Introduction

The Keeling Curve, initiated in 1958 by Dr. Charles David Keeling at the Mauna Loa Observatory, represents the ongoing record of atmospheric carbon dioxide (CO2) concentrations. Through precise measurements and the identification of atmospheric background levels, Keeling provided the first concrete evidence of the rise in CO2 levels, correlating with human activities like industrialization and fossil fuel combustion. This groundbreaking work highlighted the significant role of CO2 as a greenhouse gas contributing to global warming and climate change, marking it as one of the 20th century's pivotal scientific achievements. Our analysis aims to extend Keeling's legacy by applying models to forecast future CO2 concentrations, thereby underscoring the urgent need for discussions on mitigating climate change impacts and examining human influences on the climate. This endeavor seeks not only to predict CO2 levels but also to inform ongoing and future research within our laboratory on global climate dynamics.

Exploratory Data Analysis

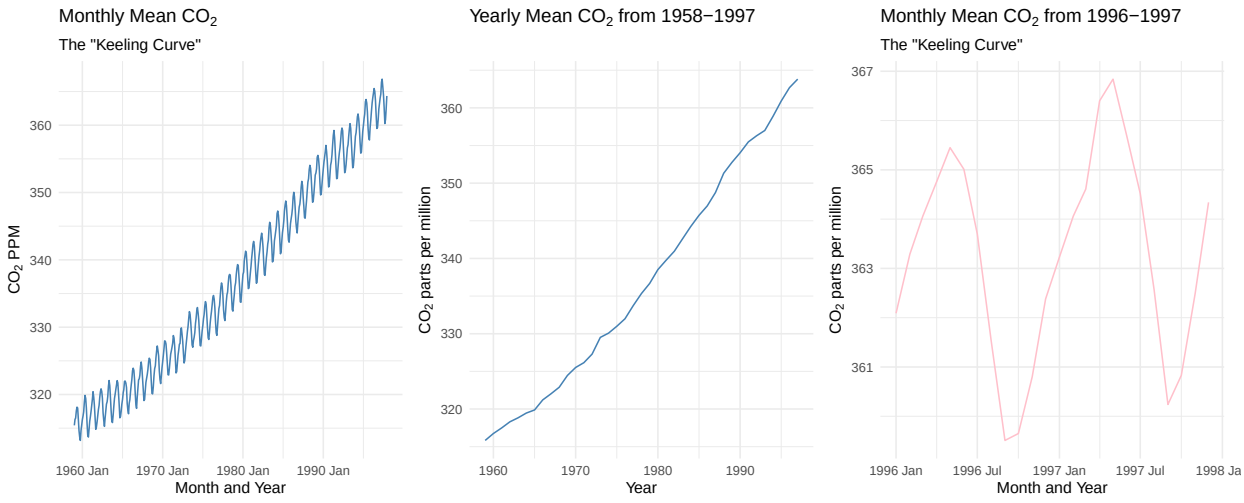
About the Data

This analysis utilizes data sourced from the National Oceanic and Atmospheric Administration's Global Monitoring Laboratory (GML). The GML is dedicated to researching significant issues such as greenhouse gases and the recovery of stratospheric ozone. As stated in the introduction, this data is gathered from the Mauna Loa observatory in Hawaii, a location favored for its ideal altitude which allows for the measurement of air masses over extensive areas. The data undergoes frequent calibration and comparison, ensuring an accuracy superior to 0.2 ppm (parts per million). The primary objective of these measurements is to determine the quantity of CO_2 that has been either added or removed from the atmosphere. These measurements reflect the mole fraction ofCO_2 in dry air, which gives us a clear picture of the changes in CO_2 levels.

Let us now investigate our data.

Graphs and Visualizations

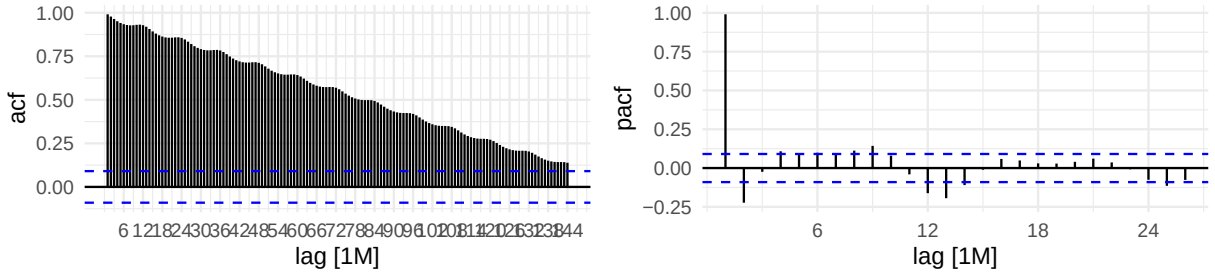
First let's visualize the entire curve up until now. We can observe what appears to be an upward trend. To better observe this, let's average the data by year and plot it. We can also see a periodic pattern that appears to be throughout the year. To illustrate this, let's look at a two-year window of the previous two years, 1996-1997.



