CS146 - Final Project

Jasen Lo

Modeling and Forecasting Atmospheric CO_2
From 1958 & Beyond

1 Introduction

Since 1958, atmospheric carbon dioxide measurements have been recorded at the Mauna Loa Observatory in Hawaii. CO_2 levels have been increasing steadily since the start of the industrial revolution in the 18th century. Since the data from Mauna Loa provide direct data on atmospheric CO_2, this data can be used for predicting future atmospheric CO_2. Thus, this report summarises the author's process of using the Bayesian Inference framework to model atmospheric CO_2, and using this model to predict future atmospheric CO_2 levels until the year 2060. Forecasts of future atmospheric CO_2 levels are an important component of guiding global climate policy since high levels of atmospheric CO_2 can severely alter natural physical processes important to preserving existing environmental conditions such as temperature and sea levels that are essential to plant and animal life on Earth.

2 Data

Figure 1 shows the overall trend of atmospheric CO_2 levels from 1958 to the present. Zooming into a 5 year stretch of measurements between 19, one can observe that there are three distinct processes that need to be modelled accurately before the model can be used to generate predictions:

1. Overall upwards trend - There is an obvious upwards trend in the CO_2 levels over time, though it is unclear if this trend is linearly, quadratically or exponentially driven.
2. Cyclical variation - CO_2 absorption by plants and emission and absorption by humans depends on the seasons. Humans burn more stuff to stay warm in winter and plants consume more CO_2 in spring and summer when they grow.
3. Random noise - Random fluctuations in data can be explained by natural fluctuations of CO_2, measurement error, or some other unexplained variables that affect can affect CO_2.

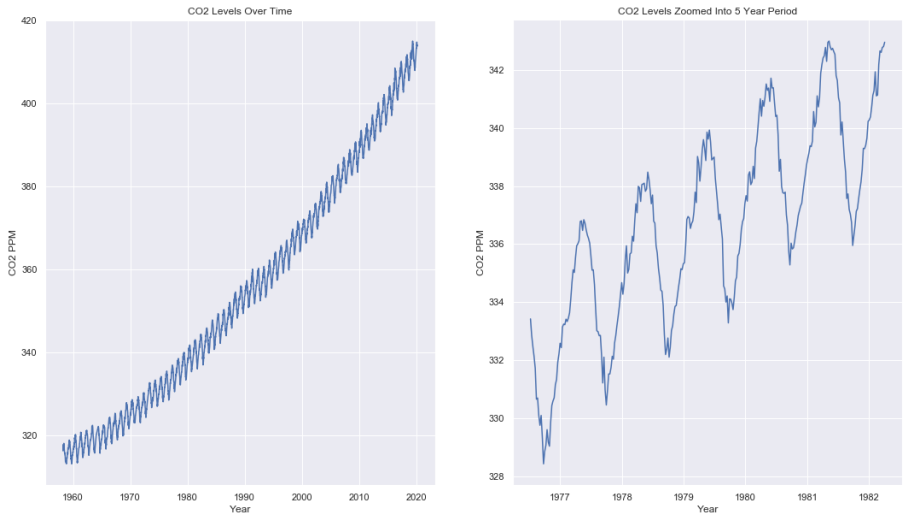

Figure 1: CO2 PPM from 1958 - present (left) and from 1978 - 1982 (right)

3 Model Example

These three forces can be modelled with a following likelihood function:

$$p\left(x_t|\theta\right) = N\left(c_0 + c_1 t + c_2 \cos\left(2\pi t/365.25 + c_3\right), c_4^2\right)$$

- where x_t represents each CO_2 measurement and θ represents the set of all unobserved parameters.
- Long-term linear trend: $c_0 + c_1 t$
- Seasonal variation (every 365.25 days): cosine, $c_2 \cos\left(2\pi t/365.25 + c_3\right)$
- Noise: Gaussian with mean 0 and fixed standard deviation, c_4
- The c_i variables are all unobserved parameters of the model.

However, this example likelihood has two major shortcomings. Firstly, t is highly unlikely that the trend of atmospheric CO_2 is linear. This modelling assumption implies a belief that atmospheric CO_2 coming from contributing forces such industrialised manufacturing and energy production increasing industrialisation in developing countries. Secondly, this model presumes that all of the noise of the model comes from measurement error, since the noise is derived from a Gaussian with a mean of 0 and a fixed standard deviation (c_4 representing measurement area. However, this fails to capture possible natural

fluctuations of CO2 that can occur from a wide variety of natural or man-made processes such as forest fires.

