A Simple Predictive Model For CO2 Emissions

Tom Kremer

CS146, Fall 2019

Prof. Scheffler

The "Example" Model

The model attempts to explain atmospheric CO_2 levels based on weekly measures taken at the Mauna Loa Observatory between 1958 and 2019 and predict CO_2 levels until 2058. The data are thus (1) a series of dates of measurement, starting from t=0 in March 1958 until 22463 in 2019, (2) the corresponding atmospheric CO_2 levels in PPM (denoted y_i). I began the modeling process by implementing a basic model dubbed' the example model' in Stan, to have an initial running model that I can improve. This model has three components, corresponding to the following parameters:

1. Long-term trend: $C_0 + C_1 * t_i$
2. Seasonal variation: $c_2 cos(\frac{2\pi t_i}{365.25 + c_3})$
3. Noise: $Normal(0, C_4)$

This leads to the following likelihood function: $p(y_t|\theta) = N(c_0 + c_1 t_i + c_2 cos(\frac{2\pi t_i}{365.25+c_3}), c_4)$.

For priors, I used the following probability distributions (all shown in Figure 1):

1. C_0 : $Normal(306, 5)$. I chose this prior to be roughly around what I knew the CO_2 levels were at the beginning of measurements, with a large enough standard deviation to allow some freedom.
2. C_1 : $Gamma(1, 0.1)$. I assumed that the slope cannot be negative, because it is clearly known that emissions are constantly on the rise. Since we are using t in days, I imagined this should be a small number, and thus most of the mass of this gamma distribution is close to zero.
3. C_2 : $Gamma(2.5, 3)$. This is the amplitude parameter of the seasonal variation, and from a rough examination of other models of CO_2 levels I set a wide prior centered around 5-7 ppm. Since negative amplitudes would just create duplicate functions I constrained it to positive values.
4. C_3 : $Gamma(2.5, 3)$. This is the phase parameter of the seasonal variation around the expected yearly rotation. I assumed that it should not be too large because we expected a yearly rotation, but I used a relatively broad prior to allow for flexibility. I constrained it to positive values for similar reasons as the amplitude.
5. C_4 : $Gamma(1, 0.03)$. This is the noise variable. It thus has to be positive, again

justifying the Gamma distribution. I assumed it would be quite small since we are modeling t by weeks, so set most of the mass to be under 1 ppm.

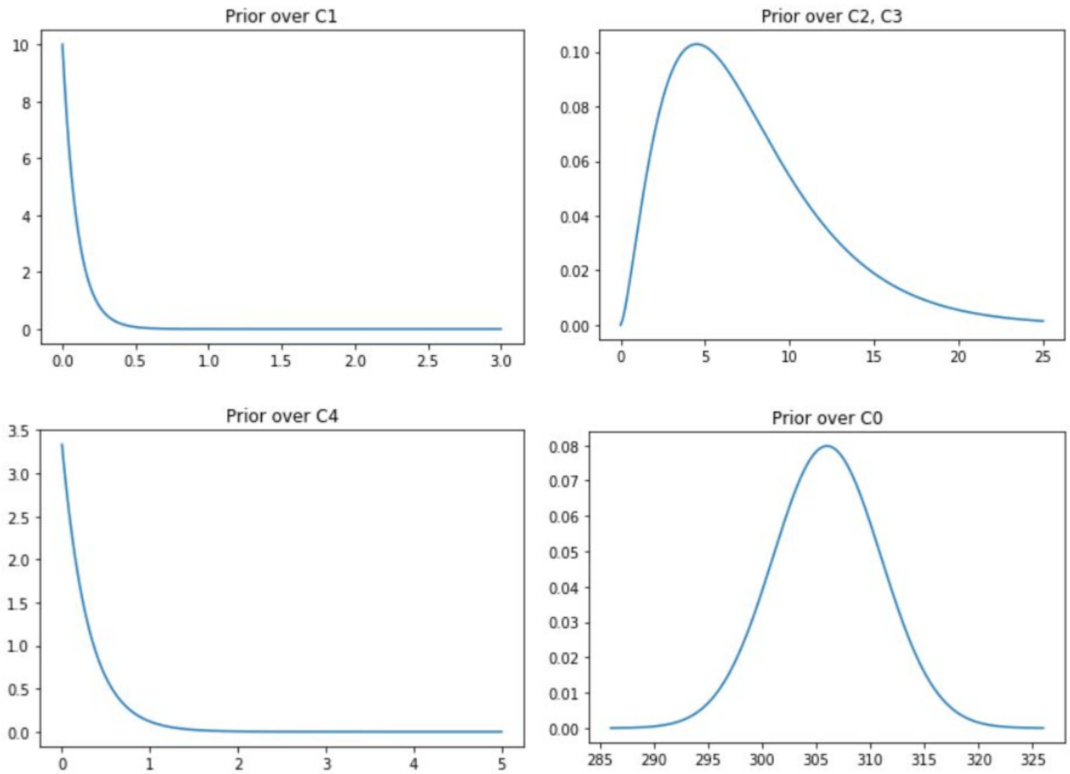


Figure 1. Priors over all parameters.

