W271 Group Lab 2

Global CO_2 concentrations in 1997

Mahesh Arumugam, Melanie Herscher, Meng-Kang Kao, Jonathan Phan

March 17th, 2023

1 Introduction

Carbon dioxide (CO_2) is a gas that we are all familiar with. Humans, or all animals rather, breathe in oxygen, and breathe out carbon dioxide. We are also familiar with photosynthesis, a process carried out by plants which takes in carbon dioxide and produces oxygen. Something may not be obvious is that carbon dioxide is also the Earth's most important greenhouse gas (Lindsey 2022). Greenhouse gas absorbs heat radiating from Earth's surface and re-releases it in all directions. Without carbon dioxide, Earth itself will not be able to keep the surface temperature above freezing. Since 18th century, human activities have raised atmospheric CO_2 by 50% (NASA Global Climate Change, n.d.), and this drastic increase of CO_2 enhanced the earth's greenhouse effect, and caused global temperature to rise.

By studying the atmosphere's CO_2 trend, we may be able to predict the CO_2 level in the future, and further predict the future global temperature and its environmental impact.

The CO_2 data we are going to analyze, come from the data collected at Mauna Loa Observatory in Hawaii since 1958. Geochemist Charles David Keeling started to observe CO_2 level in 1950s (Keeling 1998). He started to measure CO_2 level in California, and observed strong diurnal behavior of CO_2 (Wikipedia contributors 2023b), with excess level at night due to respiration by plants and soils. Seasonal effect was also observed, especially in northern hemisphere, that CO_2 level reduces in spring and summer when vegetation grows, and increases in fall when plants start to die and release CO_2 back to atmosphere. The CO_2 data collected at Mauna Loa is the longest continuous record of atmospheric carbon dioxide in the world and is considered a reliable indicator of the global trend in the mid-level troposphere (Wikipedia contributors 2023a). In addition to its continuous record, the lack of human activities as well as lack of vegetation around the observatory makes the data from Mauna Loa more consistent and reliable.

2 CO_2 Data

In this analysis, we use the time series of 468 monthly observations between 1959 and 1997 of atmospheric CO_2 concentration levels collected at Mauna Loa Observatory (C. D. Keeling and T.P. Whorf, n.d.). Figure 1 shows the Keeling curve of the series. The histogram of CO_2 concentration indicates that we are observing increased values of CO_2 concentrations. Though the distribution of such observations is not big, it is skewed towards larger values. Furthermore, the series shows higher positive autocorrelations at small lags and slowly decrease as the lags increase. We also observe that

larger autocorrelations at seasonal lags (every 12 months). The partial autocorrelation is 1 for lag 1 indicating that the value of CO_2 concentration is influenced by the value in the previous month.

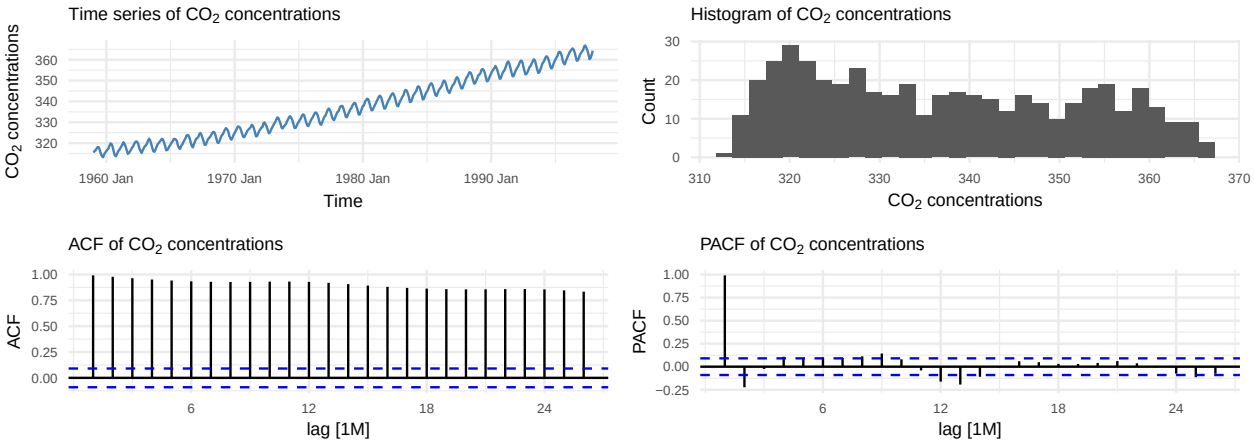

Figure 1: EDA for atmospheric concentration of CO_2. The ACF and PACF indicate a trend and a seasonal variation.

We use STL to decompose the time series into trend, seasonal and irregular components (cf. Figure 2). As expected, the annual averages trend shows CO_2 concentration increase over the years. The seasonal patterns shows a periodic (every 12 months) rise and fall of CO_2 concentration, where it slowly increases at the start of the year and slowly decreases towards the end of the year, And, the irregular component appears to have mean 0 and an almost constant variance. We did not perform a multiplicative decomposition as the variation in the data is mostly constant. In addition, we compute the annual average growth rate $x_t = \nabla log(y_t)$, where y_t is the annual average CO_2 concetration. Figure 2 shows that the growth rate appears to be a stable process. We note that CO_2 concentration tend to be mostly above the average growth rate (of 0.04%) since 1980.

