Dennis Kageni

CS146 Final Project

Fall 2019

Context[1]:

The OECD Environmental Outlook to 2050: Key Findings on Climate Change (2012) is a report by leading climate scientists that highlights the rising tension between global economic growth and the resulting increase in carbon dioxide ($C0_2$) emissions. They warn that at the current rate, we are poised to exceed the "high-risk" atmospheric carbon emissions threshold of 450ppm set by the UN Framework Convention on Climate Change (UNFCC) (Willard, 2014). That said, there is a pressing need to take decisive action and implement decarbonization initiatives and accelerate the world's transition into a low-carbon economy. However, understanding the effects of carbon emissions at a granular level and implementing the appropriate mitigation strategies is an inherently complex endeavor.

Computer models offer a promising method of understanding the effect of carbon emissions on global temperature and possibly developing novel solutions to managing it (Jose et.al., 2016). For this assignment, I will create a statistical model using carbon dioxide measurements from the Mauna Loa Observatory in Hawaii to forecast CO_2 emissions for the next 40 years (Keeling et.al., 2001). The goal is to predict when we are expected to exceed the 450ppm threshold and motivate a sense of urgency on the need to enforce policies that ensure we limit global warming to 2^{o}C (Willard, 2014).[2]

The Data:

Before building a statistical model to forecast carbon emissions 40 years from now, let's begin by making observations from the dataset. At first glance, a time series plot (Figure 1) shows that CO_2 levels have been increasing over time. This would be our first model assumption. However, on zooming in, we

[1] #context: motivated the importance of prioritizing action against carbon emissions and how statistical models can be used to understand its impacts; and hopefully help develop effective interventions

[2] #modeling: created a statistical model that explains the Mauna Loa data set and used it to predict atmospheric carbon emissions for the next 40 years, as well as when we're likely to exceed the 450ppm threshold

observe repeated small dips in the curve and random fluctuations (Figure 2). Therefore, we need to model the following components: overall trend, seasonal variations, and noise.

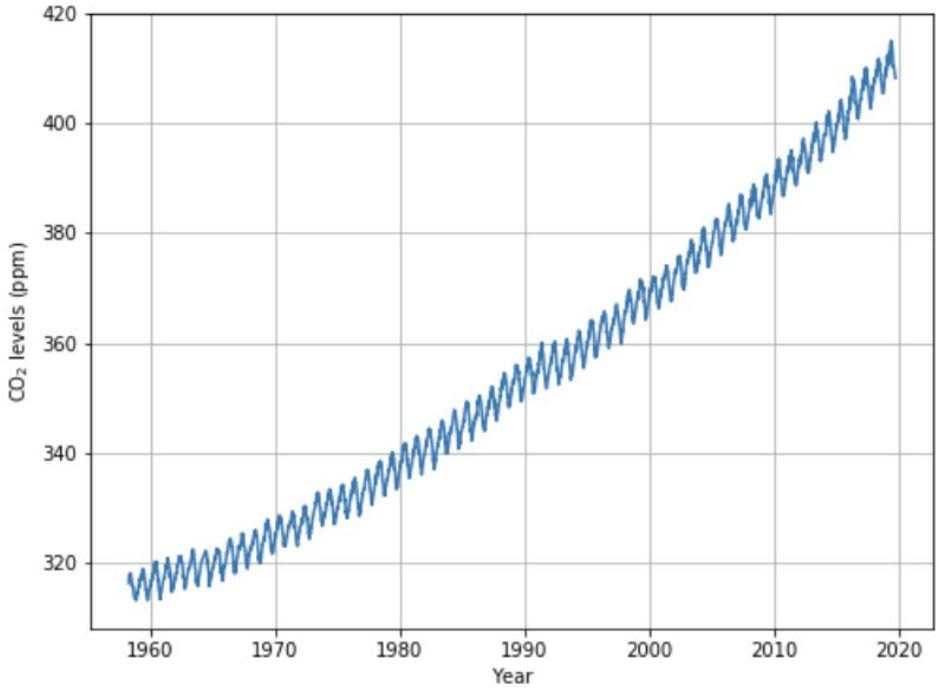

Figure 1. CO_2 levels from 1958 - present.

