

Global CO_2 Emissions 1997 and Present

Evaluating and Re-Evaluating models for the Keeling Curve

Jinsoo Chung, Priya Reddy, Yuna Kim, Jocelyn Thai

Contents

1	Background	2
2	Models and Forecasts	4
2.1	Linear Models	4
2.2	Model Forecast	7
2.3	ARIMA Models	7
2.4	Forecasting Atmospheric CO2 Growth	8
3	Report from the Point of View of the Present	10
3.1	Introduction	10
3.2	Current Trends in Atmospheric Carbon	10
3.3	Compare linear and Arima model forecasts against realized CO2	10
3.4	Re-train 1997 models on present data	11
4	Conclusions	14

Understanding a changing climate, and what it means for the earth's inhabitants is of growing interest to the scientific and policy community. Although, at this point in 1997 it is not entirely clear what the consequences of this growing awareness will be, in this report we present likely outcomes under "business-as-usual" scenarios. In doing so, our hope, is to establish a series of possible futures, and, as evidence, technology, and policy develop over the coming decades, our goal is that we can then weigh the impacts that carbon-emission reduction efforts might take. In this report we seek to understand if atmospheric CO2 levels have increased since 1997 and if we anticipate that this trend will continue. To do this we seek to answer 2 main questions: 1. What is the best model we can use to model CO2 levels over time? 2. Using this model, what do we forecast atmospheric CO2 levels will be? And how confident are we in these predictions?

1 Background

1.0.1 Carbon Emissions

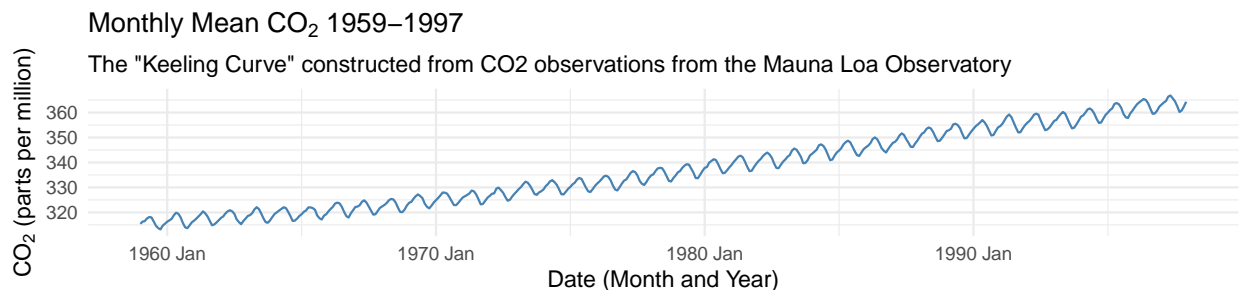
High CO₂ levels have been shown to negatively play into climate change, and the increase in atmospheric CO₂ has cascading effects from global warming to ocean acidification. Due to these effects, we may notice food supply chain disruptions, natural disasters of increasing intensity, and habitat disruption of all ecosystems and the animal species living there. In an economical sense, understanding these patterns is important so we can prepare for any shifts and uncertainties in global supply chains that might arise for these new climate patterns. In total understanding the patterns in atmospheric CO₂ concentrations is important both so that we can mitigate the existing effects of this increase and prevent continuing upwards trends in CO₂. In addition if our report findings are significant, the findings could persuade other governmental organizations to increase their CO₂ monitoring efforts and global preparedness efforts.

1.0.2 Measuring Atmospheric Carbon

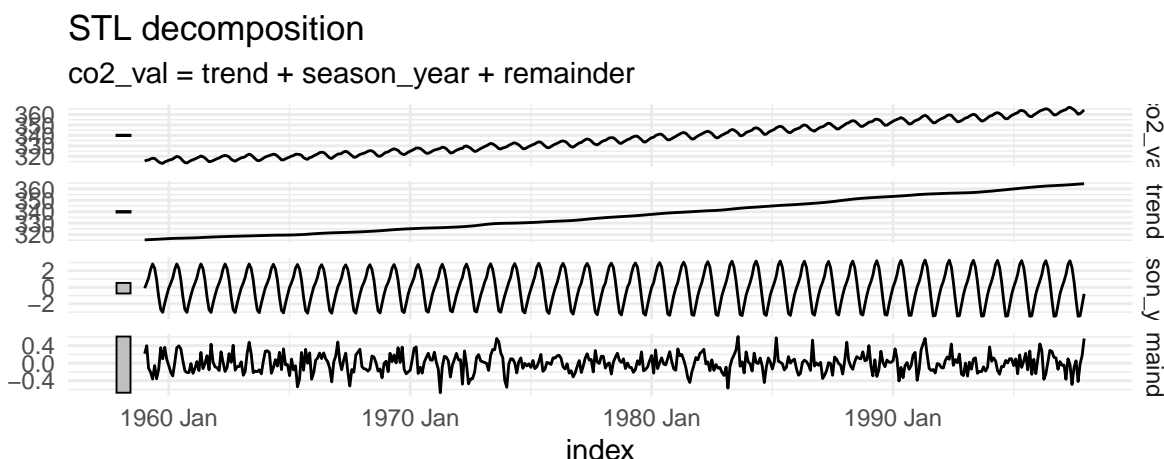
As a background on the data we are using it is important to note that this data measures the mean atmospheric CO₂ concentration and was collected at the Mauna Loa Observatory in Hawaii. Mauna Loa Observatory was chosen by Keeling to record the observations since it was far from other continents and had sparse vegetation – thereby mitigating some of the location influence on the data. The data was also normalized to remove any influence from local contamination. Carbon dioxide measurements at the Mauna Loa Observatory in Hawaii are made with a type of infrared spectrophotometer, now known as a nondispersive infrared sensor, that is calibrated using World Meteorological Organization standards [^ 1. Anal. Chem. 2010, 82, 19, 7865–7870. Publication Date: June 11, 2010. <https://doi.org/10.1021/ac1001492>].

1.0.3 Historical Trends in Atmospheric Carbon

To better understand the changes in Atmospheric Carbon we first must inspect our data. We can do so by looking at the “Keeling Curve” or the plot of the CO₂ levels over time. At first glance we can see in this curve both an overall upwards trend in CO₂ levels as the mean seems to be increasing over time. We can also see some sort of cyclic or seasonal pattern of atmospheric CO₂ fluctuations each year.



