Modeling and Forecasting Atmospheric CO_2 Levels from 1958 until 2058 at Mauna Loa

Alexandra Pukhova

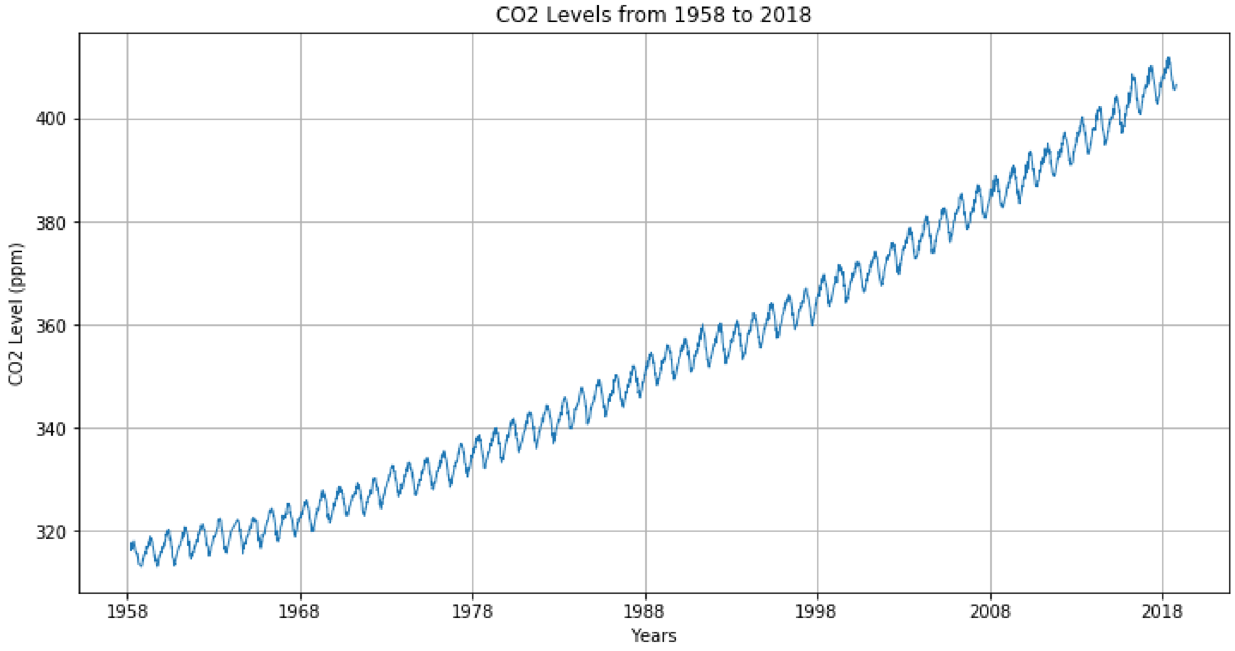
December 21, 2018

The Gist with my Python Notebook can be found here.

I. Scenario & Data Format

We are modeling atmospheric CO_2 levels based on the data that have been gathered weekly at the Mauna Loa Observatory in Hawaii since 1958. Historically, the curve, giving the model, has become known as the Keeling Curve, named after the scientist Charles David Keeling who started monitoring the program (Wikipedia, 2018).

For the purposes of my analysis, I formatted the time series data into days passed since the measurement date. Then, I converted the days into years passed, since it made sampling significantly faster, especially when the trend was given by higher degree polynomials or an exponent. The CO_2 data is provided in parts per million (ppm).



II. Model, Assumptions & Parameters

Time is the explanatory (or observed) variable in this model. The trend of the data (that I refer to as the cumulative trend in my Stan model) is an additive model that combines a trend line, a seasonal component, and noise. The unobserved quantities are those that govern this pattern. This is what we are doing inference on. I assume that all the constants but the intercept that give the trend are independent and normally distributed with a mean of zero and a standard deviation of 10. It is a fairly broad prior, given that the slopes represent the rate of change of CO_2 levels and we know that the change of above 10 ppm per week is unrealistic and equivalent to a rapid global climate change. Another way to appropriately inform our prior is to center the intercept of the trend-line around the starting observation of the CO_2 level. It does not govern the trend and ultimately only plays the role of starting our predictions from reasonable levels.

