Modelling and Forecasting CO2 Data

Hung H. Nguyen

Minerva Schools at KGI

CS146 – Spring 2020

I, Context

Global warming is now an international issue of high concern. CO2 measurements have been the key supporting evidence for the argument that we should take action to alleviate global warming now more than ever. In this assignment, I will fit a model to the weekly Mauna Loa dataset in order to predict where CO2 levels will be in the future and provide further evidence to this argument.

II, The Data

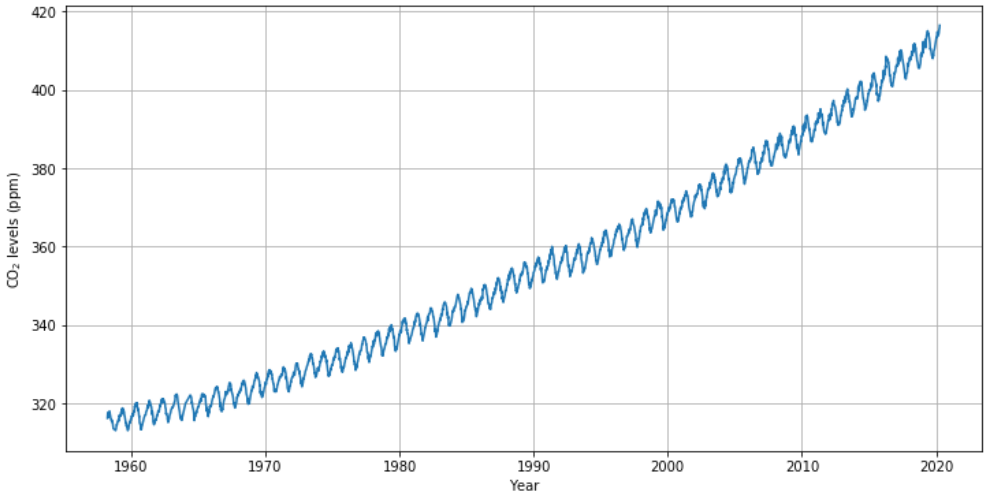


Figure 1: CO2 measurements from the Mauna Loa dataset, full date range shown.

Figure 2: CO2 measurements from the Mauna Loa dataset, from 2017 to 2020.

Observing the data (Figure 1), we see that in general, the overall trend is that CO2 levels are rising, with some curvature associated with this upward trend. Upon closer view (Figure 2), we see that CO2 measurements fluctuate seasonally: they are higher in the winter and lower in the summer. On top of this, there are random variations as well.

III. Assumptions

Because the data has three separate components (trend, seasonality, and noise), my model will also follow this pattern. Specifically, with trend and seasonality, these are the different candidate functions that I tested:

1. Trend
    a. Linear: $c_0 + c_1 t$
    b. Quadratic: $c_0 + c_1 t + c_2 t^2$
2. Seasonality
    a. Cosine: $A cos\left(\frac{2\pi}{365} t + \phi\right)$

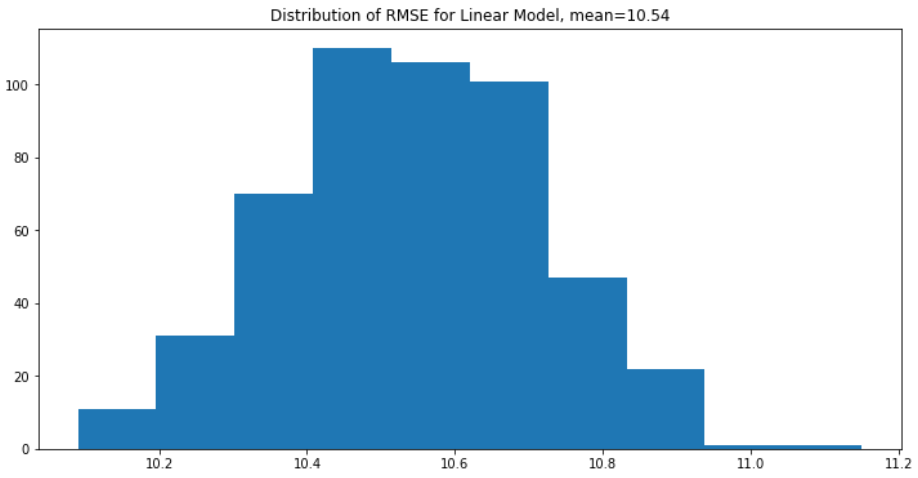
b. Double cosine: $A_1 cos\left(\frac{2\pi}{365}t + \phi\right) + A_2 cos\left(\frac{4\pi}{365}t + \phi\right)$