We seem to observe that the CO_2 levels peak in the summer and fall in the winter, a consistent seasonal effect that appears to be yearly. We can see that the monthly CO_2 levels peak around May and trough around September.

The upward trend and seasonality imply that we may have strong autocorrelation, so we're going to plot the ACF and PACF.

We see from the lagged scatterplots that although the series observations of CO2 are positively associated with their lags, the positive associations are especially strong for Lag 1 and Lag 12. This is evidence of strong seasonality with a period of 12. In this case, we would expect to see similar CO2 levels 12 months after the January of a given year - January of next year. Similarly, we could expect to see similar CO2 levels 12 months after the May of a given year - May of next year.

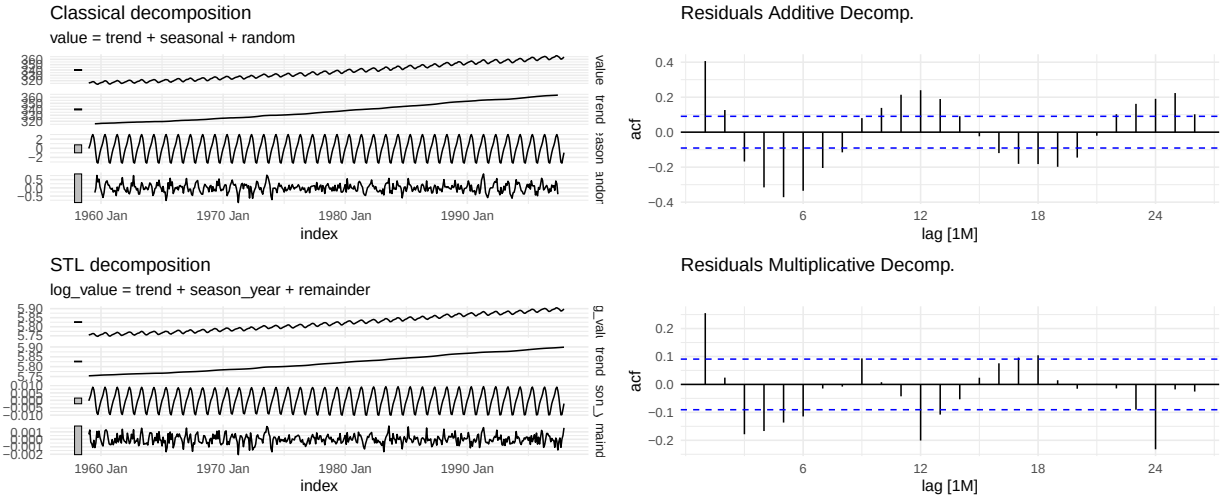


The ACF max lags has been extended to 144. This is an atypical visualization, but serves to illustrate the very slow linear decay of the ACF. This indicates a very strong autocorrelation, that past values have a significant effect on future values. It also confirms the overall upward trend we observed when visualizing the annualized averages.

We can also see a slightly 'waviness' in the ACF plot, reflective of the seasonality we see in the overall plot. From the PACF, we see that the first lag has a PACF value of nearly 1, and then it drops off very quickly. However, values are still outside of the confidence intervals. The PACF more clearly illustrates the seasonal pattern with its oscillations around 0. This PACF indicates at least a partial autoregressive component to the data.

Let's decompose the time series to examine if a multiplicative or additive model is more appropriate.

```
## Warning: Removed 6 rows containing missing values (`geom_line()`).
```



Although both residual ACF plots do not suggest perfect stationarity of the residuals, we see fewer significant autocorelations among the residuals of the multiplicative model.

