# PSTAT 174 FINAL PROJECT:
# Trends in Atmospheric Carbon Dioxide

*Trevor Rizzi*

*June 2023*

## Abstract

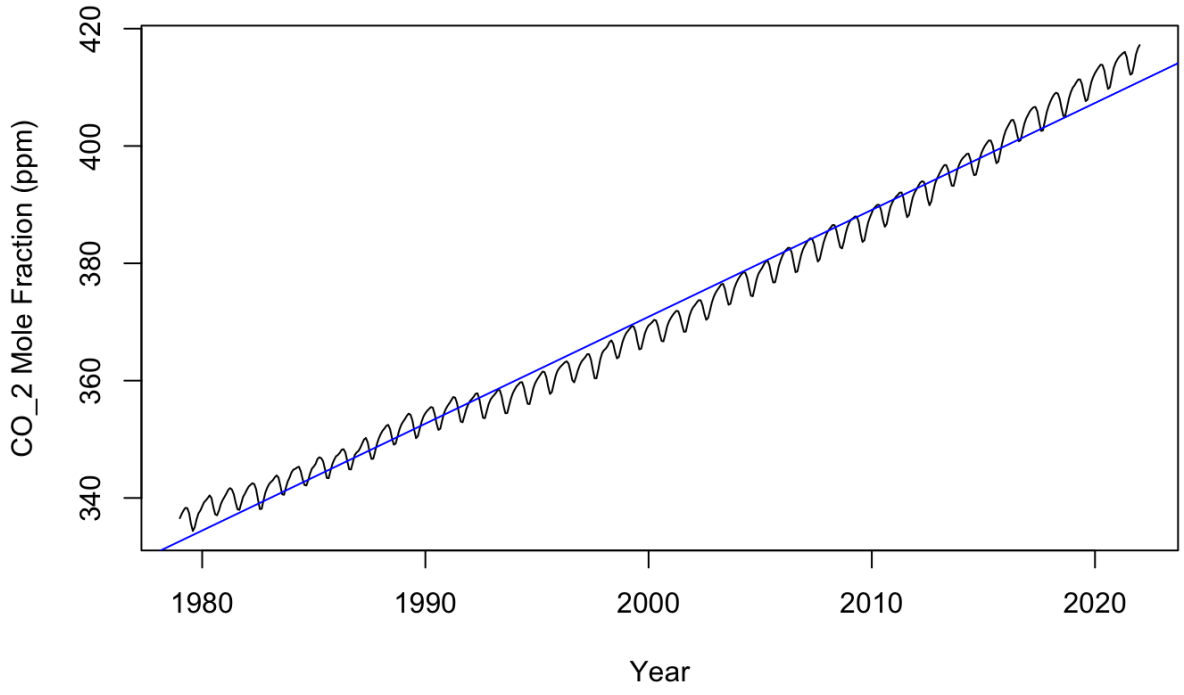
This project presents an analysis of global atmospheric carbon dioxide (CO2) concentrations over time. The goal of this time series analysis is to appropriately model the data, and forecast future CO2 concentrations. The data was split into training and testing sets, with the testing set containing the last 12 months of recorded data for comparison with the prediction. The data is not stationary, and displays clear seasonality along with a positive upward trend. Its variance however, remains fairly constant and transformations had little effect on the data's normality. The data was made stationary through differencing. It was then diagnosed with three well fitting SARIMA models. The 1st model was not invertible, but the other two models passed diagnostic testing and were judged on their AICc score. Ultimately Model C was chosen because it had the lowest AICc score, but Model B would have performed similarly. The model was then used to predict the next year's worth of data points and then compared with the actual values in the testing set. Predicted values fell within the 95% confidence interval suggesting that the selected model was an appropriate fit.

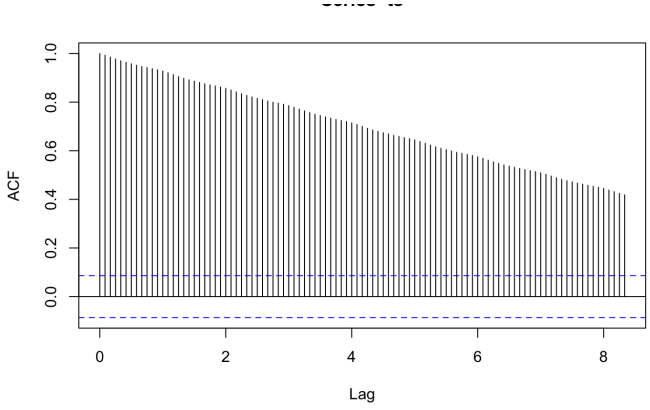
# INTRODUCTION

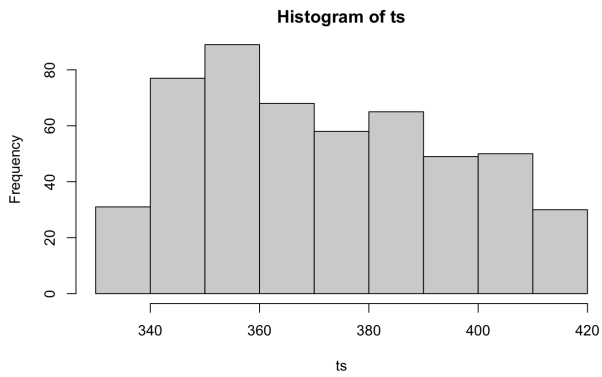
This project presents an analysis of global atmospheric carbon dioxide (CO2) concentrations over time, using time series data sourced from the Global Monitoring Laboratory (GML) of the National Oceanic and Atmospheric Administration (NOAA)( https://gml.noaa.gov/ccgg/trends/global.html). The data is recorded monthly 1979 through 2023 and provides valuable insights into the long-term trends and patterns of atmospheric CO2 levels. “Data are reported as a dry air mole fraction defined as the number of molecules of carbon dioxide divided by the number of all molecules in air, including CO2 itself, after water vapor has been removed. The mole fraction is expressed as parts per million (ppm). Example: 0.000400 is expressed as 400 ppm.”CO2 emissions have consistently increased over time, and it is very important that it is measured. Carbon dioxide plays a critical role in regulating Earth’s climate, and the increase in its concentrations is primarily attributed to human activities. Understanding the long-term trends and patterns of CO2 can help to assess the impact of humans on the climate system and evaluate mitigation efforts.