Looking at the data there are three areas of interest that we may want to look into further – the upwards trend, the seasonality, and the irregularities we observe in the data. To better visualize these we can decompose our time series data by either additive or multiplicative decomposition. If we observe multiplicative decomposition then perhaps a logarithmic transformation of our data would make sense. However, when we look at the Keeling curve above, we can see that the variance in the CO_2 values does not appear to change over time so we infer that additive decomposition is more appropriate for our data.



We can see in STL decomposition plot, that there is a clearly year-over-year upwards trend in the CO_2 levels in the data. We can also see the seasonality in this plot, and can infer that this is a yearly seasonality in CO_2 levels. It appears as though CO_2 levels peak at around May and then dip to a minimum in September and October. It is clear that our data is non-stationary and in future time series modeling exercises we will need to take steps to make it stationary.

To further explore our initial observations on seasonality we can look at our ACF, PACF, and Lag plots.

Fig 3A. ACF of Monthly CO_2

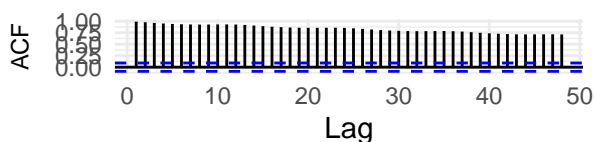


Fig 3B. PACF of Monthly CO_2

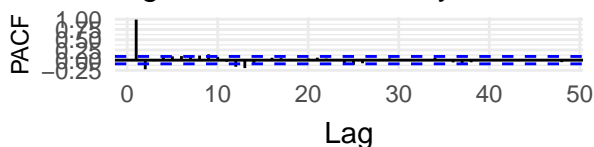
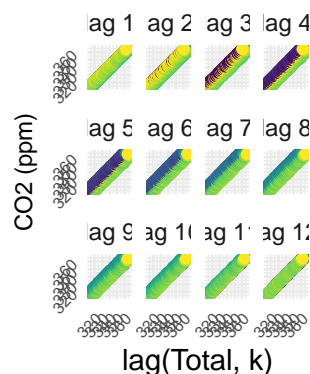


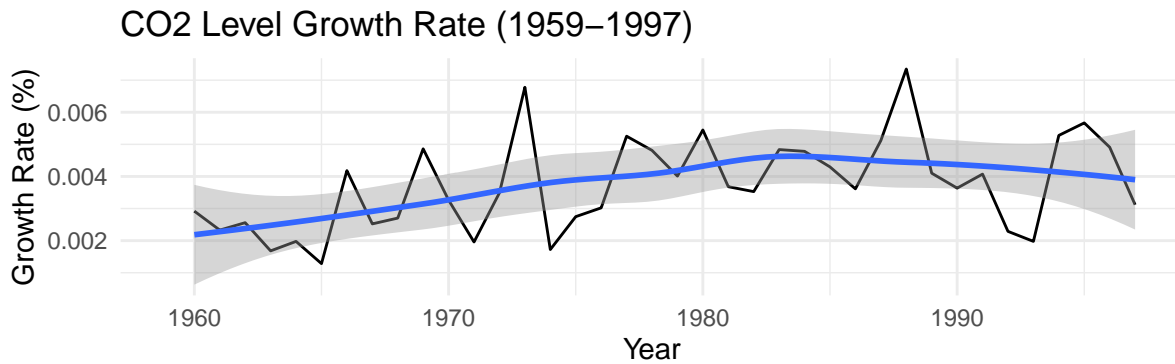
Fig 3C. CO_2 Lagged scatterplots



In Figure 3C, the lag plot, there is a very strong positive correlation across all lags, reflecting the seasonality of our data. This is also reflected in the ACF plot, Figure 3A, which decays gradually

indicating a trend in our data, and also has a subtle rise in correlation at around every 12th lag. This behavior in the ACF plot indicates a seasonal or cyclic pattern in our data, that the lags are also very significant also indicates that our data is non-stationary. These observations are in line with what we observed in our decomposition plots. The Figure 3B PACF plot cuts off to zero after lag 2 and stays that way. However, at lags 12 and 13, the values become significant again. This suggests that although the model exhibits AR model-like behavior, there's some deviances in the data.

In looking at the patterns in our data we are also interested in assessing whether the upwards trend we observed is accelerating – that is we want to look at the trend of the growth rate of CO2 levels. At a high level, knowing this could help researchers assess the efficacy of global efforts to curb CO2 emissions. Looking at Figure 4 we can see that there is some acceleration in increase in CO2 levels up until about the early 80's after which the growth rate in CO2 levels begins to drop (though overall CO2 levels continue to increase).



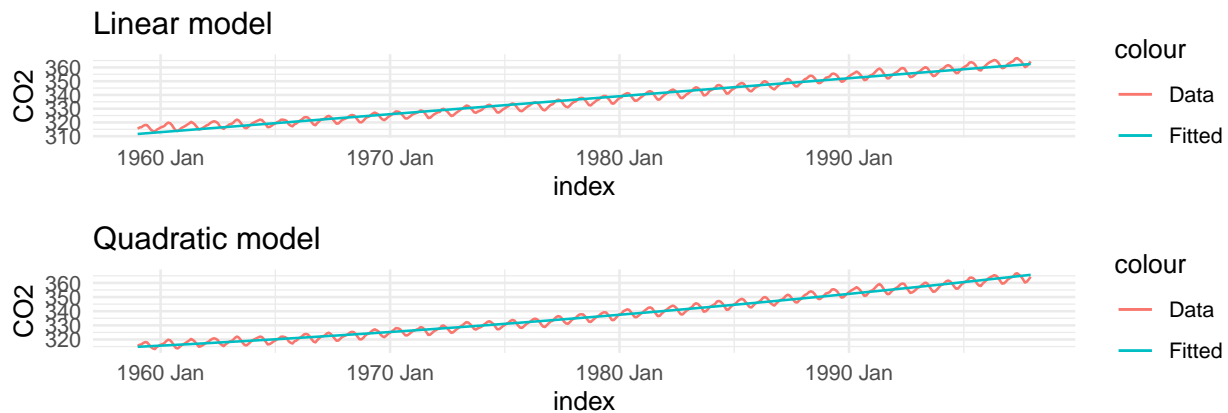
2 Models and Forecasts

In this section, we evaluate two classes of models for answering our questions – a linear time model and an ARIMA model to assess which time series model is most appropriate to use to model CO2 levels over time.