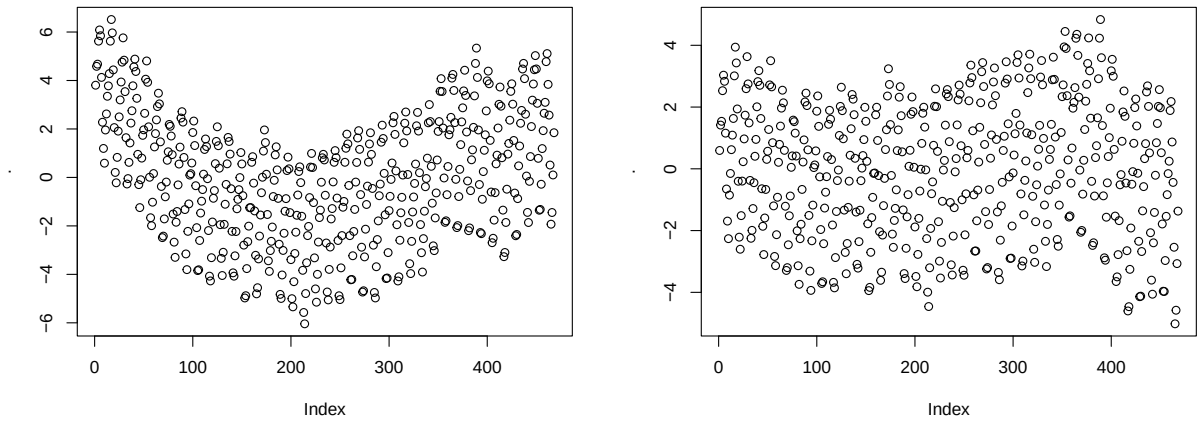
Linear Time Trend Model

Fitting a linear time trend model to the data

First let's fit a linear model to the data in an attempt to capture the underlying trend.

```
##
## Call:
## lm(formula = value ~ index, data = as_tsibble(co2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0413 -1.9469  0.0004  1.9106  6.5161
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.260e+02  1.514e-01  2153.2   <2e-16 ***
## index       3.580e-03  2.944e-05   121.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.619 on 466 degrees of freedom
## Multiple R-squared:  0.9694, Adjusted R-squared:  0.9694
## F-statistic: 1.478e+04 on 1 and 466 DF,  p-value: < 2.2e-16
```

The R-squared indicates a good fit, and there is a low standard error. Let's now plot the residuals to examine this further.

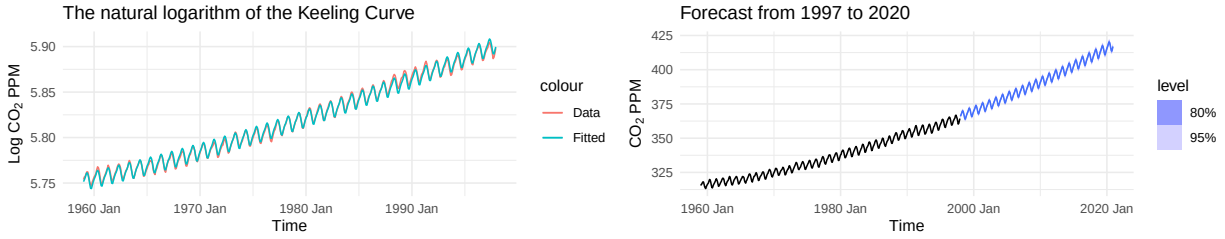


The residuals in the first plot show a clear deviation from linearity. In fact they seem to have some sort of quadratic curve, so let's next fit a quadratic time trend model using the `poly` function in R. The R^2 is slightly improved here. Let's look at the residuals and see how the model did. These residuals look much better and show less of a pattern.

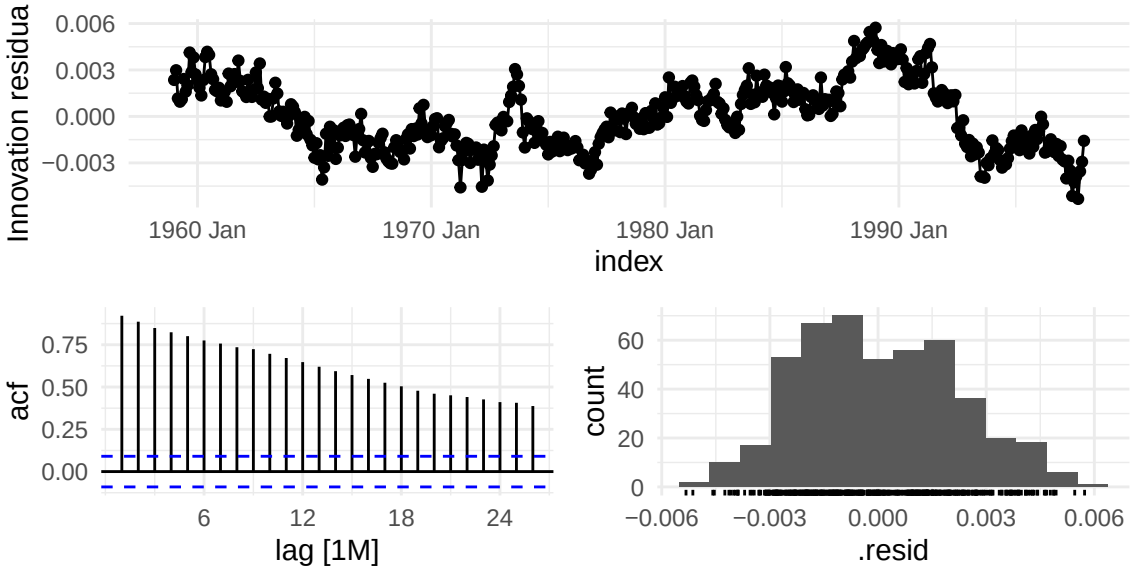
Our earlier analysis did not indicate a strict preference for a multiplicative or linear model when decomposing the data. So, we'll fit a quadratic model with seasonal values to both the raw atmospheric CO_2 ppm and the log of the value.

