CS146 CO2 Modeling

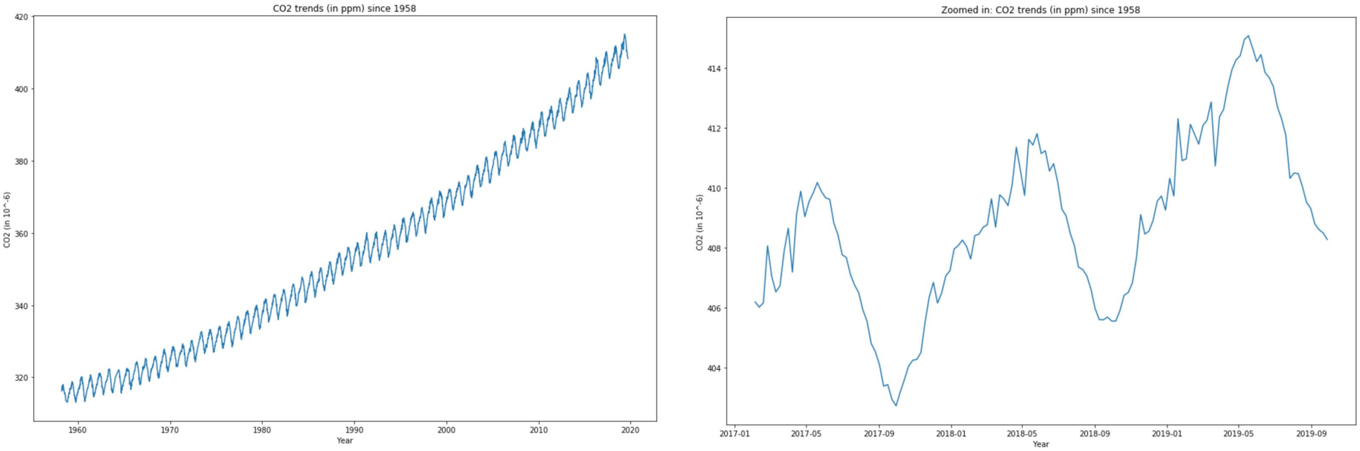
Hana McMahon-Cole

Introduction

Atmospheric carbon dioxide measurements have been recorded at the Mauna Loa Observatory since 1958, recorded according to date and CO2 ppm measurement. Given that CO2 levels at or above 450 ppm are considered “high risk” for climate change, my focus is on predicting carbon dioxide measurements for the next forty years (based on the trend seen in observed data) and identifying when we may anticipate CO2 levels to reach “high risk” levels of 450 ppm.

Process

First, I modeled the original CO2 data. From modeling the data I could see that the CO2 measurements were generally increasing over time. I also saw that it followed a quadratic, periodic trend, as the measurements were going up and down in a quadratic manner, and displaying a periodic repetition (due to seasonal variation).



Figures 1 & 2: (Figure 1) Carbon data measurements over time. (Figure 2) Close up visualization of carbon data trends.

Second, I decided to generate the suggested (linear) model in stan to approximate the observed CO2 data, despite intuitively knowing that the data could be better approximated by a quadratic approximation. This model approximated overall trend with the linear trend line, $c_0 + c_1 t$, seasonal variation, $c_2 * cos(2\Pi t/365.25 + c_3)$, and noise, c_4. As I knew very little about the model, I set broad uninformative priors. To achieve this, I used normal distributions for all of the variables ,c_0, c_1,c_2,c_3 and c_4, centered at 0 (except the y-int, c_0, which I centered at 300) with a variance of 20. I centered most of the variables around 0 as there is little change observed in individual CO2 measurements, so my prior expectation is that values will be close to 0. The y-int was centered at 300 because the initial data points for the CO2 data were around 300. Together these components output a likelihood function which broadly captures the global trend of the data. Furthermore, I constrained all of the parameters to be above zero, as I observed an *overall* increasing trend of co2 measurements from the original data. Thus, while I anticipate parameter values being close to zero, they should not be negative.