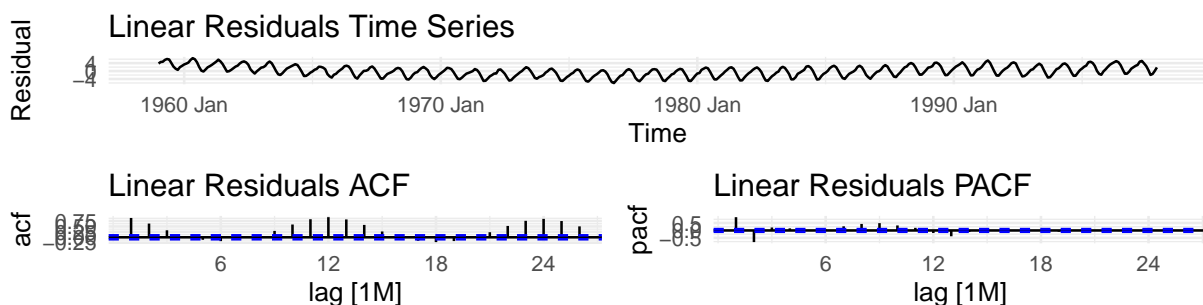
2.1 Linear Models

To fit linear time trend model to the `co2` series we will compare a regular time trend linear model to a quadratic time trend model. We will also fit a polynomial time trend model that incorporates seasonal dummy variables, and use this model to generate forecasts to the year 2020. Note that we will be evaluating using the unscaled CO2 and not the $\log(\text{CO}_2)$ value as we earlier found that an additive rather than multiplicative decomposition fits our data best.

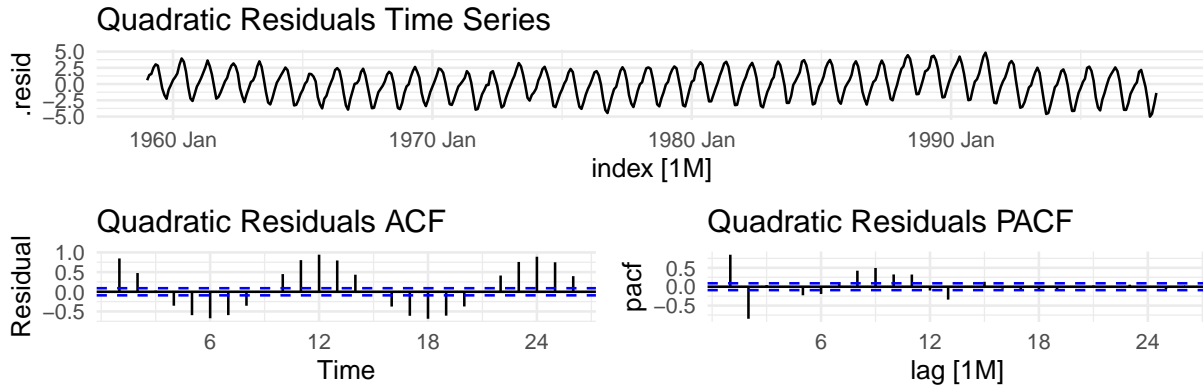
To begin our analysis, we fit a linear model of the form $\text{CO}_2 = \beta_0 + \beta_1 t$, a quadratic model of the form $\text{CO}_2 = \beta_0 + \beta_1 t + \beta_2 t^2$ to our data. We find that our linear model takes the form $\text{CO}_2 = 312 + 0.109t$ and our quadratic model takes the form $\text{CO}_2 = 315 + 0.067t + 0.0000886t^2$. For both models all terms are significant ($p < 0.001$). We can visualize these models below, against our data.



We can see that Quadratic model fits our data better as it is closer to the trend line (and mean) of our data each year. The linear model appears to underestimate the CO2 levels from 1959-1970 before overestimate the data from 1970-1982 before again underestimating the data through 1997, this appears to reflect the model being unable to capture the acceleration in CO2 growth we saw in the growth rate plo above. Most notably we can see that neither model can capture the seasonality in our data. We can further examine the fit of these models through their residuals.

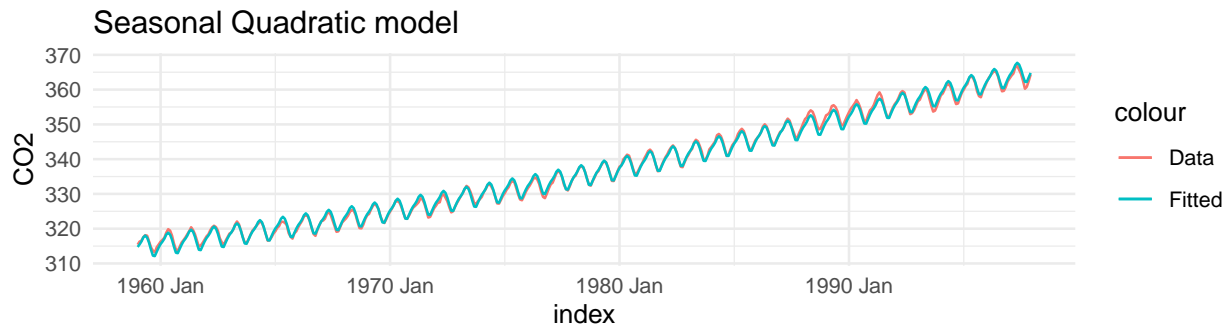


We can see that the linear residuals clearly are non-stationary as we can observe distinct seasonality in all three of the residual plots. In addition we can see in our residual plots that the seasonality appears to be additive over time rather than multiplicative. This is what we observed in the decomposition plots above and validates our decision to not apply a log transformation to our data.