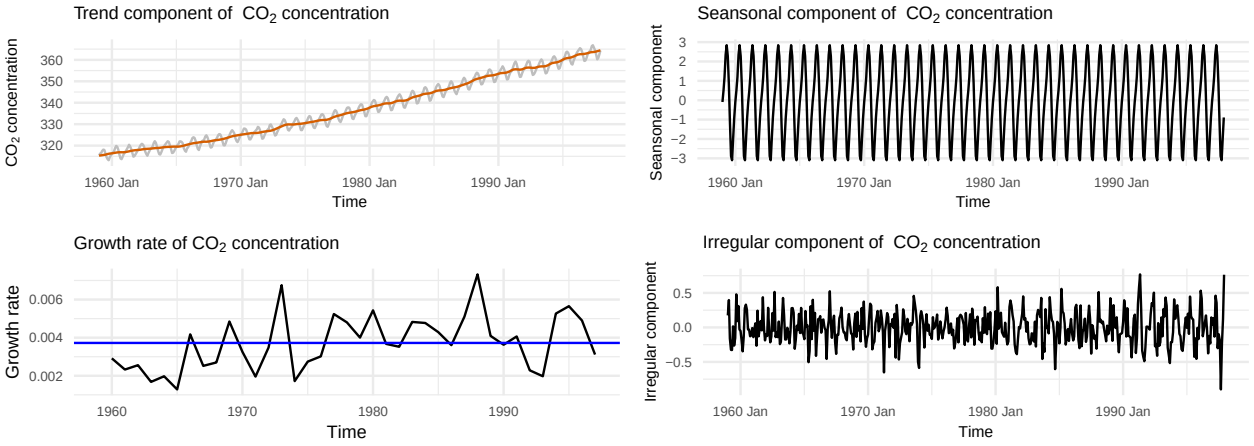

Figure 2: Components of the time series. The irregular component appears to be stationary with mean 0 and a constant variance. And, the average growth rate is close to 0.04%.

3 Linear Time Trend Model

In this section, we fit linear time trend models to the CO_2 dataset. From Section 2, we know that the variation in the data is mostly constant. STL additive decomposition also shows that series can

be decomposed into components, with the irregular component having a mean of 0 and a constant variance. Based on these observations, we believe we do not need any transformation for modeling.

Linear Model: We fit a linear model $CO_2 \sim \beta_0 + \beta_1 t + \epsilon_t$, where t is the time component and ϵ_t is the error/residual. Figure 3 shows that the model fits a straight line to the dataset that closely matches the trend. The RMSE value of the fitted data is 2.612. Residual series of the fitted model do not appear to have a mean of 0. The residual ACF plot indicates a seasonal pattern in the correlations with the lagged values. The residual PACF plot shows significance at many lags indicating the fit is not good.

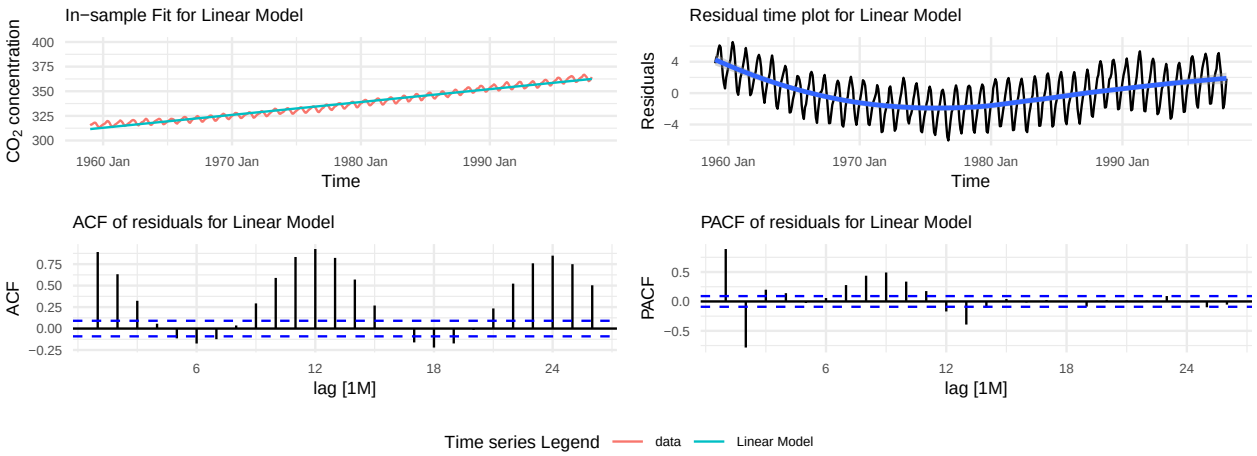

Figure 3: Linear time trend model. The residuals are not stationary.

Quadratic Model: We fit a quadratic model $CO_2 \sim \beta_0 + \beta_1 t + \beta_2 t^2 + \epsilon_t$. Figure 4 shows the fit and the residuals of the model. The RMSE value of the fitted data is 2.175. The residuals appear to have a mean of 0 and a constant variance. As we observed in the linear time trend model, we observe autocorrelations left in the residuals that show a seasonal pattern and PACF shows significance at many lags. While quadratic time trend model improved the characteristics of the residuals, both models do not fit the data well as indicated by ACF and PACF plots.

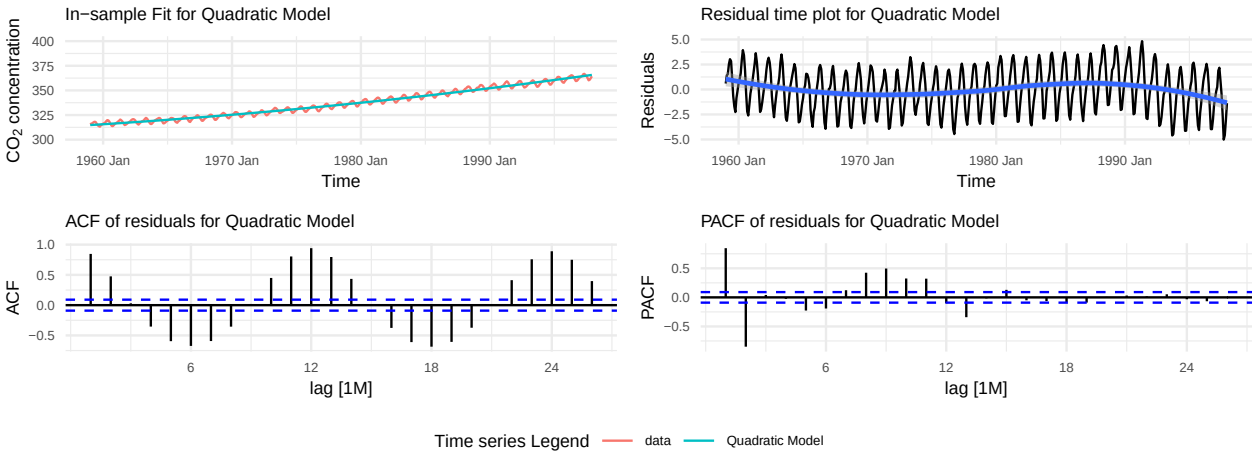

Figure 4: Quadratic time trend model. The residuals are not stationary.