Hence, the priors are of the following form:

- The intercept: $c_0 \sim N(c_0 \,|\, 300,\ 30)$
- The coefficient of the first-degree polynomial: $c_1 \sim N(c_1 \,|\, 0,\ 10)$
- The coefficient of the second-degree polynomial: $c_2 \sim N(c_2 \,|\, 0,\ 10)$
- The exponent of the exponential trend: $\lambda \sim N(\lambda \,|\, 0,\ 10)$
- The amplitude of the seasonal component: $A \sim N(A \,|0,\ 10)$
- The phase shift of the seasonal component: $\phi = \arctan\,(\frac{x}{y})$[1], where x, y are sampled uniformly.[2]
- Noise: $\sigma \sim N(\sigma \,|\, 0, 10)$

In my Stan model, I bound c_0, c_1, c_2 and λparameters that govern the slope of the trendline by 0 as the lower limit. This is a reasonable limit to set, since we know that the CO_2 levels are increasing and not decreasing over time, and this limit does not inform our prior with a numerical value. Additionally, I bound the standard deviation parameter by 0 since it is always positive.

III. Model Comparison

To observe the model fit more precisely, I modeled the general trend and the seasonal trend separately. Firstly, I built three general trend models:

- Linear: $y = c_0 + c_1 t$
- Quadratic: $y = c_0 + c_1 t + c_2 t^2$
- Exponential: $c_0 * e^{\lambda t}$

[1] I assume that the notation that gives the phase shift value could be expressed as a kind of an indicator graph in the form of I(φ=arctan(x/y)).

[2] I am not specifying a prior on these two parameters and this is equivalent to specifying a uniform prior. (Stan Development Team, 2014)

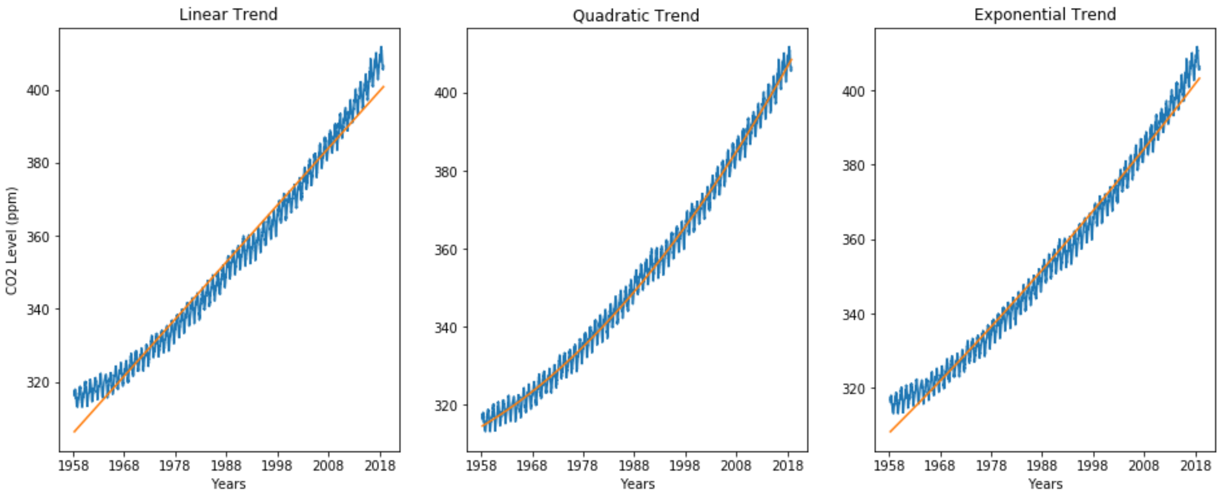


Figure 1. A linear, a quadratic and an exponential trend, respectively, in orange and CO_2 levels data in blue.