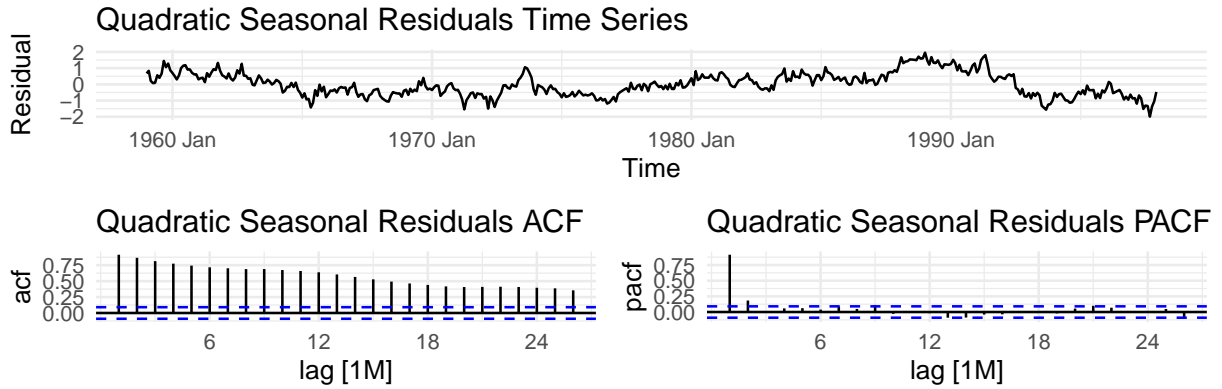


When we look at the Quadratic model residuals we can see that the data looks more stationary in that the mean varies less, but that we still are not capturing the seasonality in our data. We can see this clearly in all 3 residuals plots as the cyclic pattern in the residuals and also ACF and PACF values of the residuals.

As the linear and quadratic model are unable to account for any of the seasonality in our data, and this is reflected in their residuals, we will also fit a Polynomial (Quadratic) model with seasonal dummy variables. This model is plotted below, and we can see clearly how much better it captures our data.

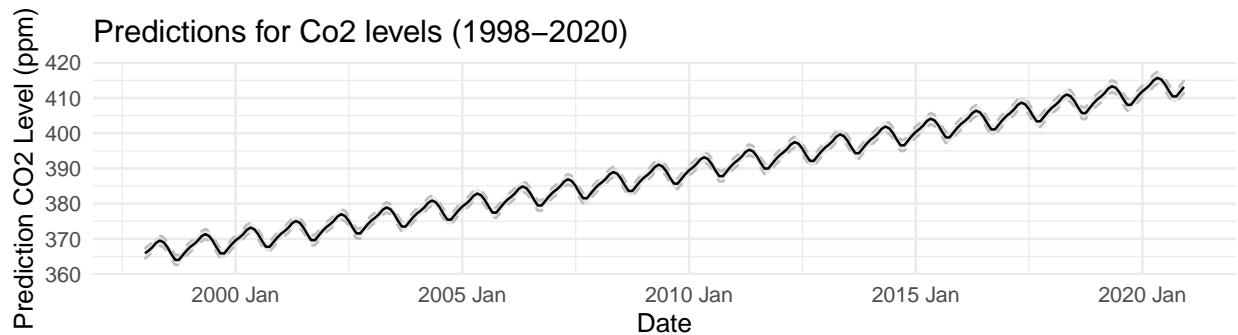


When we look at the residuals of this model we can see that our seasonal dummy variable has managed to capture quite a bit of the seasonality in the data as the residuals no longer oscillate so regularly. But in our ACF plot we can see that there is still a slight oscillation in the residual ACF, and that one of the PACF lags is still significant. This indicates that our model is still unable to capture all of the variation in the data. Perhaps fitting a higher order polynomial would help capture some of this effect, but we have elected not to in order to avoid overfitting.



2.2 Model Forecast

Seeing how well the Quadratic Seasonal Model performed in comparison to our other two models we can now generate some forecasts with this model. For instance below we have forecasted the CO₂ levels until 2020 (23 years from now) and have included the 95% confidence interval – we can see that the forecast continues both the seasonal pattern and upwards trend we observed in the original Keeling Curve. Our Quadratic Seasonal model predicts that by 2020 the atmospheric CO₂ will be at about 415 ppm.



2.3 ARIMA Models

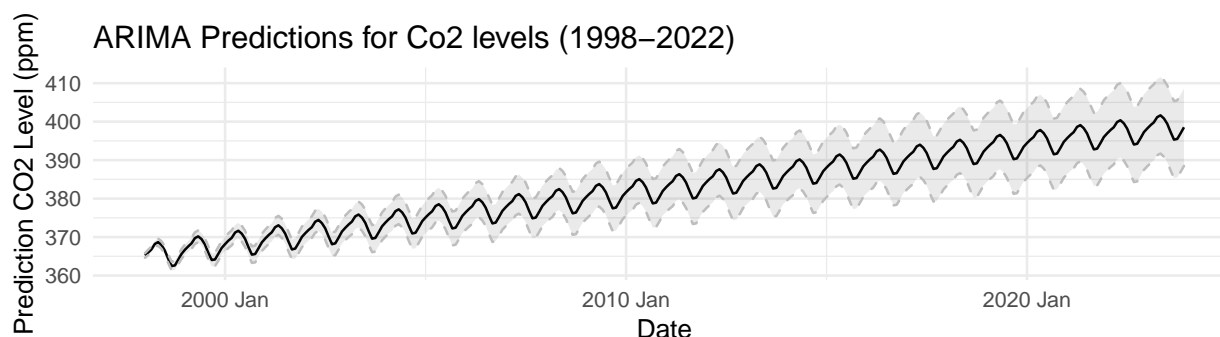
Since we earlier noted that our quadratic seasonal model, despite performing well, was still unable to encompass all of the patterns we observed in our data, we will also generate a few ARIMA models and compare them to our linear model. To select our ARIMA model we use the AIC; the AIC is a model evaluation metric that penalizes models with too many parameters, by using the AIC we hope to choose the model that best fits our data without overfitting. We know that there is a seasonal component to our data, so we will also select for seasonal AR and MA values and check if seasonal differencing will make the data stationary. We use the `ARIMA()` function to select our data. This method will determine the optimal p, d, q, P, D, Q values as well as the optimal levels of differencing for us to have stationary data and will then select a model based on the metric we have chosen as our criteria (AIC). We can interpret the resulting ARIMA model in terms of the number of (seasonal) AR terms, the level of differencing to achieve stationary data. We are

including seasonality in our model because we can see in the decomposition as well as the time series plots that there is a strong seasonal component in the CO2 levels. If adding a seasonal component improves the AIC score then such a model will be chosen.