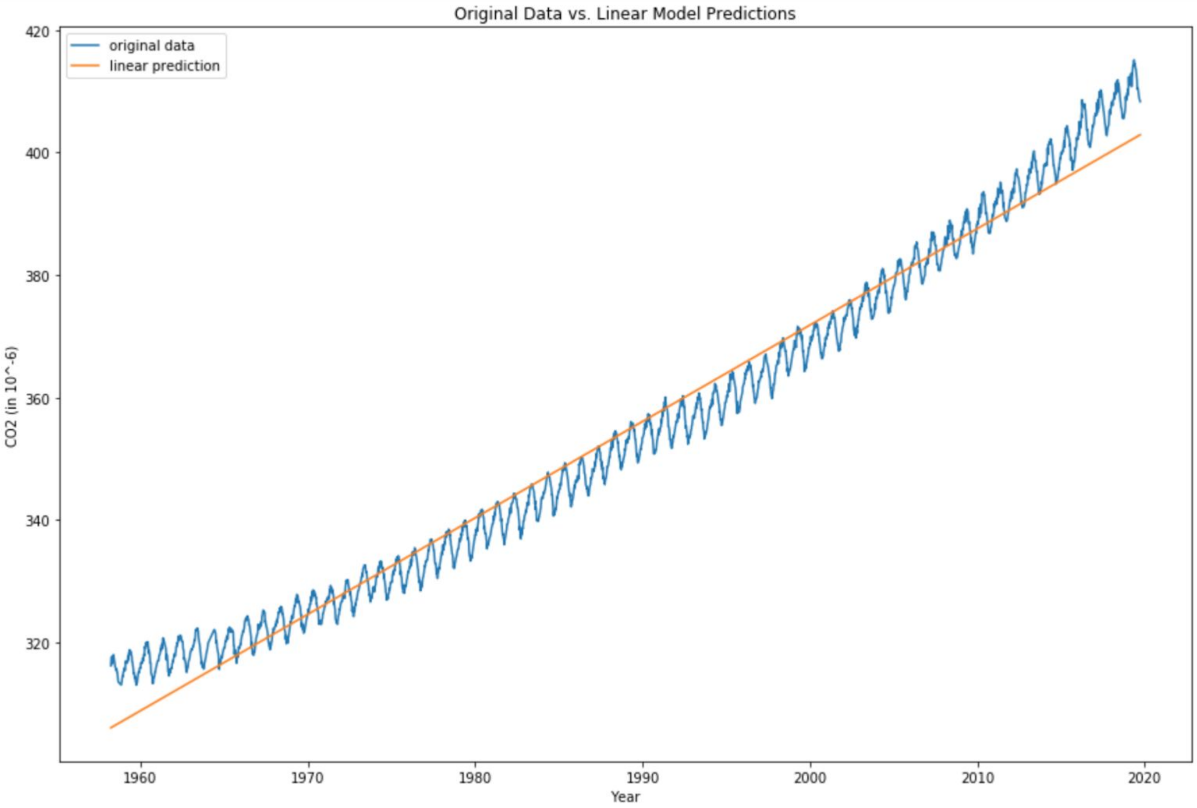


Figure 3: Visualization of original data and linear predictions.

Model Extension

Third, I decided to generate a revised model in stan. I made two substantial changes to the original linear model. I changed the model from linear, $c_0 + c_1 t$, to quadratic, $c_0 t^2 + c_1 t + c_2$ in order to better match the (generally) quadratic data trend. Second, I adjusted the season variation with a sine graph, because the sine graph approximately followed seasonal variation of the data. As a result, my improved model had 6 parameters, 3 quadratic parameters (c_0,c_1 & c_2), two seasonal variation parameters amplitude(c_4) and phase shift(c_3) and the noise parameter (c_5). Again, I set broad uninformative normal priors as done above, mean 0, variance 20 (and mean 300 for the y-int). For c_3, since it is periodic in the range of 0 to 1, it did not make sense to set such a large variance, instead, I set this parameter to have a variance of 1, centered at 0.