In each case, the ACF of the residuals still shows significant lag, indicating that they are not yet stationary. The seasonality appears to still be present in the additive residuals while the curve is more linear for the

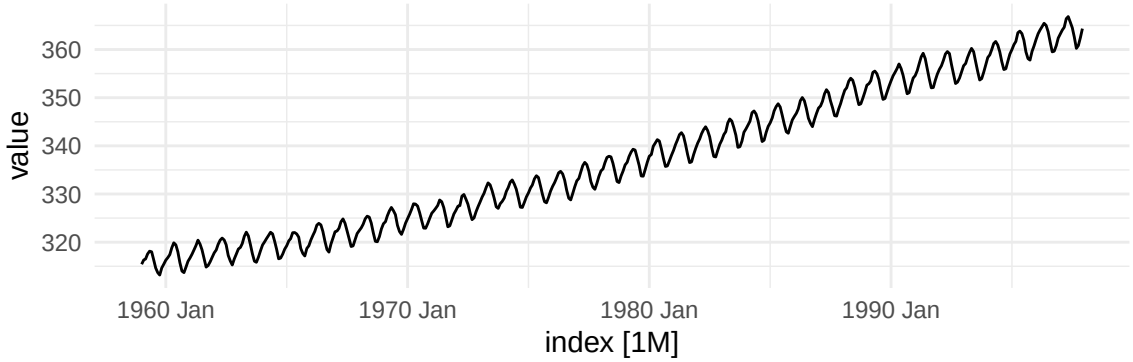
multiplicative(logarithmic) residuals. The residuals are approximately normally distributed for each model. We do not expect this to be the best model, but let's forecast through the year 2020 with the multiplicative polynomial time trend model.



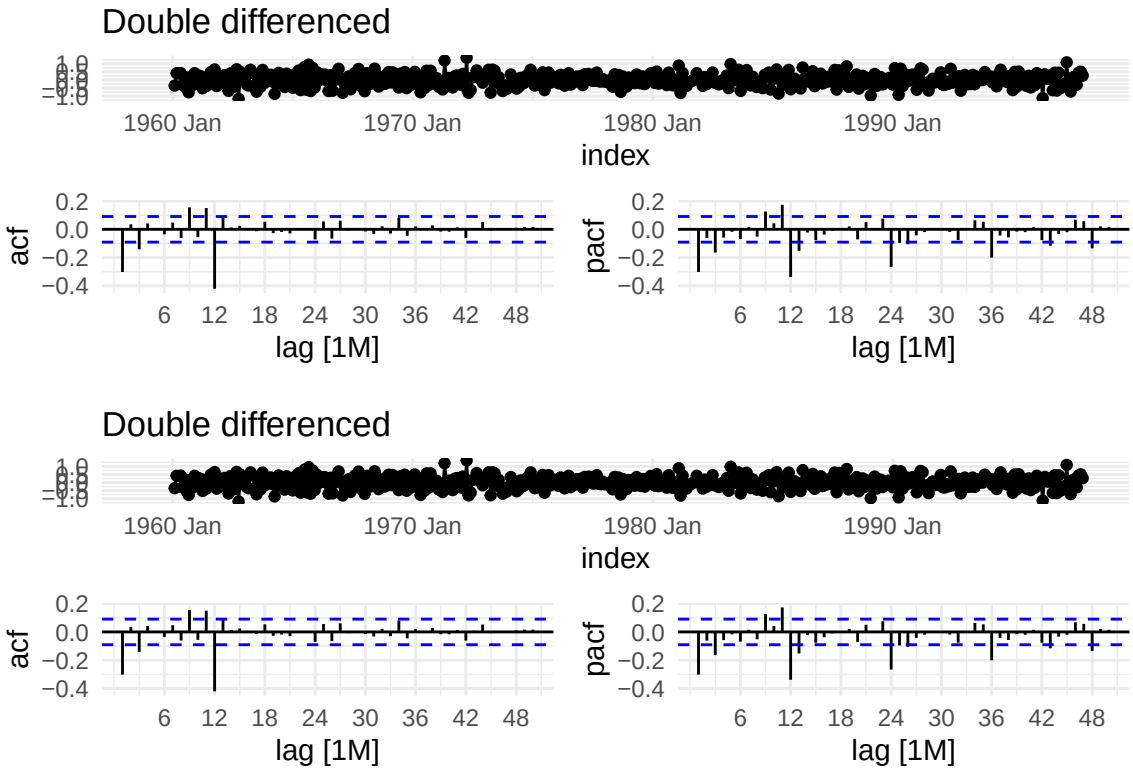
The plot above shows the forecasts produced by the quadratic time trend model. We see that the model forecasts include a consistent upward trend with consistent fluctuations.



Now that we've fit the decomposed quadratic model, let's also fit ARIMA models to the data. First let's plot the time series again for convenience.