With the AIC as our selection criteria we have estimated the model to be an $ARIMA(1,0,1)(4,1,0)[12]$ with drift. This model has an AIC of 218, since this was the model that was chosen, this must be the smallest AIC value in the models that we have compared. This $ARIMA(1,0,1)(4,1,0)[12]$ model can be interpreted by saying that there is 1 AR term, 1 MA term, 4 seasonal AR terms and that the data had to be seasonally differenced once to be made stationary. The [12] at the end of the model indicates that there is a yearly seasonality to our monthly data.

To be thorough we also repeated the above selection process but instead using BIC and AICC as our selection metrics. In general these two metrics are stricter than the AIC in penalizing additional parameters in our model, so it is possible that this selection process will result in a different model. When we repeated the above process using the BIC or the AICC as our criteria, we found that we were still selecting the exact same ARIMA model, so we will move forwards with this model.

Since we have decided on an ARIMA model we can also generate some initial forecasts for the model. In the plot below we have used our ARIMA model to predict CO2 levels until the end of 2022. We can see that the model maintains the same seasonal and upwards trend that we observed in the Keeling curve. Interestingly we can also see that the 95% confidence intervals are increasing the further out we forecast. We can also see that the ARIMA model appears to return lower estimates for the CO2 levels than our quadratic seasonal model – the ARIMA model forecasts that the CO2 levels will not have reached 410ppm by 2022 even though our quadratic seasonal model forecasted CO2 levels reaching 415ppm by 2020.



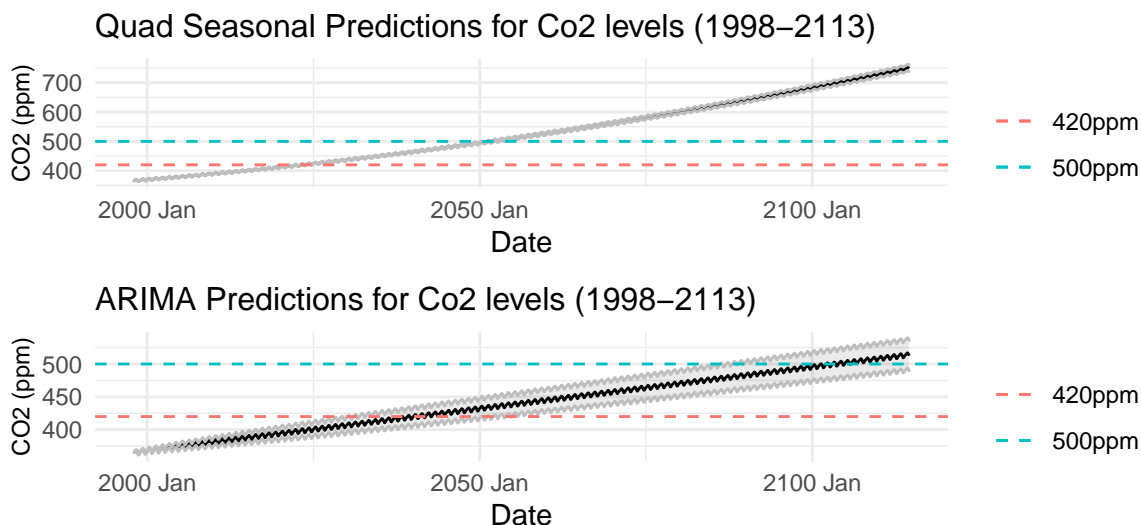
2.4 Forecasting Atmospheric CO2 Growth

Now that we have a few different models that do a decent job at modeling our CO2 data, we want to use these models to generate predictions for us.

2.4.1 When will we reach 420 and 500 ppm CO2?

First we want to see when these models forecast that the atmospheric CO2 levels will reach 420ppm and 500ppm. Since we know there is an overall upwards trend as well as a seasonal component to the CO2 levels, when we generate these predictions we will actually want to look at both the first and the last times that the model predicts the atmospheric CO2 levels to be at these values. We

should note that these are approximations as the true predictions from our model are not integers and instead are the mean of a prediction distribution. As a result we had to round these prediction values to find where the predicted CO2 will reach these levels.



Our ARIMA model predicts that the first time that the atmospheric CO2 levels will be 420ppm is April 2038 (between March and April 2039), and the last time is October 2042 (between September 2042 and October 2042). We can also see that the first time that the atmospheric CO2 levels are predicted to be 500ppm is March 2101 (between March 2101 and April 2101), and the last time is October 2105 (between October 2105 and November 2105). Our Quadratic Seasonal model predicts that the first time that the atmospheric CO2 levels will be 420ppm is April 2022 (between April 2022 and May 2022), and the last time is November 2023 (between November 2023 and December 2023). The first time that the Quadratic Seasonal model predicts atmospheric CO2 levels to be 500ppm is March 2051 (between March 2051 and April 2051), and the last time is October 2052 (between October 2052 and November 2052).

2.4.2 What do we predict CO2 levels will be in 2100 (103 years from now)?

We also want to look at the predictions for our models for the more distant future. When looking at the above chart we can see visually what the CO2 levels will be in 2100 along with the confidence intervals for these predictions, but for a quick understanding we can look at the mean CO2 value for 2100. The mean CO2 value for 2100 as predicted by the ARIMA model is 636.94 ppm and by the Quadratic Seasonal model 685.32. We can contrast this with the mean CO2 level for 1997 which is 363.82 ppm.