Polynomial Model: We fit a polynomial model $CO_2 \sim \beta_0 + \sum_{i=1}^{d} \beta_i t^i + \texttt{season}() + \epsilon_t$, where

$d \geq 1$ is the order of the polynomial and `season()` indicates seasonal dummy variables[1]. We use Akaike information criterion (AIC) to select the best degree $d, 1 \leq d \leq 10$ that gives the smallest AIC value $= -621.6144$. We select d=3 as the best polynomial model; though it does not provide the smallest AIC value, the AIC values remain closer (to -621.6144) for $d > 3$.

Figure 5 shows the in-sample fit and the characteristics of the residuals of the polynomial time trend model of order=3 with seasonal dummy variables. The RMSE value of the fitted data is 0.497. The residual series of the fitted model shows a mean of 0 and constant variance. The ACF plot shows a slowly decreasing autocorrelations, with some seasonal pattern. And, the PACF is significant at lag 1. While this model provides a better fit than the linear time trend and quadratic time trend models, it still does not capture all the patterns in the dataset.

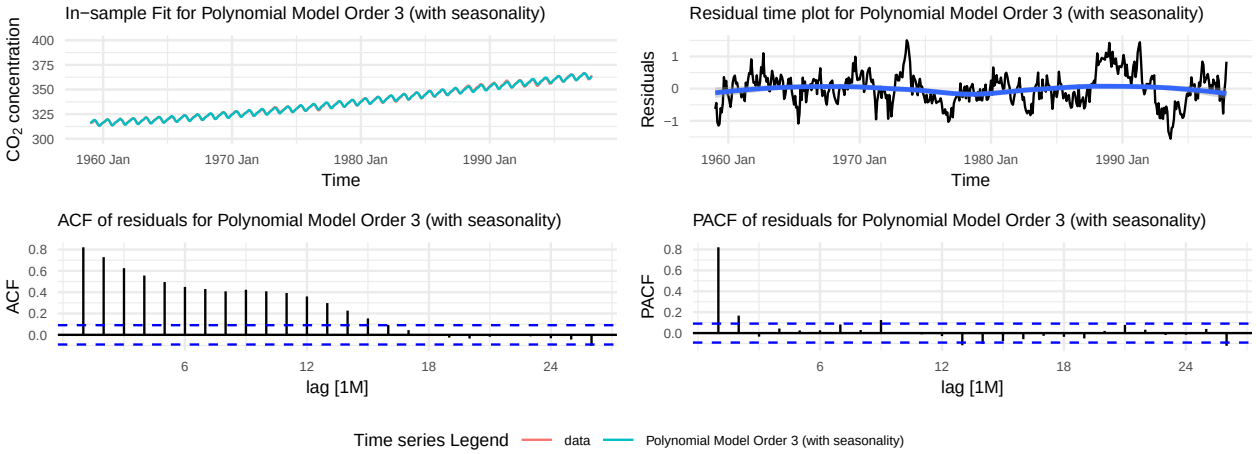

Figure 5: Polynomial time trend model of order=3 with seasonality. The residuals are not stationary.

Figure 6 shows the forecasted CO_2 concentration up to year 2020. It estimates that atomospheric of CO_2 concentration at the end of 2020 would be 382.662 ppm with 95% prediction interval [379.67, 385.655] ppm.

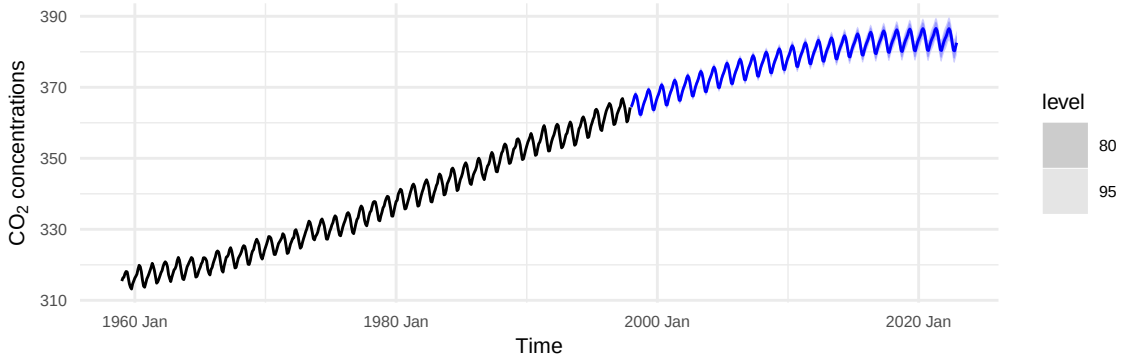

Figure 6: Forecast using polynomial time trend model of order=3 with seasonality. The estimated CO_2 concentration at the end of 2020 will be 382.66 ppm.

[1]We refer the reader to `notebooks/1997_model_selection.pdf` for polynomial model selection process.

4 ARIMA Time Series Model

In this section, we fit ARIMA time series model to the dataset[2]. Based on the discussion in Section 2, we know that differencing is required to make the time series stationary. Based on our model selection criteria[3], ARIMA $(0,1,3)(1,1,2)_{12}$ (without drift) is the best model that provides the lowest AIC ($= 179.41$), where the seasonal lag $s = 12$ months.

Figure 7 shows the residuals of the ARIMA $(0,1,3)(1,1,2)_{12}$ model. The RMSE value of the fitted data is 0.286 (which is the smallest amongst all the models we evaluated). The residual series shows a mean of 0 and constant variance. In addition, the residuals do not show significant autocorrelations at any lag. And, Ljung-Box test for stationarity of the residuals evaluates a p-values of 0.845 at lag 1 and 0.731 at lag 10. For a significance level of 0.05, Ljung-Box test fails to reject the null hypothesis of randomly distributed residuals.

Figure 8 shows the forecasted CO_2 concentration up to year 2022. The forecast estimates that atomospheric of CO_2 concentration at the end of 2022 would be 402.233 ppm with 95% prediction interval [386.378, 418.088] ppm.