If we take a seasonal difference followed by a first difference, we get the resulting time series, PACF, and ACF plots below for the raw and log of the data.



The plots above show that a seasonal difference followed by a first difference would make the series appear stationary, as the resulting series centers consistently around zero with only a few significant autocorrelations. This influences the choice of the non-seasonal difference d and the seasonal difference D taking on values of 1 for our models. We additionally see significant partial autocorrelations at roughly 3 intervals of 12 after the first lag, indicating a seasonal AR component of 3. This is how we decided for reasonable search values of also run automated model selection procedures with reasonable values of p, d, q, as well as P, D, Q. Initial EDA does not strongly suggest an additive model over a multiplicative one, so we will evaluate both.

We evaluated the residual plots and the Box-Ljung tests at 1 and 10 lags to determine whether the models were worth continuing to evaluate with. The tests are not shown for brevity in this report, but are included in the code. The Box-Ljung tests indicated that all the models generated independent residuals and could be used for forecasting.

According to our additive ARIMA model, we predict that CO_2 is expected to be at 420 ppm from April 2039 to September 2044, and at 500 ppm from April 2102 to October 2107. Our multiplicative ARIMA model predicts 420 ppm from Apr 2027 to Oct 2038, and 500 ppm from May 2064 to Oct 2084. The additive linear model predicts 420 ppm to be reached from May 2022 to Oct 2024, and 500 ppm to be present from Apr 2051 to Oct 2052. The linear multiplicative model predicts 420 ppm to be reached from May 2020 to Nov 2022, and 500 ppm to be reached from March 2045 to Oct 2046

For the year 2100, our additive ARIMA model predicts 495.1 ppm. Our multiplicative ARIMA model predicts 534 ppm. Our linear additive model predicts 686 ppm. Our linear log model predicts 855 ppm.

According to the data we have, keeping all else constant, our predictions would be somewhat valid for years nearer to the current year, 1997. A prediction on year 2100 is far too much into the future and comes with a high confidence interval and inaccuracy, which would very likely be unacceptable. Another possible approach to test the prediction performance would be to fit this model on a train-test split and evaluate the model on the test split nearer to the current date. Of course, this would be done on advice from our senior statisticians.

Introduction

Building on our 1997 report, we continue to investigate the critical question- how has atmospheric carbon dioxide (CO_2) concentrations evolved since our last analysis and what would the forecasts be?

Since 1997, the data generating process has largely remained consistent, with continuous measurements being taken at the Mauna Loa Observatory. However, the rate of CO_2 accumulation in the atmosphere has shown signs of acceleration, a concerning trend that underscores the urgency of addressing climate change.

In this report, we will delve deeper into the updated dataset, extending our previous models to capture these recent dynamics. Our aim remains to stimulate conversation about the potential impacts on Earth's ecosystem and the influence of human activities on global climate, all in the context of rising CO_2 levels. We hope this report will serve as a valuable resource for future investigations on this subject within our laboratory and beyond.

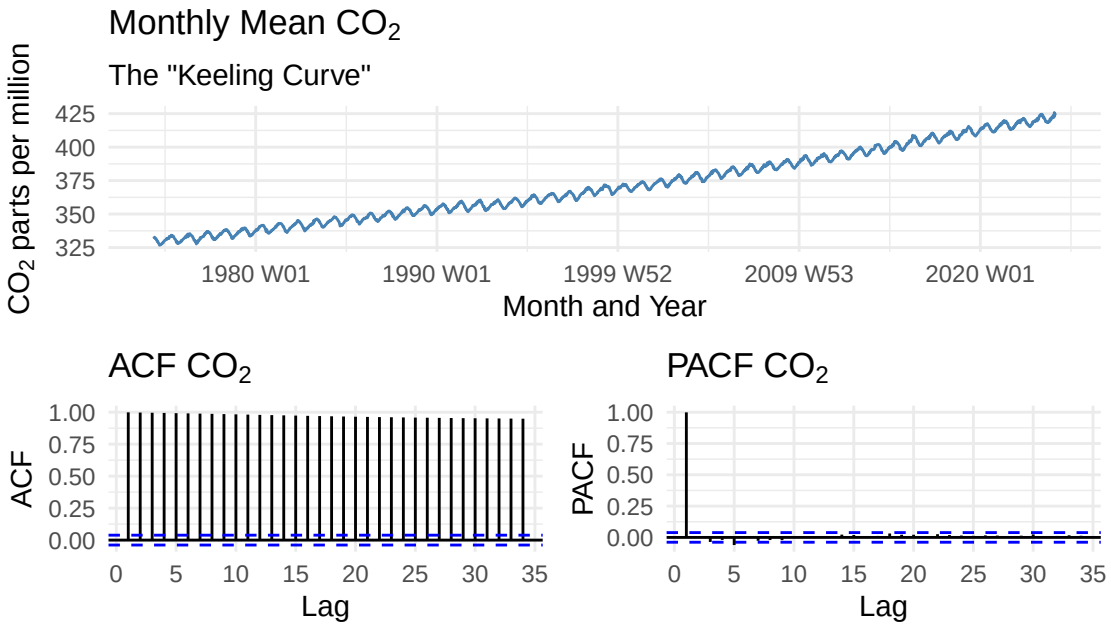
With this updated data, let us now proceed with our investigation.