Regarding trend, I tested a linear trend for completeness because a quadratic trend has a better chance of fitting to the upward curve of the data. As for seasonality, the assumption behind the double-cosine model is to add extra complexity and variability so that the resulting model exhibits a left-skew (Figure 2). Finally, I assumed that the observed data is normally distributed, so that I can easily model the noise by parameterizing σ.

IV. Model choice

From the candidate functions listed in Section III, I tested between three different likelihood functions:

1. Linear trend and Cosine seasonality
2. Quadratic trend and Cosine seasonality
3. Quadratic trend and Double-Cosine seasonality



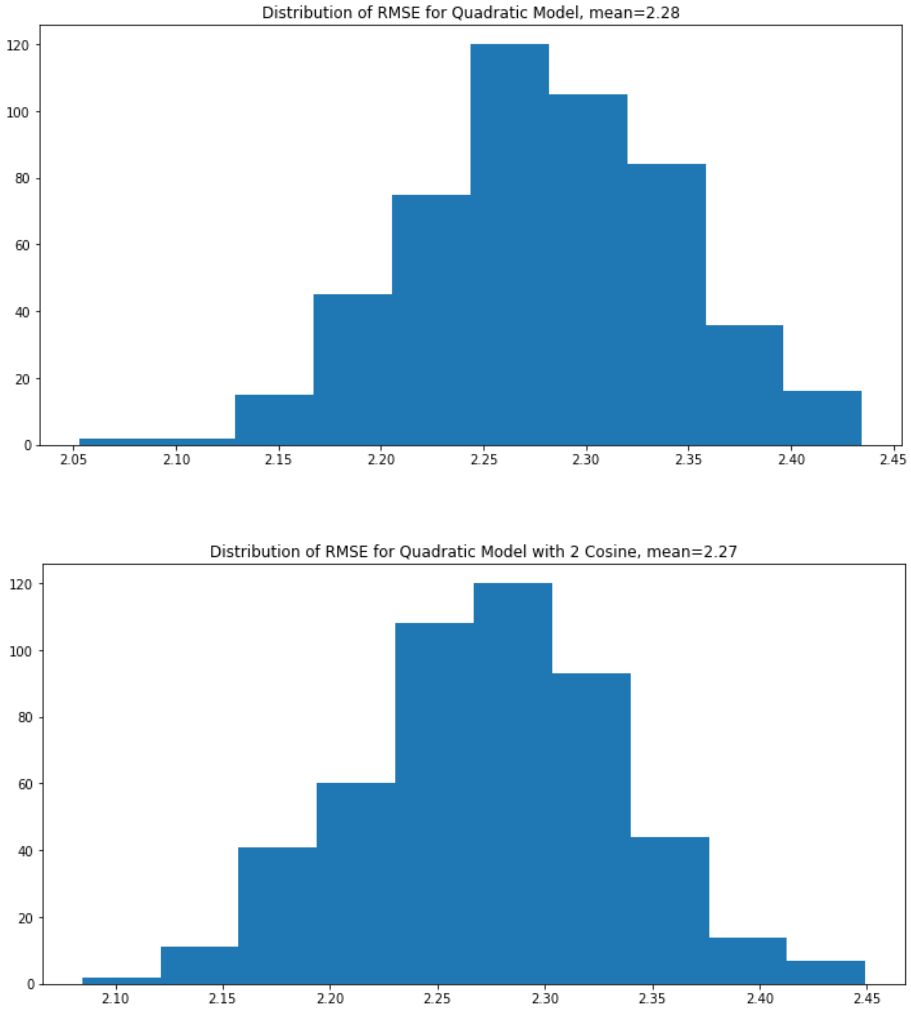


Figure 3. Distributions of RMSE values for each of the models.