4 Modeling Approach

These two shortcomings of the above model can be addressed by using Markov Chain Monte Carlo methods through Stan to investigate reasonable modelling choices. The author chose to isolate the trend of the data from the seasonal variation and model these two processes separately. Modelling choices were made by a lowest root mean squared error (RMSE) criterion, in order to maximise the model fit to the observed data. Candidate modelling choices were discarded if the Markov chains in the Stan sampling processes were not adequately mixed. The final model is summarised by the factor graph of *Figure 2*, and the final model has the following likelihood function:

$$p\left(x_t|\theta\right) = N\left(c_0 + c_1 t + c_2 t^2 + c_{3a}\cos\left(2\pi t/365.25 + c_{3p}\right), c_4^2\right)$$

- where x_t represents each CO_2 measurement and θ represents the set of all unobserved parameters.
- Long-term quadratic trend: $c_0 + c_1 t + c_2 t^2$
- Seasonal variation (every 365.25 days): cosine, $c_{3a}\cos\left(2\pi t/365.25 + c_{3p}\right)$
- Noise: Cauchy with location 0 and scale 1, c_4

The following priors were used for each of the parameters defined above with accompanying motivations:

- $c_0 \sim$ Normal(300,20). This parameter corresponds to the estimation of the starting measurement of CO_2, and thus was distributed by a normal distribution with mean of 300 and standard deviation of 20 to reflect the historical measurements of CO_2 beginning in the early 1950s. A normal distribution's relatively thin tails is used to signify high confidence in this estimate.
- c_1, $c_2 \sim$ Cauchy(0,1). There is no implicate knowledge in the data that suggests prior information on the distribution of these two parameters other than how they are likely to be positive values. Thus, the Cauchy distribution and its fat tails, allowing for a wide range of values, was chosen as the prior.
- c_1, $c_2 \sim$ Cauchy(0,1). There is no implicate knowledge in the data that suggests prior information on the distribution of these two parameters other than how they are likely to be positive values. Thus, the Cauchy distribution and its fat tails, allowing for a wide range of values, was chosen as the prior.

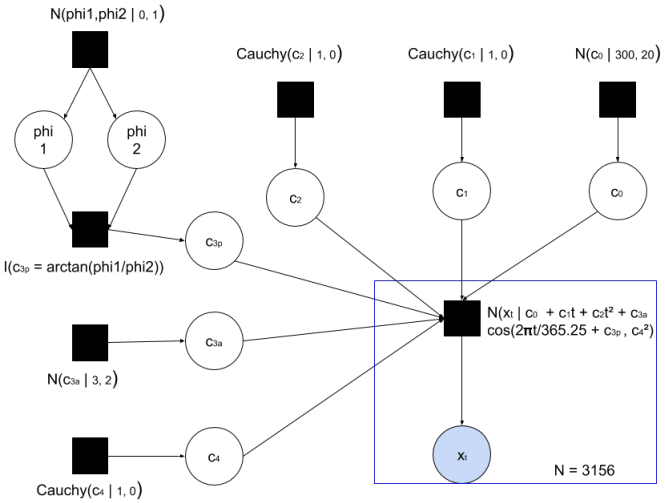

Figure 2: Full Model Factor Graph

- c_{3a}) $\sim$ Normal(3,2). This parameter corresponds to the amplitude of the seasonal cycles of measurements. Looking at *Figure 5* reveals how the amplitude peaks at an absolute value of 4, thus a normal distribution with mean of 3 and standard deviation of 2 accounts for slight increases and decreases of amplitude.

- c_{3p}) $\sim$ Normal(0,1). This parameter corresponds to the phase of the seasonal cycles of measurements. This parameter is restricted to interval [-π, π] by using a unit vector that Stan generates points at random in with independent unit normal distributions, which are then standardized by dividing by their Euclidean length before running through an arc-tangent function.

- c_4^2) $\sim$ Cauchy(0,1). This parameter corresponds to the noise of the data. However, because of the possibility of natural fluctuations, a Cauchy distribution was chosen to represent the prior beliefs on the data noise because of its fat tails, which allow for samples further away from the mean.