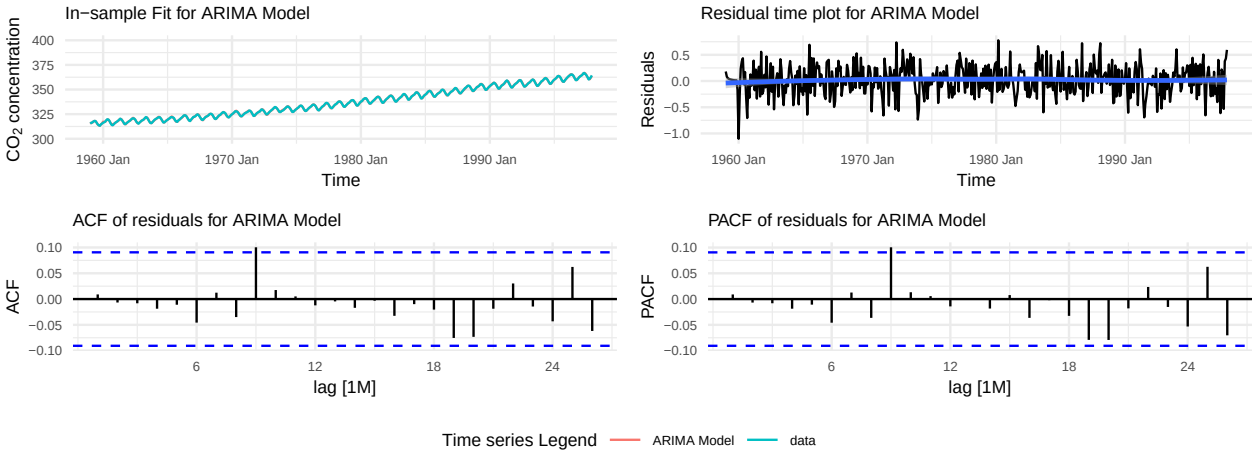

Figure 7: ARIMA $(0,1,3)(1,1,2)_{12}$ model. The residuals are stationary.

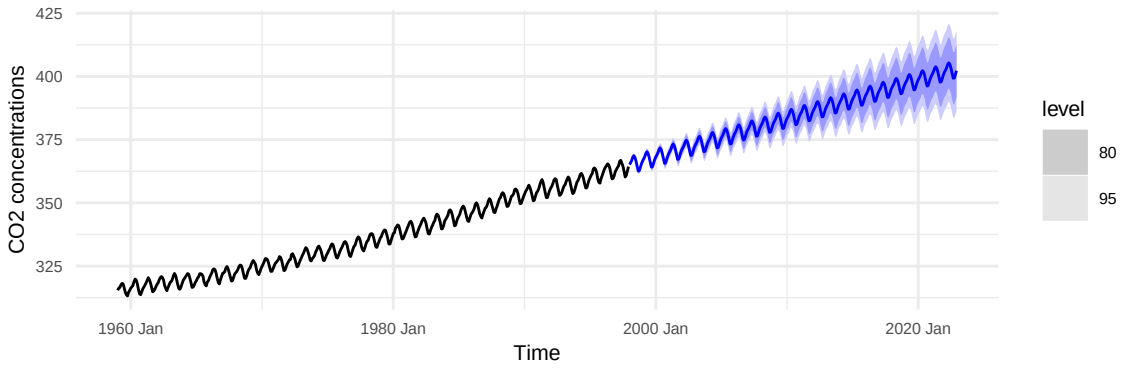

Figure 8: Forecast using ARIMA$(0,1,3)(1,1,2)_{12}$ model. The estimated CO_2 concentration at the end of 2022 will be 402.23 ppm.

[2] $CO_2 \sim (p,d,q)(P,D,Q)_s$ with and without drift, p auto-regressive (AR), d differencing, q moving-average (MA), P seasonal AR, D seasonal differencing, and Q seasonal MA terms with s seasonal lags.

[3] We refer the reader to `notebooks/1997_model_selection.pdf` for ARIMA model selection process.

5 Forecast Atomospheric CO_2 growth

Figure 9 shows the forecasted CO_2 concentration up to year 2100. CO_2 concentration of 420 ppm corresponds to reaching the halfway point towards doubling the concentration of pre-industrial revolution era (Stanch 2021). The ARIMA model projects that we are expected to hit 420 ppm on 2032 Apr. The 95% prediction interval indicates that the we might cross this limit for the first time on 2022 Apr and the last time on 2084 May. And, CO_2 concentration of 500 ppm corresponds to hitting a critical milestone that puts us on the path to global temperature rise of $3°C$ that could lead to extreme weather events (Jones 2017). The ARIMA model projects that we are expected to hit 500 ppm on 2084 May. The 95% prediction interval indicates that the we might cross this limit for the first time on 2055 Apr and the last time beyond year 2100. In summary, though it appears that we have time to get our acts together to reduce CO_2 emissions, unless we act, we will eventually hit the threshold and a point of no return.