I trained each model with 90% of the available data and used the other 10% to calculate an RMSE for each model. Based on the histograms of RMSE (Figure 3), both quadratic variants performed better than the linear variant. Between the quadratic variants, the double-cosine variant has a slightly better average RMSE than the single cosine variant. For the sake of simplicity, I decided to pursue the model with a quadratic trend and single cosine seasonality to conduct further inference. The chosen model is as follows:

$$P(x_t|\theta) = N\left(c_0 + c_1 t + c_2 t^2 + A\cos\left(\frac{2\pi}{365}t + \phi\right), \sigma^2\right)$$

a, Parameters and priors

1. All c_i variables, ϕ, and σ are unknowns. We only know the CO2 measurements from the dataset.
2. c_0, c_1, and c_2 are coefficients in the quadratic formula. The prior for c_0 is $N(315, 2)$ because we know with high certainty where the intercept is, but some uncertainty is still allowed. The priors for c_1 and c_2 are all $Cauchy(0, 2)$ because we want these coefficients to be small yet broad.
3. σ is the noise, and the prior for this parameter is $N(2, 1)$, constrained to be larger than 0 so that noise is positive and small.
4. For the cosine portion of the model, A is the amplitude variable and ϕ is the phase variable. The period (2π) is divided by 365 so that the time unit accepted here are days. Both variables have a broad prior of $Cauchy(0, 1)$.

b, Factor Graph

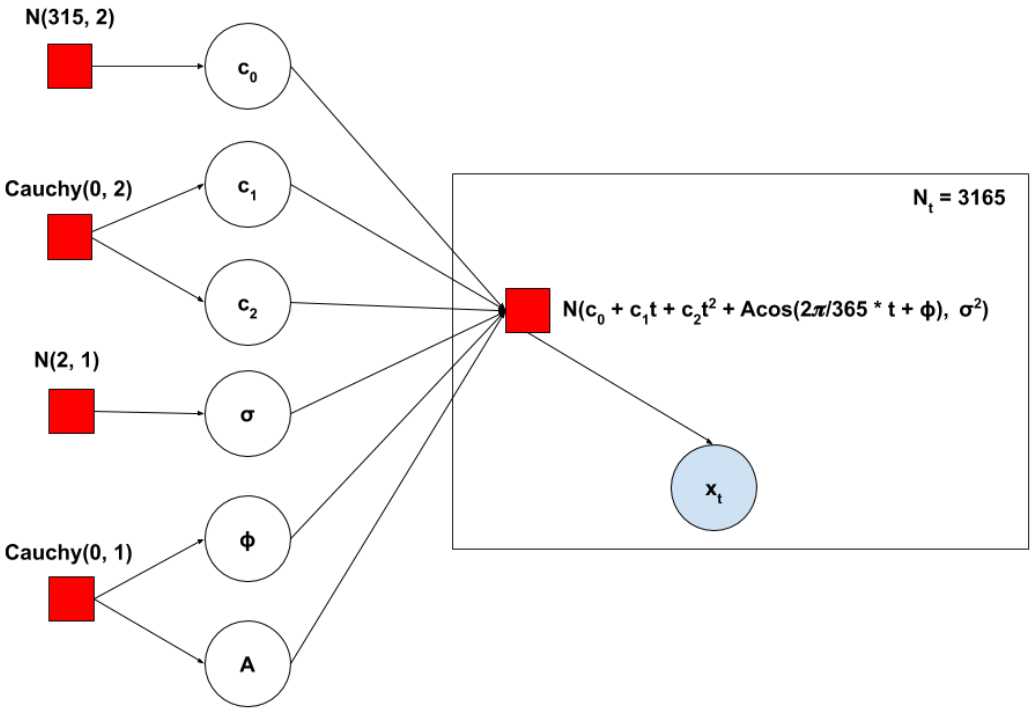


Figure 4. Factor Graph describing the chosen model.[1]