4.1 Modeling Trend

Three candidate option were attempted to model the trend of the data: linear, quadratic and exponential. While the exponential model failed to converge, the linear model fit the data worse than the quadratic model, with a higher RMSE. *Figure 3 & Figure 4* shows why the quadratic modelling trend was a much

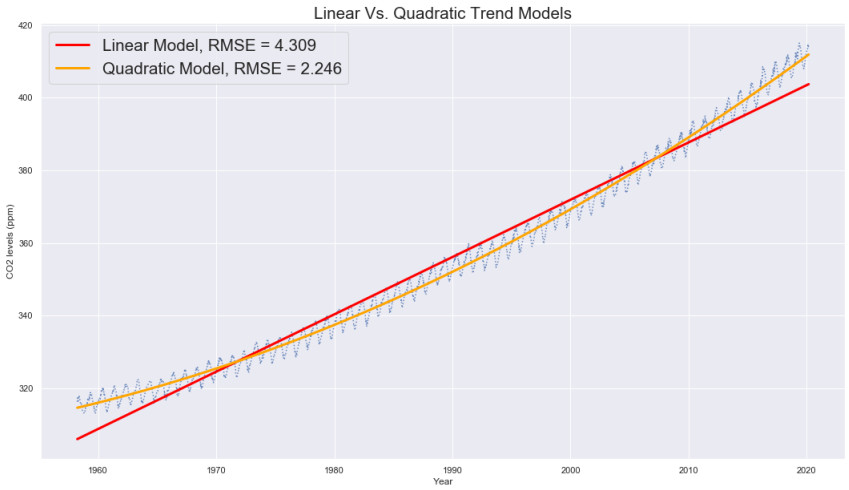

Figure 3: Modeling Trend, Linear vs. Quadratic Model

more suitable choice to model the data compared to the linear modelling trend. *Figure 3* shows how the linear model underestimates data measurements from the earliest and latest dates while overestimating the middle portion of the data. On the other hand, the quadratic model fits the model well. The difference in the two candidate models are highlighted when comparing their predictions in *Figure 4*, as the linear model fails to capture the upwards trajectory of the rising CO_2 levels.

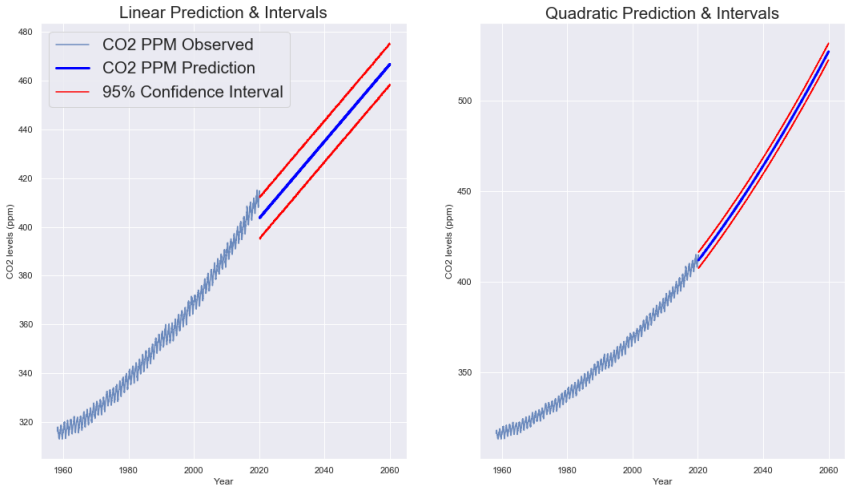

Figure 4: Predictions of the Linear vs. Quadratic Model

4.2 Modeling Seasonal Variation

Figure 5 shows the seasonal variation and noise data that remains after subtracting the quadratic trend data from the observed measurements. The author attempted the following models to model this variation: cosine, sine, and multiple different double sine and cosine models with varying frequencies in order to capture different levels of seasonal variation. However, only the sine and cosine models, which are mathematically equivalent since they only differ by a phase factor, managed to have converged samples. *Figure 6* shows how well the cosine and sine models fit the data, and it is clear they are identical given their similar RMSE and the shape of their estimates. *Figure 7* shows model predictions and their accompanying confidence intervals.