The following analysis was conducted in R, using RStudio. The data was split into training and testing data. The training data contains points from 1974-2022 which is used for model-fitting, and the testing data contains points from 2023 which is used for prediction checking. There is a clear upward trend in the data which can be explained by increasing industrialization and human activity. A yearly seasonal pattern can also be seen, explained by the National Science Foundation as such: “Levels of carbon dioxide in the atmosphere rise and fall each year as plants, through photosynthesis and respiration, take up the gas in spring and summer, and release it in fall and winter. ” The data has a very consistent variance, and a strong right skew. All of the common transformations were applied to the data to try to normalize it, but they did not have much effect so the analysis proceeded with the original data. To make it stationary, it is differenced first at lag 1 to remove trend and then again at lag 12 to remove seasonality. It now resembles a stationary process which can be modeled by SARIMA. Using the ACF and PACF, multiple potential candidates were identified. They were then compared by their AICc scores, where the lowest three were chosen. SARIMA(4,1,1)x(1,1,1)_12, SARIMA(4,1,2)x(1,1,1)_12, SARIMA(0,1,7)x(0,1,1)_12 . Diagnostic checking was then performed to ensure that these models were appropriate for forecasting. The first model is not invertible, and thus a poor choice for fitting the data. The other two models passed all of their diagnostic tests except for the shapiro test for normality. Looking at the histograms and QQ-plots, however, suggest that they are both very close to normal. Ultimately Model C was chosen because it had the lowest AICc, but Model B would have performed very similarly. The model was then used to forecast 1 year of future data, which was then compared to the actual values in the testing data set. Prediction points fell within the prediction interval, suggesting that the forecast was successful.

**Plotting time series:**

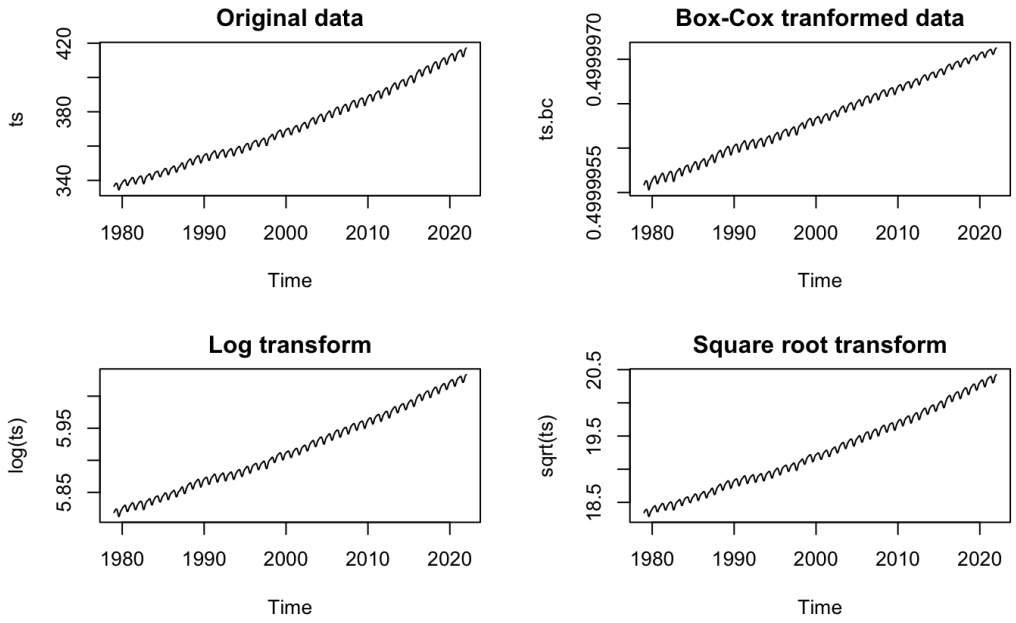


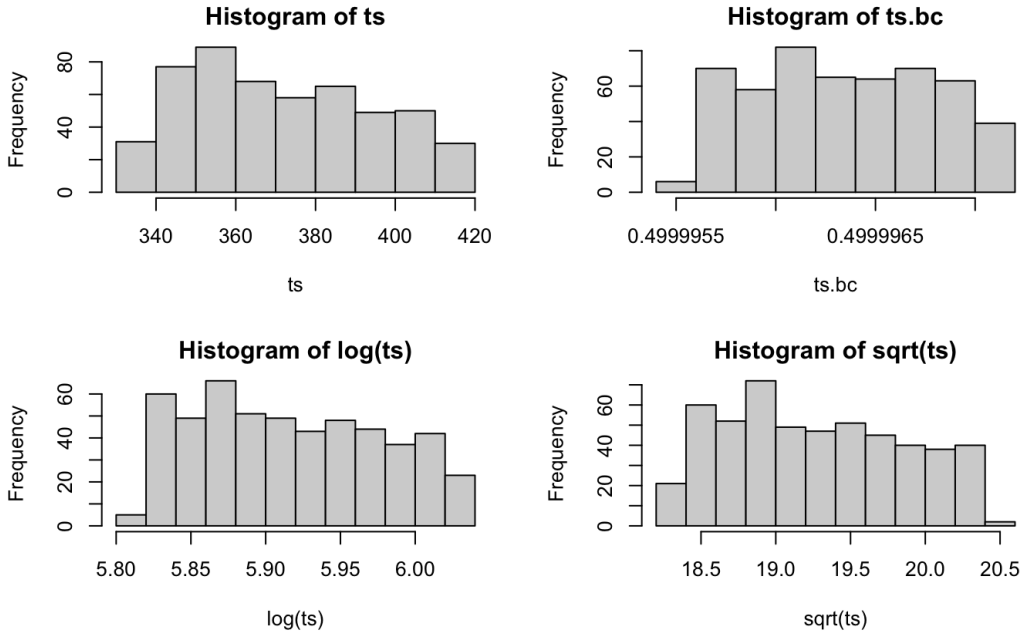




The data has both a clear upward trend and strong seasonality, suggesting that the data is non-stationary. The chart in the lower right hand corner shows the fit of our trend, which is very linear. The variance = 522.3376 appears to be pretty constant over time, indicating a transformation may not be necessary. The ACF is very large, and has very slight periodicity. The histogram looks to be close to normal, but is badly right skewed. To address this, some transformations will be applied to the data.