The full model is thus the product of these five priors and the above-mentioned likelihood function for each of the 3139 observations (I write the full joint distribution for the model I eventually use).

I used 90% of the data to generate a posterior over its parameters and predict the last 10% observed data that can serve as a benchmark for its performance. Since the sample size is big, I expected that this should provide a good indication of its performance in predicting the next 40 years as requested.

The top part of Figure 2 presents the Example Model prediction vs the observed last 10% after the observed data (green) used for fitting it. The horizontal axis is the time in days from the first measurement as instructed, and the vertical red line marks the end of the 'training'. It's clear that the model is not a good fit. Since the model assumes the long-term trend is linear, I plotted the line of best fit using the mean values of the posteriors over the intercept C_0 and slope C_1. A linear trend does not seem to capture the curvature in the long-term trend.

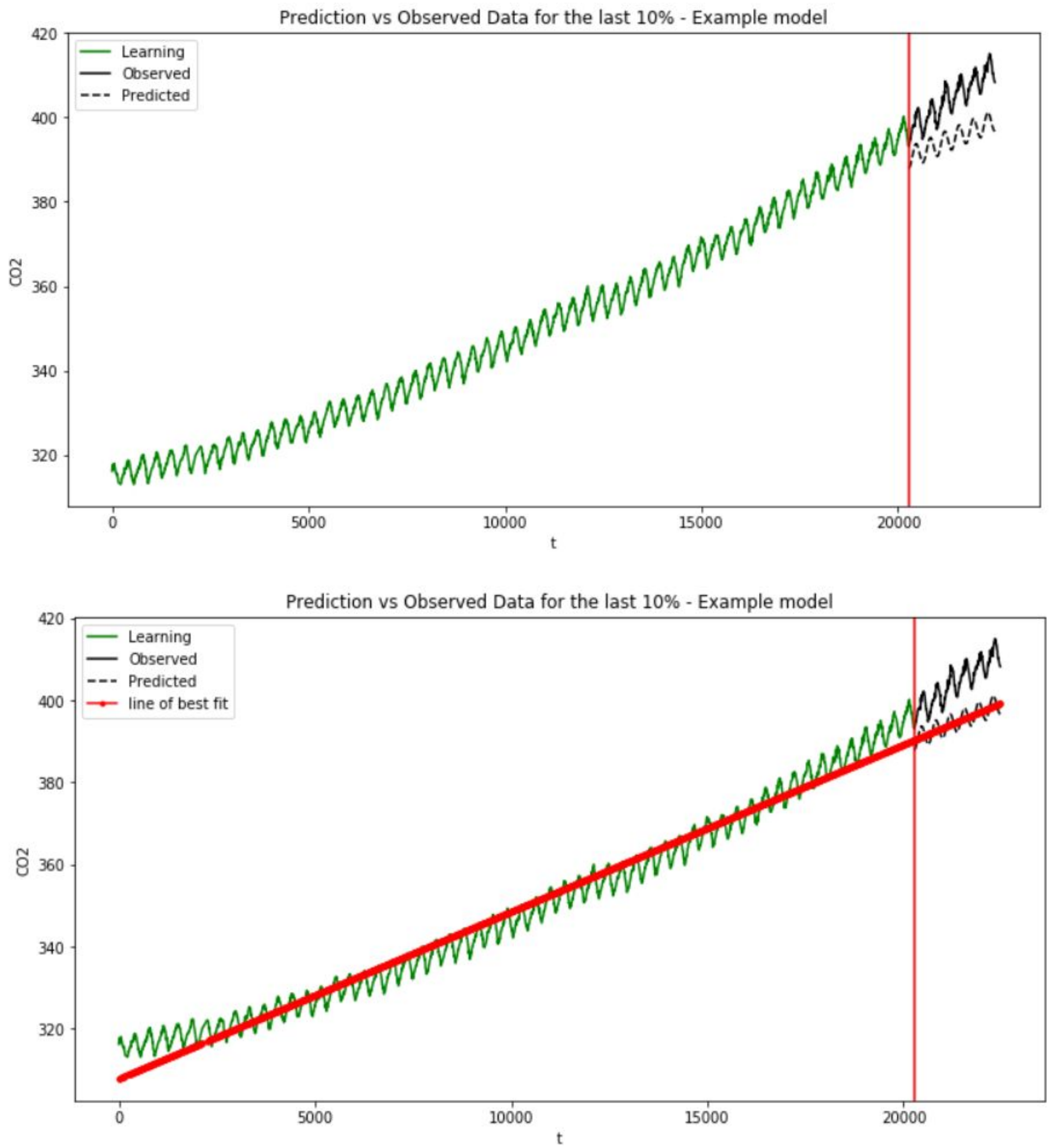


Figure 2. Predictions from the Example Model.