To determine the best model fit, I calculated the Root Mean Square Error (RMSE) for each model and obtained the following results:

RMSE for the linear trend = 4.127395100974672.
RMSE for the quadratic trend = 2.2416621985822.
RMSE for the exponential trend =3.379175999081375.

The quadratic model results in the least standard deviation of the residuals and I will move forward with this trend.

As the second step, I un-trended the data and fit a seasonal component.

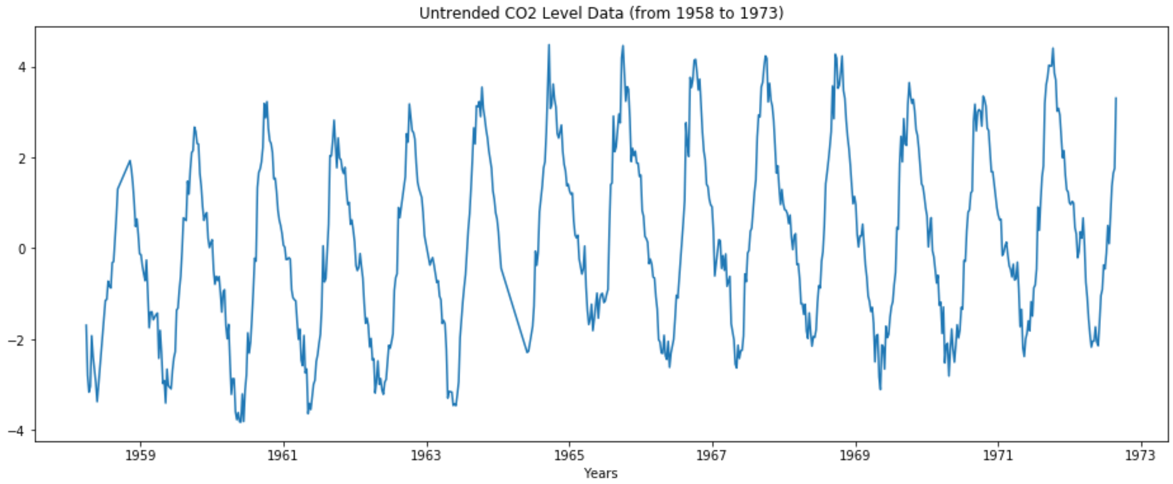


Figure 2. Untrended CO2 level data (from 1958 to 1973).

The first model I fit was a sine trend, given by a sinusoidal trigonometric function:

$$y = A\sin(2\pi t + \phi)$$

Note that I do not need to adjust my period by 365.25 since I had previously converted my data into years.

