I, Introduction

Climate change is one of the most urgent issues we are facing right now as a society. One strong indicator of global warming is the carbon dioxide level in the atmosphere. If we can model the CO_2 level and estimate what it looks like in the future, we can predict the risk of global warming and have a general idea of when we reach an irreversible and life-threatening level.

This paper looks at the weekly atmospheric carbon dioxide measurements from 1958 until now at Mauna Loa Observatory in Hawaii, models the CO_2 behavior and aims to give an estimate for that for the next 40 years.

II, Model description

1, Statistical model

- **Bayes' Theorem**

$$P(\theta|x) = \frac{P(\theta)P(x|\theta)}{P(x)}$$

x is the observed data, θ is a set of unknown parameters that we are trying to estimate that can help us predict future data. In this case, our data is the CO_2 level that depends on θ and time t (days from the start of the measurements).

- **Likelihood function**

Looking at the data, we can see an increasing trend in the CO_2 level. In this model, I assume that it is a quadratic trend, i.e., x can be modeled by $c + c_0 t + c_1 t^2$.

There are also seasonal variations that are periodic, hence a periodic function such as sine or cosine can be a good choice, e.g., $Acos(\frac{2\pi t}{365.25} + \phi)$, with A being the amplitude of the fluctuations, $\frac{t}{365.25}$ corresponding to a year (because the period for the seasonal change in CO_2 level is one year) and ϕ reflecting the phase – where the measurement lies in the beginning. Finally, there is random noise that follows a normal distribution $N(0, \sigma^2)$.

Therefore, a good distribution that the CO_2 level follows is: $x \sim N(c + c_0 t + c_1 t^2 + Acos(\frac{2\pi t}{365.25} + \phi), \sigma^2)$

- **Priors**

There are 6 parameters (unobserved quantities) $\theta = \{c,\ c_0,\ c_1,\ A,\ \phi,\ \sigma\}$. For $c,\ c_0,\ c_1$, since I have very little information about what they are, their priors should be broad, so Cauchy distribution is a good choice. Because the CO_2 level increases over time, c_1 should be non-negative. To avoid negative CO_2 levels, just to be sure, I also put $c,\ c_0$ to be non-negative.

For ϕ, since the cosine function has 2π as its period, ϕ can be constrained between 0 and 2π. For any α, $cos(\alpha) = cos(2\pi - \alpha)$. Therefore, to avoid symmetric modes, ϕ can be constrained even further between 0 and π. ϕ could equally likely take any value within that range so its prior can be a uniform distribution.

I let A and σ follow normal distributions. Because A is the amplitude, it should also be nonnegative. σ is standard deviation, so it should also be non-negative.

- **Data**

There are two known variables – time t and CO_2 level x. I divided the dataset into a training set and a test set which will be used later to evaluate the accuracy of the model.