Extension 1: A quadratic term

I added a quadratic term (*'quad'* parameter) to better model the apparent curvature. The model remained the same, except that the component of the long-term variation $C_0 + C_1 * t_i$ turned into $C_0 + C_1 * t_i + quad * t_i^2$. Since *t* is in days I imagined this term would be very small, so the prior I chose was *Gamma(1,0.01)* shown in Figure 3.

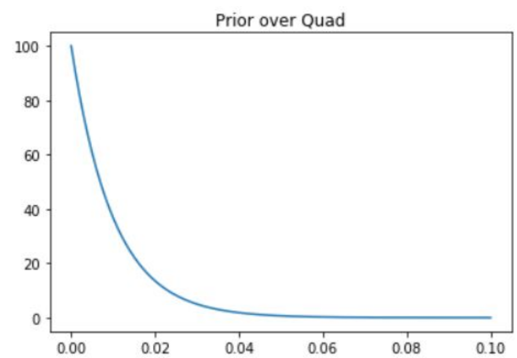
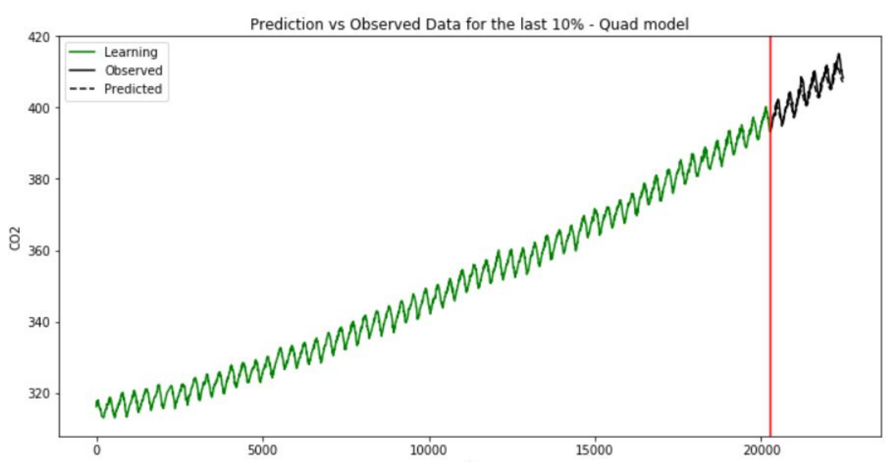


Figure 3.

I followed the same 90/10 procedure as before, and the results are shown in Figure 4. The top part shows that the long-time trend looks accurate this time. However, it is hard to actually see what the prediction is like, so I produced a 'zoomed-in' plot around the 90% cut-off point (bottom part). The smooth curve is the mean of all predictions, that were produced by the 'generated quantities' part of the model. One problem that stood out to me was that the predicted start and end-points of the cycles match the data, but within the cycles, the data is actually not symmetric - the midpoint is skewed.



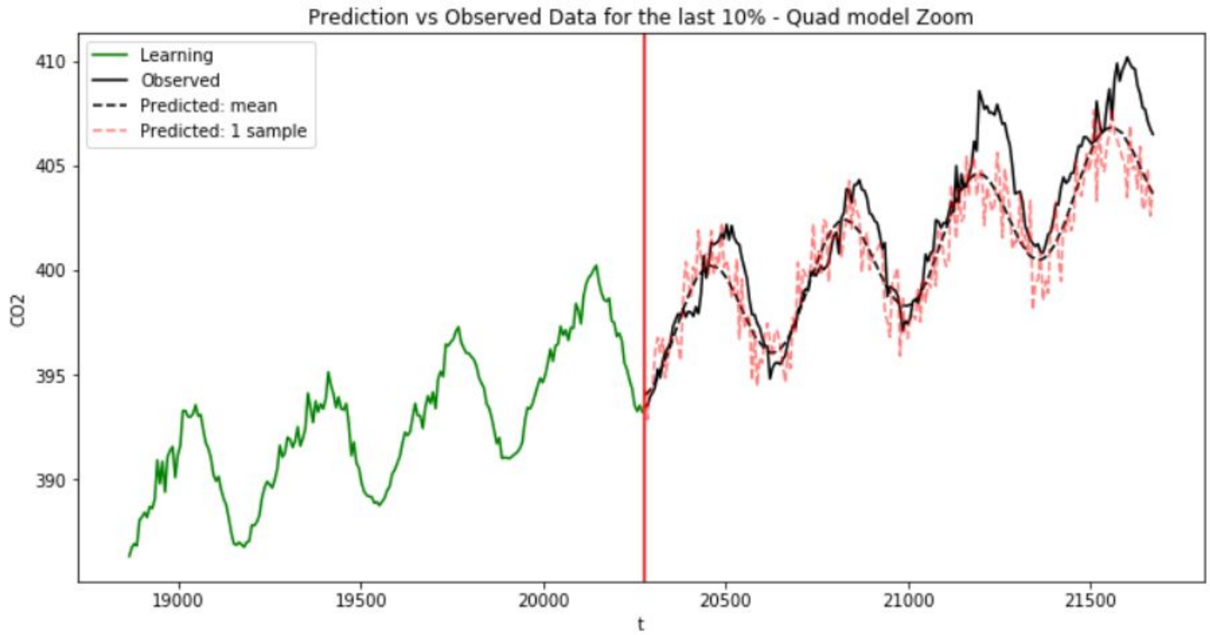


Figure 4. Predictions from the Quad Model.