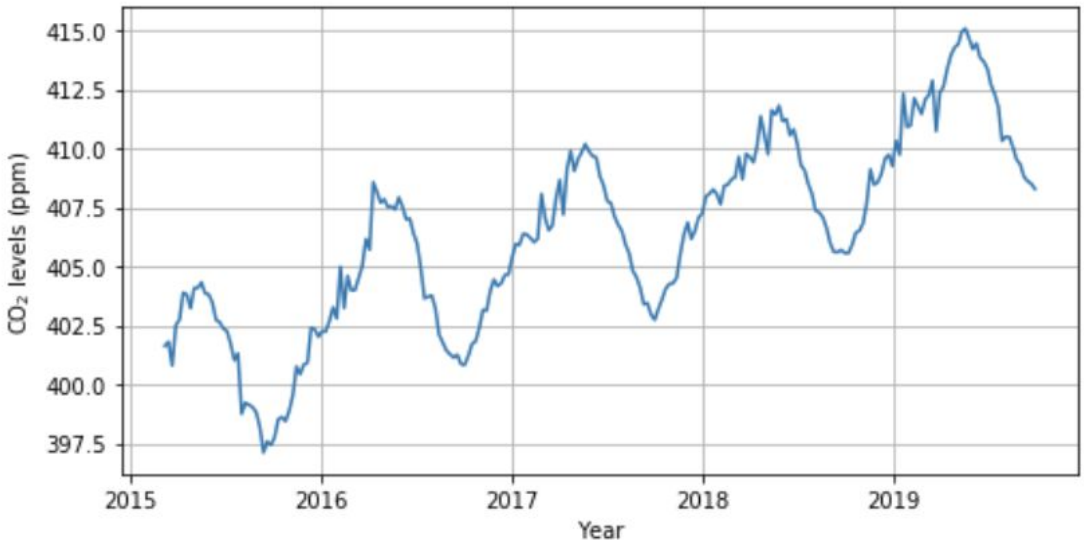

Figure 2. CO_2 levels from 2015-present

Critiquing the Simple Model

The basic model that captures the observed component assumes a linear trend, a sinusoidal seasonal variation that is symmetric and a normally distributed noise:

$$p(x_t \mid \theta) = N(c_0 + c_1 t + c_2 cos(\frac{2\pi t}{365.25} + c_3),\ c_4^2)$$

where c_i are unobserved parameters and $p(x_t \mid \theta)$ is the likelihood function.

The linearity assumption for the long term trend generates a poor fit for the data as it has a higher RMSE than the quadratic and exponential comparative models (Figure 4). In addition, we can't assume linearity because the data on global GDP output (where the resulting industrialization contributes to carbon emissions) is not linear (Figure 3). Modeling the seasonal variation using the cosine also generates a poor fit since it is symmetric and therefore does not capture the "right-skew" in Figure 2. In addition, the peaks and trough of the dataset are jagged as opposed to the cosine's which has smooth curves (Figure 4).

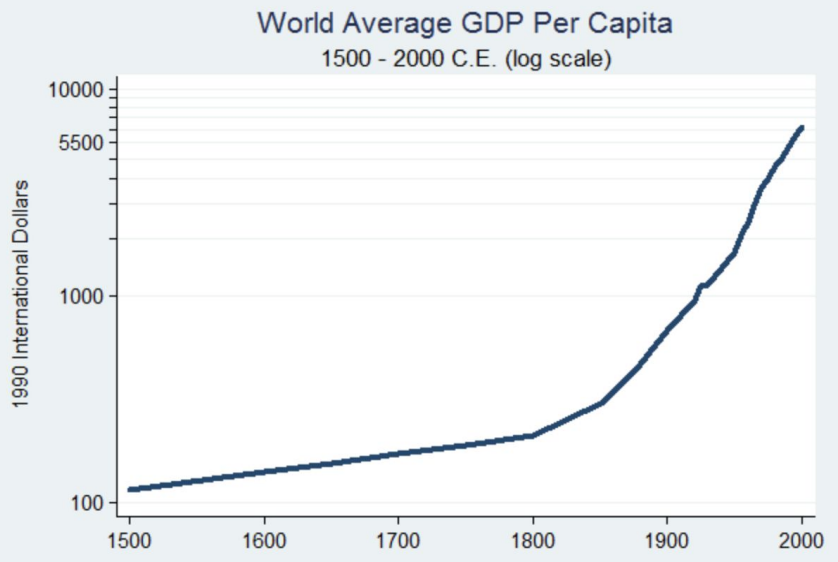

Figure 3. Average world GDP Per Capital Data (1500 - 2000) that shows the exponential increase in industrialization in the past 170 years