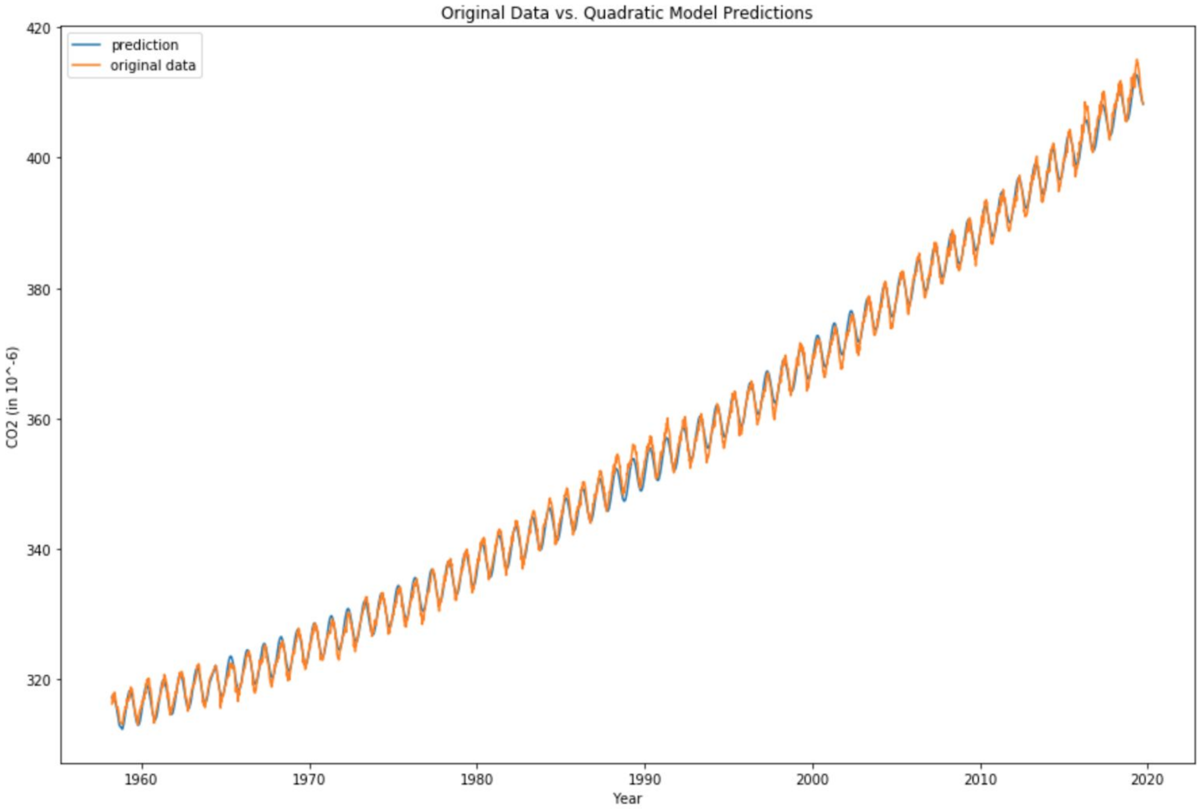


Figure 4: Visualization of original data and quadratic predictions.

Although the second model visually seemed to match the original CO2 data better, I calculated the RMSE (root mean squared error) in order to understand how much better the second model was. RMSE calculates the standard deviation of the prediction errors, basically measuring how spread out the predictions errors, residuals, are. The lower the RMSE score, the more closely the predictions match actual data.

I found that the RMSE score of my model was significantly improved, 0.97 vs 4.27, compared to the linear model (shown below).