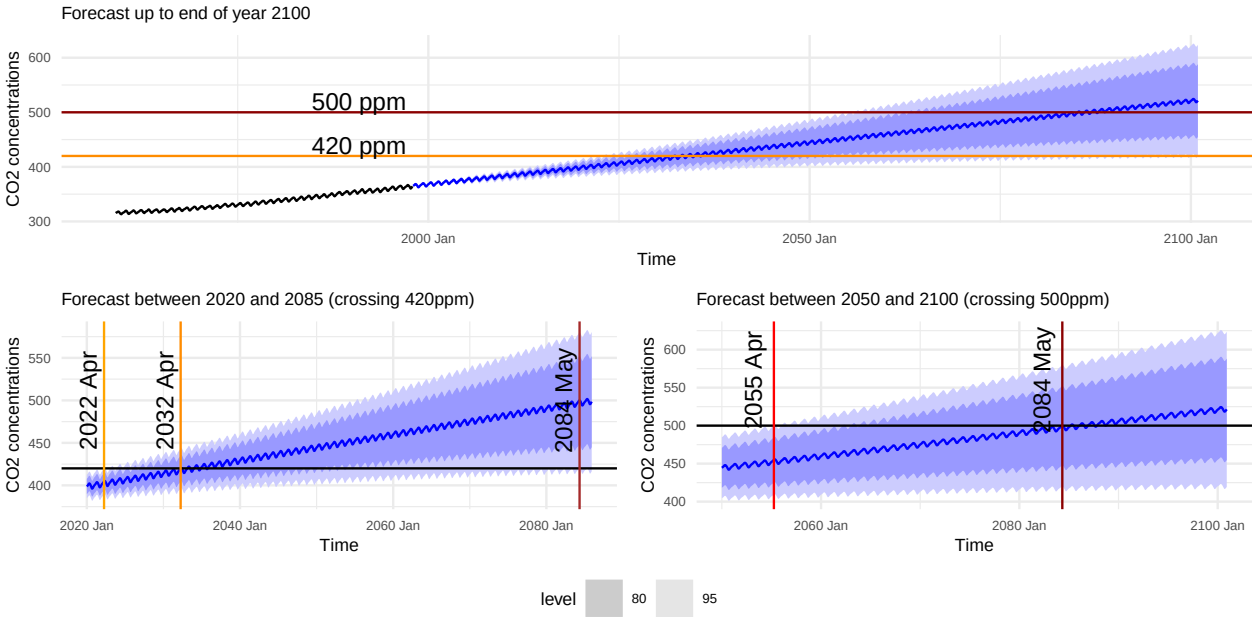

Figure 9: Forecast up to 2100 using ARIMA$(0,1,3)(1,1,2)_{12}$ model. We are projected to cross 420ppm between [2022, 2084] and 500ppm between [2055, >2100].

References

C. D. Keeling and T.P. Whorf. n.d. "Atomospheric CO2 Data." https://scrippsco2.ucsd.edu/data/atmospheric_co2/.

Jones, Nicola. 2017. "How the World Passed a Carbon Threshold and Why It Matters." https://e360.yale.edu/features/how-the-world-passed-a-carbon-threshold-400ppm-and-why-it-matters#:~:text=At%20the%20current%20rate%20of,global%20food%20supplies%2C%20cause%20disruptive.

Keeling, Charles D. 1998. "Rewards and Penalties of Monitoring the Earth." *Annual Review of Energy and the Environment* 23 (1): 25–82. https://doi.org/10.1146/annurev.energy.23.1.25.

Lindsey, Rebecca. 2022. "Climate Change: Atmospheric Carbon Dioxide." https://www.climate.gov/news-features/understanding-climate/climate-change-atmospheric-carbon-dioxide.

NASA Global Climate Change. n.d. "Carbon Dioxide Concentration." https://climate.nasa.gov/vital-signs/carbon-dioxide/.

Stanch, Kenny. 2021. "'We Have to Act': Atmospheric CO2 Passes 420 PPM for First Time Ever." https://www.commondreams.org/news/2021/04/06/we-have-act-atmospheric-co2-passes-420-ppm-first-time-ever.

Wikipedia contributors. 2023a. "Charles David Keeling." https://en.wikipedia.org/w/index.php?title=Charles_David_Keeling&oldid=1131804887.

———. 2023b. "Keeling Curve." https://en.wikipedia.org/w/index.php?title=Keeling_Curve&oldid=1136157103.

W271 Group Lab 2

Evaluating Time Series Techniques on Present Day Carbon-Dioxide Data

Mahesh Arumugam, Melanie Herscher, Meng-Kang Kao, Jonathan Phan

March 17th, 2023

1 Introduction

Over 25 years ago, our colleagues analyzed carbon dioxide (CO_2) level to create models that could forecast future CO_2 levels. However, since their original study, the carbon emissions for the world dramatically changed trajectory with the rise of globalization through the 2000s into the 2010s (Zabarenko 2007). As such, we utilized data from the Mauna Loa Observatory to evaluate the performance of the 1997 models and revise the models to conduct a new forecast.

2 Exploratory Data Analysis on 2022 Weekly Dataset

For our present day analysis (as of 2022), we will be pulling the atmospheric Carbon dioxide data from Global Monitoring Laboratory[1]. The data is aggregated in various time scale, and for our analysis, we will be using the weekly data.

We will be utilizing all data points, between 1974 to 2022, observed at Mauna Loa Observatory. Note that we only used data until November 2022, since starting December 2022, the observation was collected at a nearby site, approximately 21 miles north of of Mauna Loa[2].

Upon our preliminary data analysis, we observed 18 rows with negative CO_2 value. We will treat these rows as NA, and apply arima estimation to fill the gaps for missing value for our explanatory data analysis (EDA) purpose. For our model building, we will continue to use the original dataset in order to build the best models that solely rely on available data points.

The measurements and models created at Mauna Loa Observatory up until 1997 could not have possibly captured the slight quadratic trend the Keeling Curve takes on. The plot also shows little to no changes in magnitude of its variation over time. This is clear evidence that suggests using an additive decomposition is appropriate for this series.

Similar to what was seen in the 1997 report, we observe that there are high positive autocorrelations for small lags that slowly decrease as weekly lags increase, suggesting that there is a strong trend within this data that makes it non-stationary. Additionally, we do see the peak and valley behavior (peaking at 52 week, or one year, increments) that suggest that there's some level of seasonality within the weekly data as well. This makes sense as we saw similar yearly seasonality signs within

[1] https://gml.noaa.gov/ccgg/trends/data.html

[2] We refer the reader to `notebooks/co2_datapipe_2022.pdf` for more details.

the monthly version of this dataset and we would expect to see similar behavior in more minute aggregations of this data as well.