Source: DeLong (2014)

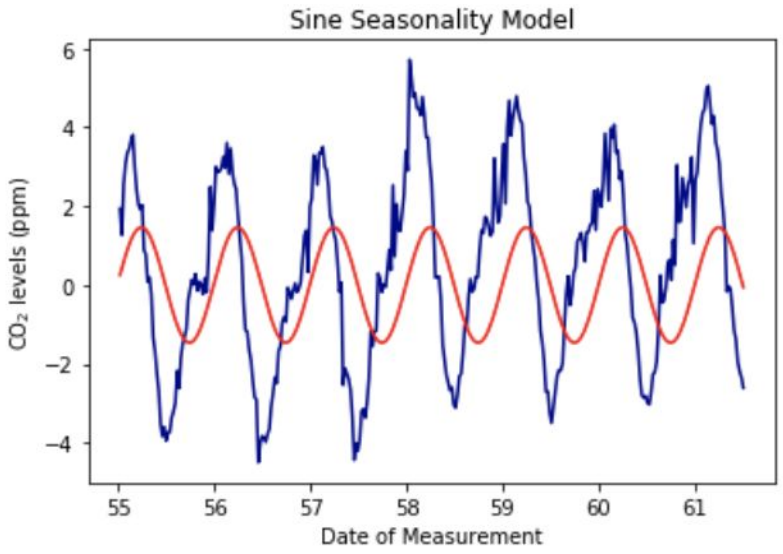

Figure 4. The Differentiated Sequence of CO_2 levels overlayed with a sinusidal graph

Model Extension

Trend

The candidate models that I tested were:

1. Linear trend: $c_0 + c_1 t$
2. Exponential trend: $c_0 e^{c_1 t}$
3. Quadratic trend: $c_0 + c_1 t + c_2 t^2$

where c_i are unobserved parameters

Visually, we can observe from Figure 5 that of the three candidate models, the quadratic trend model fits the data best. This can also be inferred from Figure 6 where we observe that the quadratic trend has the lowest RMSE value.