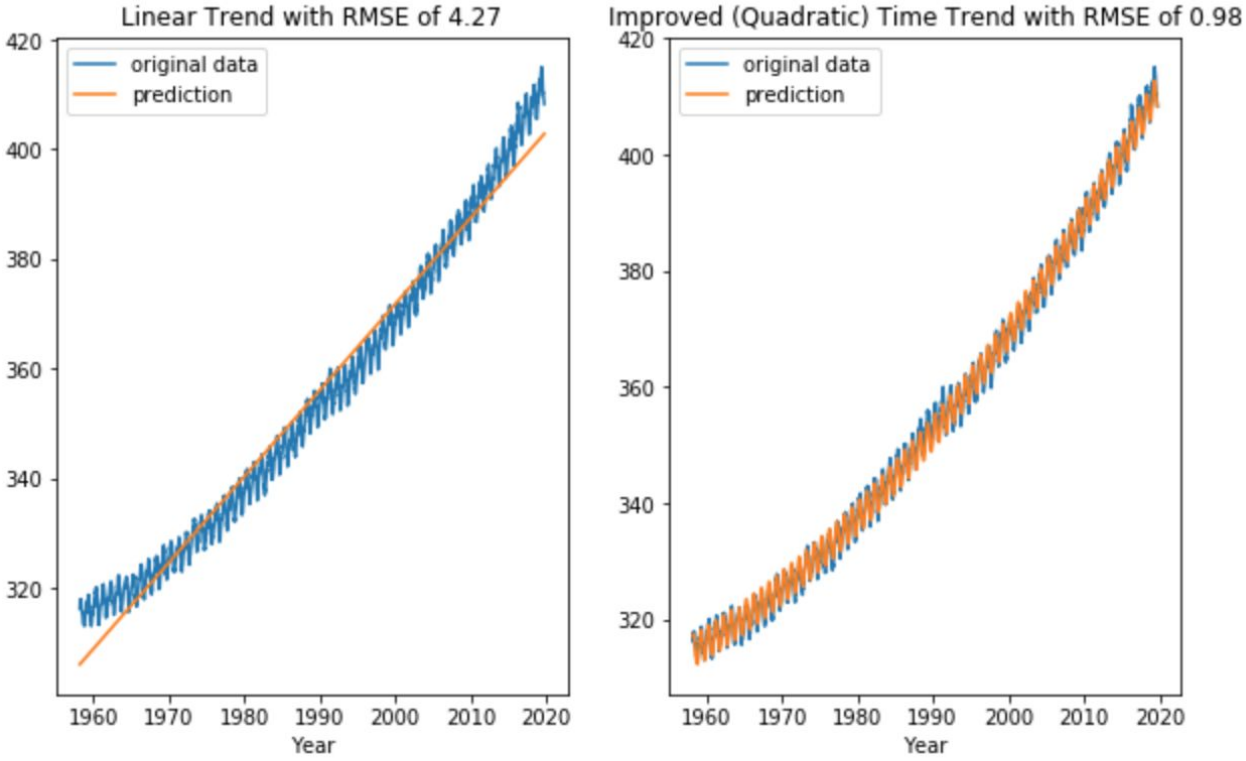
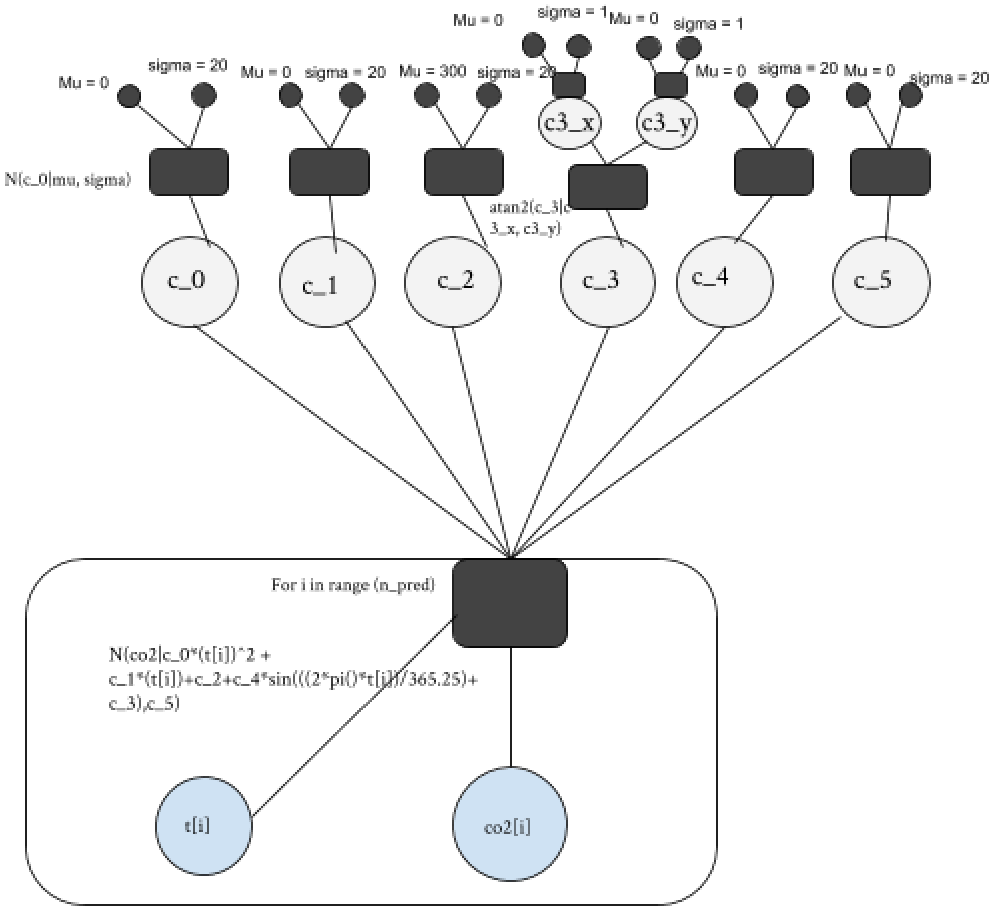