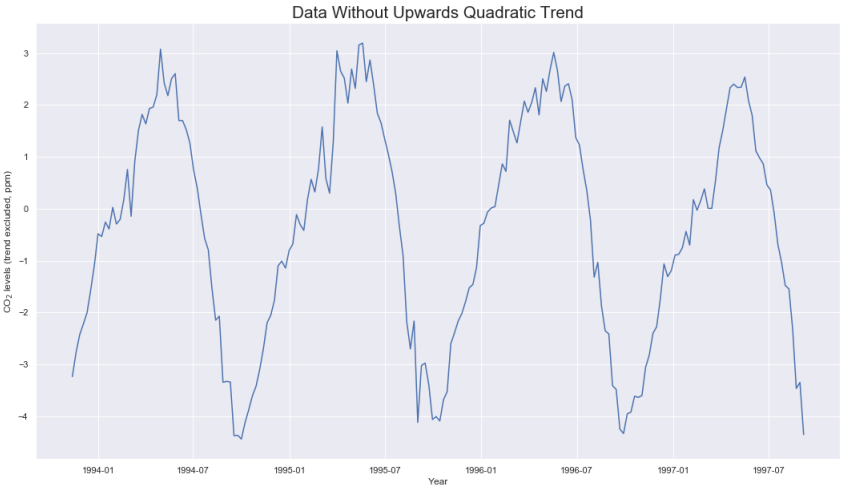

Figure 5: CO2 Data Without Quadratic Upwards Trend

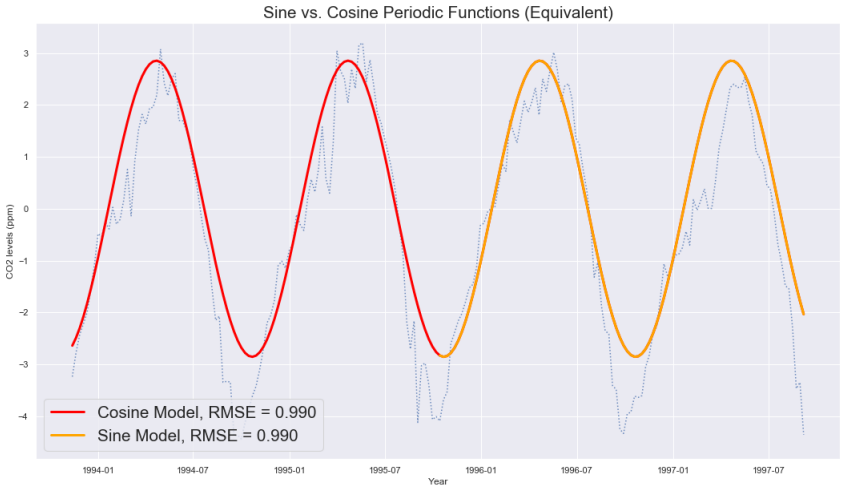

Figure 6: Fit of Cosine/Sine Models onto Untrended Data

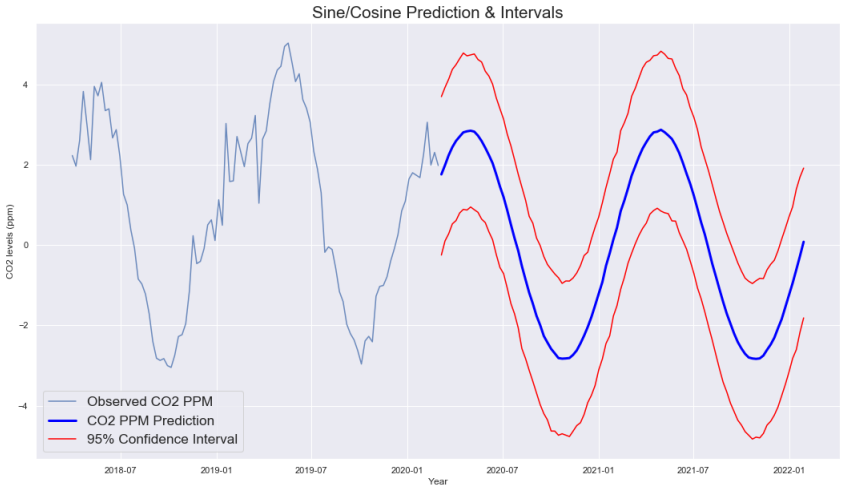

Figure 7: Predictions and confidence intervals of Cosine/Sine Models

5 Predictive Inference Using the Full Model

Figure 8 shows the predictions of the full model up until the year 2060. The model predicts that we will reach an atmospheric CO_2 level of approximately 525 PPM in 2060. CO_2 levels of 450 ppm are considered high risk for dangerous climate change, and the right plot of *Figure 8* shows the dates when the model predicts atmospheric CO_2 level will reach 450 ppm:

- There is a strong possibility (lower bound estimate) that atmospheric CO_2 level will reach 450 ppm in April 2036.
- There is a moderate possibility (mean estimate) that atmospheric CO_2 level will reach 450 ppm in March 2034.
- There is a slight possibility (lower bound estimate) that atmospheric CO_2 level will reach 450 ppm in May 2031.

Given the current data, and reasonable modelling assumptions that were used in this model, activists, policy-makers, government officials and other climate change related individuals must note that there are only between 11 to 16 years before atmospheric CO_2 level will reach critical levels. Thus, it is imperative that policies to slow down the emission of atmospheric CO_2 be put in place and attitudes towards climate change be changed if the world is to reverse the quadratically driven increase in atmospheric CO_2 levels.

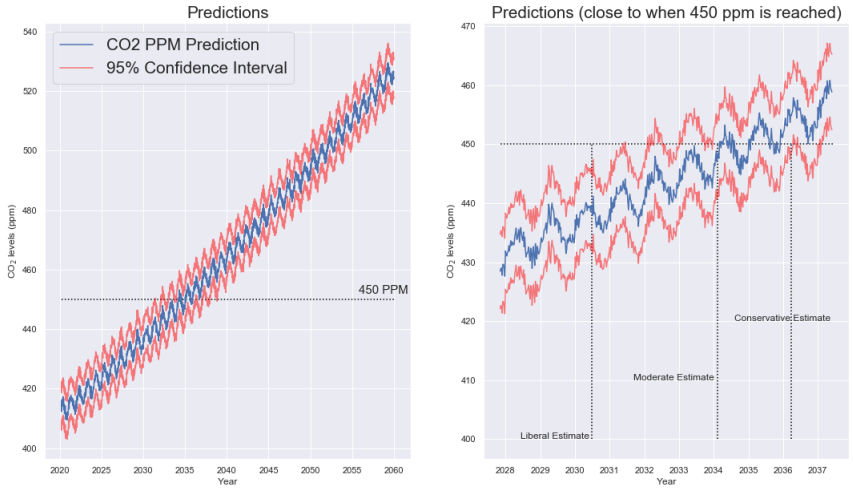

Figure 8: Predictions of the atmospheric CO_2 from 2020 - 2060 (left) and from 2028 - 2038 (right)

6 Shortcomings, Difficulties & Complications

6.1 Unreasonably High Degrees of Confidence in Later Predictions

One shortcoming with this predictive model is the consistent levels of uncertainty that the confidence intervals reflect even for predictions that are 50 years later. Intuitively, confidence in later predictions should decrease due to the increasing degree of possible interventions and changes that can occur after long stretches of time. This is due to the fact that the predictive samples that Stan drew from the posterior are auto-correlated, since each previous data point is highly correlated to the next. Perhaps exploring other predictive inference tools such as Gaussian processes, that naturally account for increasing uncertainty in areas of low probability density, can model the future uncertainty more accurately.

6.2 Periodicity, Multiple Modes & Sampling Time

A significant challenge of working with periodic data is the inherent multimodal nature of the posterior space that MCMC sampling methods that Stan uses struggles with. If the phase is not declared as a periodic parameter, then sampling diverges, becoming trapped in one of the may periodic modes, and produces unreasonable results that are highly auto-correlated. Worse of all, is

that there is no way to effectively debug the sampler when it is diverged until after the Stan has finished sampling, even though sampling can take a lot of time. A solution the author used to some degree of effect was lowering the iteration count to 500 to quickly test out different models, however, this still took a lot of time when Stan had to sample from a posterior that included periodic functions. Despite declaring multiple intermediate unit vectors to restrict the value of the phase variables to be between [-π, π], the author was still unable to create a model that incorporated different periodic functions in order to capture different levels of seasonal variation beyond the yearly level.