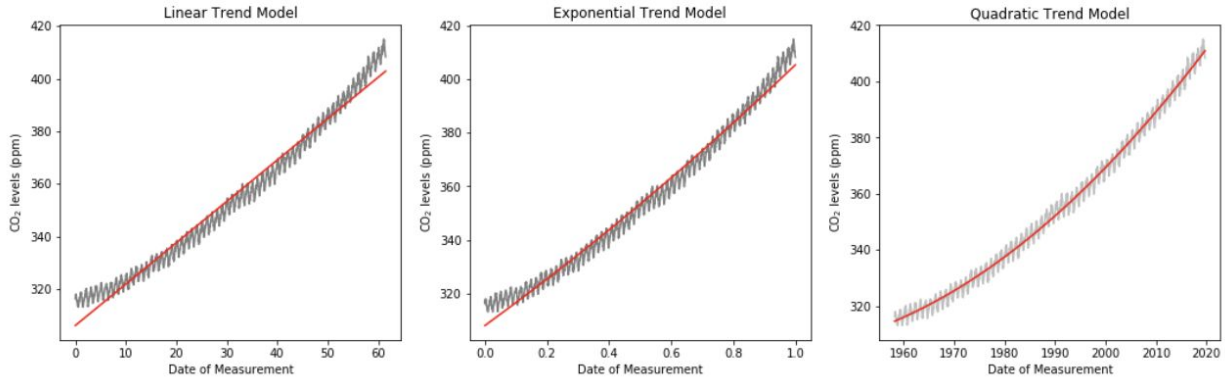

Figure 5. Long-term trend curves for the three models I tested

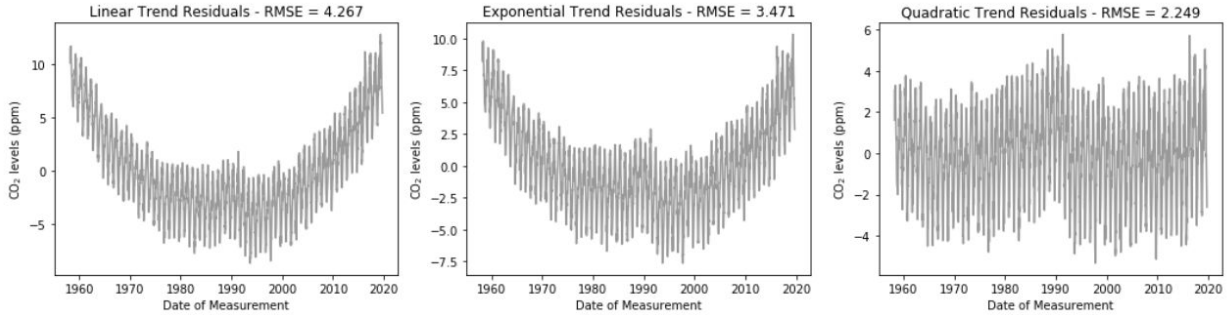

Figure 6. Residuals and RMSE of the three models I tested

Seasonal Variation

To have a better understanding of the seasonal variation, we first need to remove the influence of the long-term upward trend. This is done by taking the first difference of the values (i.e, subtracting each value from the preceding one) so that we can observe the month-to-month changes in carbon dioxide levels. Figure 5 shows a portion of the differentiated sequence from 2013-2019. We can also observe that the amplitude is between 6 and -4. Given that $c_2 sin(x) = -c_1 sin(x + \pi)$, our priors for any hyperparameter for the amplitude needs to be restricted to only positive or negative values

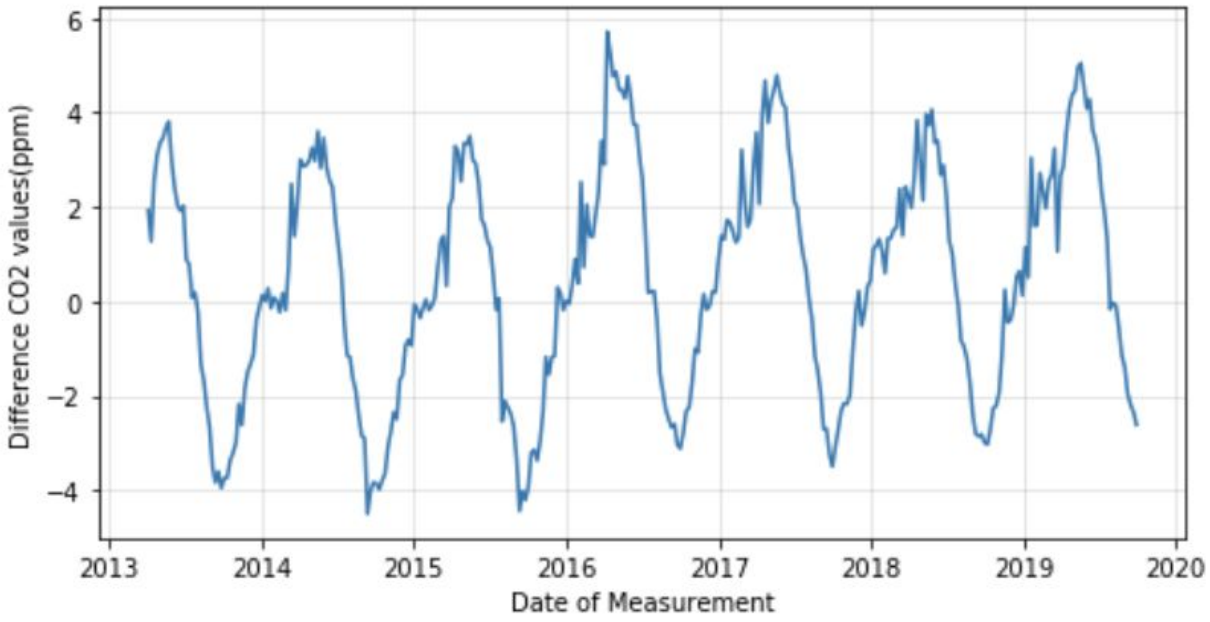

Figure 7. Differentiated Sequence of CO_2 levels from 2013-2020

The candidate models that I tested were:

1. Sine: $c_0\, sin(\frac{2\pi}{365.25}t\ + c_1)$
2. Cosine: $c_0\, cos(\frac{2\pi}{365.25}t\ + c_1)$
3. Aperiodic Sine: $-(1+e^{\,c_3 t + c_4})\ arctan\ (\frac{sin(c_3 t + c_4)}{1+e^{\,-c_1} - cos(c_3 t + c_4)})\ + c_2)$
4. Double Sine: $c_0\, sin(\frac{2\pi}{365.25}t\ + c_1) + \ c_2\, sin(\frac{2\pi}{365.25}2t\ + c_1)$

where c_i are unobserved parameters. They all successfully converged. The trend and seasonality components converged when receiving time data in years while the exponential trend model required us to normalize time. Also, note that the sine and cosine models have the same explanatory power but are both included in the analysis for completeness.

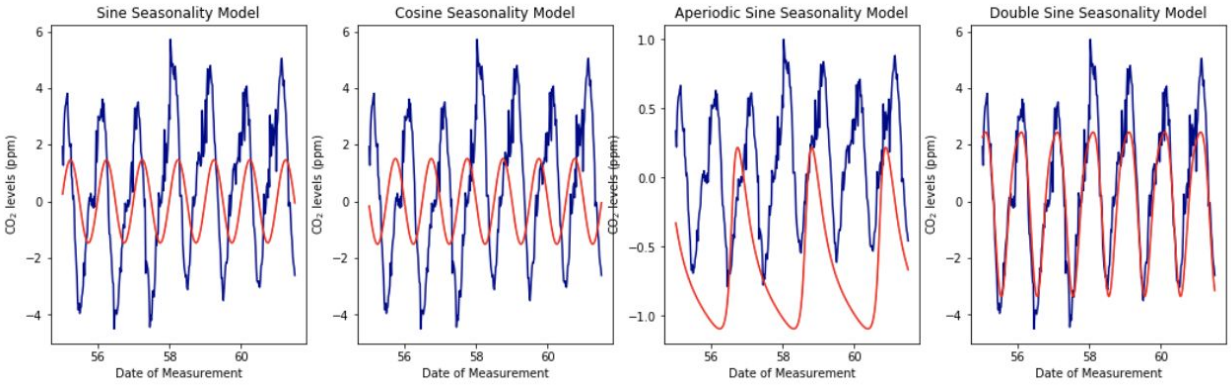

Figure 8. Seasonality curves of the four models I tested

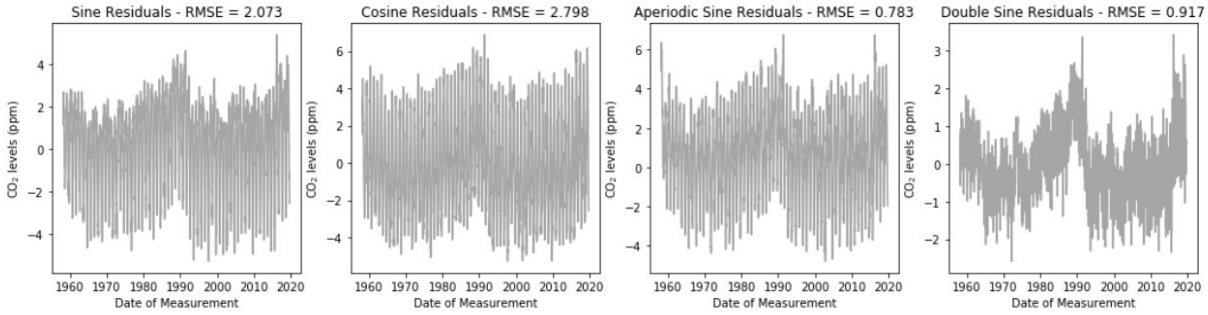

Figure 9. Residuals and RMSE of the four models I tested for the seasonality curves