Figure 5: Comparison of Linear Model vs. Quadratic Model, RMSE T-Statistic

Since the second model had a significantly improved RMSE, I proceeded to evaluate the sampling of my improved model. All of the final rhat values were 1.0, which shows that the sampling chains had converged on parameter values in Pystan. The effective sample sizes for each parameter were also good, the smallest effective sample size, n_eff, was 1680, which is (very) good as more than 300 samples indicate a low degree of correlation. I also generated auto-correlational plots in order to check the degree of independence between samples of a given parameter and found all the parameters to output a value of approximately 1 at x = 0, indicating independence, and a value of approximately 0 at all other x values. I also generated pair plots to assess the relationships between parameters, as it is important to check that the variables do not indicate a strong relationship between other variables. All of my pair plots did not indicate a relationship between variables except for c_0 and c_1 and c_2. Since these are the variables of the quadratic function we would anticipate some relationship between them, so these results are in line with my expectations. So overall, the sampling in my improved model went well. (See pair plots, autocorrelation plots and parameter posterior distributions in code)

Factor Graph

In order to visualize the relationship between hyperparameters and observed and unobserved quantities in my model, I created a factor graph, shown below.



First of all, the unobserved quantities are displayed as open, white circles, c_0, c_1, c_2, c_3, c_4, c_5, c3_x and c3_y. The observed quantities, time and co2, are displayed as blue circles. The functions are shown as black squares, the hyperparameters as black circles. The observed quantities and main function are displayed within the plate to indicate that this loops through all the values of t, time, that are predicted in the model.

It was slightly challenging to fit all of the information into the visual, so I left out some of the functions for the sake of visibility. Although only the left most factor box has (N(c_0|mu, sigma)) written next to it, this same relationship exists between the remaining unlabeled parameters, c_1, c_2, c_4 and c_5, except with c_i instead of c_0. c_3, applies atan2 to incorporate information from c3_x and c3_y, while the factor boxes between c3_x and c3_y exhibit the same factoring relationship: N(c3_x|sigma,mu) & N(c3_y|sigma, mu).

Results

Given that my model was a significant improvement compared to the original linear model, I proceeded with the revised model to generate predictions for the next 40 years. I needed to collect 2080 data points since measurements were taken weekly, therefore: 40*52 = 2080. I also plotted confidence intervals in order visualize the uncertainty in my predictions. Since 450 ppm CO2 is the 'high risk' level of CO2 I also plotted when my model predicts a level of 450 pm will be reached so that it easy to visualize for a non-statistician audience.