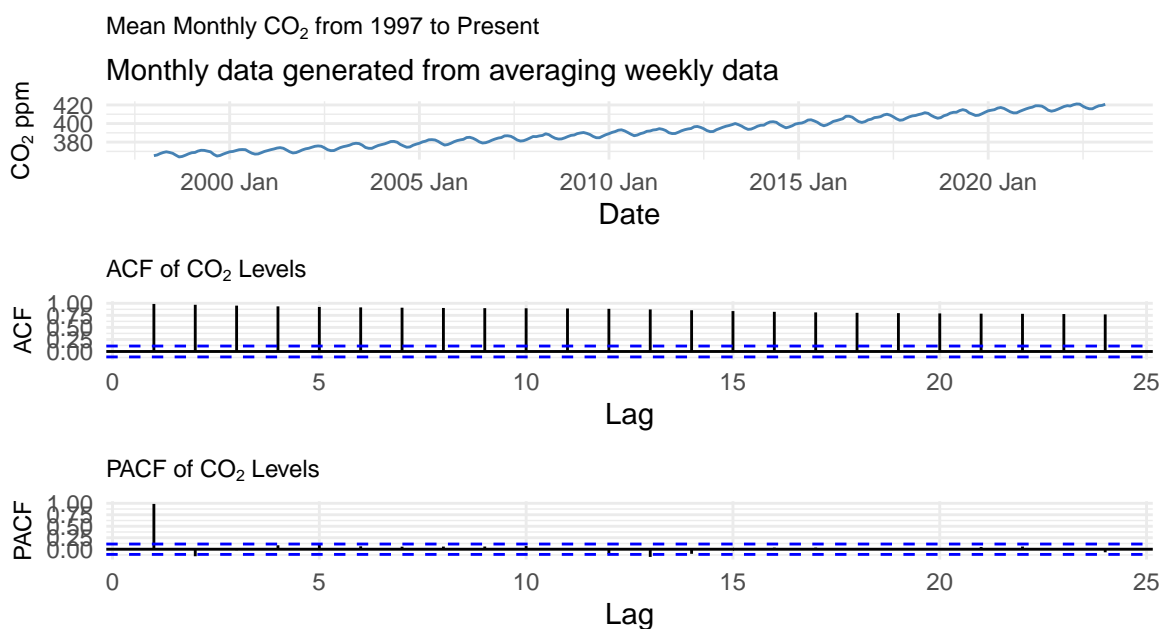
3 Report from the Point of View of the Present

3.1 Introduction

In this follow up report to our 1997 report evaluating CO₂ trends in the atmosphere we seek to understand what, if anything, has changed in the Keeling curve given the addition of almost 26 years of CO₂ data.

3.2 Current Trends in Atmospheric Carbon

To begin this report we can take an initial look at the modern data to compare with our 1997 data.

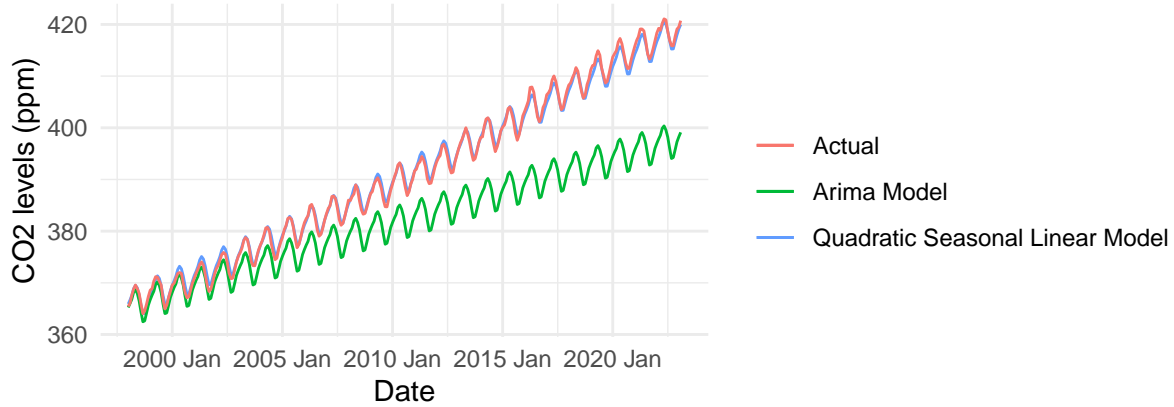


We can see in the above plots that the current data appears to reflect the same patterns we saw in the dataset from 1997. There is still an observable upwards trend and yearly seasonality in the timeplot. Both of which are reflected in the slow decay across lags in the ACF and the slight peaks in ACF values at about every 12th lag. The PACF plot also looks similar to that of the 1997 data with a sharp peak at the first lag and then sudden decay. Overall these plots indicate to us that our current data appears to be following the same pattern as the data from 1997. In the interest of space in this report we have omitted the decomposition plots for the current data, but in our analysis we found them to indicate similar trends and seasonality that we observed in the 1997 dataset.

3.3 Compare linear and Arima model forecasts against realized CO₂

Now that we have conducted our initial analysis of the current data we can compare the forecasts from our 1997 models to the realized data by plotting these predictions against the current data. For visual clarity we have omitted the confidence intervals for our forecasts from our plots.

Quadratic Seasonal Model Predictions vs Actual Data for 1998–2023



We can see that our 1997 Quadratic Seasonal model did a fairly good job at predicting the CO2 levels through 2023, though it does seem that from 200-2005 this model has a tendency to slightly overestimate the CO2 levels, and then from 2015 to 2020 it appears to slightly underestimate the CO2 levels. By contrast, the ARIMA model deviates severely from the actual values. The ARIMA model values are similar until 2000, though the model does tend to underestimate the CO2 levels even during this time period. The model really deviates from the actual values at around 2003, as it starts to underestimate the CO2 levels. It is possible that the keeling curve has shown some acceleration in the increase of CO2 over time, and this might explain the gap between the ARIMA model performance and the true CO2 levels as we move away from 1997.

In looking at our predictions we can also see that in 1997 our ARIMA model predicted that the CO2 levels would cross 420ppm at around April 2038, but from the actual data we can see that the CO2 levels cross 420ppm by February 2022, which is an about a 15 year difference. Our Quadratic Seasonal model prediction of April 2022 was much closer. These results are in line with what we see in the graph above – the quadratic seasonal model does a fairly good job at forecasting, while the ARIMA model tends to increasingly underestimate the CO2 levels over time.