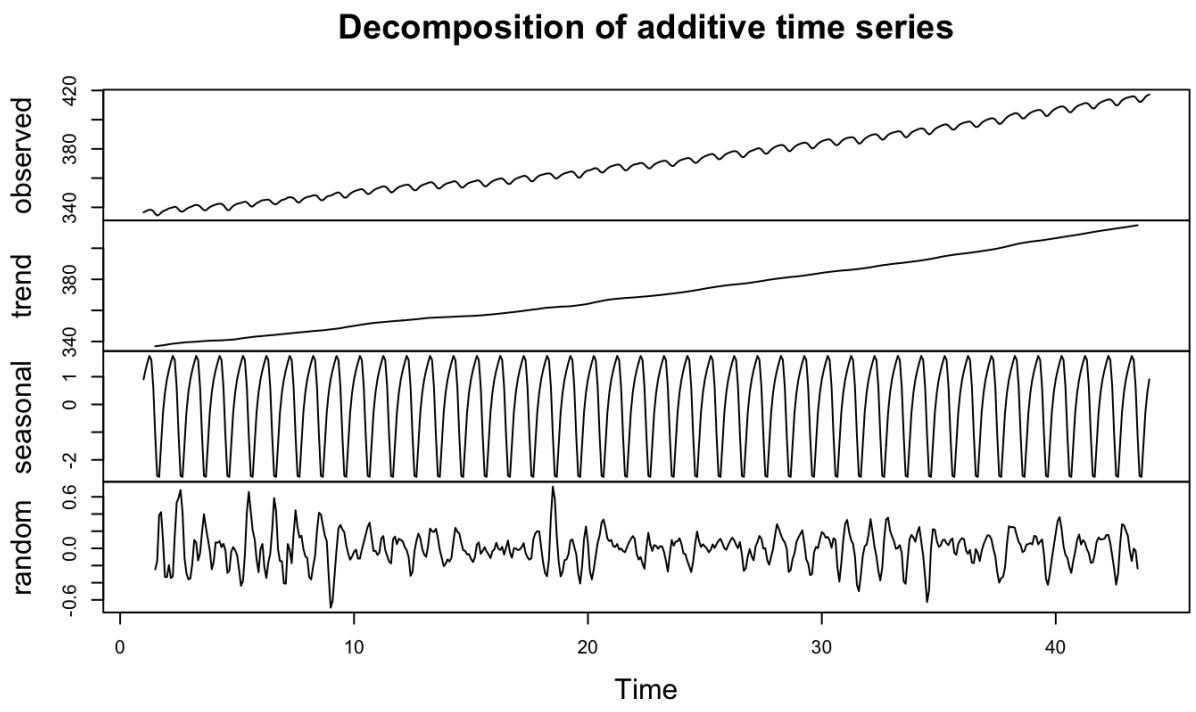
## Transformations




Three transforms were attempted on the data, a box-cox with lambda = 2, a log transform, and a square root transform. The box-cox and log transform do seem to have slightly improved the linearity of the data.

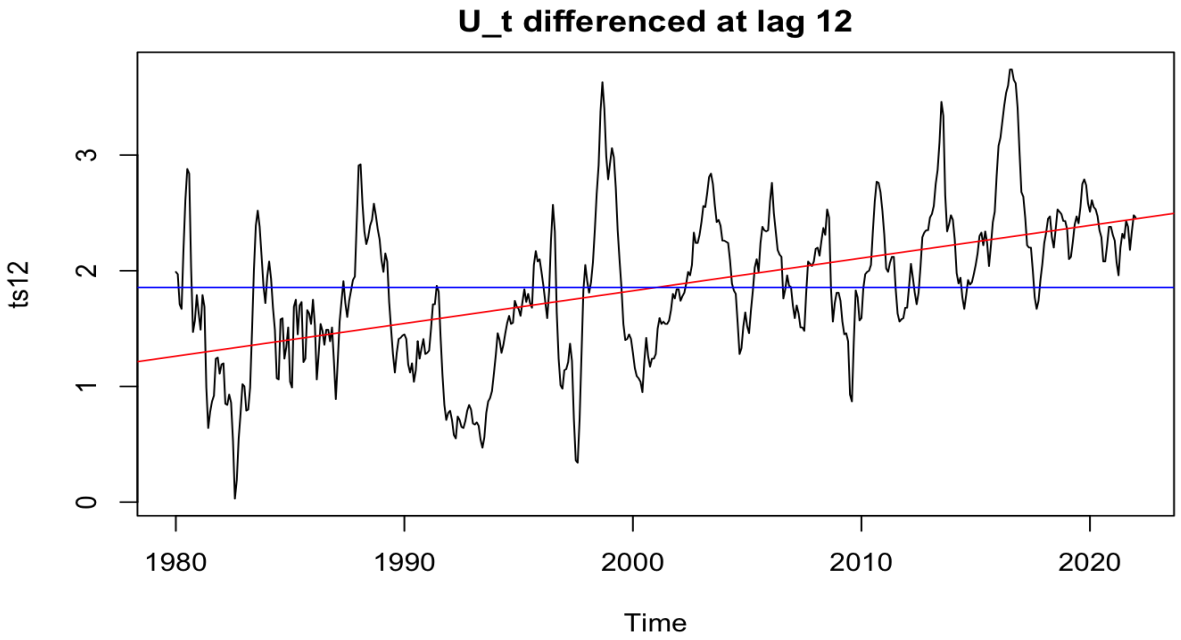



None of the transforms effectively made the data more normal than it already is. Initially, I chose the box-cox transform to proceed with the project, but R kept throwing errors when I used it to fit any SARIMA model. Due to this, I chose to proceed with the original data as if it were normal.