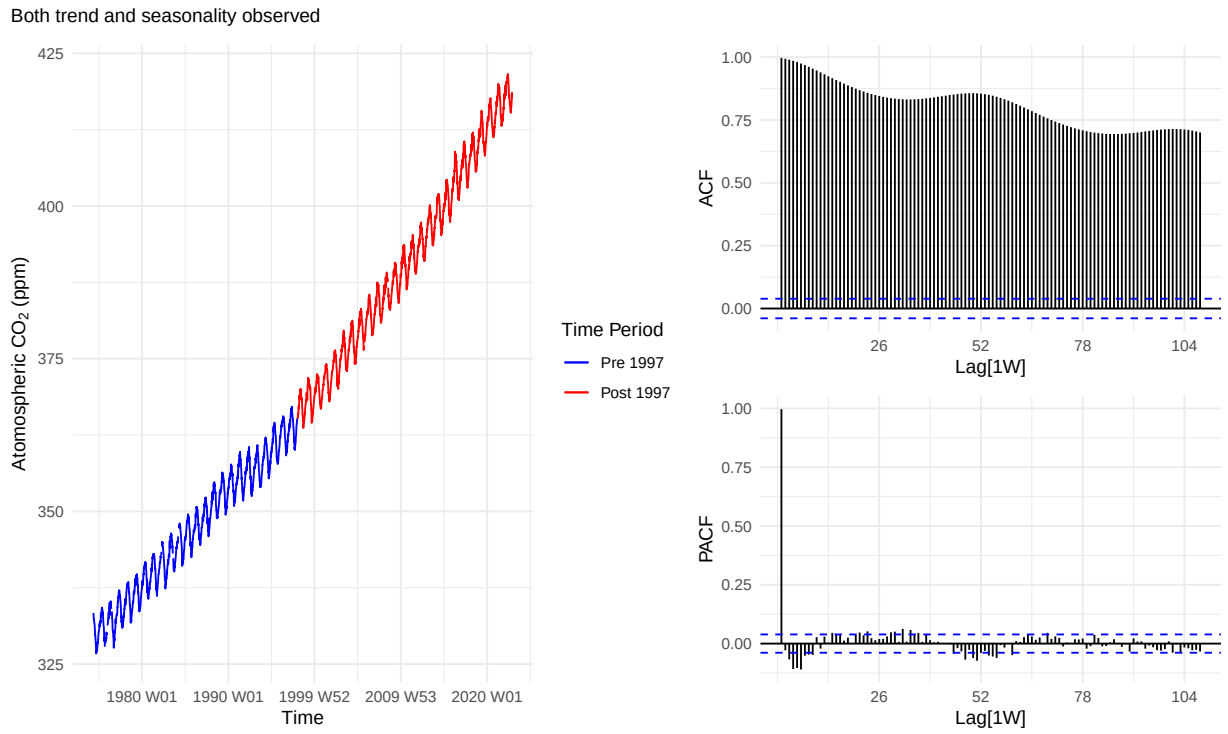

Figure 1: EDA for CO_2 level (Autoplot, ACF, and PACF)

The partial autocorrelations (PACF) suggests that there is a trend within the weekly data. We specifically see that there is a partial autocorrelation is 1 for lag of 1 week. For lags 3-5m, we also see a high level of significance for partial autocorrelation, which suggests that there maybe some amount of trend that is affected by CO_2 emissions 3-5 weeks previously. Similar to the monthly version of the data as well, we do see additional significant correlations further in larger lags, especially around the already identified seasonal lags.

3 Evaluating Previous Modeling from 1997 CO_2 Atmospheric Concentration Findings

The best linear model created using data of CO_2 levels from up to 1997 resulted in a cubic model. This model was concluded to be the "best" model out of its counterparts based on information criteria. Specifically, the cubic model had the lowest AIC value, making it the most desirable linear model. We similarly used AIC as a metric in determining the best ARIMA model.[3]

Both of these models perform reasonably well when forecasted until the year 2000. Their short-term forecasts indeed capture the upward trend of CO_2. However, it is clear that these models do not accurately forecast the realized post-1997 CO_2 values long term. The cubic model diverges further

[3]We refer the reader to `report/co2_1997.pdf` for clarity on the selection of these prior models

away from realizations than the ARIMA model, where its trend flattens out around 2015 and fails to capture the upward CO_2's upward trend.

In 1997, it was predicted that the first time that CO_2 would cross 420 ppm was around the year 2040. It is staggering to see how this level CO_2 is already present in our atmosphere since early 2022. To put into perspective, the predicted year of reaching 420 ppm was overshot by almost 20 years.

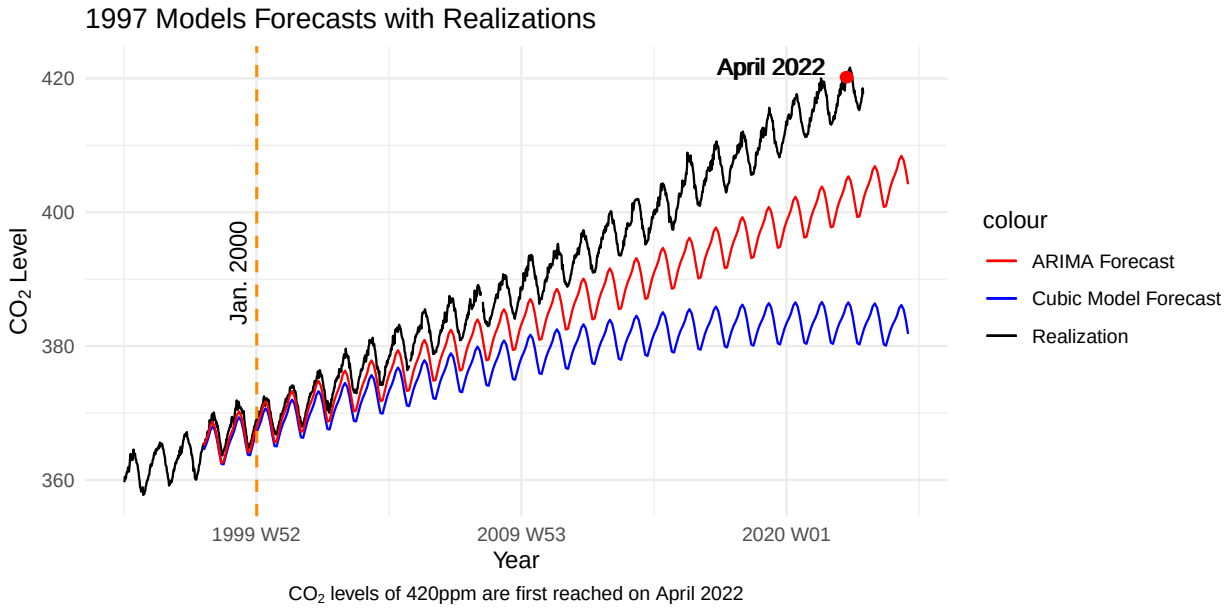

Figure 2: 1997 Model Forecasts against Realization

To better quantify the fit of the 1997 cubic and ARIMA modelling for more current the CO_2 emissions, we generated and reviewed the residuals between the forecasted values of the 1997 modeling to the actual results seen from Mauna Loa from 1998 to 2022. As we saw within the 1997 reporting, the residuals for before 1997 held a normal variance across time periods for the original dataset.However, the residuals from the forecasting quickly devolve to high differences progressively through time. By looking at the innovation residuals and seeing the heavy skew of residuals away from a normal distribution of the residuals that we have seen within the previous report, we can affirm that the trend that was posited within the previous report no longer reflects the trend that we see within current time periods.