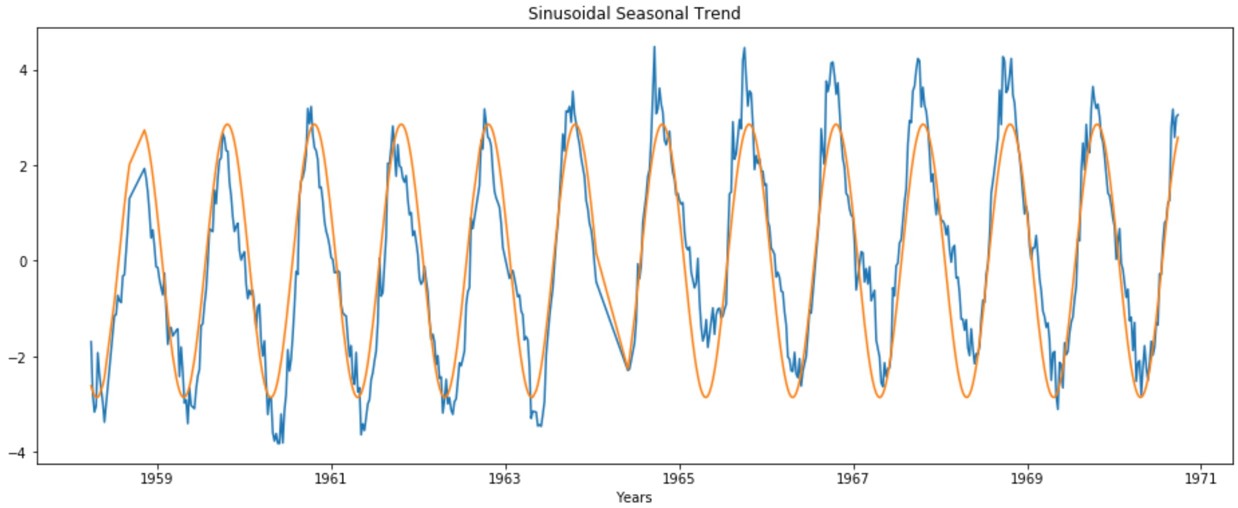


Figure 3. A sinusoidal seasonal trend in orange and un-trended residuals of the quadratic trend line in blue.

As we can see in Figure 3, the sine is slightly left-tilted. To account for that, I applied a tilted trigonometric function in my model, given by $\sin(x + \frac{\sin(x)}{2})$(Stackexchange, 2017).

$$y = A\sin(2\pi t + \frac{\sin(2\pi t+\phi)}{2} + \phi)$$

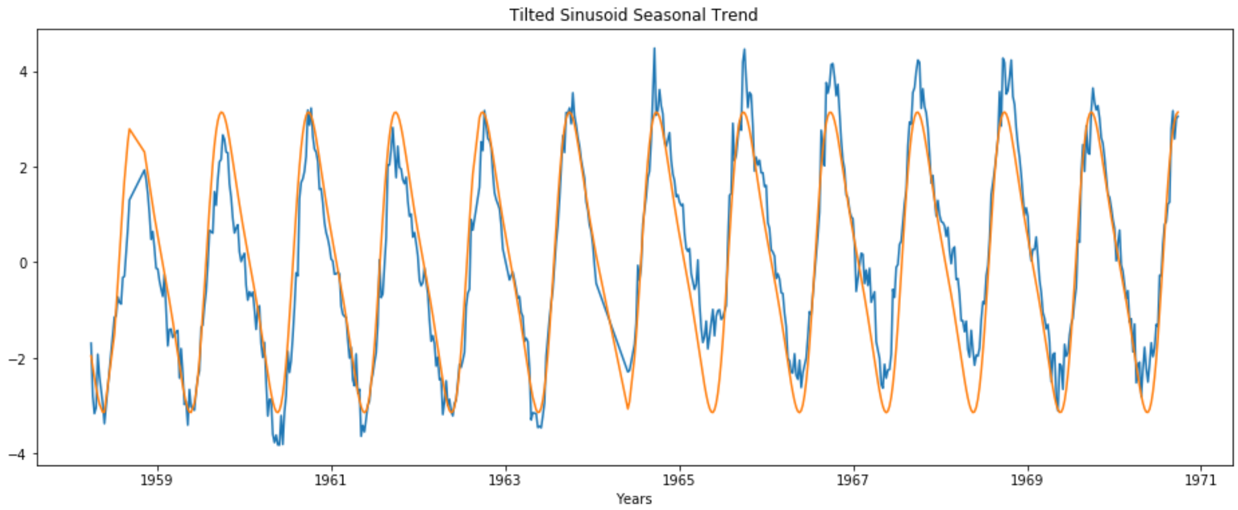


Figure 4. A tilted sinusoid seasonal trend in orange and untended residuals of the quadratic trend in blue.

In Figure 4, we can see that the new tilted sinusoidal model fits the peaks better. To confirm that, I compared the RMSEs for the two seasonal trends:

RMSE for the sinusoidal trend is equal to 0.9714038071387873.
RMSE for the tilted sinusoidal trend is equal to 0.80703599944415.

