.48  < 2e-16 ***
## season()year5   3.02e+00   1.64e-01   18.40  < 2e-16 ***
## season()year6   2.35e+00   1.64e-01   14.36  < 2e-16 ***
## season()year7   8.33e-01   1.64e-01    5.08  5.5e-07 ***
## season()year8  -1.23e+00   1.64e-01   -7.53  2.7e-13 ***
## season()year9  -3.06e+00   1.64e-01  -18.66  < 2e-16 ***
## season()year10 -3.24e+00   1.64e-01  -19.78  < 2e-16 ***
## season()year11 -2.05e+00   1.64e-01  -12.53  < 2e-16 ***
## season()year12 -9.37e-01   1.64e-01   -5.72  2.0e-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: 0.998,   Adjusted R-squared: 0.998
## F-statistic: 1.53e+04 on 13 and 454 DF, p-value: <2e-16
```



The model is a good fit. We see that the predicted values follow the original data quite faithfully. Given that only the quadratic term is included we do not believe that this model is an over- fit. The residuals appear normally distributed around mean 1.215×10^{-15}. Let us do a forecast and see what we get.

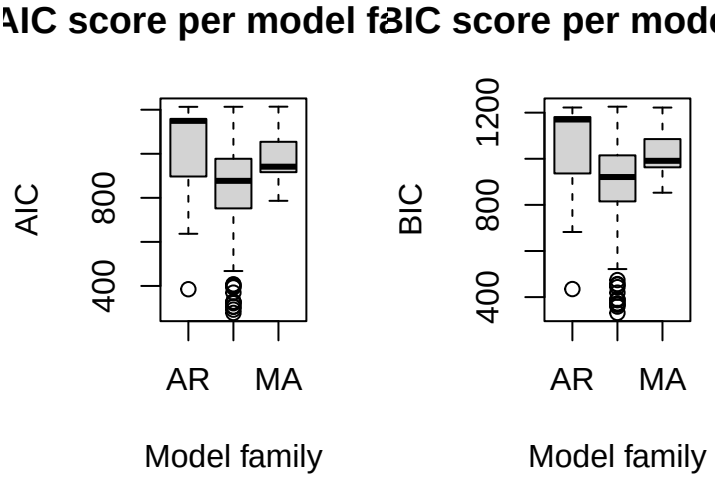
```
quad_model_forecast <- forecast(co2_1997_quad_Final, h=294)
```



The forecast appears in line with what we may expect given the trend and seasonality of the observations we have seen thus far. We believe that this is a reasonable forecast. The last six values from the forecast are 416.675, 417.541, 418.486, 419.82, 420.501, 420.041. With forecast set till June 2022 (h=294, being the number of months from Janurary 1998 on), we see the linear model predicting a value of 420 for the first time in April 2022. We need to match with observations that are being measured in 1998 and beyond to validate the effectiveness of the model.

ARIMA Models