Create a modern data pipeline (1b)

```
# Load data
co2_present <- read.csv("co2_weekly_mlo.csv", comment.char = "#")

# There are no "missing" values but there are values where the average is -999
co2_present <- co2_present[co2_present$average >= 0, ]

# Index and convert to tsibble
co2_present <- co2_present %>%
                mutate(date=make_date(year, month, day)) %>%
                mutate(index=yearweek(date)) %>%
                as_tsibble(index=index) %>%
                mutate(value=average)
```



The Keeling Curve hasn't evolved drastically from 1997. The trend looks very similar as earlier but there seems to be a very slight dip around the years 1997-1999. There seems to be some variation in the crests of the seasonality but overall the periods of seasonality remain the same. In conclusion, the curve is largely similar to the data evaluated in 1997. Let us now explore the autocorrelation. The ACF and PACF plots show that the current values of lag are significantly correlated to the previous ones. There is no constant mean and variance.

Compare linear model forecasts against realized CO2 (2b)

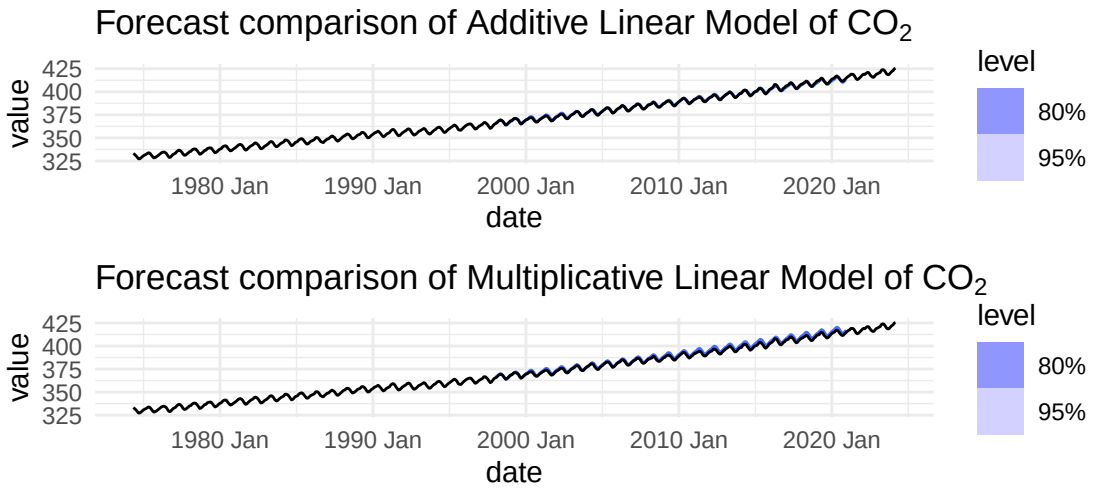
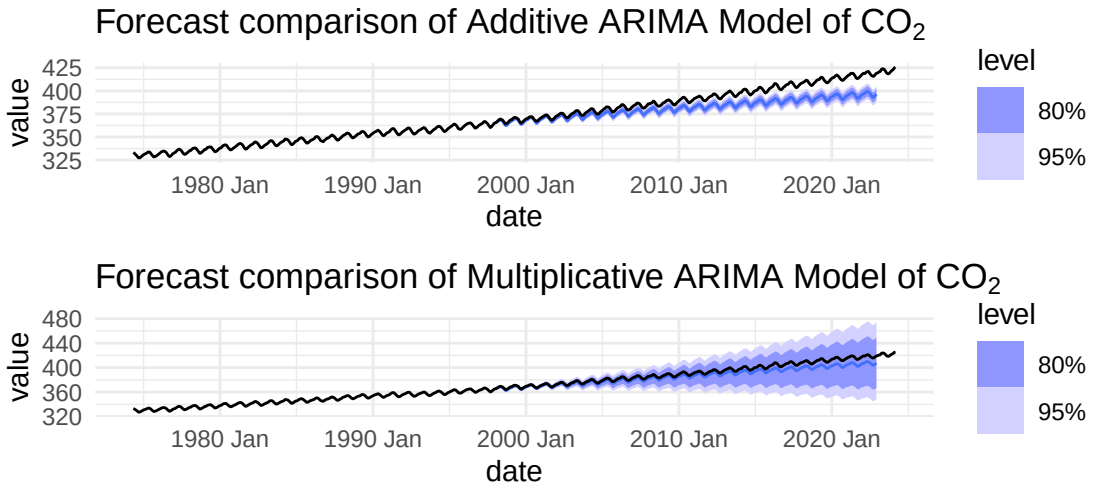


Figure above shows forecasts of both additive and multiplicative decomposed linear models fitted in 1997, layered to the realized data. We can see from the first plot which is forecasting `linear_add_forecasts` aligns very closely and is similar to the realized data. The `linear_mult_forecasts` predictions align well with the realized data in the early years of the forecast, but then slowly deviates as time progresses.