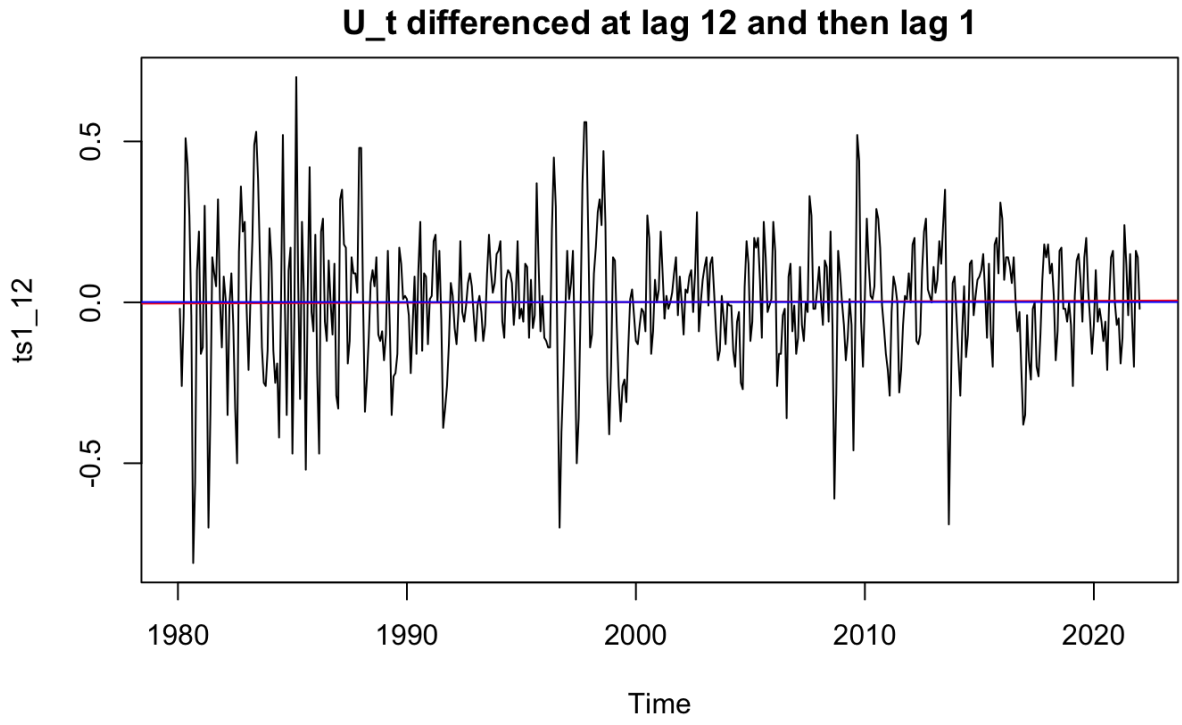



The decomposition shows the data's strong positive linear trend, and the consistent monthly seasonal component. We will have to use differencing to remove trend and seasonality to create a stationary series

**Differencing**

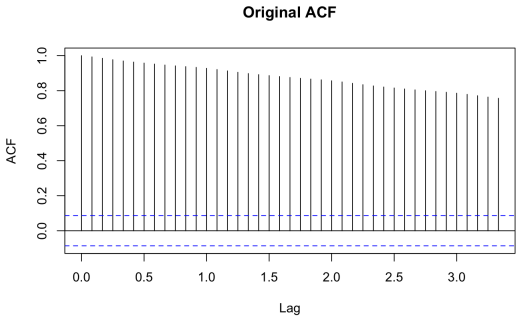



Differencing at lag 12 appears to have reduced the seasonality, but the trend is still strong. The variance was greatly reduced to 0.4232723.

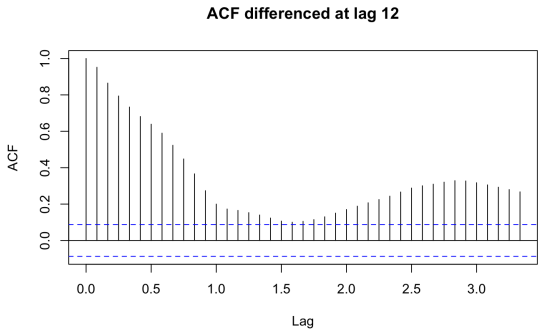



Differencing again at lag 1 successfully removes the trend, leaving us with what appears to be stationary time series data. The variance was further reduced to 0.04008959.

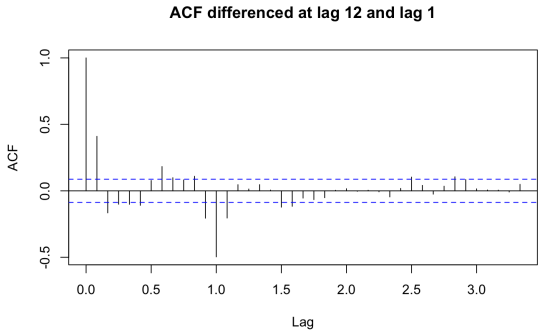
## Acf of Differenced Data



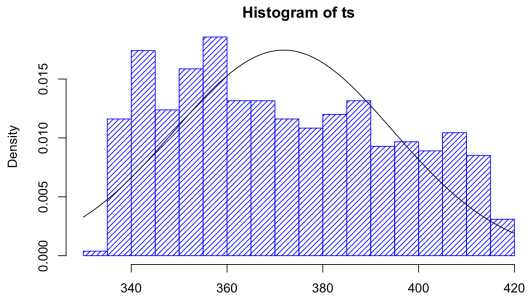

The ACF of the original data has a slow decay and a seasonal trend, though it is hard to see on the graph. This is not stationary.

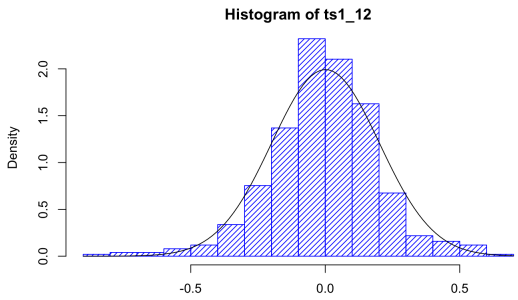



The ACF after differencing at lag 12 no longer shows seasonality, but the slow decay is still present. This is not stationary.