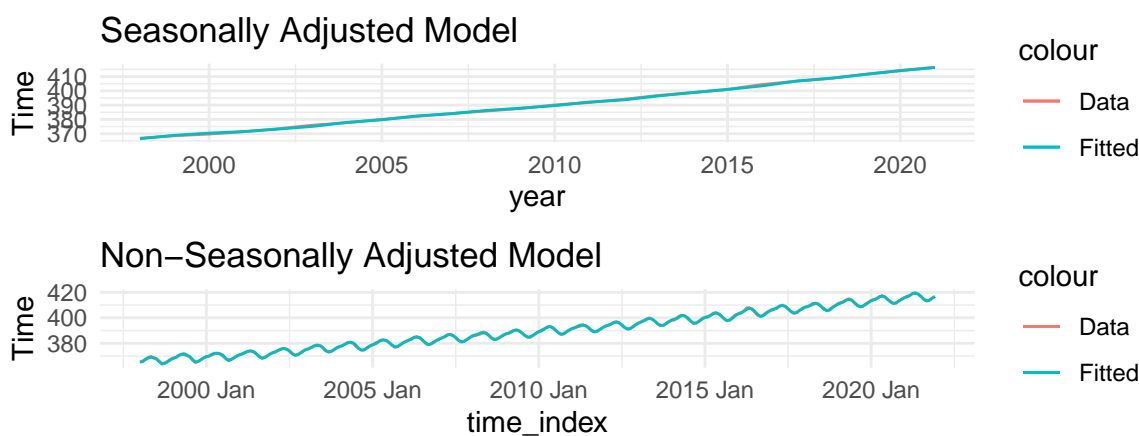
We can also assess the overall performance of our model using a test metric such as the RMSE. The RMSE is the root mean squared error and is a metric that captures on average how far our predictions are from the actual values. It is calculated with the formula $\sqrt{\sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{N}}$ where N is the number of observations, y_i is the i th actual CO2 value and \hat{y}_i is the i th CO2 prediction – in essence we are squaring the sum of the differences between the actual and predicted values, and then dividing by the number of observations before taking the square root. When we calculate the RMSE for our two models of interest we can see that the ARIMA model performs worse than the quadratic seasonal model by an order of roughly 10. The Quadratic Seasonal model's RMSE is 0.79235, meaning that on average the Quadratic Seasonal model prediction is off by about 0.79235 ppm. By contrast the ARIMA model RMSE is 11.301 meaning that on average the ARIMA model is off by about 11.301 ppm.

3.4 Re-train 1997 models on present data

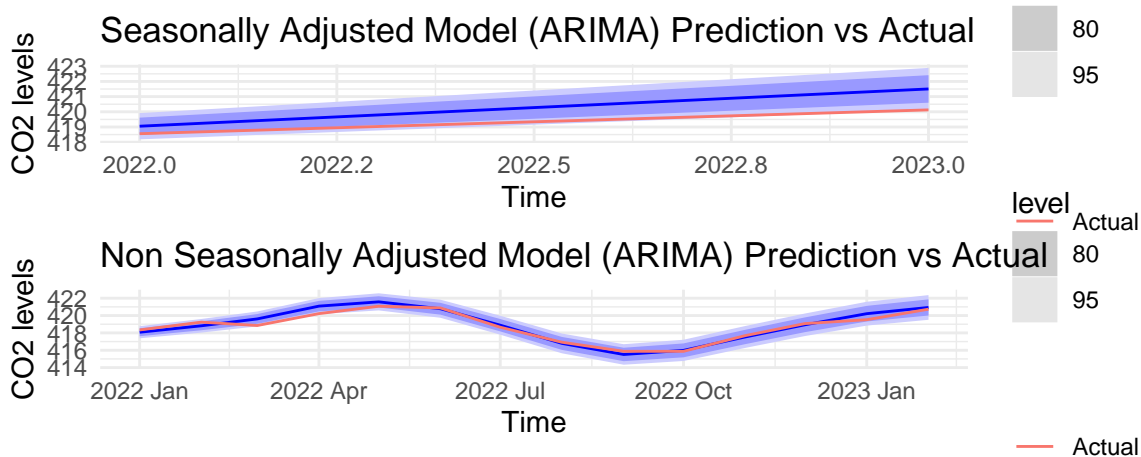
Now that we have an additional 25 years worth of data we can train our models on it. We begin by creating a seasonally adjusted and non-seasonally adjusted dataset and then splitting each of

these datasets into a training and test set where the test set is the last two years of data. We will reserve the test sets to test the performance of our models and will only use the training data to train our models. Best practice would have been to create the train-test split before conducting any sort of EDA in order to preserve the purity of the test set. But since we wanted to evaluate the old model performance and in a time series split the test split must be sequential to the training set we needed to examine all of the data.

For both the seasonally adjusted (SA) and non-seasonally adjusted (NSA) series, we fit ARIMA models. To create the ARIMA models we follow the same methodology as in our 1997 report. With the AIC as our selection criteria we have estimated the seasonally adjusted ARIMA model to be $ARIMA(3,0,1)$ with drift. This $ARIMA(3,0,1)$ model can be interpreted by saying that there is 3 AR term, and 1 MA term and then, of course since the data is seasonally adjusted, that the data had no seasonal terms. For our non-seasonally adjusted data we found the ARIMA model to be $ARIMA(1,0,1)(2,1,2)[12]$ with drift. This can be interpreted as a model that has 1 AR term, 1 MA term, 2 seasonal AR terms, 2 seasonal MA terms, and that the data had to be seasonally differenced once to be stationary. This is notable in that this non-seasonal model is different than the original model that we found in 1997, and this in and of itself may be evidence of a change in the pattern of CO2 levels.