2, Factor graph

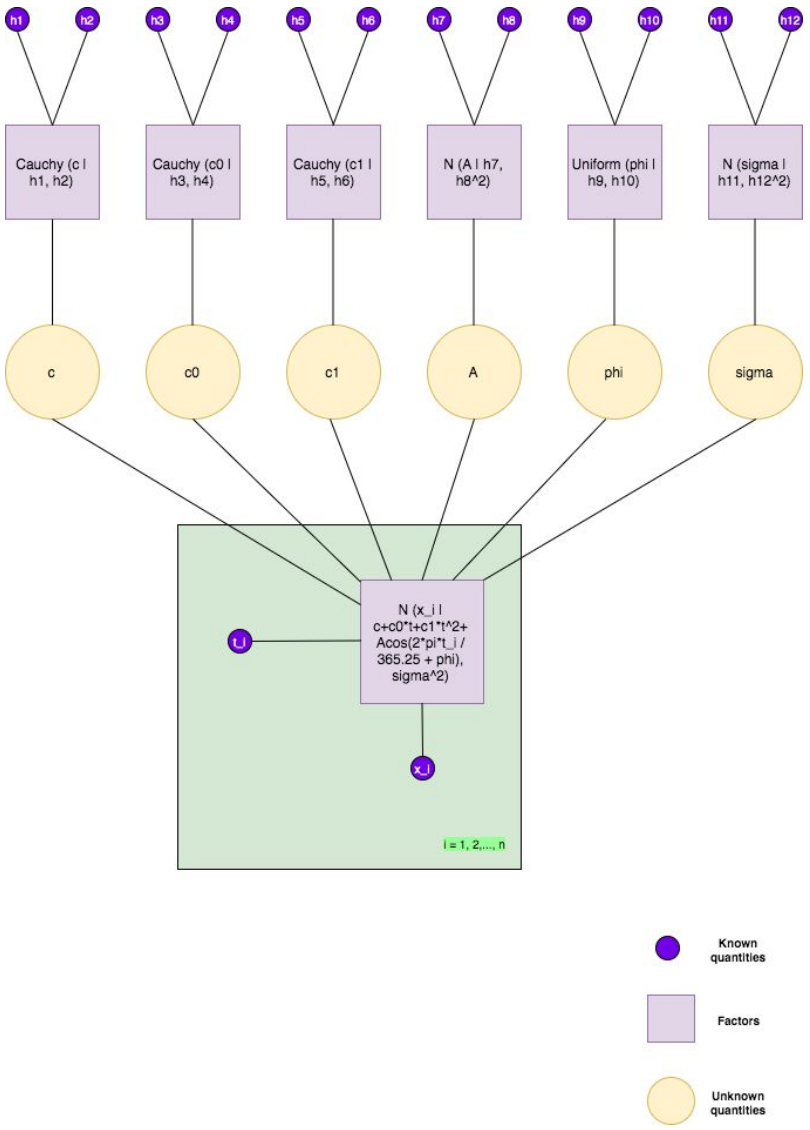

Figure 1. *Factor graph that represents the statistical model chosen.*

III, Results and evaluation

1, Model comparisons and evaluation

Before coming up with the chosen model described above, I tested different models that yielded different results. Below is a summary of why I did not choose them – there were two main reasons: non-convergence and bad prediction.

(i) Linear trend model

Apparently, the linear trend model does not do a good job at sampling, with very low n_eff scores. n_eff scores reflect how good the sampling is – ideally, we want Stan to sample a lot of independent samples, so n_eff scores

represent the number of independent samples equivalent to the 4000 samples drawn by Stan. Furthermore, this model also does not replicate the data correctly. If we were to choose this model to predict future data, we would underestimate the CO_2 level and be over-optimistic about climate change.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
c	310.56	0.01	0.09	310.4	310.5	310.55	310.61	310.73	58	1.05
c0	3.6e-3	8.4e-7	9.1e-6	3.6e-3	3.6e-3	3.6e-3	3.6e-3	3.6e-3	118	1.04
A	2.59	0.01	0.05	2.47	2.55	2.59	2.61	2.7	27	1.1
phi	1.6e-3	2.5e-4	5.6e-4	6.6e-4	1.1e-3	1.5e-3	2.2e-3	2.4e-3	5	1.8
sigma	2.05	0.01	0.03	1.99	2.03	2.05	2.08	2.09	7	1.28

Figure 2. *The convergence of model 1.*

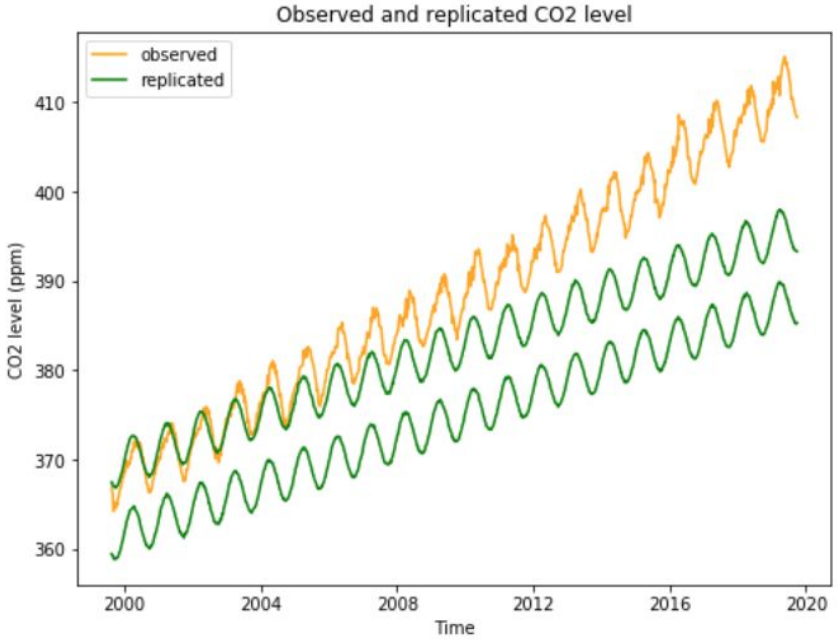

Figure 3. *Observed data for the test set and replicated data from model 1.*

(ii) Quadratic trend model with $\phi \in [0, 2\pi]$

The second model is pretty similar to the chosen one – the only difference is that I had $\phi \in [0, 2\pi]$ instead of $\phi \in [0, \pi]$. This causes the divergence for ϕ with a very big Rhat value. Looking at the pair plot, we can see that ϕ has two modes that actually sum up to 2π, indicating that these two modes have the same effect on the data (due to the periodic characteristic of the cosine function). This multi-modality might have caused A and σ to diverge and have two modes as well. It is interesting to see that the lower mode of A corresponds to the higher