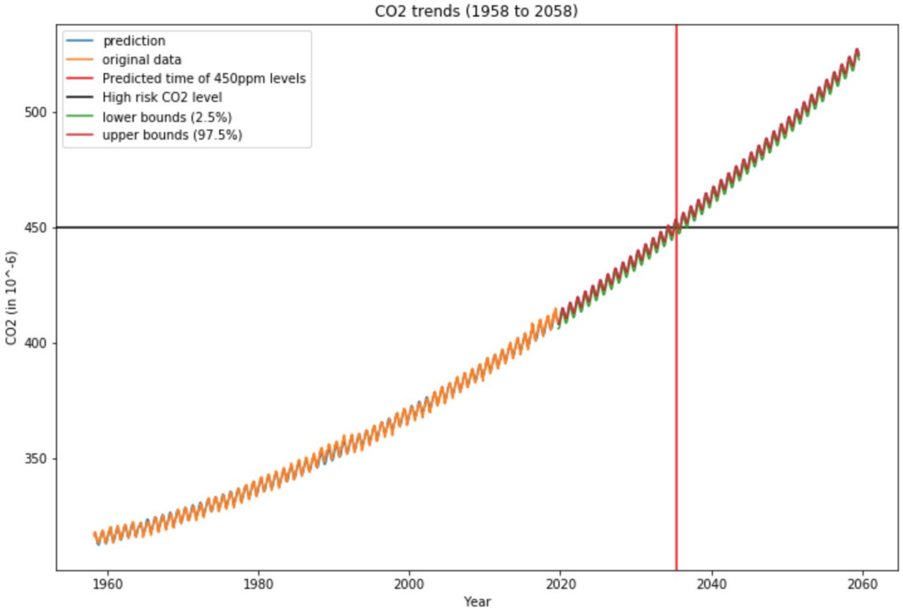
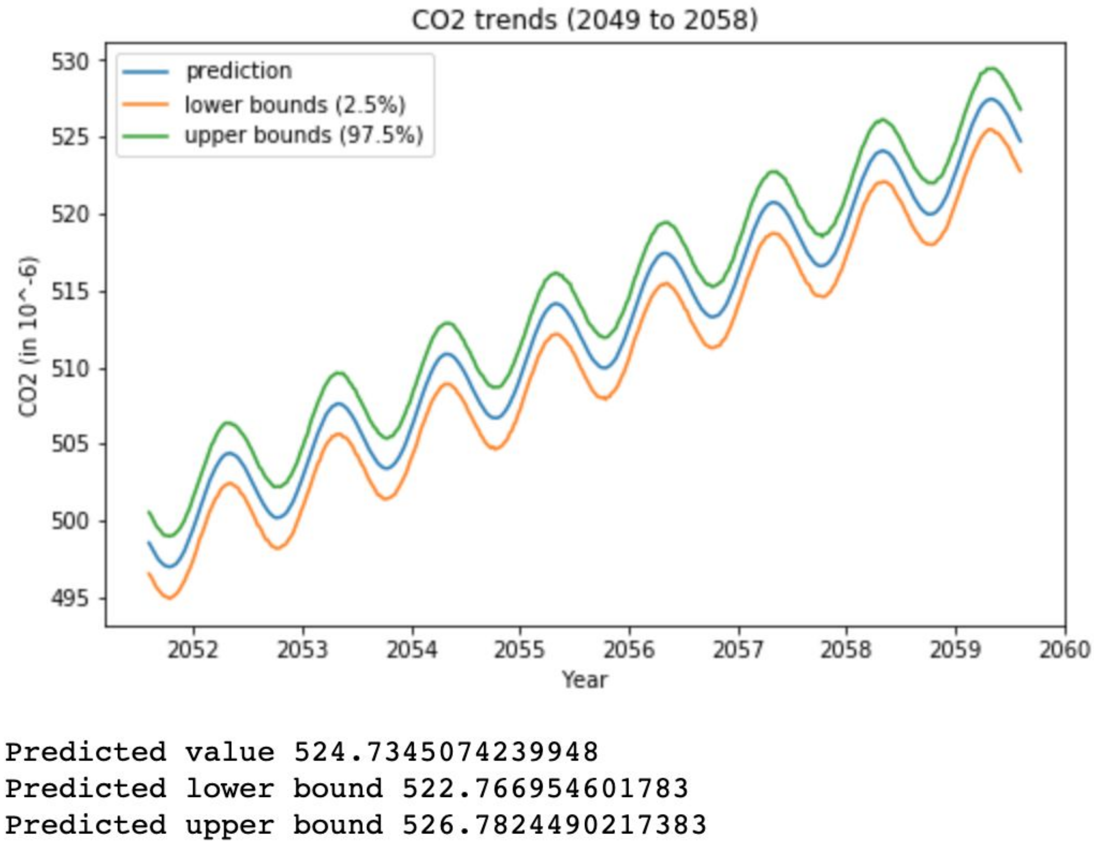


Figure 6: Visualization of original data, predicted data and confidence intervals.

In order to better visualize the predictions for the last 10 years, I created an additional visualization and output prediction values for the last data point, in 2058.



```
Predicted value 524.7345074239948
Predicted lower bound 522.766954601783
Predicted upper bound 526.7824490217383
```

Figure 7: Visualization of original data, predicted data and confidence intervals for the last 10 years.

In the last year predicted, 2058, my model predicts a mean atmospheric CO2 level of 524.7 and a confidence interval of [522.75, 526.74].

Uncertainty

While this model captures the local trends of the data and is an improvement compared to the linear approximation, it does not fully approximate noise in the original data (see below).