The previous conclusion that the cubic model performs visually worse than the ARIMA model is proven when observing the RMSE of each model. Between the two, the RMSE of the ARIMA model is 8.63 and the cubic model is 17.5. Clearly, this formal test concludes that the 1997 ARIMA model is "the best" model out of other candidate 1997 models.

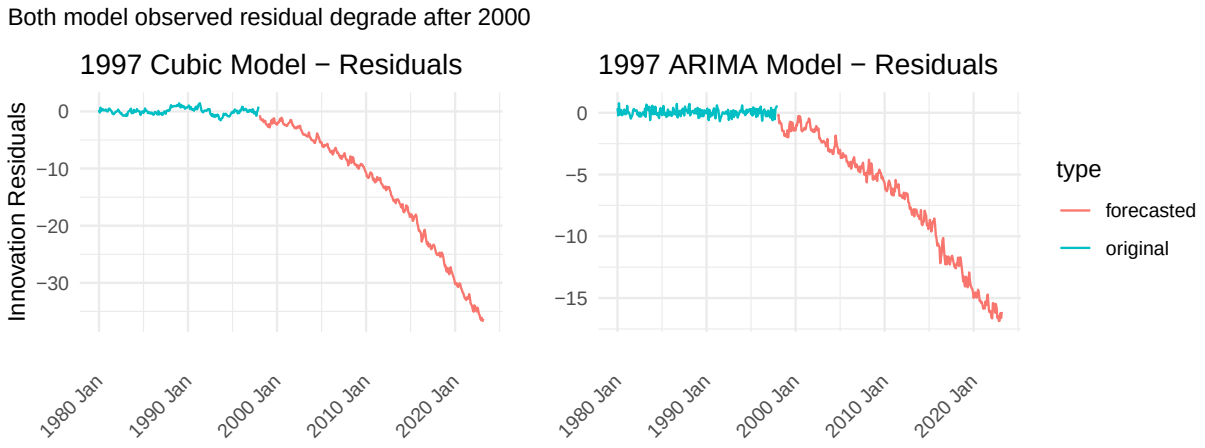

Figure 3: Evaluating 1997 Modeling on Atmospheric Concentration of CO_2

4 Best models on present data

In order to train the best model on present day data, we first create a new column as seasonally adjusted CO_2 (SA) to distinguish from non-seasonally adjusted CO_2 (NSA) by taking the seasonal differencing. Then we split the entire dataset to training and test set, with training data from 1974 to 2020, and test data from 2021 to November 2022 (end of data points).

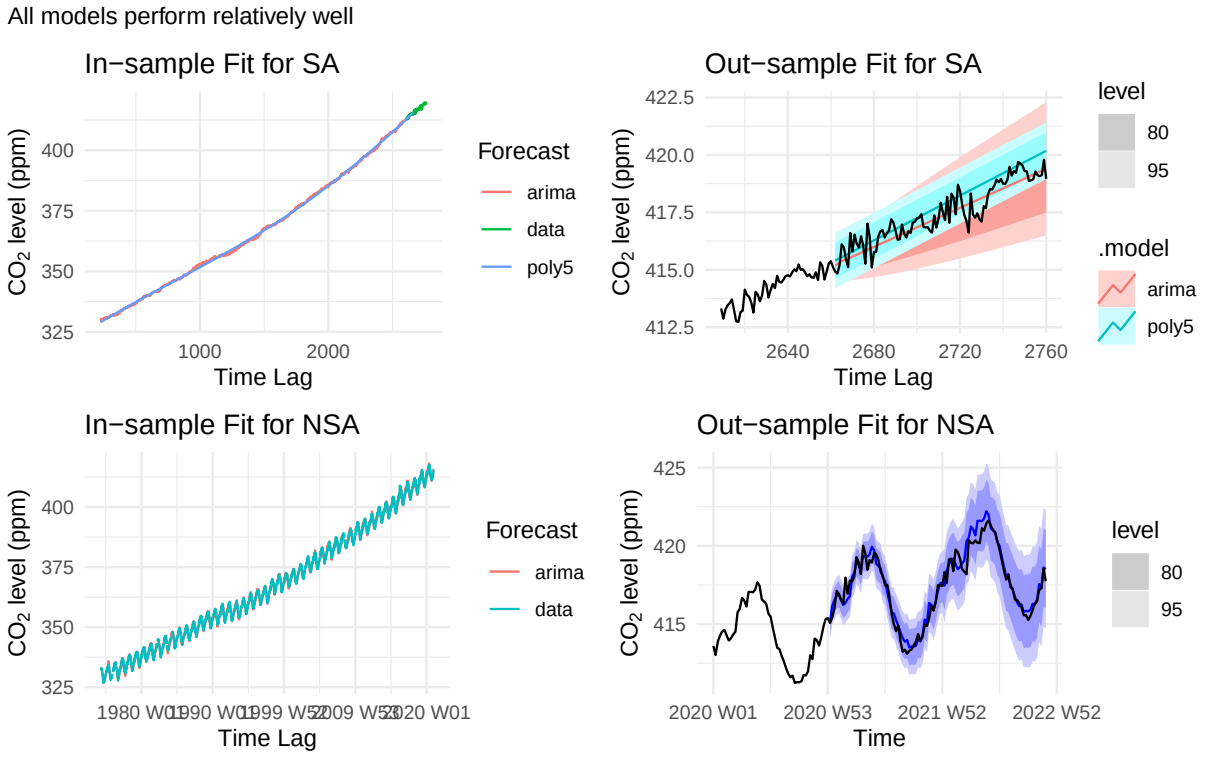

Figure 4: Atmospheric CO_2 level Model Performance

For our model selection, our method is to choose the model with lowest AIC score[4]. Our best model for NSA data, is Seasonal ARIMA $(5,1,1)(0,1,0)_{52}$ (AIC = 4049.96). Our best model for SA data is ARIMA $(2,1,1)_{52}$ with drift (AIC = 1936.79). For SA data, we further selected the best performing polynomial model (order = 5) for comparison.