Although the aperiodic sine has the least residuals overall (Figure 8), we can see that it falls out of sync with the data (Figure 7). For this reason, we select the double sine model instead. Compared to the sine and cosine models, it has a higher amplitude and is in better sync with the original data. It is worth noting, however, that its peaks are lower than those in the data. Finally, I used a normally distributed noise to model the random fluctuation

Full Model

$$p(x_t \mid \theta) = N(c_0 + c_1 t + c_2 t^2 + c_3 \sin(\tfrac{2\pi}{365.25} t + c_4) + c_5 \sin(\tfrac{2\pi}{365.25} 2t + c_4)),\ c_6^2)$$

The priors over all model parameters are as follows:

1. $c_0 \sim$ N(310, 30). Inferred using the minimum value in the dataset (313.04ppm) and the standard deviation (27.9945). A normal distribution is used to communicate high confidence that the CO_2 concentration at the start of the measurements should be around 300ppm.
2. $c_1 \sim$ N(0.5, 0.5) and $c_2 \sim$ N(0.1, 0.1): We have no prior knowledge of the coefficients other than the fact that they should be small. The values for the quadratic term have a smaller standard deviation to ensure that we avid having unreasonable values for t^2
3. $c_3, c_5 \sim$ N(0, 4): The absolute value of the trend residual commonly peaks at 4 (Figure 7). Thus, the amplitudes of the periodic function should be close to this value
4. $c_4 = arctan(\frac{d_0}{d_1})$ where $d_0, d_1 \sim$ (0, 1). The noise is centered around 0 to give it a low standard deviation
5. $c_6 \sim$ InvGamma(3, 2). We use the inverse gamma to ensure that c_6 is positive. The chosen hyperparameters ensure that the residual error is small