mode of σ and vice versa, which makes sense because if we increase A, we need to decrease σ to obtain the same data.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
c	314.34	2.6e-3	0.08	314.19	314.29	314.34	314.39	314.49	873	1.02
c0	2.2e-3	1.6e-6	2.3e-5	2.1e-3	2.2e-3	2.2e-3	2.2e-3	2.2e-3	223	1.04
c1	9.5e-8	1.6e-10	1.5e-9	9.2e-8	9.4e-8	9.5e-8	9.6e-8	9.8e-8	87	1.05
A	2.67	0.09	0.14	2.47	2.54	2.68	2.8	2.85	2	4.12
phi	2.92	2.06	2.92	2.1e-5	3.3e-4	2.9	5.84	5.86	2	417.87
sigma	1.11	0.11	0.16	0.93	0.95	1.11	1.27	1.3	2	10.17

Figure 4. *The convergence of model 2.*

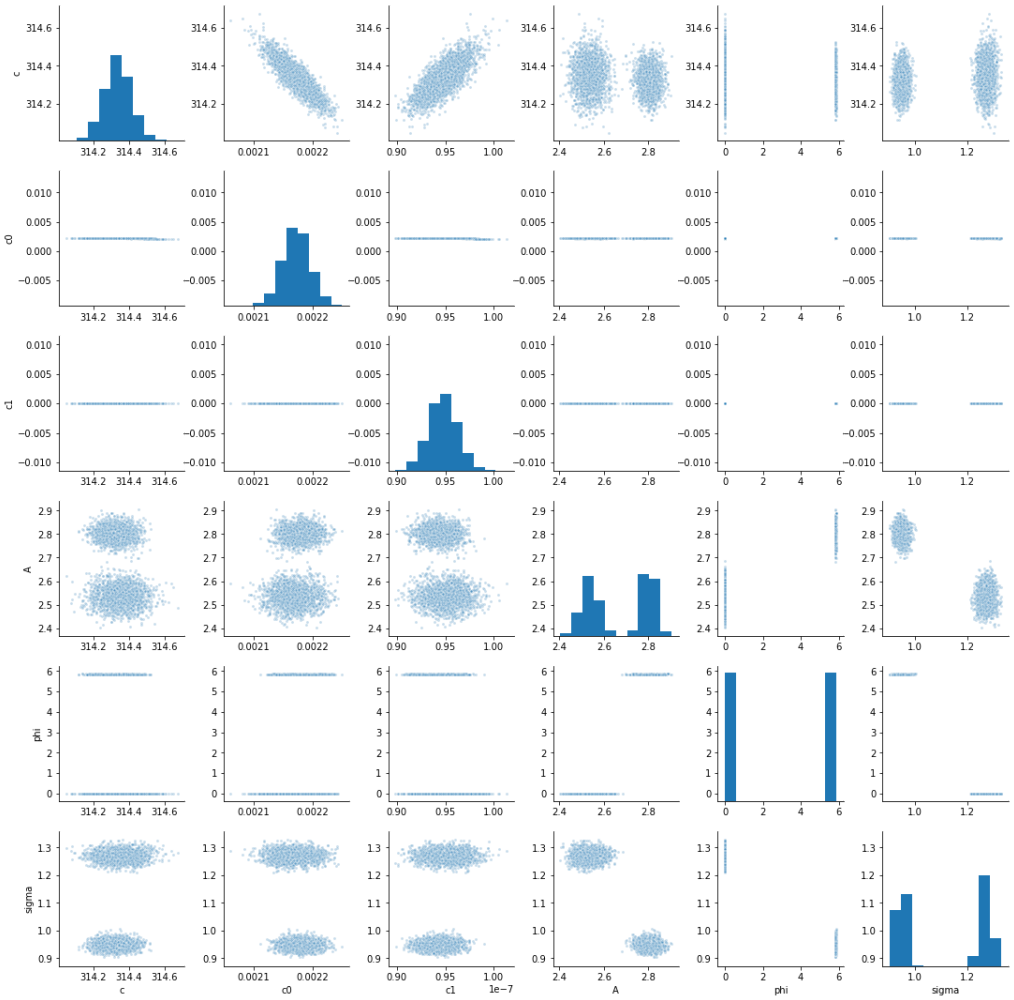

Figure 5. *Pair plots for model 2.*

(iii) Quadratic trend model with priors being normal distributions

For this model, the constraints for all the parameters are the same as for the chosen model, but the priors for all the parameters are normal distributions. The model converged pretty well; however, its replication for observed data is terrible – it overestimates the CO_2 level and the confidence interval is too big. This results from the estimates for σ, the standard deviation for noise. The mean of σ is 39.29, which is very big.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
c	23.31	0.01	0.51	22.32	22.95	23.31	23.66	24.31	2471	1.0
c0	0.03	2.0e-6	1.1e-4	0.03	0.03	0.03	0.03	0.03	3133	1.0