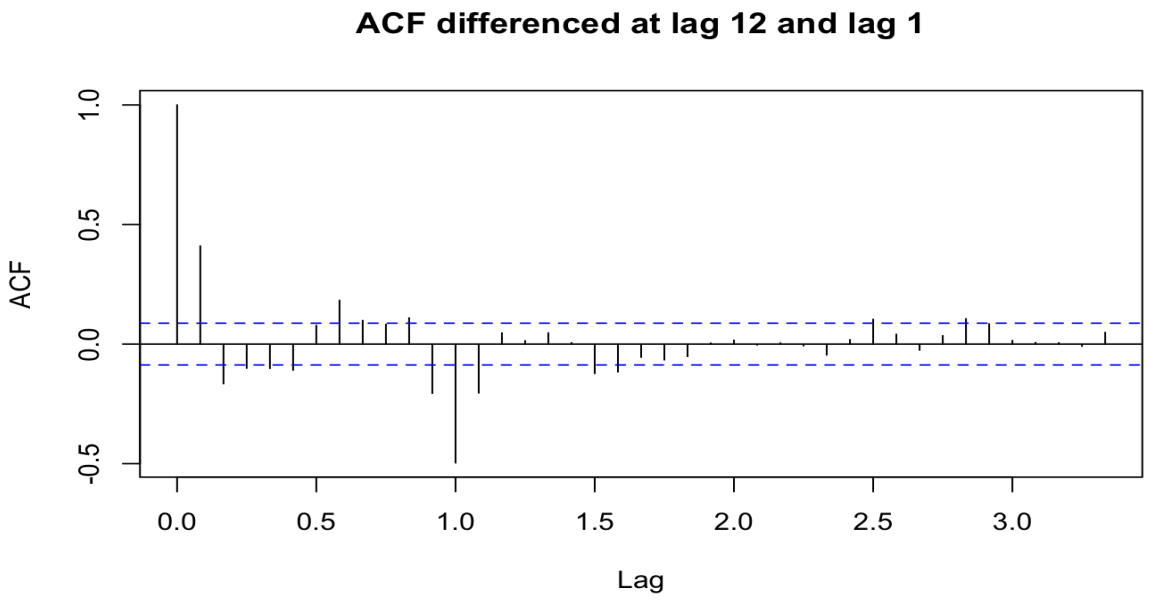



After differencing at lag 12 and then lag 1, the data resembles a stationary process.

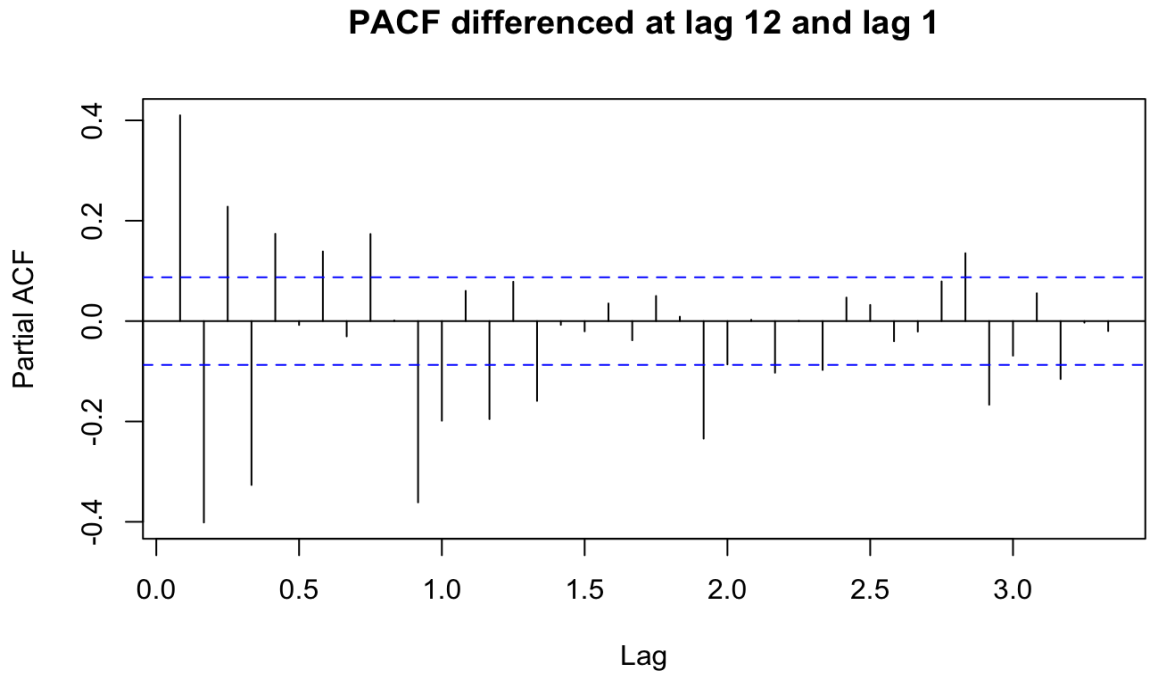



We can see that the differencing has fixed the right skew, making the data normal and almost Gaussian.




The ACF is outside confidence intervals at lags 1, 2 , 7, 11, and 12. The large spike at 1 = lag 12 is indicative of a seasonal moving average component. So we will consider d = 1, D = 1, s = 12, Q = 1, q = 1, 2, 7, 11. We will now look at the PACF to judge auto regressive components next.




The PACF is pretty strong for the first 4 lags and for lag 11/12. There is also a considerable spike at lag 9. We can consider p = 4, 9, 11 ; P = 0 , 1.

A total of 13 models were run, and they can be seen on PAGE 3 of the APPENDIX. Of all of these models, the three with the lowest AICc were taken.

Model A: SARIMA(4,1,1)x(1,1,1)_12

Model B: SARIMA(4,1,2)x(1,1,1)_12

Model C: SARIMA(0,1,7)x(0,1,1)_12

Model A: SARIMA(4,1,1)x(1,1,1)_12

```
arima(x = ts, order = c(4, 1, 1), seasonal = list(order = c(1, 1, 1), period = 12),
    method = "ML")
Coefficients:
         ar1      ar2     ar3      ar4     ma1     sar1     sma1
      0.1897  -0.2963  0.0890  -0.1330  0.6277  -0.0418  -0.8413
s.e.  0.1067   0.0893  0.0804   0.0621  0.1026   0.0543   0.0320

sigma^2 estimated as 0.01308:  log likelihood = 369.35,  aic = -722.69
[1] -722.4023
```

We can take ar1, or ar3 as 0 since 0 falls within their coefficients plus or minus 2 * SE.

```
arima(x = ts, order = c(4, 1, 1), seasonal = list(order = c(1, 1, 1), period = 12),
    fixed = c(NA, NA, 0, NA, NA, NA, NA), method = "ML")
Coefficients:
         ar1      ar2  ar3      ar4     ma1     sar1     sma1
      0.1025  -0.2269    0  -0.0902  0.7106  -0.0457  -0.8416
s.e.  0.0588   0.0559    0   0.0488  0.0481   0.0540   0.0316

sigma^2 estimated as 0.01311:  log likelihood = 368.68,  aic = -723.36
[1] -723.1316
```

I tried multiple combinations of fixing these coefficients to 0. Ultimately, taking the third coefficient as 0 effectively reduced the AICc, so this is the version of the model that will move forward.