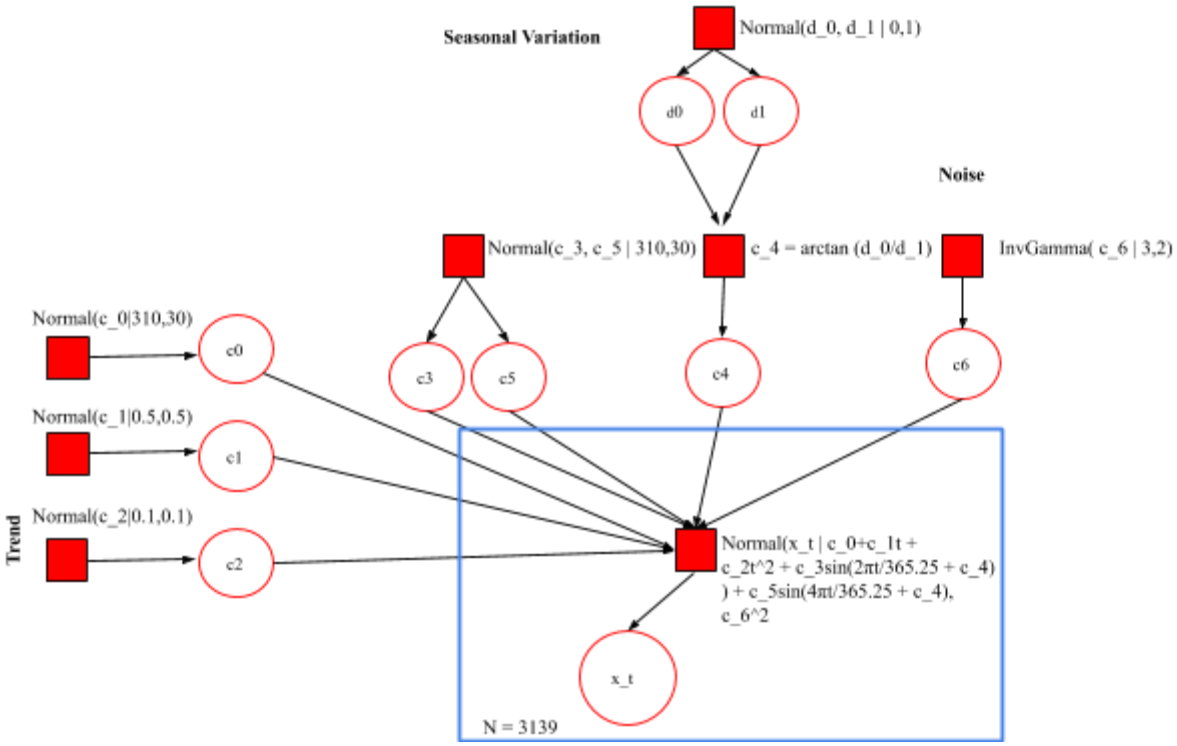

Figure 10. Factor Graph of the full model[3]

Model Critique and Reflection

Challenges

The most difficult part of this assignment was defining a function that best captured the seasonality component of the dataset. In Figure 2, we can see that our function roughly needs to be periodic, increasing, right-skewed and jagged. Whereas the sine and cosine functions capture the periodicity, they fail to represent the jagged crests and troughs. Most importantly, they also fail to capture the "right-skew" and the varying amplitudes the data exhibits. Ideally, we would want to combine a sawtooth (this would help with the right skew) and sine function but my attempts at implementing a "tilted sine" function did not exhibit model convergence

[3] #evidencebased: Provided detailed justification for the final model selected, as well as the choice of priors

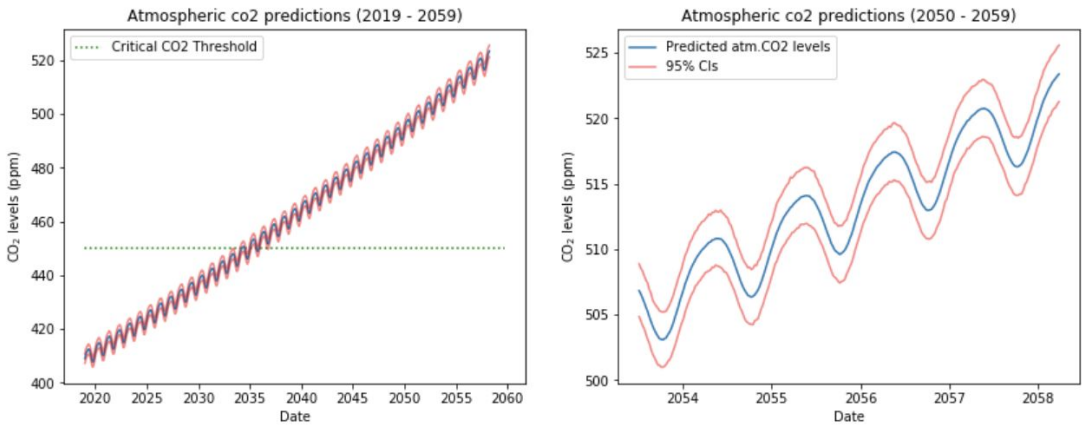

Figure 11. CO_2 level predictions until 2059. The figure on the right zooms in on the data to see the 95% confidence intervals[4]

Critique

The main problem with this model is that it does not account for the right skew in the data. In addition, one would expect the confidence intervals of the model's predictions to decrease with time. However, Figure 11 shows that our model has the same confidence intervals between 2019 and 2059 and can be explained by our assumption that the data points are independent and uncorrelated. While this assumption makes it easier to develop a model, it does not represent the true nature of the phenomena we're modeling. One way to address this issue is by implementing autoregressive models where future predictions can "inherit" the uncertainty of past predictions (Kalekar, 2004).

[4] #dataviz: I used various visualizations through the report to (i) make sense of the data, (ii) support claims about the data and the predictive model, and (iii) show model convergence for the models used for the long term trend and seasonality curves

Inference

The model predicts that at the current rate, we will exceed 450ppm with high confidence (97.5%) on February 23rd, 2035. As of 29th March 2058, carbon emissions will be at 523ppm with a 95% confidence interval of [521.2, 525.5]. We have a reason to believe the predictions since posterior predictive checks show several test statistics (mean, standard deviation, maximum value, and skewness) have similar characteristics to the actual data (Figure 12).