Therefore, I move forward with the tilted sine model to combine it with the quadratic trend and the arrive at the following cumulative trend:

$$y = c_0 + c_1 t + c_2 t^2 + A\sin(2\pi t + \frac{\sin(2\pi t+\phi)}{2} + \phi)$$

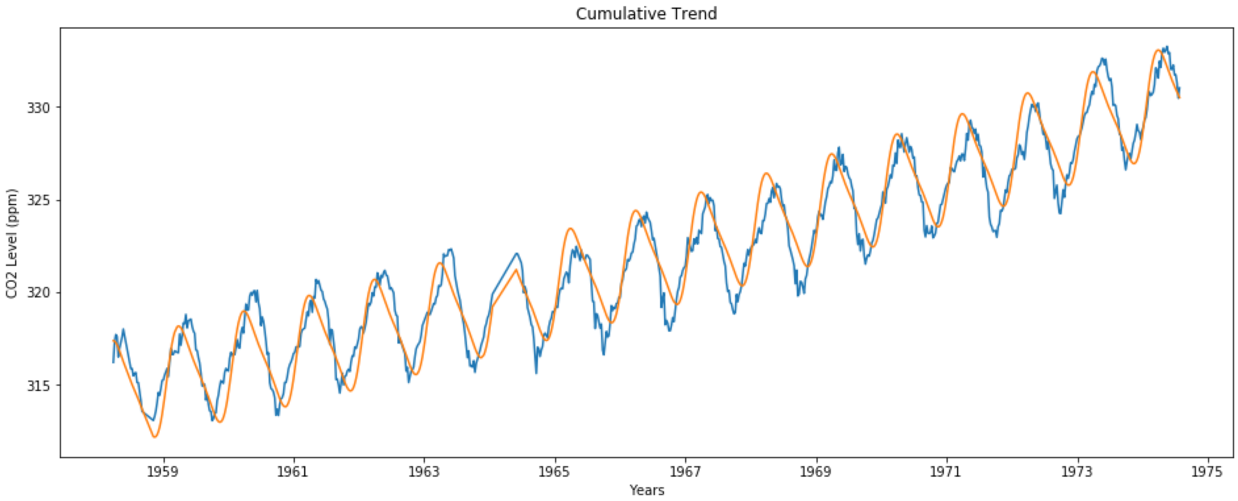


Figure 5. A cumulative trend model in orange and CO₂ levels data in blue.

As for my noise model, the only modeling technique that allowed me to achieve growing confidence intervals was making the standard deviation in my normal likelihood function a function of time as the rest of the parameters. However, I believe such a noise model is not justifiable. My model is not autoregressive, meaning that the samples are independent of each other. For example, a sample from 50 years from now has not been obtained using a sample from 5 years from now at any stage, and a sample from 50 years from now is just as uncertain as one 5 years from now. Hence, I choose to leave my noise model to be independent of time, and sample the sigma from a normal distribution, centered at 0 with a standard deviation of 10.