Model B: SARIMA(4,1,2)x(1,1,1)_12

```
arima(x = ts, order = c(4, 1, 2), seasonal = list(order = c(1, 1, 1), period = 12),

    method = "ML")

Coefficients:

         ar1      ar2     ar3      ar4     ma1      ma2     sar1     sma1

      0.3164  -0.2983  0.1078  -0.1293  0.5003  -0.1029  -0.0403  -0.8425

s.e.  0.2659   0.0881  0.0853   0.0611  0.2651   0.2080   0.0544   0.0320


sigma^2 estimated as 0.01307:  log likelihood = 369.47,  aic = -720.94

[1] -720.5743
```

In this model, ar1, ar3, ma1, ma2, sar1 can be taken as 0.

```
arima(x = ts, order = c(4, 1, 2), seasonal = list(order = c(1, 1, 1), period = 12),
    fixed = c(0, NA, 0, NA, NA, NA, NA, NA), method = "ML")
Coefficients:
      ar1      ar2  ar3      ar4     ma1     ma2     sar1     sma1
        0  -0.2364    0  -0.0965  0.8150  0.0924  -0.0461  -0.8412
s.e.    0   0.0608    0   0.0522  0.0452  0.0519   0.0540   0.0317

sigma^2 estimated as 0.01311:  log likelihood = 368.71,  aic = -723.43
[1] -723.2003
```

The lowest AICc was produced by fixing ar1 and ar3 to 0.

Model C: SARIMA(0,1,7)x(0,1,1)_12

```
arima(x = ts, order = c(0, 1, 7), seasonal = list(order = c(0, 1, 1), period = 12),
      method = "ML")

Coefficients:
         ma1      ma2      ma3      ma4      ma5     ma6      ma7     sma1
      0.8066  -0.1465  -0.1686  -0.0478  -0.0685  0.0301  0.0701  -0.8492
s.e.  0.0462   0.0599   0.0592   0.0537   0.0581  0.0615  0.0525   0.0274

sigma^2 estimated as 0.01309:  log likelihood = 369.09,  aic = -720.18
[1] -719.8183
```

Can take ma4, ma5, ma6, ma7, as 0 .

```
arima(x = ts, order = c(0, 1, 7), seasonal = list(order = c(0, 1, 1), period = 12),
    fixed = c(NA, NA, NA, 0, 0, 0, NA, NA), method = "ML")

Coefficients:
         ma1      ma2      ma3  ma4  ma5  ma6     ma7     sma1
      0.8155  -0.1519  -0.1852    0    0    0  0.0217  -0.8539
s.e.  0.0448   0.0593   0.0473    0    0    0  0.0290   0.0259


sigma^2 estimated as 0.01314:  log likelihood = 367.96,  aic = -723.91
[1] -723.7441
```

Taking ma4, ma5, ma6 as zero resulted in the lowest AICc

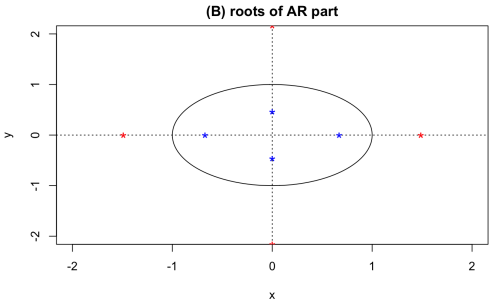
**Stationarity / Invertibility checking.**

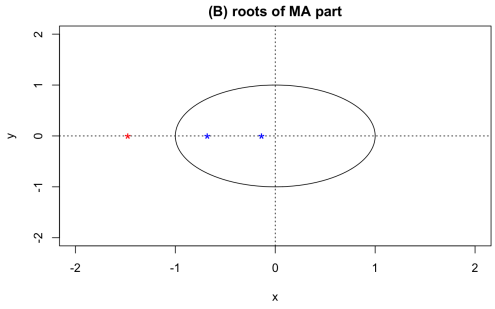
**Model A:** Coefficients: ar1 ar2 ar3 ar4 ma1 sar1 sma1

0 -0.1785 -0.0221 -0.0610 0.7687 -0.0428 -1.1895

This model has a sma1 coefficient of -1.1895, whose absolute value is greater than 1 and thus the model is not invertible. Due to this, we will not proceed with Model A.

**Model B:**



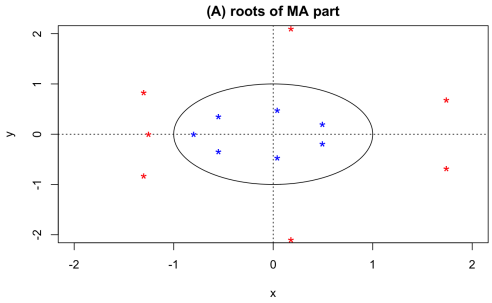




Coefficients: ar1 ar2 ar3 ar4 ma1 ma2 sar1 sma1

0 -0.2364 0 -0.0965 0.8150 0.0924 -0.0461 -0.8412

The autoregressive and moving average coefficients of model B both lie outside of the unit circle (inverse of the blue points). This along with the fact that sar1 and sma1 absolute values are less than 1 suggests that this model is both stationary and invertible.

**Model C:**




Coefficients: ma1 ma2 ma3 ma4 ma5 ma6 ma7 sma1

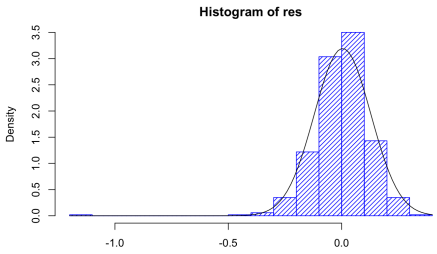
0.8155 -0.1519 -0.1852 0 0 0 0.0217 -0.8539

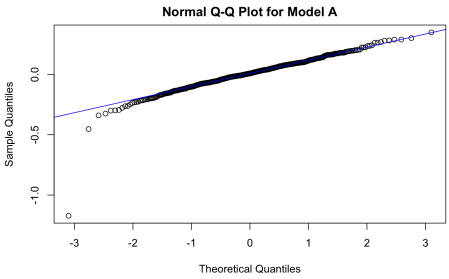
This is a pure moving average model, so it is automatically stationary. The roots of the non-seasonal moving average component lie outside of the unit circle, and the absolute value of the sma1 coefficient is less than 1, so the model is invertible.