Before moving on to making further a extension, I produced a quantitative estimate of the Example and Quad models by computing the out-of-sample MSE for the two models based on their mean prediction values for the last 10% of the data (I also calculated the 'mean MSE' from all 4000 predictions for both, the results are qualitatively similar). The Example Model MSE was 9.59 and the Quad Model was 1.486 — a significant improvement.

The Inner Cosine Model

To try accounting for the asymmetric cycles, I added an inner cosine term to the current periodic function, that has the same content, but a weight parameter 'inner'. Thus now the seasonal variation component was: $cos(\frac{2\pi t_i}{365.25+c_3} + Inner * cos(\frac{2\pi t_i}{365.25+c_3}))$. I put the standard Gaussian prior over *Inner* by observing that positive and negative values deviate the symmetry to different directions (and thus both relevant) and that any value beyond 2 or -2 makes the cycle extremely wiggly, which would not fit this data, thus 95% of the mass is between these values, with values closer to 0 being the most probable.

Figure 5 shows the zoom plot compares the predictions obtained from the Inner and Quad models around the cutoff point. Although the posterior mean for Inner was -0.25, which is far enough from zero to show an effect, it is not so clear in this plot. Computing the MSE of the Inner model predictions yielded a result of 1.46, just a little better than before.

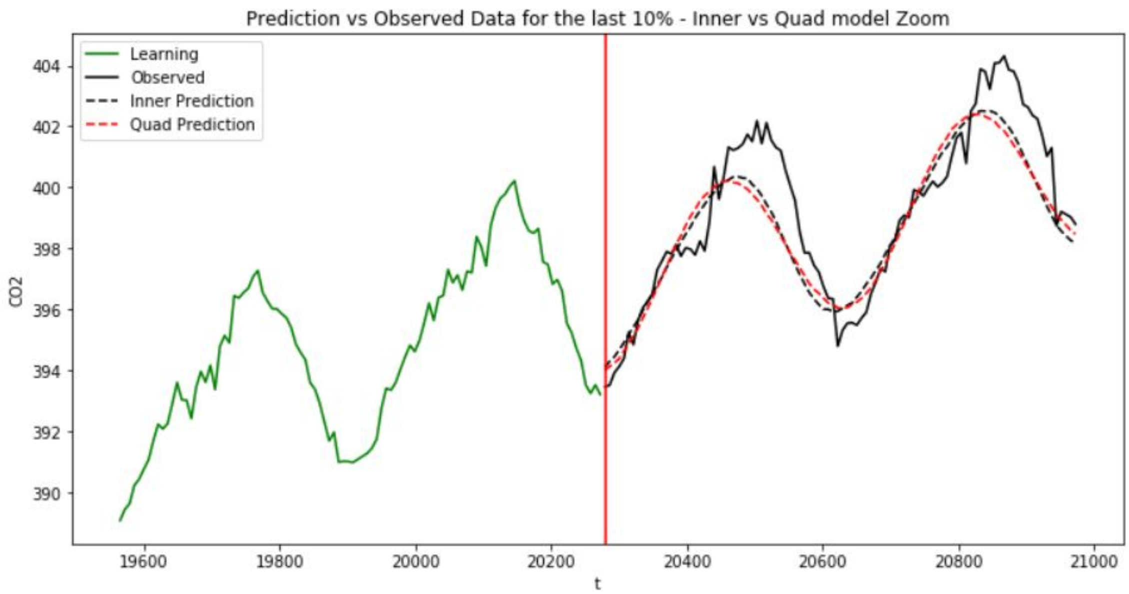


Figure 5. Inner vs Quad model predictions.

For time limitations, and since the predictions were decent, though not excellent, I proceeded with this model. Before moving to do so, I shall specify further apparent limitations and explicate the full form of the model. First, the amplitude seems to be lower than that of the original data. Second, The individual predictions from both quad and inner models are much noisier - the posterior over the noise parameter was 1.26, which is higher than I expected, and also seems to be higher than what the data justified. Finally, closely examining the data, it seems like the noise is smaller in Summer, i.e. after the peak of the seasonal variation than before the peak. Thus, a better model might incorporate two separate noise terms based on the observation's time within the seasonal variation. In general, this model assumes that the parameters are consistent from the beginning of measurement until now, while perhaps amplitude or noise in the variation differed. Another improvement could thus place some weight on these parameters based on *t*. Finally, another assumption of this model is that the current long-term trend will remain the same, which does not leave much room for scenarios of higher emissions or absorptions. Thus, a better model might also incorporate some parameter that relates to the rate of human emissions, as the last few years of data in which humans started addressing CO_2 levels seriously might not have been generated by the same underlying process as when observations began.