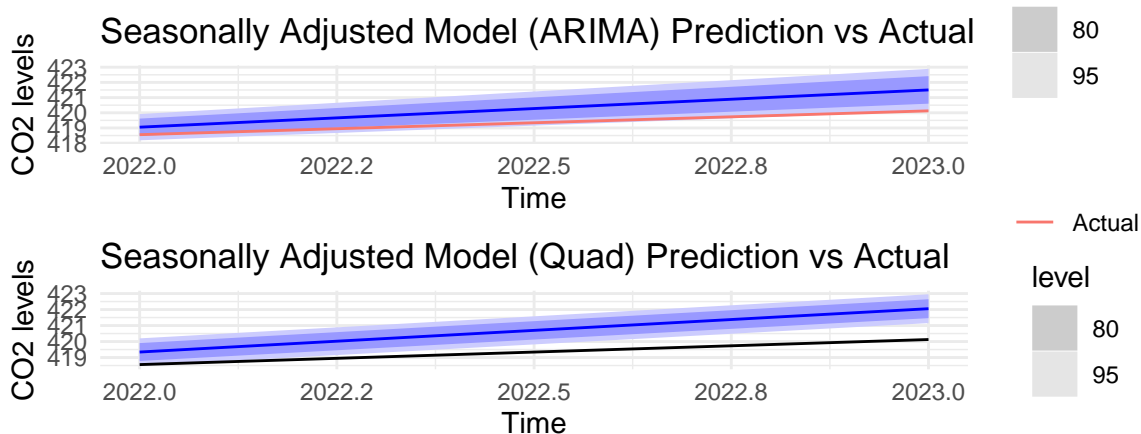


We can see from the graphs above that both models appear to fit the respective datasets fairly well, and when we calculate the training RMSE for each model we get that the SA RMSE is 0.03918 and the NSA RMSE is 0.34434. But to truly evaluate their performance we can generate forecasts from each model and compare these forecasts to the test sets that we set aside earlier.



Visually, when we look at the plots we can see that both the SA and NSA models performed fairly well in the forecasting task but that the SA model was slightly overestimating the CO2 levels. However we can also see that for the SA model the actual data fell within the 95% confidence interval of the SA forecasts. By contrast the NSA model appears to be estimating the CO2 levels much more consistently, with my systematic over or underestimations. We formalizes these observations by calculating the test RMSE for both models and found that the SA test RMSE was 1.0351 and the NSA test RMSE was 0.43192. We can see here that the NSA model did better at forecasting the CO2 levels than the SA model.

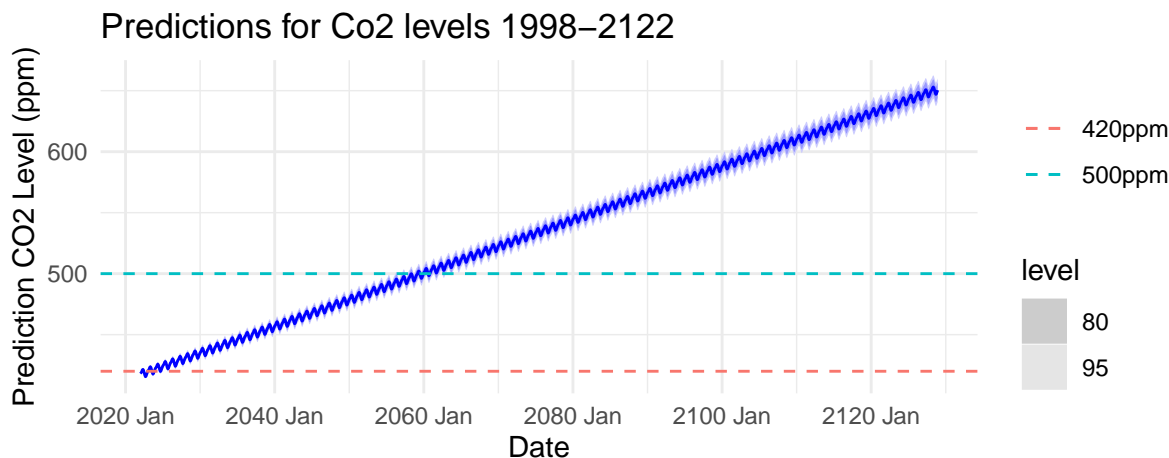
We observed in 1997 that the quadratic seasonal dummy model performed better than the ARIMA models, so we can also construct a seasonally adjusted polynomial time trend model. In this case, just as in 1997, we will stick to a quadratic model of the form $CO_2 = \beta_0 + \beta_1 t + \beta_2 t^2$ to our data to avoid overfitting issues. We find that the SA quadratic model takes the form $CO_2 = 365 + 1.66t + .0205t^2$ where all terms are significant ($p < 0.001$).



Here we see that the SA quadratic model performs worse in comparison to the SA ARIMA model, as overestimates the CO2 levels more than the SA ARIMA model (the actual data is not even contained in the 95% confidence interval). This observation is validated by our formal tests as the SA test RMSE was 1.0351 and the Quad SA test RMSE for 1.4717, meaning that the Quadratic SA model tends to overestimate CO2 values more than the SA model.

3.4.1 Generate NSA Predictions for 420ppm and 500ppm

Since we have determined that the NSA model is the best model for the current data (of the models we tested) we can continue the same prediction tasks we did in our 1997 report. Namely we can generate predictions for when the CO₂ levels will reach 420ppm and 500ppm as well as a forecast for the CO₂ levels 100 years from now in 2122. These results are plotted in the graph below.



We can see in the graph above that the CO₂ levels are predicted to reach 420ppm fairly quickly, for the first time in April 2022 and for the last time in October 2024. We can also see that the CO₂ levels are predicted to reach 500ppm by March 2059 for the first and to be at 500ppm for the last time in September 2061. Note that these are point predictions, and that it may be more accurate to say that it is predicted to reach 420ppm for the first time sometime in between March 2022 and April 2022 and for the last time sometime in between September 2024 and October 2024. Similarly it may be more accurate to say that CO₂ levels are predicted to reach 500ppm for the first time sometime in between February and March 2059 and for the last time sometime in between September and October 2061. In addition we can see that the predicted average co₂ level in 2122 would be around 637.42 ppm..

4 Conclusions

In the original 1997 report and the subsequent follow up report our team was able to answer our initial questions. We found that in 1997 the best model for the CO₂ levels was a quadratic time-series model with seasonal dummy variables and in our follow up analysis we found that a non-seasonally adjusted ARIMA model was the best fit given the new data we had. In addition we were able to generate predictions for what CO₂ levels would be in the future, which we were then able to assess. And when we found some of them lacking we were able to create new models and generate new predictions. In sum though what we found from the data was that there has been a continued increase in CO₂ levels over time, and that this increase is separate from a coincidental seasonal pattern in CO₂ levels. In both our initial report and the follow up we found that CO₂ levels are predicted to continue increasing over time. These findings are just the start of efforts to understand and then curb the rising CO₂ levels globally and to us indicate that we should continue to invest resources and maintain our efforts in collecting this data.