We can proceed to diagnostic checking with models B and C.

**Model B:**

$$(1 + 0.2364_{(0.0608)}B^2 + 0.0965_{(0.0522)}B^4)(1 + 0.0461_{(0.0540)}B^{12})\nabla_1\nabla_{12}X_t =$$
$$(1 - 0.8150_{(0.0452)}B - 0.0924_{(0.0519)}B^2)(1 + 0.8412_{(0.03170)}B^{12})Z_t$$

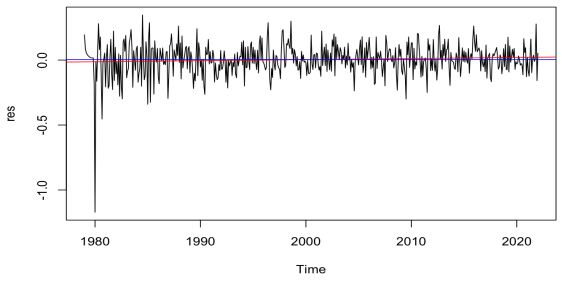



Shapiro-Wilk normality test
data: res
W = 0.91473, p-value < 2.2e-16

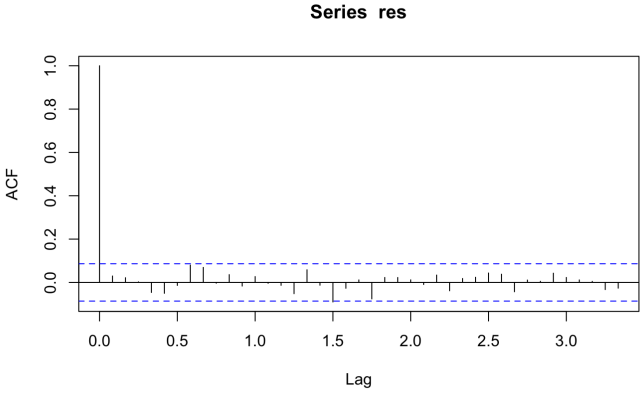
Box-Pierce test
data: res
X-squared = 10.208, df = 8, p-value = 0.2507

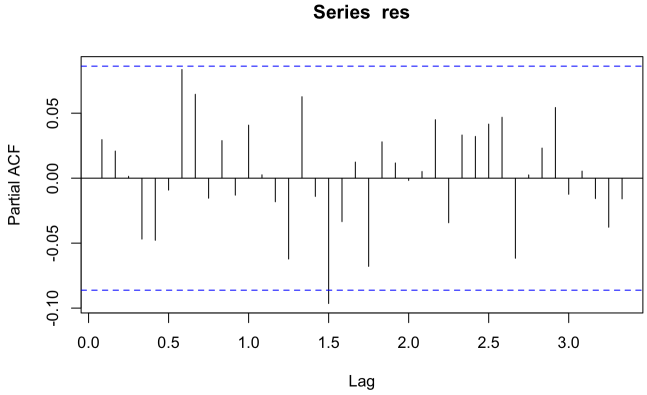
Box-Ljung test
data:
resX-squared = 10.383, df = 8, p-value = 0.2392




Box-Ljung test
data: res^2
X-squared = 10.515, df = 12, p-value = 0.5708

ar(x = res, aic = TRUE, order.max = NULL, method = c("yule-walker"))
Order selected 0 sigma^2 estimated as 0.01561



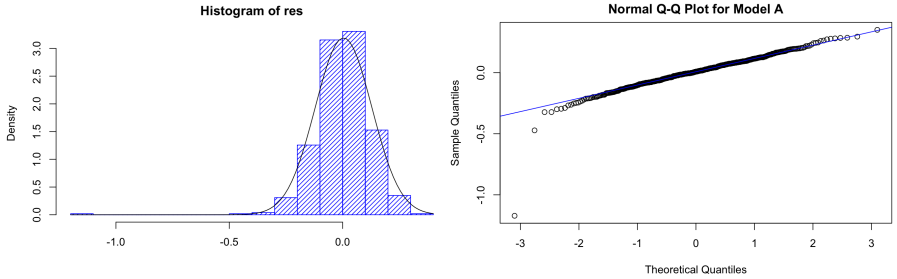




The residuals for the ACF and PACF are all within the confidence interval and can be taken as 0. The sample mean of the residuals is = 0.0033 which is almost 0. The Shapiro-Wilk test for normality has a p-value lower than .05, suggesting that the data is not normal. If we look at the histogram and Q-Q plot, however, the data does appear to be pretty normal so I will choose to accept it as normal. All of the other tests have p-values greater than .05 and as such will be accepted. The residuals were fitted to AR(0), which is white noise. Diagnostic checking is therefore passed by this model.

**Model C:**

$$\nabla_1\nabla_{12}X_t =$$

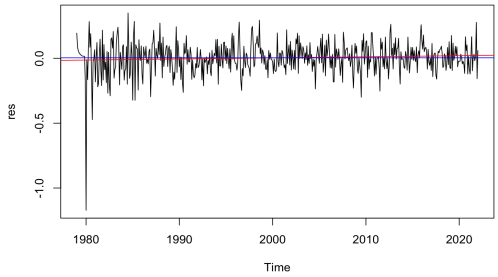
$$(1 - 0.8155_{(0.0448)}B + 0.1519_{(0.0593)}B^2 + 0.1852_{(0.0473)}B^3 - 0.0217_{(0.0290)}B^7)(1 + 0.8539_{(0.0259)}B^{12})Z_t$$




Shapiro-Wilk normality test
data: res
W = 0.91531, p-value < 2.2e-16

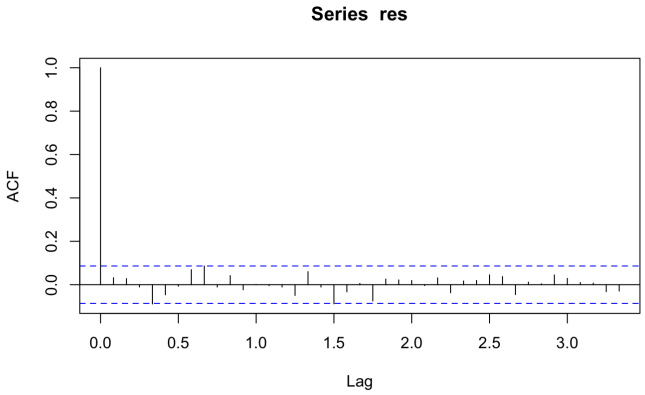
Box-Pierce test
data: res
X-squared = 13.79, df = 7, p-value = 0.05504