```
for (i in 1:nrow(aic_bic_scores)) {
  p <- aic_bic_scores$p[i]
  q <- aic_bic_scores$q[i]
  aic <- c(aic, try_default(AIC(Arima(co2_tsb$co2_ppm,
                                      order = c(p, 1, q),
                                  seasonal=list(order = c(0, d ,0), 12))),
                            default = NA, quiet = TRUE))
  bic <- c(bic, try_default(BIC(Arima(co2_tsb$co2_ppm,
                                      order = c(p, 1, q),
                                  seasonal=list(order = c(0, d, 0), 12))),
                            default = NA, quiet = TRUE))
}
```



We ran an ARIMA model with difference (d) set to zero. We iterated over several values of AR order (p) and MA order (q). The box plot above shows the AIC and the BIC values from the iterations. We clearly see that the ARMA model gives the optimal fit.

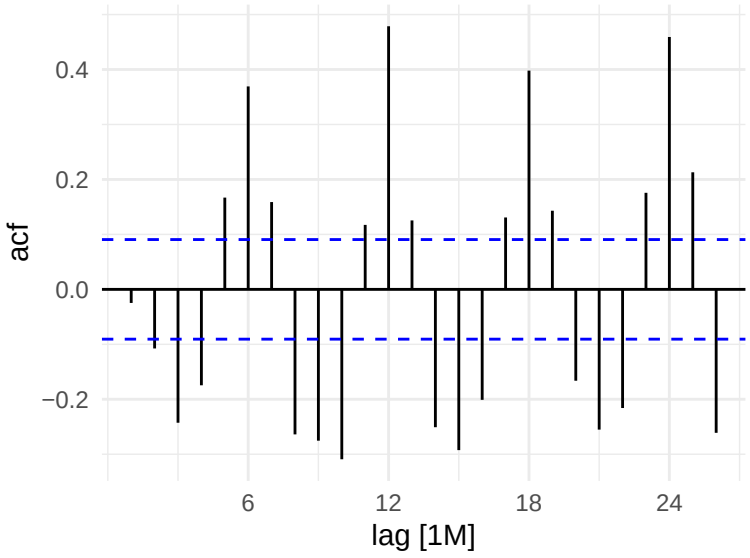
```
## # A tibble: 10 x 5
## # Rowwise:
##        p     q family   aic   bic
##    <int> <int> <chr>  <dbl> <dbl>
##  1     5     7 ARMA    278.  331.
##  2     7     8 ARMA    293.  359.
##  3    13     1 ARMA    305.  367.
##  4    12     3 ARMA    307.  373.
##  5    10     5 ARMA    321.  388.
##  6     6     9 ARMA    325.  392.
##  7     4     5 ARMA    328.  369.
##  8     6     7 ARMA    329.  387.
##  9     7     4 ARMA    370.  419.
## 10    11     0 AR      385.  435.
```

As is evident from the AIC and BIC scores reported ARIMA(6,1,9) scores the best. In addition, we'll consider the next two model also - ARIMA(13,1,1) and ARIMA(12,1,3). We see data for each month are well correlated across years. The earlier plot titled **Seasonal Sub-series for each month** shows that when we consider data for a given month across all years we see an upward trend.

```
model_aic<-co2_tsb %>%
  model(ARIMA(co2_ppm ~ 1 + pdq(0:14,0:2,0:5) + PDQ(0,0,0), ic="aic",
              stepwise=F, greedy=F))

model_aicc<-co2_tsb %>%
  model(ARIMA(co2_ppm ~ 1 + pdq(0:14,0:2,0:5) + PDQ(0,0,0), ic="aicc",
              stepwise=F, greedy=F))

model_bic<-co2_tsb %>%
  model(ARIMA(co2_ppm ~ 1 + pdq(0:14,0:2,0:5) + PDQ(0,0,0), ic="bic",
              stepwise=F,greedy=F))
```

```
## -----AIC Model Report-----------------------------------------------

## Series: co2_ppm
## Model: ARIMA(2,1,4) w/ drift
##
## Coefficients:
##          ar1      ar2      ma1     ma2     ma3     ma4  constant
##       1.6886  -0.9587  -1.3228  0.1540  0.1374  0.1909    0.0286
## s.e.  0.0137   0.0134   0.0481  0.0749  0.0902  0.0563    0.0039
##
## sigma^2 estimated as 0.2901:  log likelihood=-373
## AIC=763   AICc=763   BIC=796

## -----AICc Model Report----------------------------------------------

## Series: co2_ppm
## Model: ARIMA(2,1,4) w/ drift
##
## Coefficients:
##          ar1      ar2      ma1     ma2     ma3     ma4  constant
##       1.6886  -0.9587  -1.3228  0.1540  0.1374  0.1909    0.0286
## s.e.  0.0137   0.0134   0.0481  0.0749  0.0902  0.0563    0.0039
##
## sigma^2 estimated as 0.2901:  log likelihood=-373
## AIC=763   AICc=763   BIC=796