IV. Factor Graph

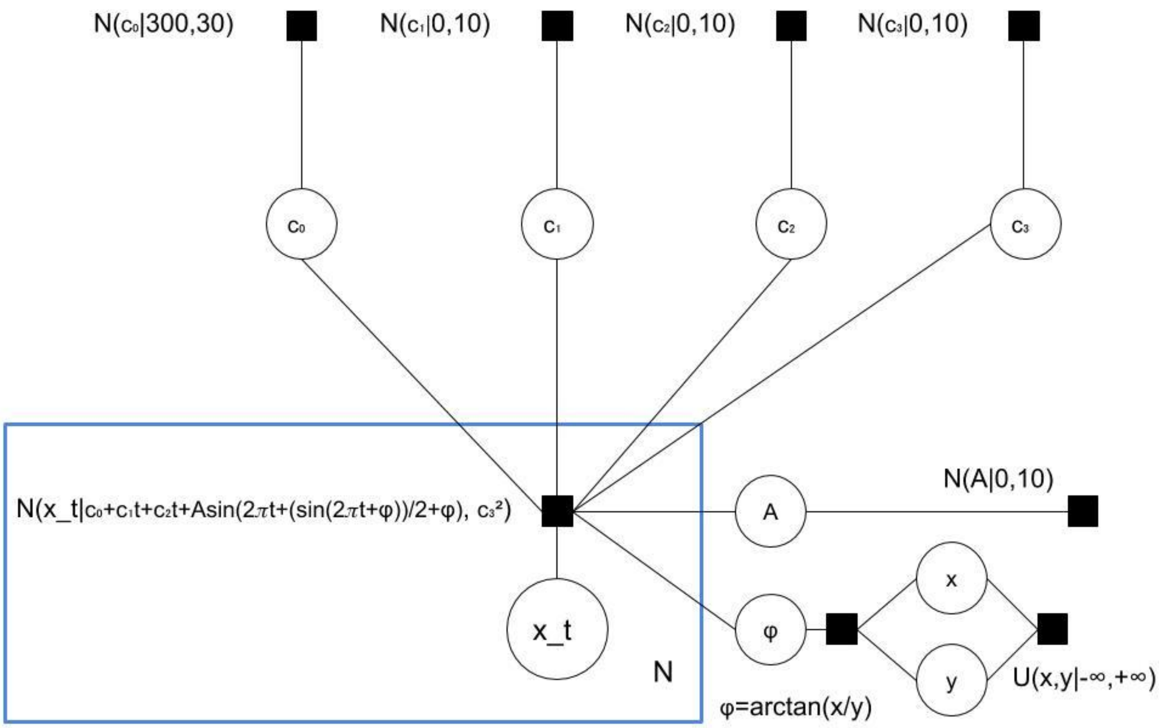


Figure 6. A factor graph of the CO_2 levels model.

V. Predictions

In Figure 7, we can see where the true observations meet the prediction.

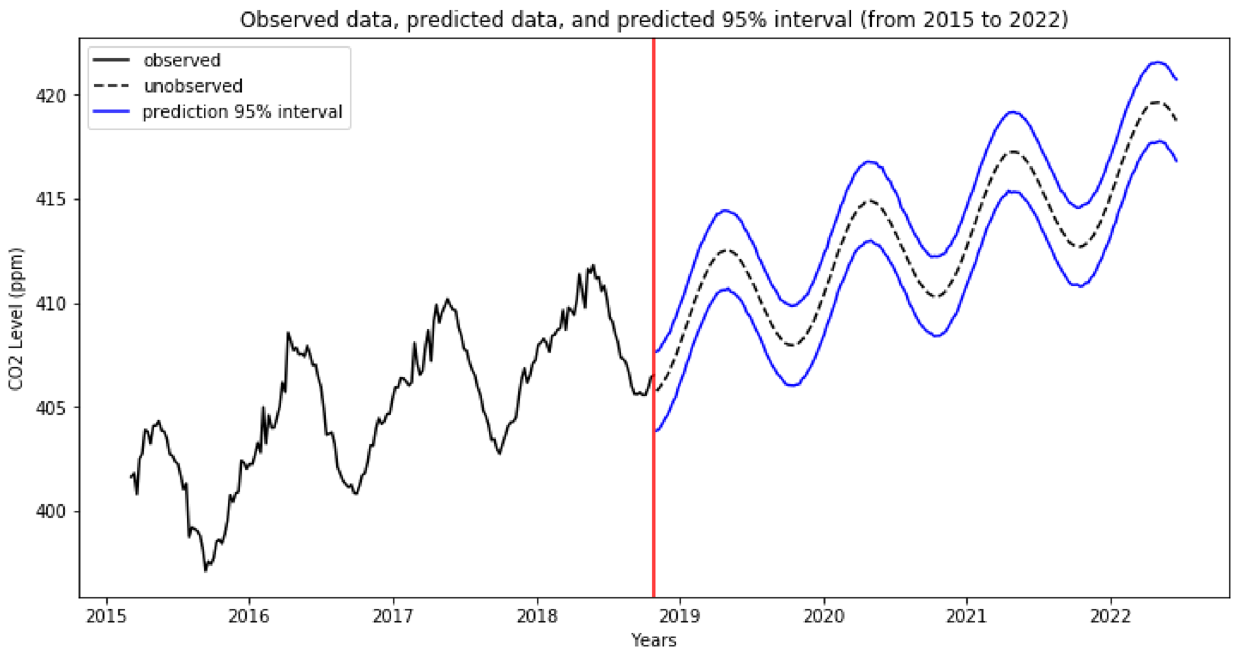


Figure 7. Observed data, predicted data, and predicted 95% interval (from 2015 to 2022).