V. Inference

In the process of training the different models with 90% of the data, the main complication comes from modelling the seasonality of the data. For the cosine variant, if there is no upper bound for the ϕ variable, sometimes the Markov chains would not converge because the sample distribution for ϕ would have two modes due to the periodic nature of the function. After constraining the upper bound to π, the Markov chains converges as expected, with good Rhat values and decent effective sample sizes. The double-cosine variant also had the same

[1] #modelling: I created a statistical model, with explanations and justifications, to describe current CO2 data and predict future levels.

issue, but even after constraining ϕ to an upper bound of π, ϕ samples still often exhibited two modes. Changing the upper boundary to $\frac{\pi}{2}$ fixes the issue.

Using the quadratic and single cosine model, the predictions for CO2 levels up until the start of 2060 are shown in Figure 5 and Figure 6.

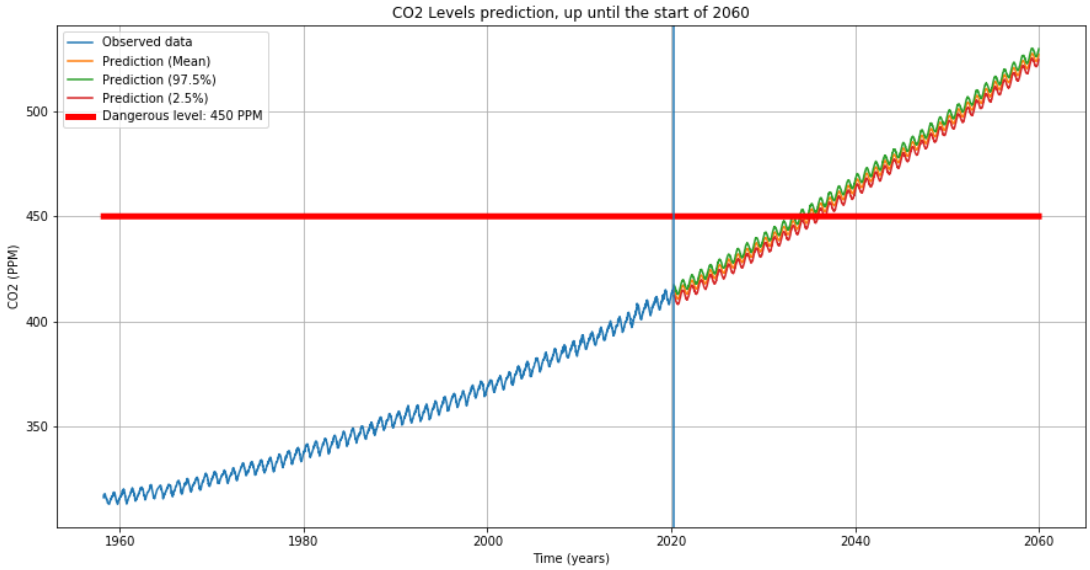


Figure 5. Line graph containing observed data and future predictions (mean and 95% percentile) based on proposed Bayesian model.