c1	2.7e-10	4.6e-12	2.6e-10	7.9e-12	8.2e-11	1.9e-10	3.7e-10	9.9e-10	3263	1.0
A	0.63	0.01	0.48	0.02	0.24	0.52	0.9	1.76	2318	1.0
phi	0.67	0.01	0.56	0.02	0.23	0.54	0.99	2.07	2950	1.0
sigma	39.29	2.3e-3	0.13	39.04	39.2	39.29	39.38	39.54	3336	1.0

Figure 6. *The convergence of model 3.*

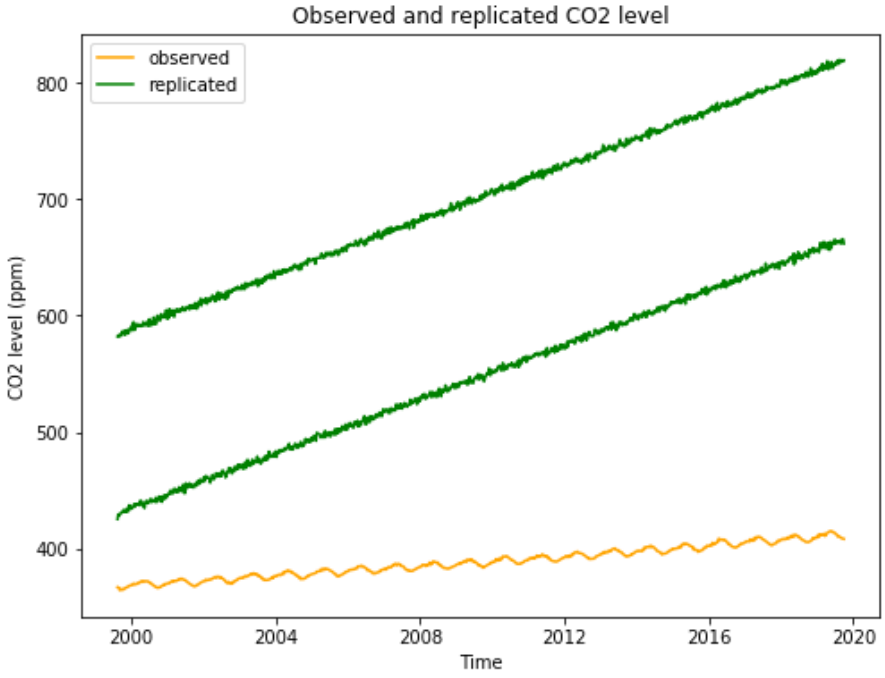

Figure 7. *Observed data for the test set and replicated data from model 3.*

(iv) Quadratic trend model with priors being normal, uniform and Cauchy distributions (chosen model)

The chosen model converges and samples very well and gives a good replication for observed data and good prediction for future data. Specifically, all the n_eff values are bigger than 1000, all the Rhat values are between 1.0 and 1.1. There is very little autocorrelation among the samples for all the parameters, and the pair plot shows the unimodality of all the parameters.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
c	314.35	1.7e-3	0.09	314.18	314.29	314.35	314.41	314.52	2747	1.0
c0	2.2e-3	5.5e-7	2.6e-5	2.1e-3	2.1e-3	2.2e-3	2.2e-3	2.2e-3	2225	1.0
c1	9.5e-8	3.5e-11	1.6e-9	9.2e-8	9.4e-8	9.5e-8	9.6e-8	9.8e-8	2192	1.0
A	2.54	7.2e-4	0.04	2.46	2.51	2.54	2.56	2.61	2899	1.0
phi	4.9e-4	1.5e-5	5.0e-4	7.4e-6	1.3e-4	3.3e-4	6.8e-4	1.8e-3	1097	1.01
sigma	1.27	3.2e-4	0.02	1.23	1.26	1.27	1.28	1.31	3649	1.0