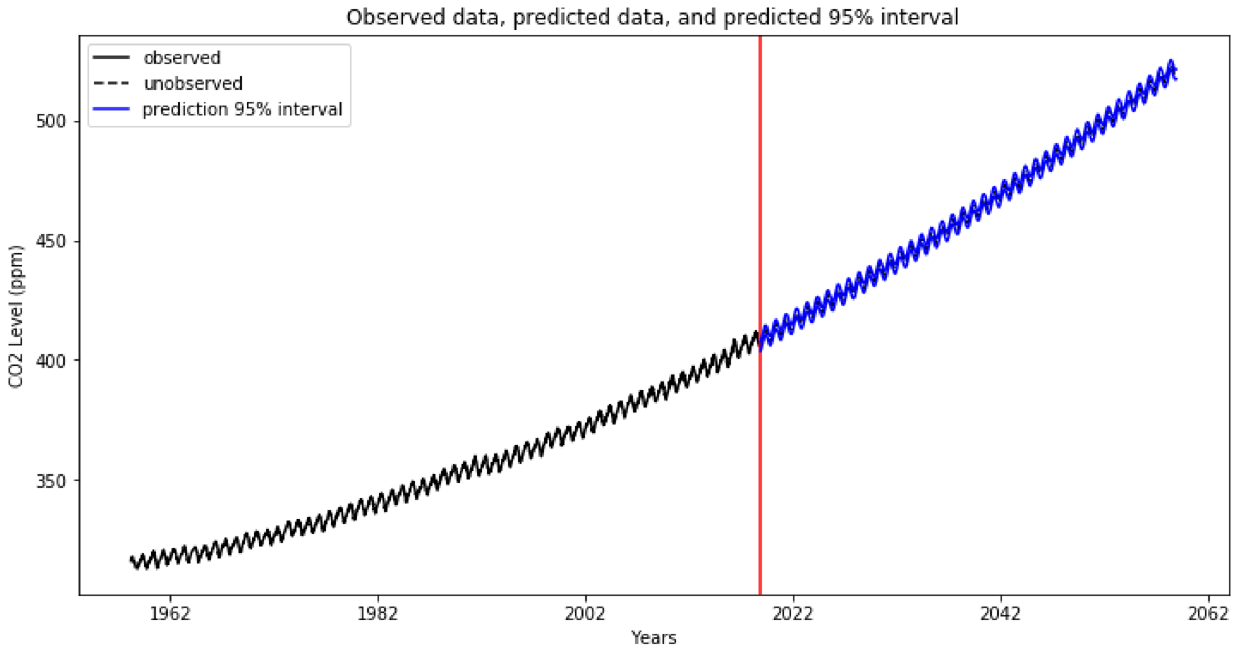


Figure 8. Observed data, predicted data, and predicted 95% interval from 1958 until 2058.