Compare ARIMA models forecasts against realized CO2 (3b)



From the image above, the predictions from both ARIMA models fitted in 1997 initially align with the actual data but eventually deviate. Specifically, after a few years post-1997, the actual data falls outside the prediction interval of the `add_arima_forecasts`. On the other hand, the `mult_arima_forecasts` manages to capture the actual data within the 95% prediction interval for the entire duration.

It is evident from these plots above that the Additive Linear Model of CO_2 aligns best with the actual observed data, clearly defying the expectations in 1997. The Keeling Curve continues to grow with the with little change in trend and seasonality. The accuracy for these models are compared in the next section.

Evaluate the Performance (4b)

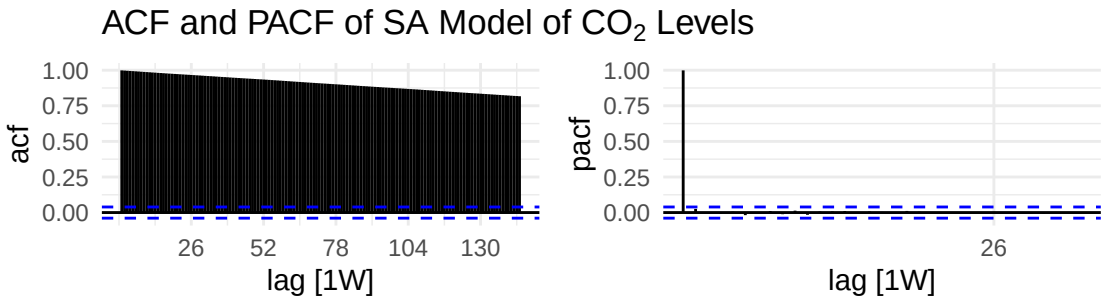
The prediction made in 1997 estimated that CO_2 levels would hit 420 ppm by April 2039. However, when compared with the realized data, this CO_2 level was already reached in April 2022, indicating that the 1997 prediction was not very accurate, as it underestimated the rate of CO_2 increase by a margin of 17 years!

Name	RMSE
Linear Additive	0.7612847
Linear Multiplicative	1.9696454
ARIMA Additive	11.8729033
ARIMA Multiplicative	5.9665362

To check the performance of the models, we perform accuracy tests on each of them for the entire forecasted period. We compare the `RMSE`, which is the measure of the average deviation from the actual values, between these models. We see that the Linear Additive model has the lowest `RMSE` of `0.76` indicating a better fit to the realized data compared to the other models.

Train best models (5b)

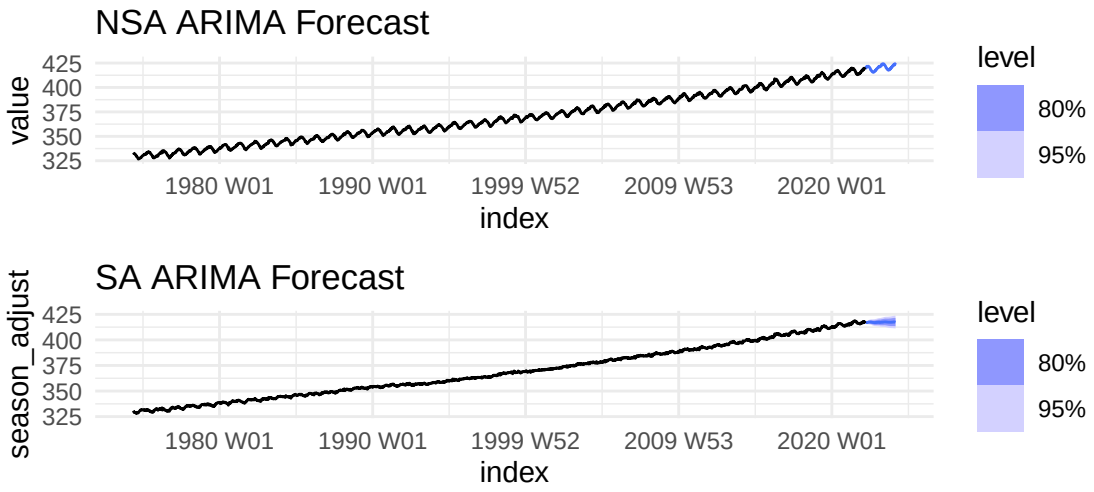
```
# Seasonally Adjust
co2_present_sa <- data_interpolated %>%
  model(STL(value ~ season(window = "periodic"))) %>%
  components() %>%
  select(index, trend, season_adjust = season_adjust)
```



On the seasonally adjusted data, the ACF and PACF plots show the presence of unit root. The original data (Non-seasonally adjusted (NSA)) and the seasonally adjusted (SA) data are both split into train and test set, and are used to fit ARIMA models.