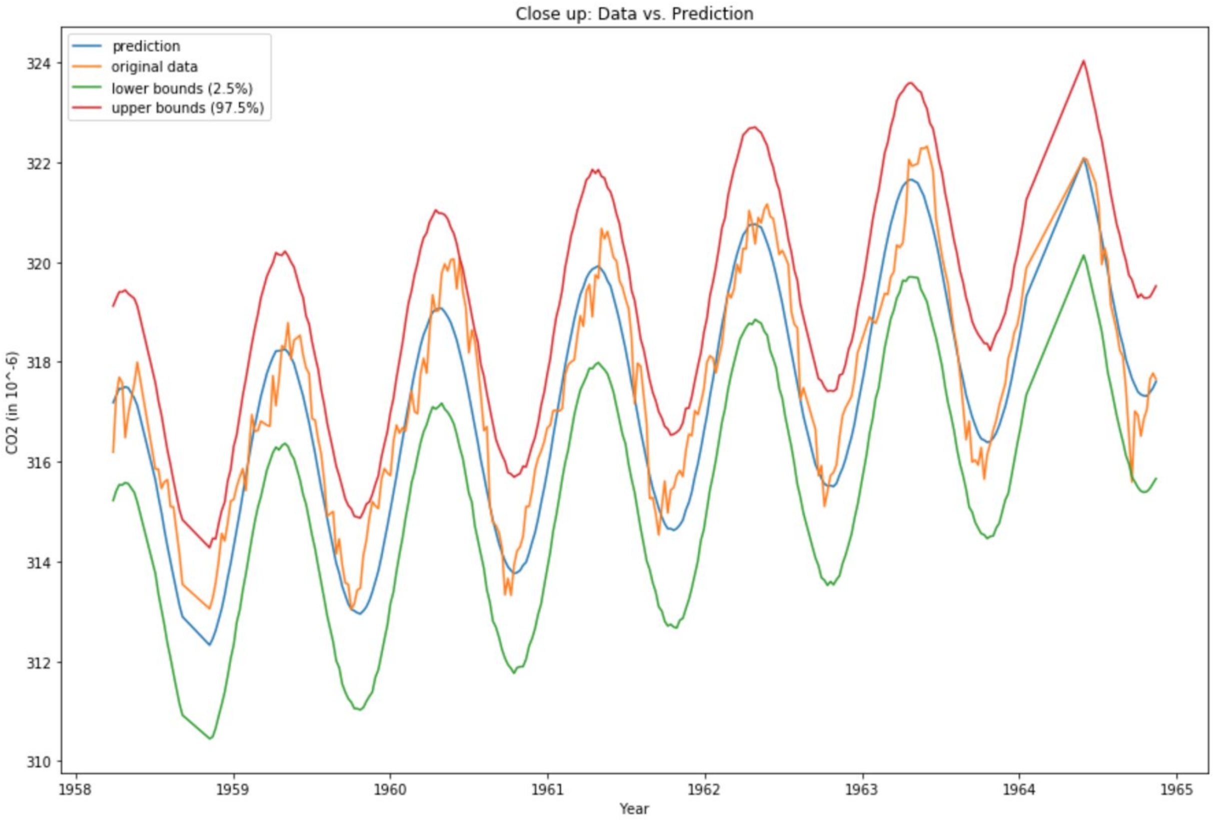


Figure 8: Close up visual of discrepancy between original data and predicted data.

Moreover, even though our model does a good job of *approximating* CO2 data, we cannot be sure that present data is a good proxy for *future* CO2 levels. The present data could either underestimate or overestimate future CO2 measurement trends. It is plausible that the data could underestimate future trends given the increasing industrialization, and therefore increase in CO2 release, observed worldwide. Conversely, it is possible that this data could overestimate future CO2 measurements if many countries may come together and decide to limit reliance on fossil fuels and hence CO2 emissions.

However, overall, the model well reflects observed CO2 measurements, meaning that the prediction for 2058 is reasonable. While there are improvements that can be made, the model predictions seem plausible, based on the observed CO2 levels. One way that this could be accounted for in our model would be to marginally widen the confidence intervals over time to account for higher predictive uncertainty the farther away from observed measurements.

References

C. D. Keeling, S. C. Piper, R. B. Bacastow, M. Wahlen, T. P. Whorf, M. Heimann, and H. A. Meijer, Exchanges of atmospheric CO2 and 13CO2 with the terrestrial biosphere and oceans from 1978 to 2000. I. Global aspects, SIO Reference Series, No. 01-06, Scripps. Institution of Oceanography, San Diego, 88 pages, 2001.

Pre-Class 14.1 Solutions

People to thank for debugging code: Gelana, Nikesh <3