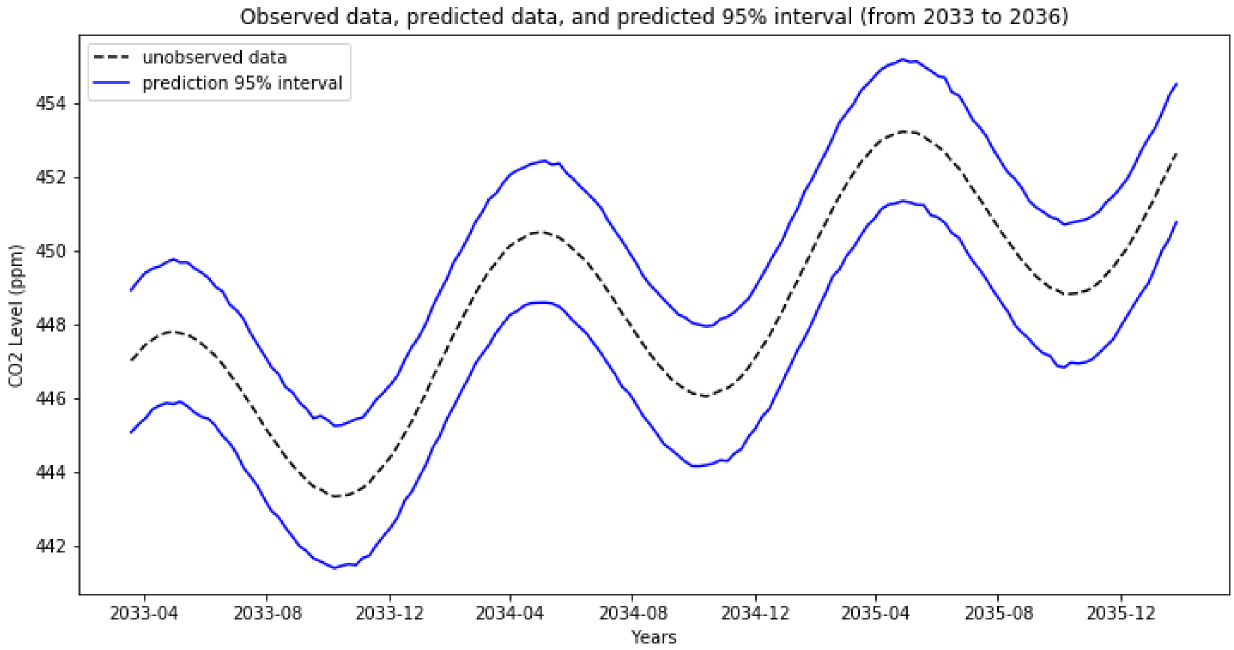


Figure 9. Predicted data and predicted 95% interval from 2033 until the start of 2036.

The estimate for the atmospheric CO_2 levels until the start of 2058 is 518.5 with the 95% confidence interval projected is between 516.5 and 520.5 ppm. The CO_2 level of 450 ppm that is considered high risk for dangerous climate change will be approximately reached between the end of February, 2034, and the middle of March, 2035, with the highest likelihood of occurring on April 1, 2034. If I were to use a wider confidence interval of 99%, the threat of hitting a 450ppm mark would be present starting July-August, 2033.

VI. Complications

I encountered a few difficulties while building the model. Firstly, when modeling the seasonal component, I was faced with a multimodality problem. Because of the trigonometric identity that $\sin(x + \pi) = -\sin(x)$, the amplitude and the phase shift, modeled as a periodic variable, were compensating for each other which required me to restrict the amplitude to only the positive values.

Additionally, the x and y values that were used to build the phase were very heavily correlated. However, this correlation is expected, since ϕ is ultimately modeled as a constant, given by the arctangent of the ratio of x and y to produce the principal value, therefore, we can tolerate their correlation.

VII. Model Improvements

Firstly, I did not achieve widening confidence intervals to account for the uncertainty that comes with estimations overtime. However, as I discussed in the Model Comparison section, I did not find a way to incorporate such a noise component that would be consistent with the properties of the independent sampling. Additionally, I could not account for the sharp peaks of the seasonal component, as well as small fluctuations within a year.

To conclude, the CO_2 levels are increasing quadratically and are likely to increase to a dangerous mark of 450 ppm by mid-2034. This should serve as a warning and should inform environmental policies around the world.

References

Jaideep Khare. Equation of a "tilted" sine. (2017). Retrieved from https://math.stackexchange.com/questions/2430564/equation-of-a-tilted-sine/2430662

Keeling Curve. (2018). Retrieved from https://en.wikipedia.org/wiki/Keeling_Curve

Stan Modeling Language User's Guide and Reference Manual, Version 2.2.0. (2014). Retrieved from http://www.datascienceassn.org/sites/default/files/Stan%20Modeling%20Language%20User%27s%20Guide%20and%20Reference%20Manual%2C%20Version%202.2.0.pdf