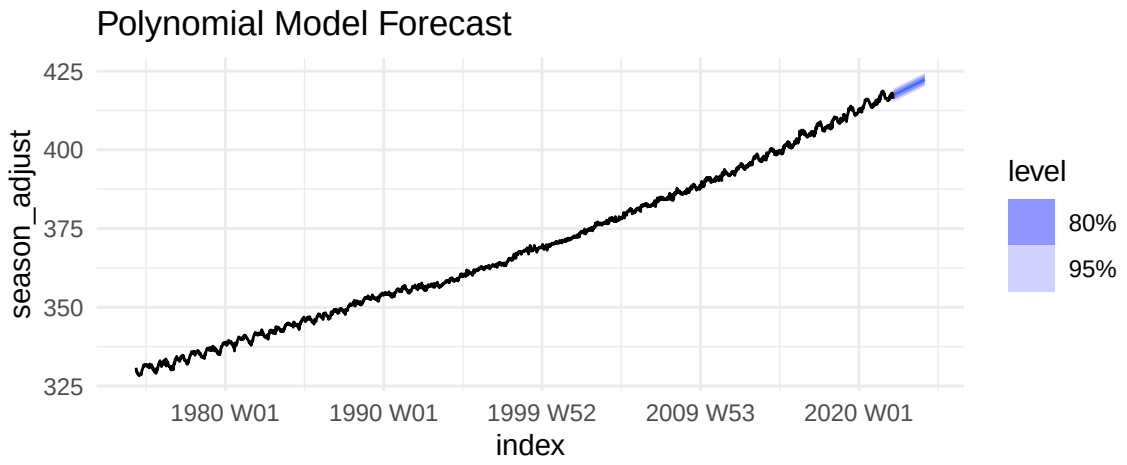
```
fit_arima_wo_log <- training %>%
                  model(auto_mod_aic = ARIMA(value ~ 0 + pdq(0:5, 0:2, 0:5) +
                                                PDQ(0:5,0:2,0:5), ic="aic",
                                                stepwise=F, greedy=F))
fit_arima_sa_test2 <- training_sa %>%
                  model(auto_mod_aic = ARIMA(season_adjust ~ 0 + pdq(0:5, 0:2, 0:5) +
                                                PDQ(0:5, 0, 0:5), ic="aic",
                                                stepwise=F, greedy=F))
```

We use the auto ARIMA function to find the best models in a limited search space. The function suggests a non-seasonal ARIMA model `ARIMA(0,1,3)(2,1,0)[52]` and a seasonal ARIMA model `ARIMA(1,1,3)(2,0,0)[52]`. The residuals of both these models posses a normal distribution and the Ljung-Box test provides a non-significant p-value, failing to reject the null that the data are IID.



The NSA model has a lower RMSE of `0.78` compared to the RMSE of the SA model, `3.50`. Which means the NSA ARIMA model `ARIMA(0,1,3)(2,1,0)[52]` fits the data better and would be our preferred choice of model.

```
# Fit ploynomial
fit_poly <- training_sa %>%
  model(model = TSLM(season_adjust ~ trend()+I(trend()^2)))
# Forecast with polynomial model
forecast_poly <- forecast(fit_poly, new_data = test_sa)
forecast_poly_plot <- forecast_poly %>% autoplot(training_sa) +
  labs(title="Polynomial Model Forecast")
forecast_poly_plot
```



The forecast plot for the polynomial model is as above. We get an `RMSE` of 1.39 from the polynomial model fitted on the seasonally adjusted data. In comparison, our NSA `ARIMA(0,1,3)(2,1,0)[52]` has a lower RMSE of `0.78`, indicating a better fit, therefore a better choice (unless we're over-fitting).

How bad could it get? (6b)

```
# Forecast from NSA Model
extended_forecast <- fit_arima_wo_log %>%  forecast(future_years_2130, 20)
```

According to our model, we predict that CO_2 is expected to be at 420 ppm from 12th week of 2022 to the 42nd week of 2024, and at 500 ppm from the 6th week of 2057 to the 36th week of 2059.

The CO_2 level prediction for the year 2122 is 651.89 ppm. Again, similar to the 1997 model and report, our predictions would be somewhat valid for the nearer future. A prediction on year 2122 is far too much into the future and comes with a high confidence interval and inaccuracy, which would very likely be unacceptable. Therefore, we are not really confident with the output especially considering current geopolitical efforts to reduce carbon emissions, meaning the current trend may not continue to grow.