Conclusion

The results suggest that **immediate** action is required if we are to avoid catastrophic consequences related to increasing carbon emissions.

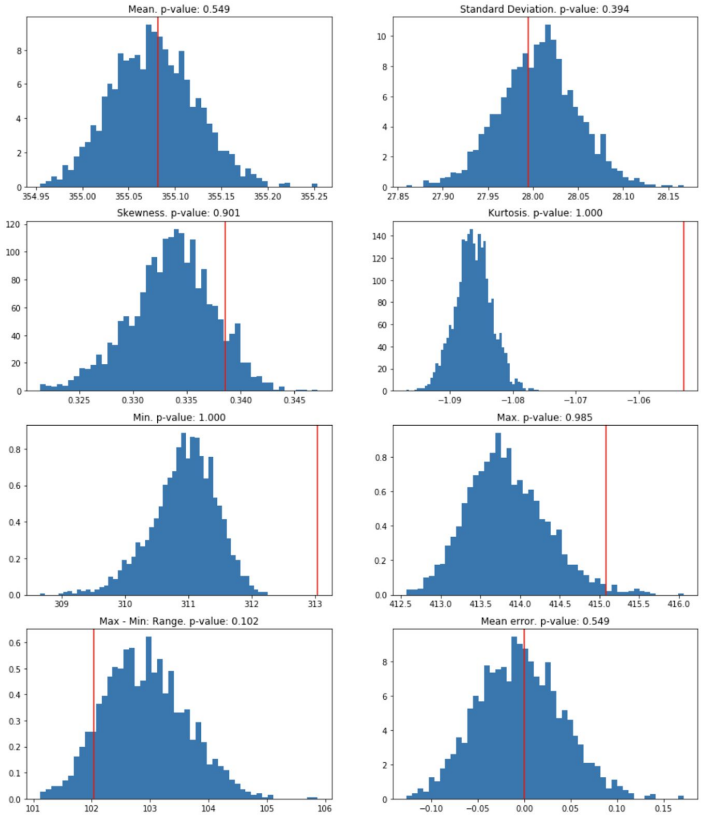

Figure 12. Posterior predictive checks on the model

References

Delong, B. (2014) Estimates of World GDP, One Million B.C.-Present [personal blog]. Retrieved from https://delong.typepad.com/sdj/2014/05/estimates-of-world-gdp-one-million-bc-present-1998-my-view-as-of-1998-the-honest-broker-for-the-week-of-may-24-2014.html

Jose, V. S., Sejian, V. P., Bagath, M. M., Ratnakaran, A. A. S., Lees, A. B., Al-Hosni, Y., Sullivan, M., Bhatta, R., Gaughan, J. (2016). Modeling of Greenhouse Gas Emission from Livestock. Frontiers in Environmental Science, 4. doi: 10.3389/fenvs.2016.00027

Kalekar, P. S. (2004). Time series forecasting using holt-winters exponential smoothing. Kanwal Rekhi School of Information Technology, 4329008, 1-13.

Keeling, C. D., Piper, S. C., Bacastow, R. B., Wahlen, M., Whorf, T. P., Heimann, M., and Meijer, H. A. (2001) Exchanges of atmospheric CO2 and 13CO2 with the terrestrial biosphere and oceans from 1978 to 2000. Global Aspects, SIO Reference Series, No. 01-06. Scripps Institution of Oceanography, San Diego. Retrieved December 14, 2019 from https://scrippsco2.ucsd.edu/data/atmospheric_co2/mlo.html

OECD Environmental Outlook. (2012). *OECD Environmental Outlook*. doi: 10.1787/9789264188563-en

Willard, B. (2015, May 26). CO2–Why 450 ppm is Dangerous and 350 ppm is Safe. Retrieved Fromhttps://sustainabilityadvantage.com/2014/01/07/co2-why-450-ppm-is-dangerous-and-350-ppm-is-safe