Figure 8. *The convergence of model 4.*

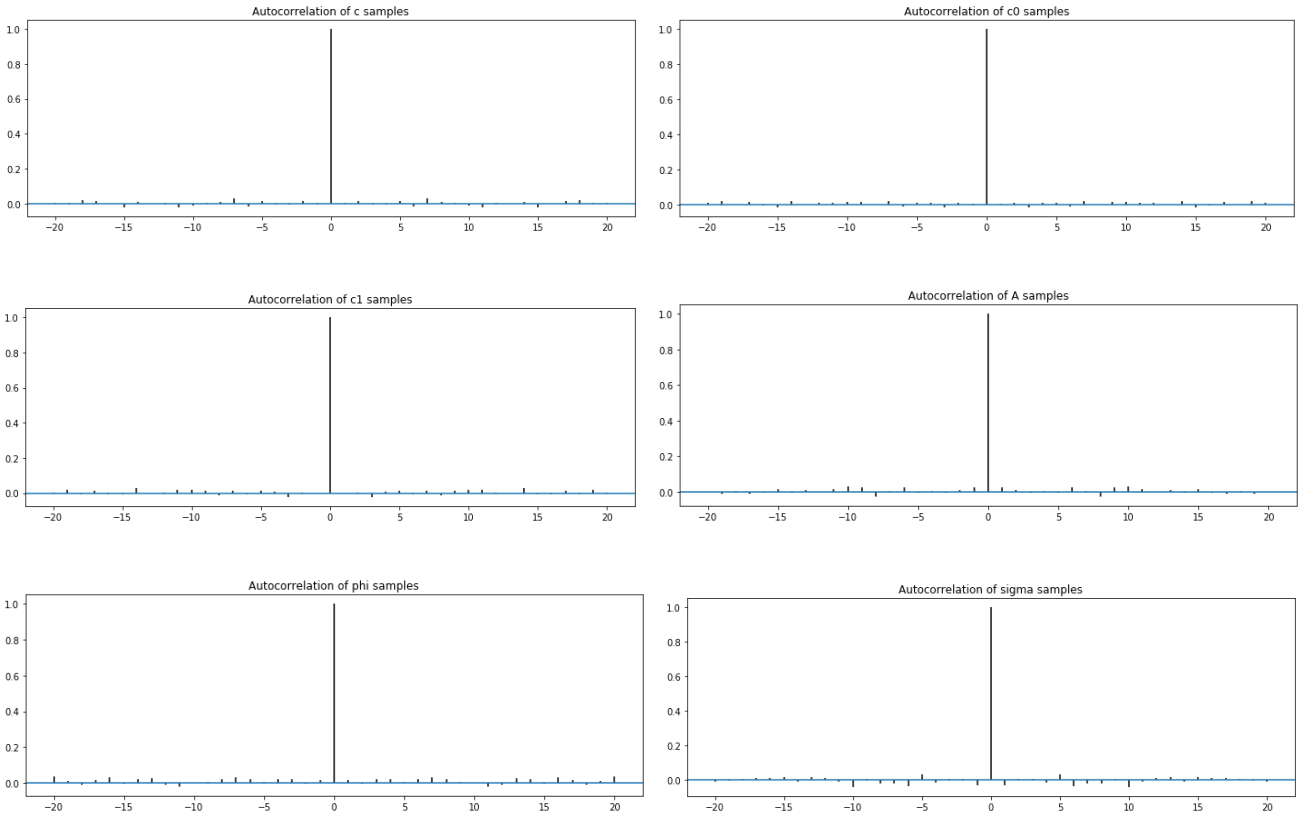

Figure 9. *Autocorrelation among samples for each parameter generated by model 4.*