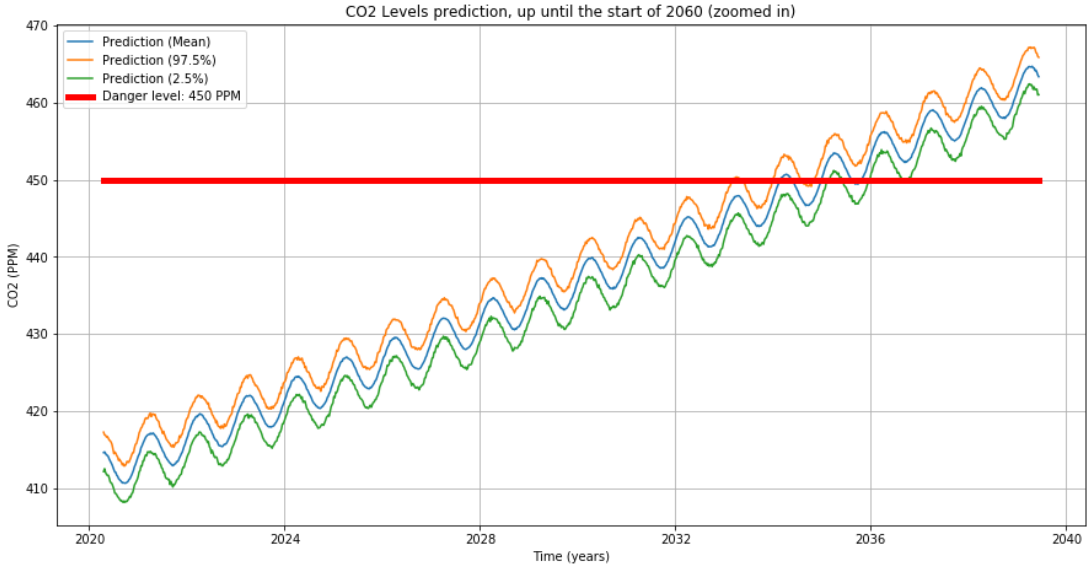


Figure 6. Line graph containing observed data and future predictions (mean and 95% percentile) based on proposed Bayesian model, zoomed in to show data from 2020 to 2040.[2]

The data generation is as follows. For each day, 500 data samples (predictions) are generated using the mean posterior results for all parameters. Then, the mean and 95% confidence interval for that day are calculated and recorded using these 500 samples. The 97.5% percentile can be interpreted as the "worst-case scenario", the 2.5% percentile represents the "best-case scenario", which leaves the mean prediction to represent the "most-likely scenario". It is believed that CO2 levels of around 450 PPM is dangerous for the environment (Willard, 2014). Based on the predictions shown in Figure 5 and Figure 6, we can calculate, with a level of certainty, the different dates at which CO2 levels will reach this threshold:

- **Most likely**, CO2 levels will reach the 450 PPM mark at around **2034-03-03** (intersection with the blue line).

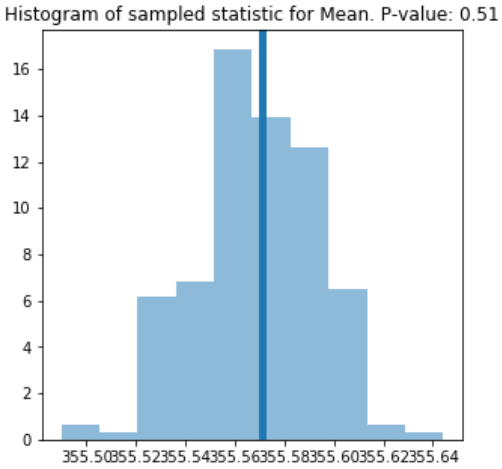
[2] #dataviz: I effectively visualize the results of my analyses in an easy-to-understand format.

- **In the best case**, CO2 levels will reach the 450 PPM mark at around **2035-02-23** (intersection with the green line).
- **In the worst case**, CO2 levels will reach the 450 PPM mark at around **2033-03-11** (intersection with the orange line).