## -----BIC Model Report-----------------------------------------------

## Series: co2_ppm
## Model: ARIMA(2,1,4) w/ drift
##
## Coefficients:
##          ar1      ar2      ma1     ma2     ma3     ma4  constant
##       1.6886  -0.9587  -1.3228  0.1540  0.1374  0.1909    0.0286
## s.e.  0.0137   0.0134   0.0481  0.0749  0.0902  0.0563    0.0039
##
## sigma^2 estimated as 0.2901:  log likelihood=-373
## AIC=763   AICc=763   BIC=796
```

When we include the difference term we find that the model is tuning itself to ARIMA(2, 1, 4). This likely happens because when we introduce difference term it tends to take away the trend and/or seasonality. The ARIMA(2, 1, 4) is a good fit for the difference time-series for all three information criteria.



SARIMA Model

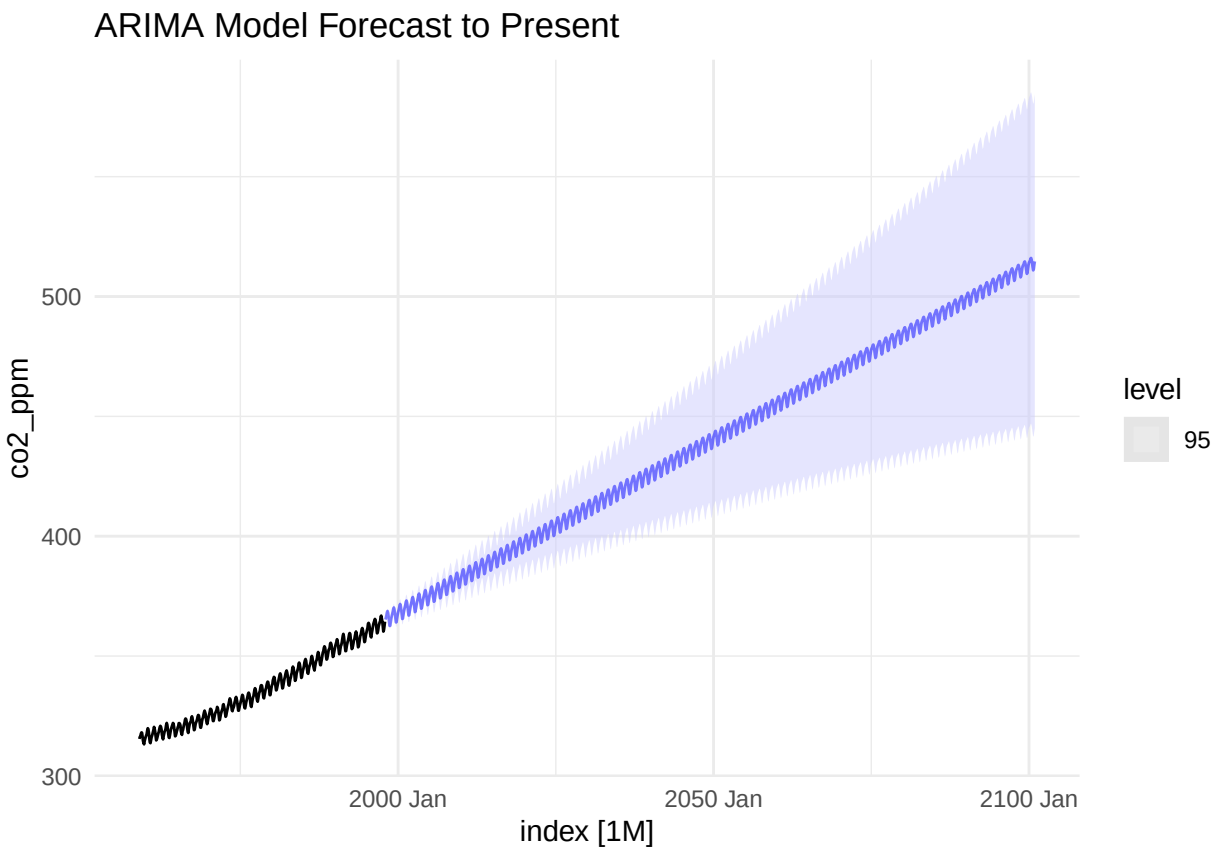
Here we consider a full SARIMA model with seasonality and iterate through the values for parameters, keeping the difference parameters (d and D) as 1. We then want to consider models with highest p-value from the Ljung-Box test first, and then choose those models which have lower AIC/BIC values.