Figure 4 shows different models' in-sample and out-sample performance, and they all perform pretty well visually, so we will proceed to compute their RMSE and compare the performance quantitatively.

Since we are comparing model performance between seasonally adjusted and non-seasonally adjusted data, it is not appropriate to simply compare the AIC score for each model. Instead, we select to use Root Mean Square Error (RMSE) to measure the model performance, on both training and test dataset. In the following table, we can see the ARIMA model against seasonally adjusted data (SA) produces the lowest RMSE, and therefore we believe this is the best model that can be used for the CO2 prediction.

Table 1: RMSE for seasonally and non-seasonally adjusted models

Dataset	ARIMA		Polynomial (Order=5)	
	In Sample Fit	Out Sample Fit	In Sample Fit	Out Sample Fit
SA	0.3649621	0.5040973	0.6108	0.6481
NSA	0.562408	0.5721576	NA	NA

5 How bad could it get?

Using our in-sample forecast, our model predicts that with 95% confidence interval, the expected atmospheric CO_2 level could reach 420ppm as early as 2021, and the expected CO_2 will reach 420 ppm in 2022. Compared with the test dataset, the first time earth reach atmospheric CO2 level of 420 ppm was exactly April 2022.

Using the same model, with the same 95% confidence interval, we predict the earth could reach 500 ppm CO_2 level as early as 2035, and the expected CO_2 value would reach 500 ppm in 2057. Note that the gap between expected value and 95% confidence interval grows significantly compared to the prediction at 420 ppm. This is expected due to the non-stationary nature of our time series data, that the variance will grow larger over time.

Now if we use the same model to predict the CO_2 level in January 2122, a hundred years from now, We estimate the mean value of the CO_2 level would be 649.01 ppm. However, with weekly time series data, the 95% confidence interval will be extremely large and even put the lower bound below 0, which is obviously unrealistic. This is due to the time frame we are forecasting, for 100 years, we will need to forecast over 5000 lags into the future, the variance will be too far for this model to be useful.

Alternatively, if we use monthly, or even year data to build the model and perform prediction, then we will have a much more reliable confidence interval. The following graph shows the same prediction out in 100 years, but with different data aggregation method, and their corresponding

[4]We refer the reader to `notebooks/2022_model_selection.pdf` for more details.

confidence interval. When comparing with the three aggregation method, we believe if we want to forecast out 100 years, the monthly data is the most suitable model.

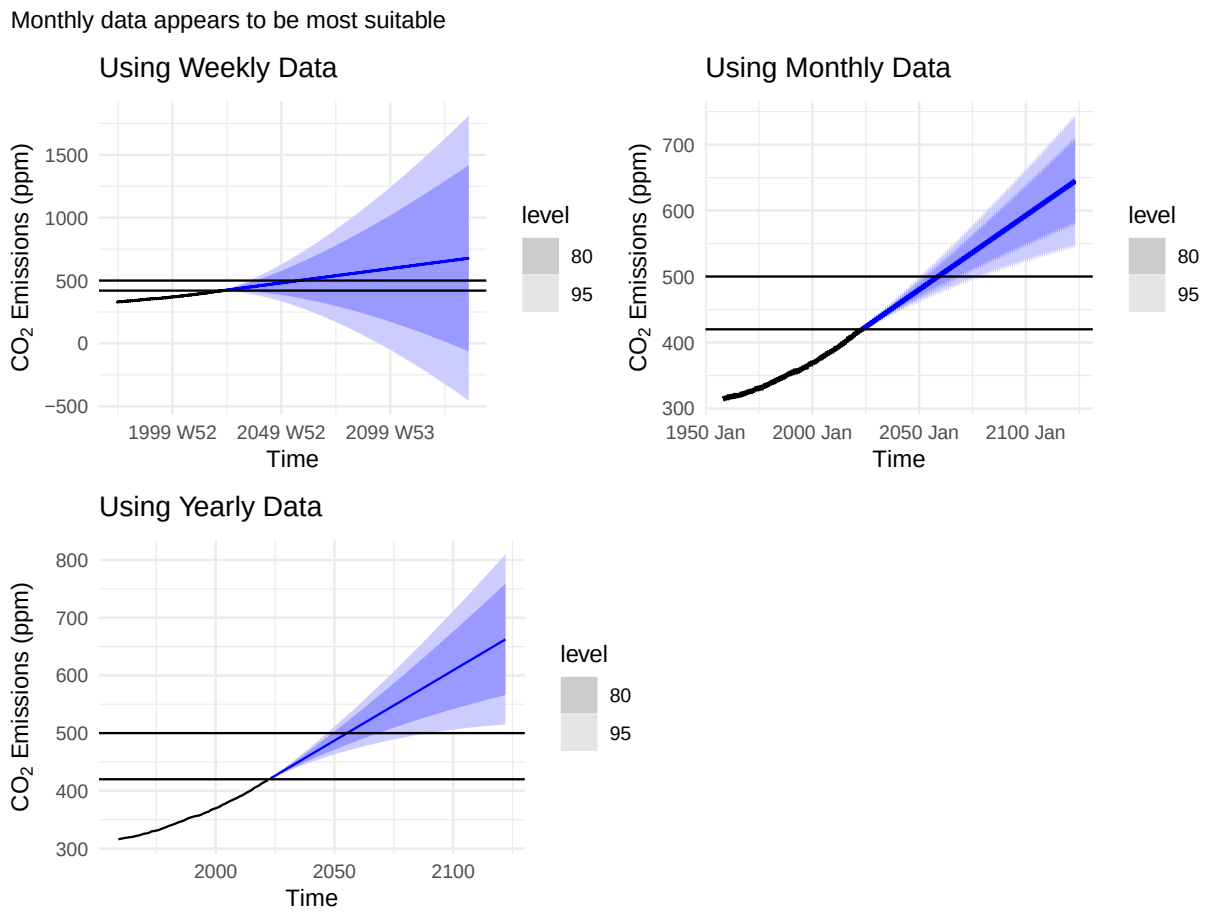

Figure 5: Forecast CO_2 Emissions Out 100 Years

This concludes our team's analysis for atmospheric CO_2 level. We look forward to hearing your feedback to further fine tune our model and utilize such information.

References

Zabarenko, Deborah. 2007. "World CO2 Emissions Speed up Since 2000." *Reuters*, May.