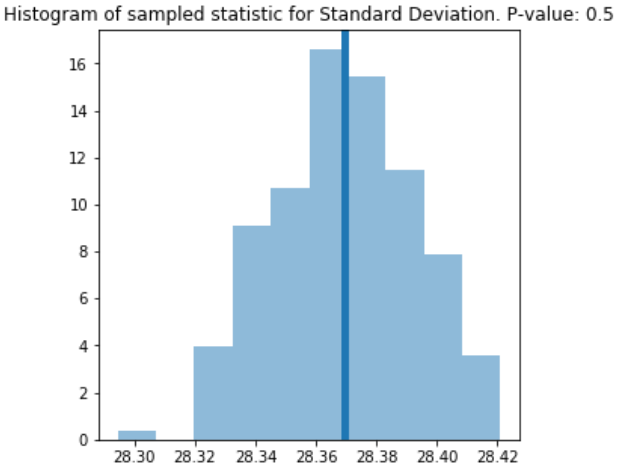
Based on my model, we can reasonability conclude that humanity has around 13 to 15 years until the amount of CO2 in the atmosphere causes serious damage to the environment. This certainly means that we must act now, even if it might be too late in retrospect. Immediate action is better than no action at all.

VI. Model Limitations

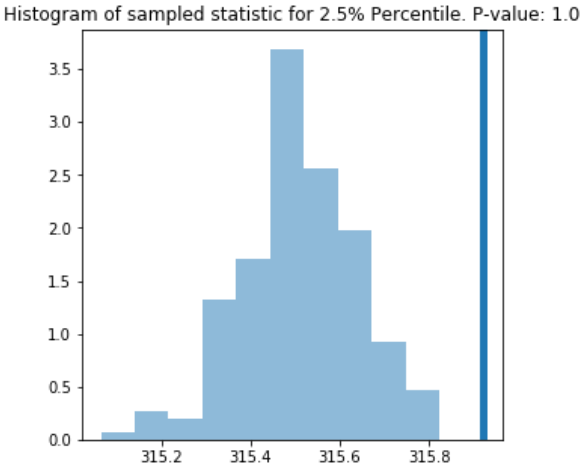
The main limitation of this model is that it does not capture the unique left-skew nature of the seasonality very well, despite my efforts to use a more complicated cosine function. If this issue was fixed, we can achieve a better RMSE value for the test set, which mean better predictive power. A potential solution is to figure out a periodic function that has a left skew within each period. Another limitation is that the 95% confidence interval is quite narrow and inaccurate, which could be the result of the relatively small number of parameters in our current model. Figure 7C and 7D show that the sampled percentiles did not include the actual percentiles. A potential fix is to reparameterize and/or add extra parameters and priors/hyperpriors to include more complexity into the model.



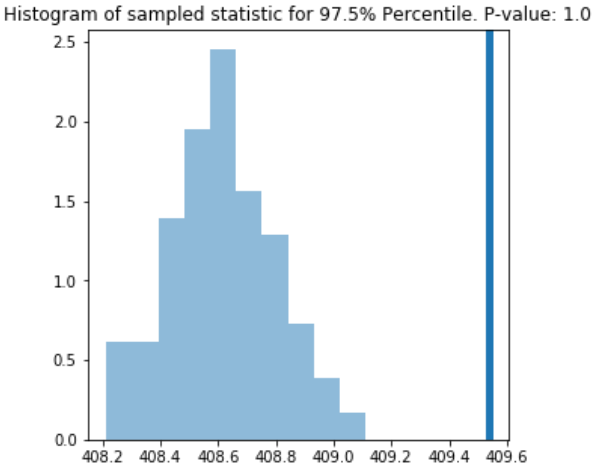
A. Mean



B. Standard Deviation



C. 2.5% Percentile



D. 97.5% Percentile

Figure 7. Posterior predictive checks using different test statistics.

Reference

Willard, B. (2014, January 7). *CO2 – Why 450 ppm is Dangerous, and 350 ppm is Safe*. Sustainability Advantage. https://sustainabilityadvantage.com/2014/01/07/co2-why-450-ppm-is-dangerous-and-350-ppm-is-safe/