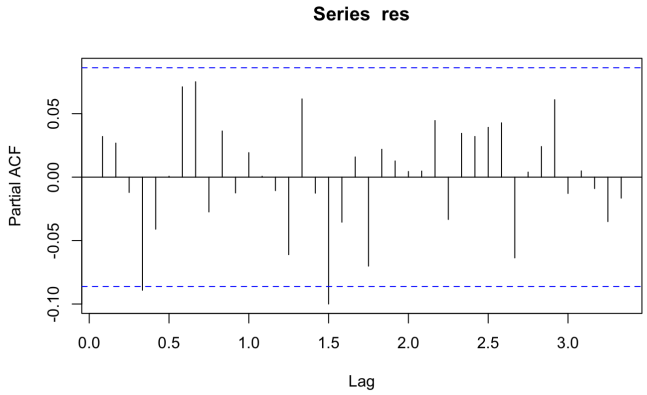
Box-Ljung test
data:
resX-squared = 14.01, df = 7, p-value = 0.05101




Box-Ljung test
data: res^2
X-squared = 11.73, df = 12, p-value = 0.4676

ar(x = res, aic = TRUE, order.max = NULL, method = c("yule-walker"))
Order selected 0 sigma^2 estimated as 0.01564



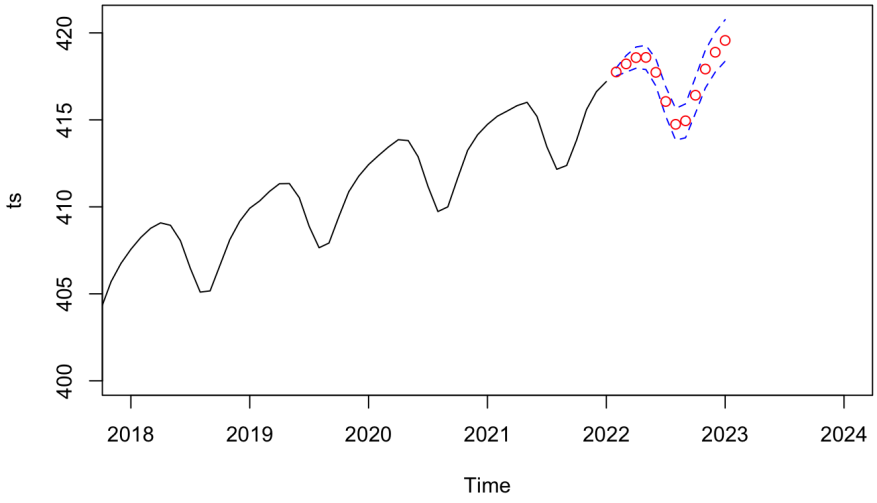



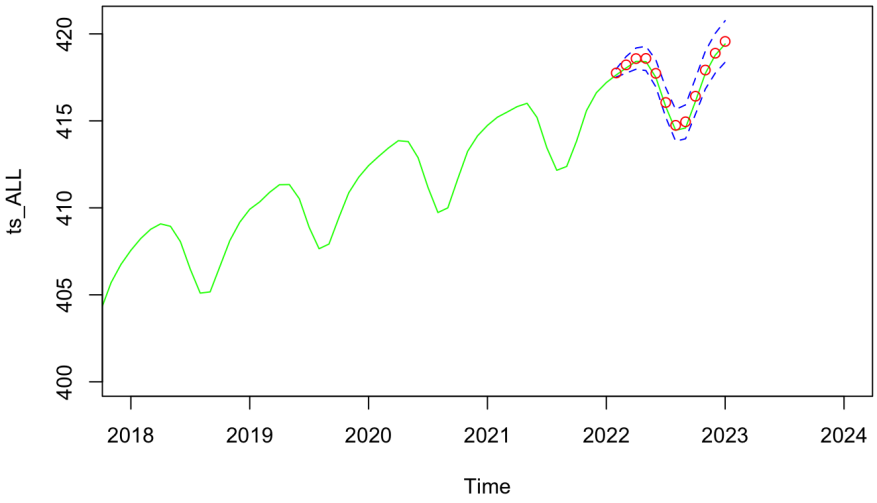

The residuals for the ACF and PACF are all within the confidence interval and can be taken as 0. The sample mean of the residuals is = 0.0041 which is almost 0. The Shapiro-Wilk test for normality has a p-value lower than .05, suggesting that the data is not normal. If we look at the histogram and Q-Q plot, however, the data does appear to be pretty normal so I will choose to accept it as normal. All of the other tests have p-values greater than .05 and as such will be accepted. The residuals were fitted to AR(0), which is white noise. Diagnostic checking is therefore passed by this model.

**FORECASTING:**

Since both model B and model C passed the diagnostic tests, I will simply choose model C for forecasting because it has the lower AICc score. Both models, however, would perform similarly for forecasting.

We will predict the next 12 months of CO2 data, and plot them with confidence intervals on the original data.




The testing data is clearly within the confidence intervals, so the prediction was successful!

**CONCLUSION:**

In conclusion, the goal of the project was to develop a forecasting model for carbon dioxide levels based on historical data and evaluate its performance. The project successfully achieved its goals by developing a SARIMA model for carbon dioxide forecasting, demonstrating a good fit to the data and accurate predictions for the testing period. The findings contribute to understanding the trends and patterns of carbon dioxide levels, which can aid in assessing the impact of industrialization and human activity on the environment.

Final model: **SARIMA(0,1,7)x(0,1,1)_12**

$$\nabla_1\nabla_{12}X_t = (1 - 0.8155_{(0.0448)}B + 0.1519_{(0.0593)}B^2 + 0.1852_{(0.0473)}B^3 - 0.0217_{(0.0290)}B^7)(1 + 0.8539_{(0.0259)}B^{12})Z_t$$

**REFERENCES:**

Professor Feldman's lectures, slides, and notes.

Global Monitoring Laboratory for the data set.