The joint probability over all parameters and given the time data t and CO_2 PPM measurements y was as follows: $p(c_0, c_1, c_2, c_3, c_4, Quad, Inner \mid t, y) =$

$$N(C_0|306, 5) * Gamma(c_1|\alpha = 1, \beta = 0.1) * Gamma(c_2|\alpha = 2.5, \beta = 3) * Gamma(c_3|\alpha = 2.5, \beta = 3) *$$
$$Gamma(c_4|\alpha = 1, \beta = 0.3) * Gamma(Quad|\alpha = 1, \beta = 0.01) * N(Inner|0, 1) *$$
$$\prod_{i=1}^{3139} Normal\ (y_i|\mu = c_0 + c_1 t_i + Quad * t_i^2 + c_2 cos(\tfrac{2\pi t_i}{365.25+c_3} + Inner * cos(\tfrac{2\pi t_i}{365.25+c_3})), \sigma = c_4)$$

A factor graph expressing this model:

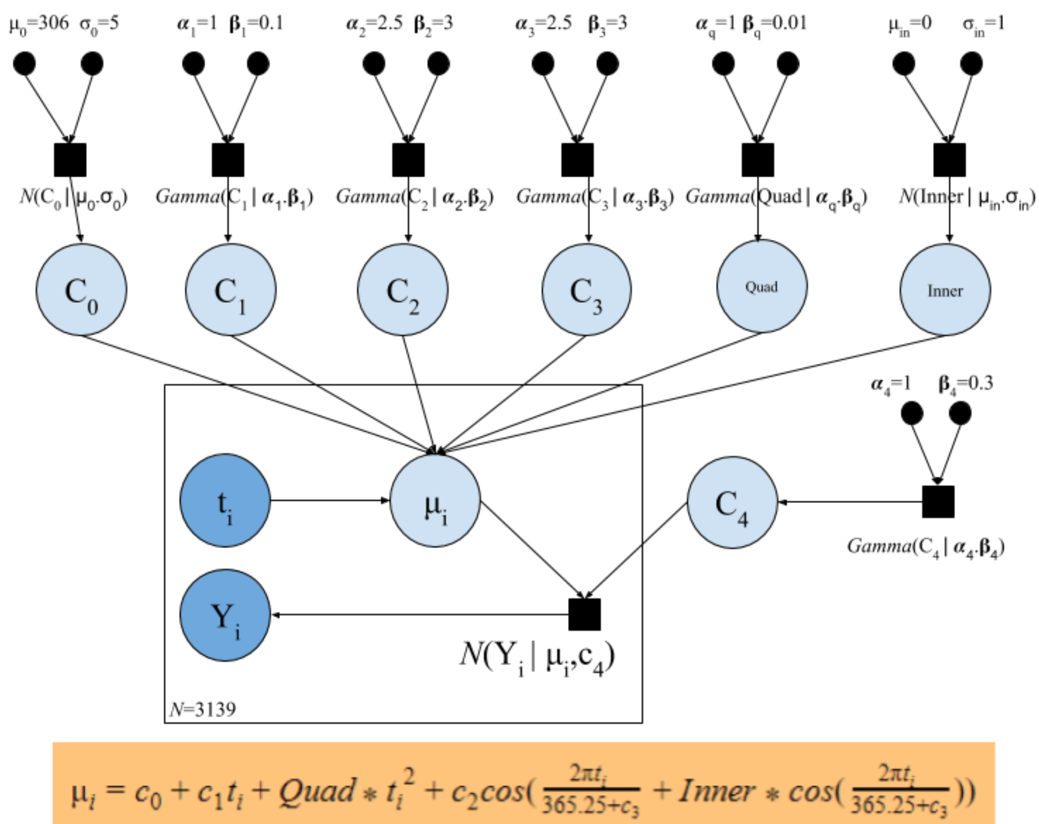
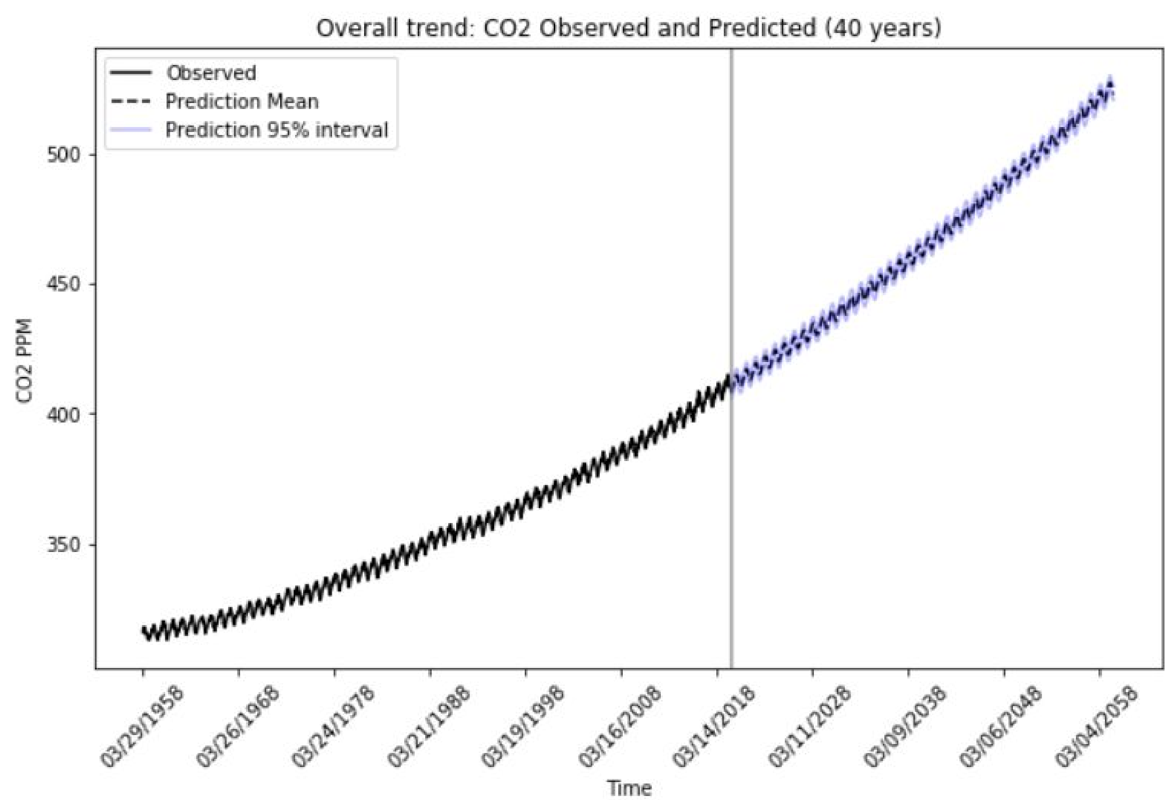