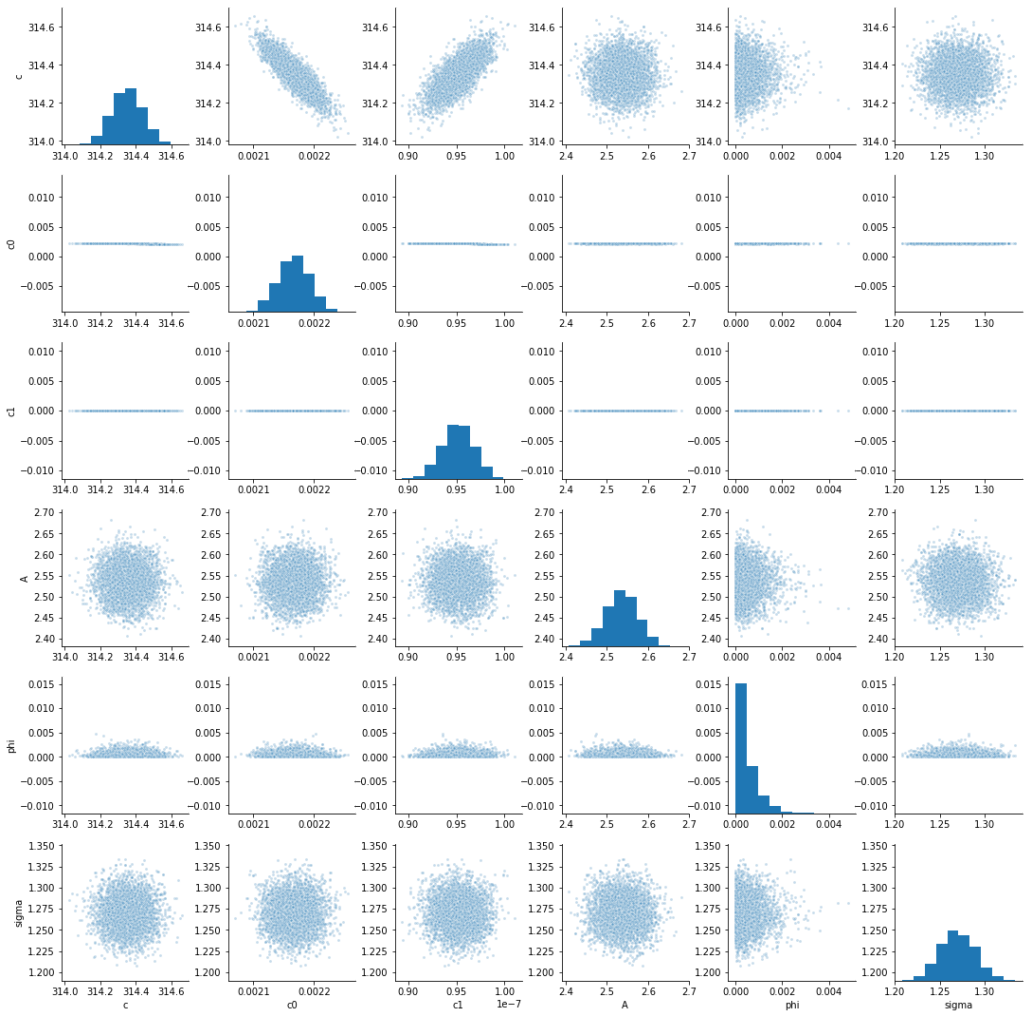

Figure 10. *Pair plots of the parameters for model 4.*

This model also gives a reasonable 95% confidence interval for the replicated data and compared to other models, the error is much smaller. However, it is not perfect – the confidence interval is much smoother than the actual data, which is also indicated in the small value for the mean of the noise standard deviation.

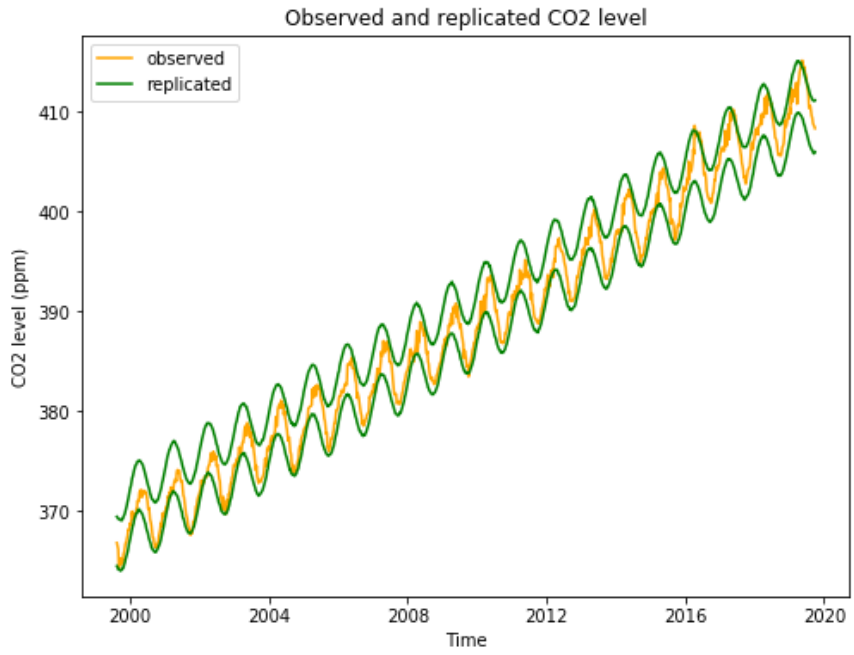

Figure 11. *Observed data for the test set and replicated data from model 4.*

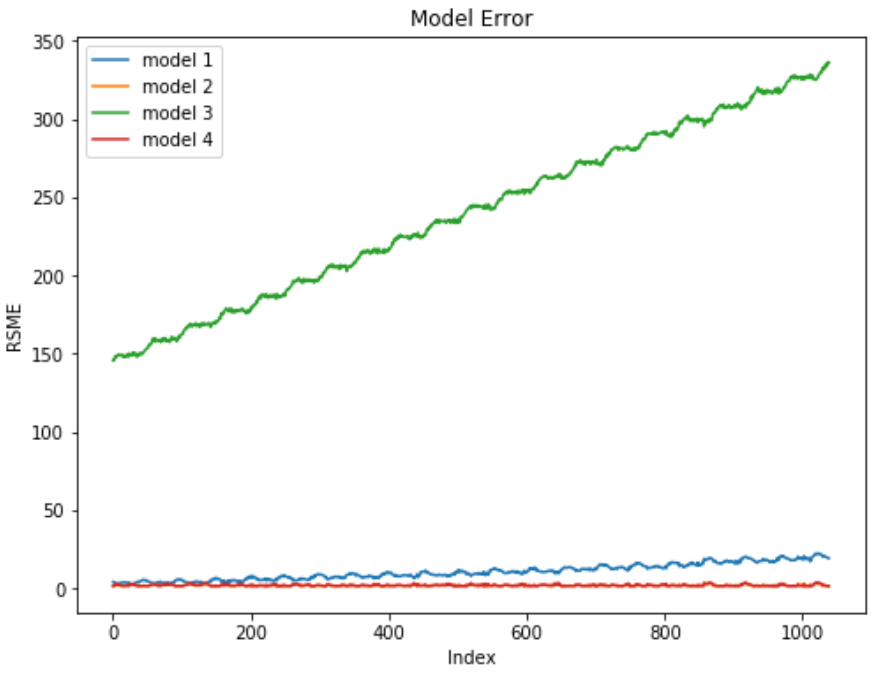

Figure 12. *Observed data for the test set and replicated data from model 4.*

2, Posteriors over the parameters

Figure 8 gives us an overview of the posteriors from the chosen model over the parameters. The corresponding pair plot (**Figure 10**) also includes the visualized posterior distributions.

IV, Prediction

1, CO_2 level for the next 40 years

Based on our posteriors for the parameters, we could predict the carbon dioxide levels for the next 40 years. From the estimation, we can see that at the beginning of 2058, 95% of the time, the CO_2 levels will fall into the range $[516.3,\ 523.6]$, which is really bad because it will pass the safe threshold of 450 ppm.

	Time	t	2.5%	50%	97.5%
0	2019-10-06	22471	406.050618	408.596559	411.148661
1	2019-10-13	22478	406.072312	408.690201	411.227928
2	2019-10-20	22485	406.261699	408.825368	411.230880
3	2019-10-27	22492	406.475821	408.973243	411.658277
4	2019-11-03	22499	406.702131	409.211055	411.868138

	Time	t	2.5%	50%	97.5%
1992	2057-12-09	36415	515.161587	518.645129	522.159974
1993	2057-12-16	36422	515.355586	518.960300	522.505106
1994	2057-12-23	36429	515.624004	519.337323	522.990192
1995	2057-12-30	36436	516.007535	519.718957	523.371680
1996	2058-01-06	36443	516.301202	520.110811	523.618031

Figure 13. *Predicted CO_2 levels for the future generated by model 4.*

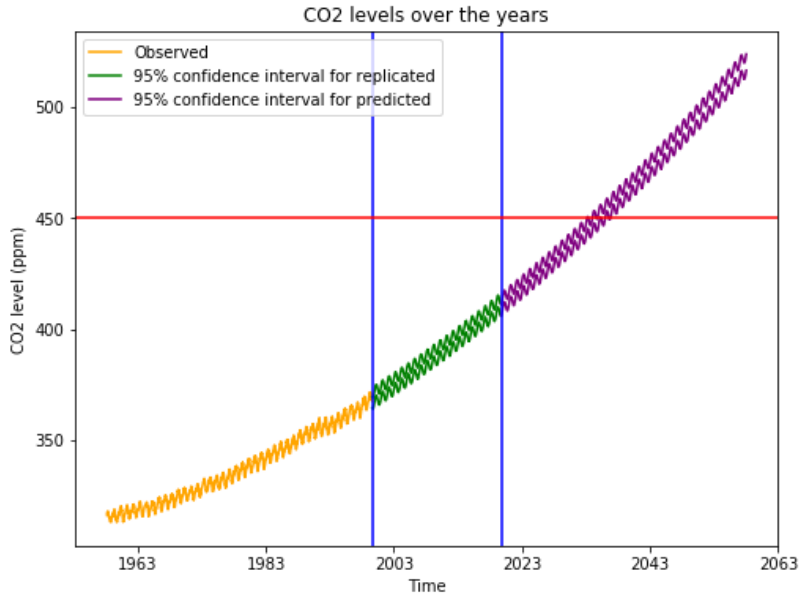

Figure 14. *Observed, replicated and predicted* CO_2 *levels for the future generated by model 4.*

2, High-risk time

We can also figure out when the CO_2 level will likely be high-risk, i.e., 450ppm. From the prediction, it looks like 2034 will be the year when there is a good chance that we will reach that number. For example, on March 12, 2034, the median for the prediction is 450.08 which, in simple words, means that 50% of our predicted values for that particular day are above 450.08 and 50% are below. The situation deteriorates very quickly, as a year after that, on March 18, 2035, the 95% confidence interval is $[450.1,\ 455.9]$ which means that for 95% of the time, the CO_2 level will be in this range. In other words, we can be very certain that 2035 is the beginning of dangerous climate change.

	Time	t	2.5%	50%	97.5%
749	2034-02-12	27714	446.431932	449.304473	452.152199
750	2034-02-19	27721	446.572689	449.544792	452.273668
751	2034-02-26	27728	446.821899	449.720788	452.544951
752	2034-03-05	27735	447.203147	449.960553	452.757756
753	2034-03-12	27742	447.250089	450.081540	452.870199
754	2034-03-19	27749	447.397352	450.203000	452.958130
755	2034-03-26	27756	447.520084	450.325974	453.150093
756	2034-04-02	27763	447.582369	450.316130	453.078404
757	2034-04-09	27770	447.506649	450.353947	453.184948
758	2034-04-16	27777	447.604363	450.335958	453.143422

Figure 15. *Predicted CO_2 levels for the future generated by model 4 for 2034.*

	Time	t	2.5%	50%	97.5%
802	2035-02-18	28085	449.362167	452.221061	454.999059
803	2035-02-25	28092	449.530847	452.449082	455.282692
804	2035-03-04	28099	449.802648	452.637642	455.511654
805	2035-03-11	28106	449.926910	452.787015	455.623615
806	2035-03-18	28113	450.142758	452.949979	455.853080
807	2035-03-25	28120	450.228778	452.964924	455.882407
808	2035-04-01	28127	450.187944	453.092061	455.941712
809	2035-04-08	28134	450.275390	453.123132	455.878199
810	2035-04-15	28141	450.210750	453.098330	455.802632
811	2035-04-22	28148	450.191453	453.075421	455.771362

Figure 16. *Predicted CO_2 levels for the future generated by model 4 for 2035.*

3, Uncertainty and practical implications

The model being discussed gives us deep insights into the situation of climate change that we are facing right now. Being able to predict the rise over time of the carbon dioxide levels based on past data helps us see not only how badly the situation has been deteriorating but also in specific details when we should expect to ruin the planet completely if we do not make changes. It is particularly vital that environmentalists, policy-makers and big corporates apply this model and things alike to take necessary actions in a timely manner. As an individual, during

the course of training and testing the model, I was also shocked at how close we are to the edge of serious irreversible climate change.

There are some uncertainties in the model, of course. For example, we cannot really give concrete numbers for the carbon dioxide levels for the next 40 years – hence the use of confidence intervals. However, these intervals are very tight, usually around 5 ppm, so they still give us good predictions.

V, Critique

I identified several shortcomings of the model. First, it could be very fragile. In Bayesian statistics, I learned that as we have a lot of data, the priors should not have too big of an impact on the posterior distributions of the parameters. However, in this case, changing the priors changed the whole results (model 3 versus model 4). Therefore, there is no guarantee that changes to the hyperparameters of the priors despite keeping the same distributions will not change the results. Second, as noted earlier, the replicated data is less prone to noise than the actual data, i.e., the estimated values for σ might not be very accurate. In fact, it is smaller than the fluctuations from the periodic function seen on the plot for the real data. However, in general, I believe we have a decent statistical model for the description and prediction of CO_2 levels at Mauna Loa Observatory in Hawaii.

APPENDIX

Link to Github