```
##    order_pdq order_PDQ  aic  bic ljung_box
## 1      2-1-3     0-1-0 1026 1051     1.000
## 2      0-1-3     2-1-0 1026 1051     1.000
## 3      3-1-9     1-1-1  833  895     0.998
## 4      1-1-9     0-1-0 1053 1098     0.996
## 5      0-1-9     1-1-0 1053 1098     0.996
## 6      6-1-6     2-1-0  406  469     0.990
## 7      0-1-6     0-1-2  940  977     0.987
## 8     0-1-12     2-1-0  835  897     0.986
## 9      2-1-9     1-1-1  831  889     0.985
## 10     3-1-9     0-1-1  831  889     0.985
```

Among the models, the one with highest p-value from the Ljung-Box test is the $SARIMA(6,1,3)(1,1,1)_{12}$ model. We'll use this model to run the forecast.

Forecasts

```
Model_Forecast <- co2_tsb %>%
  model(ARIMA(co2_ppm ~ 0 + pdq(6,1,3) + PDQ(1,1,1), stepwise=FALSE,
              approximation=FALSE)) %>%
  forecast(h = 1236)
```



Given that we have fitted a model, we can make predictions from that model. Our preferred model, named in Equation 1 is quite simple, and as you might notice, does not in fact match up with the model that we have fitted.

The forecast is done up to December 2100. The 1st time we hit 420 ppm, is 435, which corresponds to year 2034 and month March. We hit 500 ppm in 1096, which corresponds to August of 2089. We do hit 420 ppm a few months later too. Similarly, we hit 500 ppm September of 2089 and a few months after that too. It is worth noting that the distribution of CCO_2 in March 20234 is normal.

```
Model_Forecast$co2_ppm[match(c(420), round(Model_Forecast$.mean))]
```

```
## <distribution[1]>
## [1] N(420, 94)
```

The model predics that CO_2 is a Normal Distribution with a mean of 420 and standard deviation (σ) of 9.3. Thus, the actual value is likely between 401 to 439 with 95% confidence $(420 - 2 * 9.3,\ 420 + 2 * 9.3)$.

Conclusions

In this report we started with an analysis of the data (commonly referred to as Exploratory Data. Analysis, or EDA) to see what we can learn of the observations from MLO. We then progressively added additional parameters to fit a model and do the forecast. We started with a naive model of CO_2 being a linear function of time. We abandoned this idea because we clearly see seasonality. Then we added time and season parameters to the linear model and used the time-series version (TSLM function) to get a model. We got

a linear model that provided a decent approximation to the observations we had. We could get the model fit and forecast. Subsequently, we modeled as an ARIMA model with difference of first order. Then, we introduced seasonality and created a SARIMA model.

While we can look at the metrics of the model (AIC, BIC, Ljung-Box test etc.) and try to refine the model we don't have a pragmatic way as yet to check the forecast. The one take away is that, at least qualitatively, we see a linear trend in the amount of CO_2 in our atmosphere. This alone should give us a cause for concern. We just can't continue business as usual. In a larger sense the exact amount of CO_2 that we'll see in the atmosphere is less important than the qualitative trend line. We also see the role the plants and forest play in absorbing CO_2. If this report is able to encourage readers to reduce the amount of CO_2 we emit we believe that the large goal is achieved. We can and will continue to refine the model as more data comes in to get a good enough model, which is what is realistically achievable.

Appendix: Model Robustness

While the most plausible model that we estimate is reported in the main, "Modeling" section, in this appendix to the article we examine alternative models. Here, our intent is to provide a skeptic that does not accept our assessment of this model as an ARIMA of order (1,2,3) an understanding of model forecasts under alternative scenarios.