Figure 6. Factor graph representing the Inner model.

Finally, in my initial runs of the Inner model, it seemed to converge well: Rhat was 1.0 for all parameters, there were at least a 1000 samples for each, auto-correlation plots were all around zero for all parameters and all lags, and the pair-plots of the posterior showed smooth, unimodal distributions. However, another run in which I changed nothing did not mix at all. To address this instability, I increased the warm-up period to 3000, still generating 1000 effective samples for each. This seemed to have solved the problem, as far as I could tell (10 successful runs). Whenever the model mixed well, the posterior over all parameters were consistent. Images of the Stan output, autocorrelation plots, and pair plots are in the appendix. I now proceed to discuss the results using this model.

Results

The top part of Figure 7 shows the overall prediction for the next 40 years of CO_2 levels measurements from the Observatory. The horizontal line marks the end of observed data, and the blue lines indicate the 95% confidence intervals, which appear clearer in the zoom-in plot (bottom).



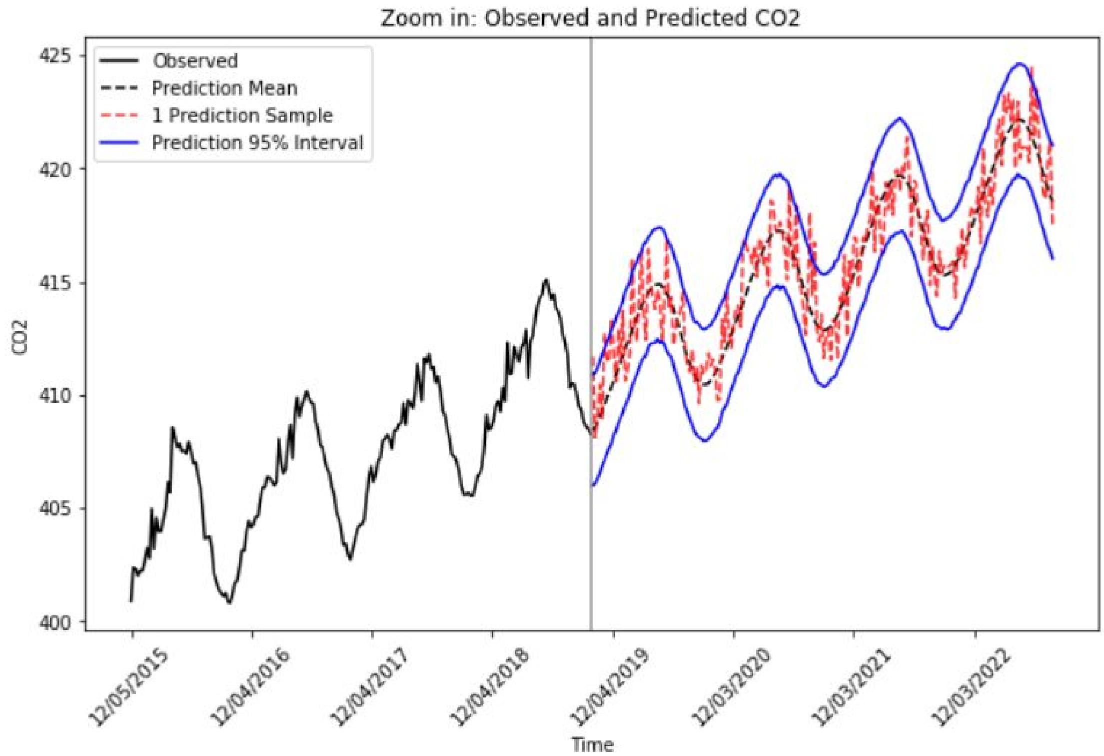


Figure 7. Overall and zoom-in plots of predicted CO_2 levels using the Inner model.

Based on the mean prediction values (of all 4000 'generated quantity' future value samples generating using the posterior samples), the best estimate for CO_2 level at the first expected measurement of 2058 (1/5/2058) is 520.63 ppm, with the 95% confidence interval at [518.09, 523.19].

References

Crippa, M. et. al (2019). Fossil CO2 and GHG emissions of all world countries - 2019 Report. In *Publications Office of the European Union*, retrieved from doi:10.2760/687800.

Hansen, J., Sato, M. & Ruedy, R. (2012). Perception of Climate Change. *Proceedings of the National Academy of Sciences, 109*(37), E2415–E2423.

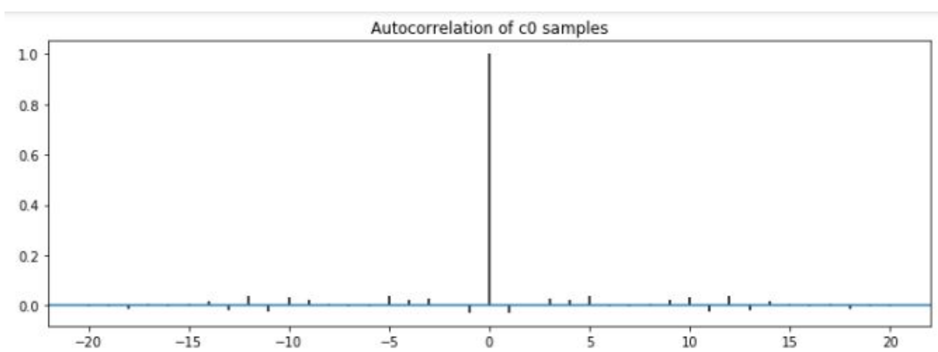
Williams, A. P., Abatzoglou, J. T., Gershunov, A., Guzman‐Morales, J., Bishop, D. A., Balch, J. K., & Lettenmaier, D. P. (2019). Observed impacts of anthropogenic climate change on wildfire in California. *Earth's Future*, *7*, 892–910.

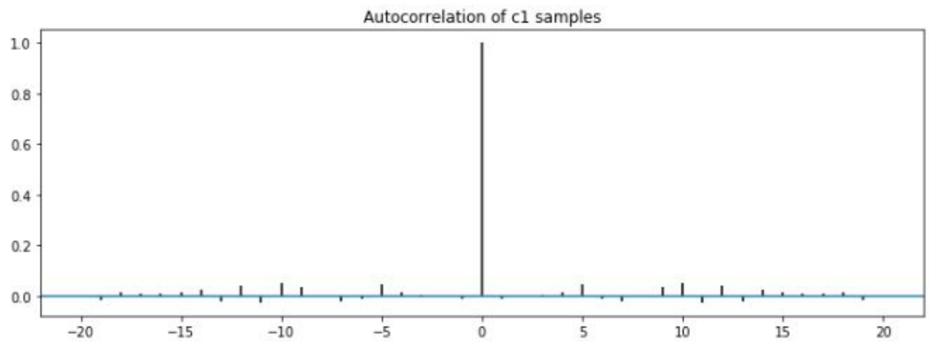
Appendix

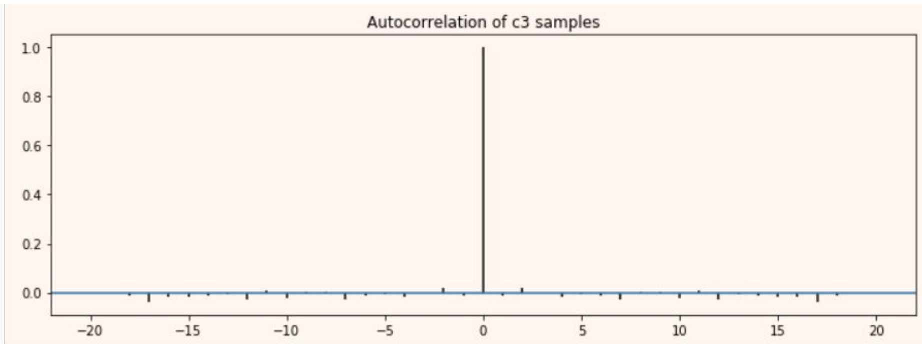
Stan metrics for the final Inner model used to produce the results:

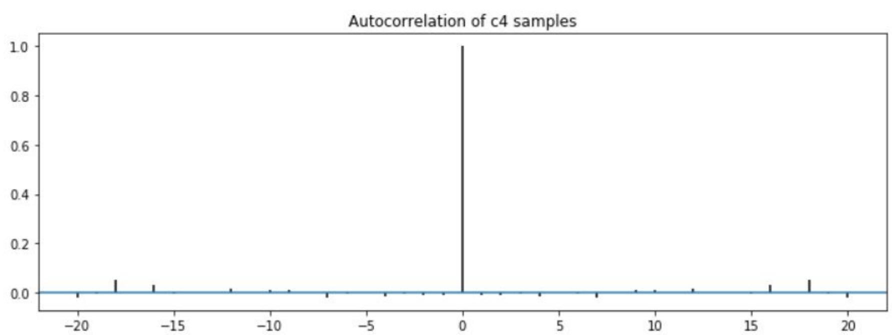
	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
c0	314.58	1.7e-3	0.07	314.44	314.53	314.58	314.62	314.71	1670	1.0
c1	2.1e-3	3.7e-7	1.4e-5	2.1e-3	2.1e-3	2.1e-3	2.1e-3	2.1e-3	1462	1.0
c2	2.69	5.9e-4	0.03	2.62	2.66	2.69	2.71	2.75	3116	1.0
c3	7.0e-4	7.4e-6	4.6e-4	1.2e-4	3.7e-4	6.1e-4	9.3e-4	1.8e-3	3809	1.0
c4	1.26	2.9e-4	0.02	1.22	1.25	1.26	1.27	1.29	3027	1.0
quad	9.7e-8	1.5e-11	6.0e-10	9.6e-8	9.7e-8	9.7e-8	9.8e-8	9.8e-8	1572	1.0
inner	-0.27	3.8e-4	0.02	-0.31	-0.28	-0.27	-0.25	-0.22	3349	1.0

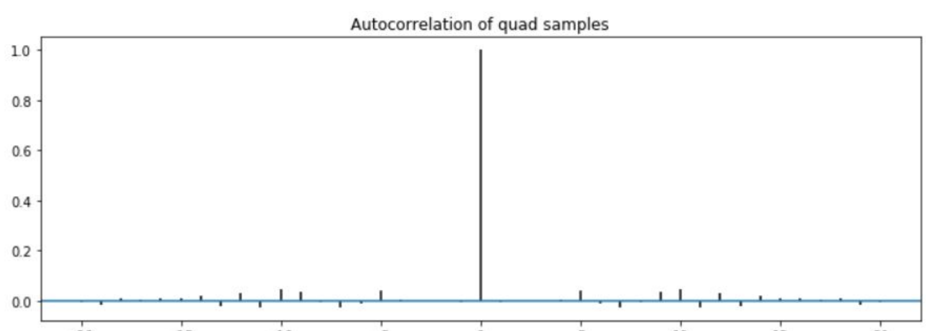
Autocorrelation plots for the same model:

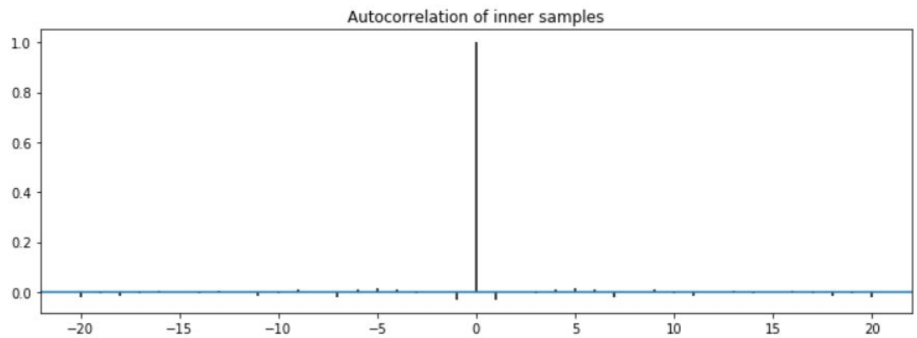



Autocorrelation of c1 samples


Autocorrelation of c3 samples


Autocorrelation of c4 samples


Autocorrelation of quad samples



Pair plots for the same model: