Modeling and Forecasting Atmospheric CO_2 from 1958 into the Future

-- CS146 Final Project

Executive Summary

This project models atmospheric CO_2 level with regards to time using data recorded from 1958 to 2021.

The key findings are:

1. CO_2 level is estimated to reach 535.50 ppm at the start of 2060 which is an 28.98% increase since the start of 2021 (415.19 ppm -> 535.50 ppm), implicating that the CO_2 level is increasing at a high rate and actions needs to be taken or it would be out of control in 2060.

2. It's predicted that by Dec, 2032, CO_2 level would reach 450ppm which is considered as high risk for dangerous climate change assuming current projection. Actions need to be taken to prevent that from happening in 10 years.

Introduction

In this project, with the data of atmospheric CO_2 level recorded at the Mauna Loa Observatory in Hawaii since 1958 (Keeling & et.al, 2001), a statistical model is created to explain the data and forecast the CO_2 level between now and the start of 2060. We also predict when it would reach a high risk for dangerous climate change. Uncertainty of the result and shortcomings of the model are also analyzed.

The code used to build the model, conduct sampling, and generate visualizations can be found here.

Model Description

To demonstrate the model structure, a factor graph is provided below. Choices of distributions and parameter values will be explained in the next section.

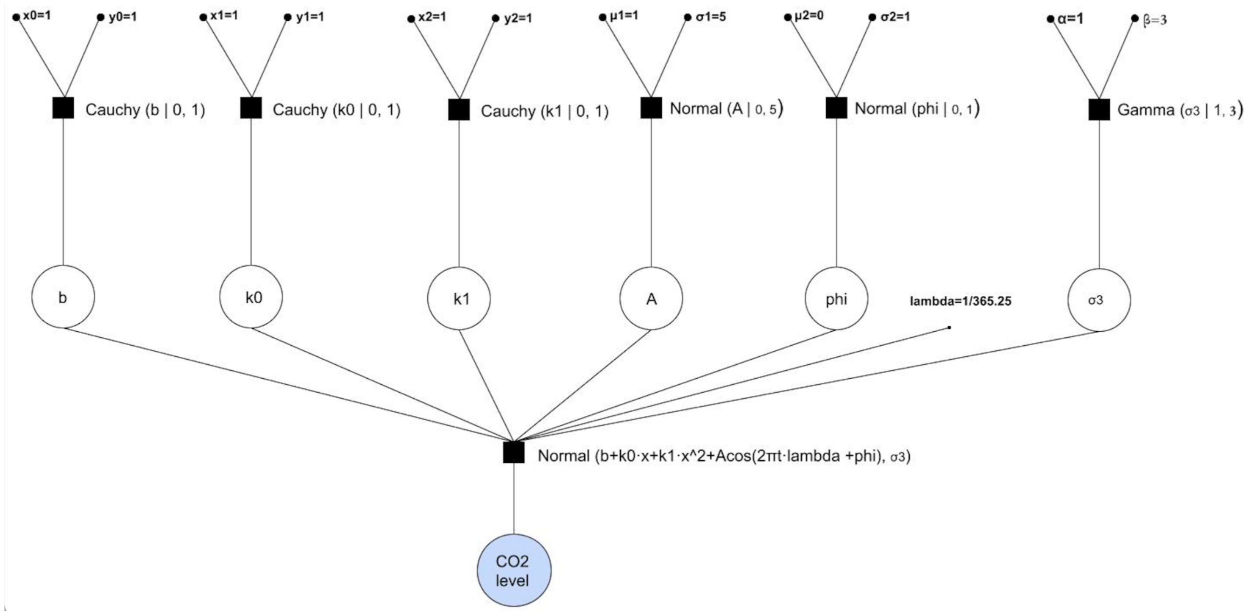


Figure 1. Factor Graph of the Model

Choice for the likelihood function

$$CO2\ level \sim Normal\ (b\ +\ k_0x\ +\ k_1x^2\ +\ Acos(2\pi t\cdot\lambda\ \ +\ \ \phi),\ \sigma_3)$$

Observing the data, there are three components which motivates the design of the model:

1. There's overall a quadratic relationship between time and CO_2 level (as shown in Figure 2), which is modeled by $b\ +\ k_0x\ +\ k_1x^2$ in which b is the intercept and k_0 & k_1 are the two slopes for the original term and the squared term. Variable x represents the independent variable - the number of days since the date of the first measurement (1958-03-29).
2. Zooming in, the data also shows seasonality, with a yearly interval (as shown in Figure 3), which is modeled by $Acos(2\pi t\cdot\lambda\ \ +\ \ \phi)$ in which A is the amplitude (function value vary from [-A, A], λ is the known frequency (1/365.25) and ϕ is the phase offset of the function (periodic between [0, 1]).
3. The data also has a lot of noise, so normal distribution is chosen for the likelihood function to account for the variations. The CO_2 level is positive real numbers with a unimodal symmetric distribution.

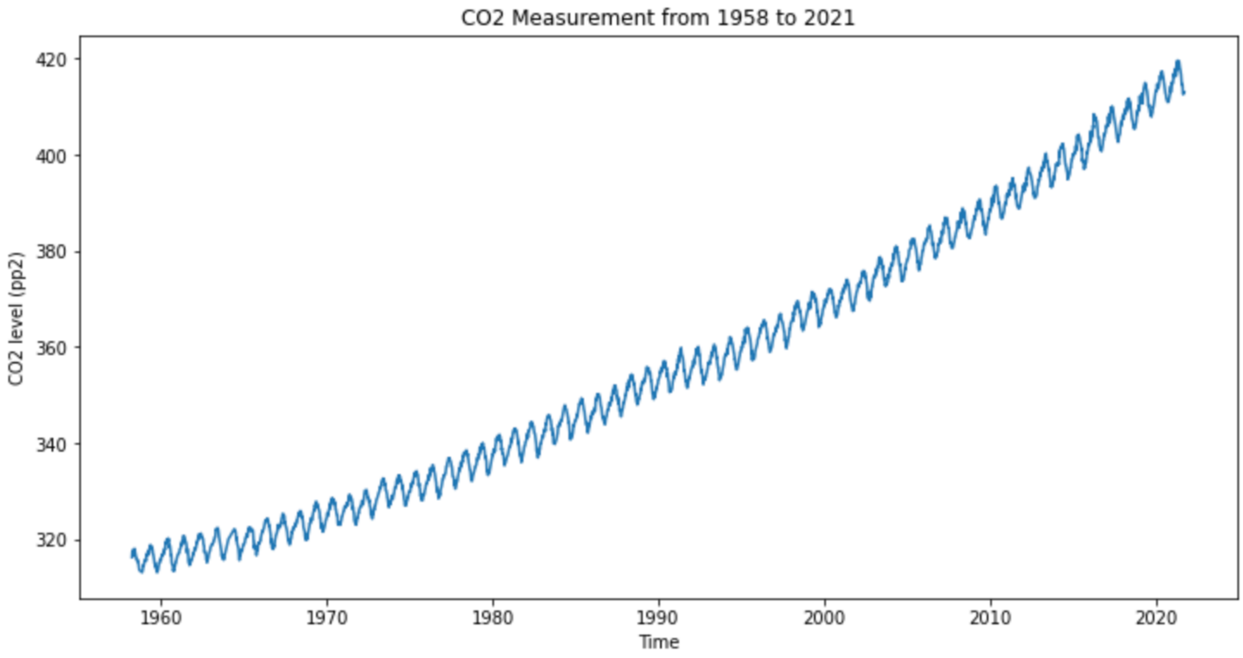


Figure 2. CO_2 Measurement from 1958 to 2021

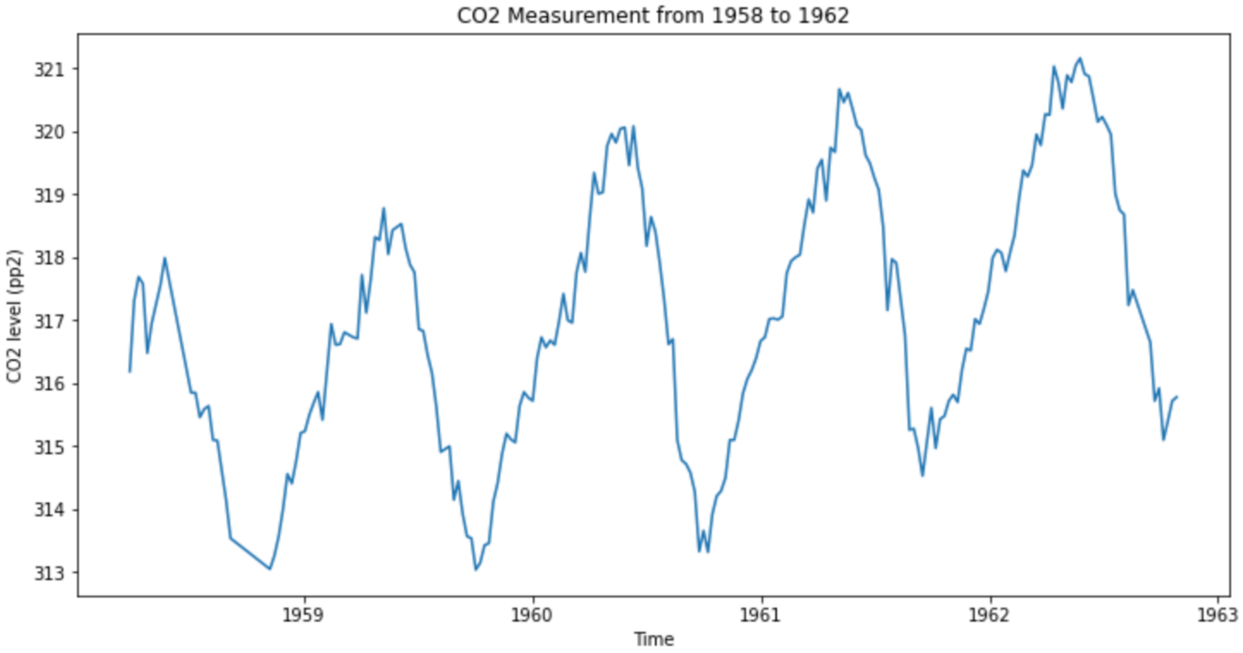


Figure 3. CO_2 Measurement from 1958 to 1962, showing seasonal trend

Choices for prior distributions

1. **Parameters for the quadratic relationship (b, k_0, k_1)**

$$b \sim Cauchy\,(0,\ 1)$$

$$k_0 \sim Cauchy\,(0,\ 1)$$

$$k_1 \sim Cauchy\,(0,\ 1)$$

Cauchy distribution with a location of 0 and scale of 1 is chosen for the prior distribution to estimate the parameters for the quadratic relationship (the intercept and slopes) because we have no prior knowledge of the trend of the data and Cauchy (0, 1) allows us to have a wide range and heavy tails covering extreme values. Since they are constrained to be positive real numbers based on that we are assuming an increasing trend in CO_2 over the years and a positive starting point, we are actually using a half positive Cauchy distribution.

2. **Multiplier for each country (m1) and each store type (m2)**

$$A \sim Normal\,(0,\ 5)$$

$$\phi \sim Normal\,(0,\ 1)$$

A is the amplitude of the function and controls the range of the function value to be [-A, A].

Phi is the phase offset of the function, controlling how many waves the function would move to the left. Because it's periodic in the range [0, 1], it's constrained to be [0, 1].

Lambda is the frequency of the function. Because from the data we can see that the variation has a yearly interval, we set lambda to be 1/365.25.

Normal distributions are chosen for the prior distribution to estimate them because we would expect the values to be centered around 0 because we generally have few knowledge of them but they are expected to be centered around 0. A has a larger variation than phi because of the parameter range.

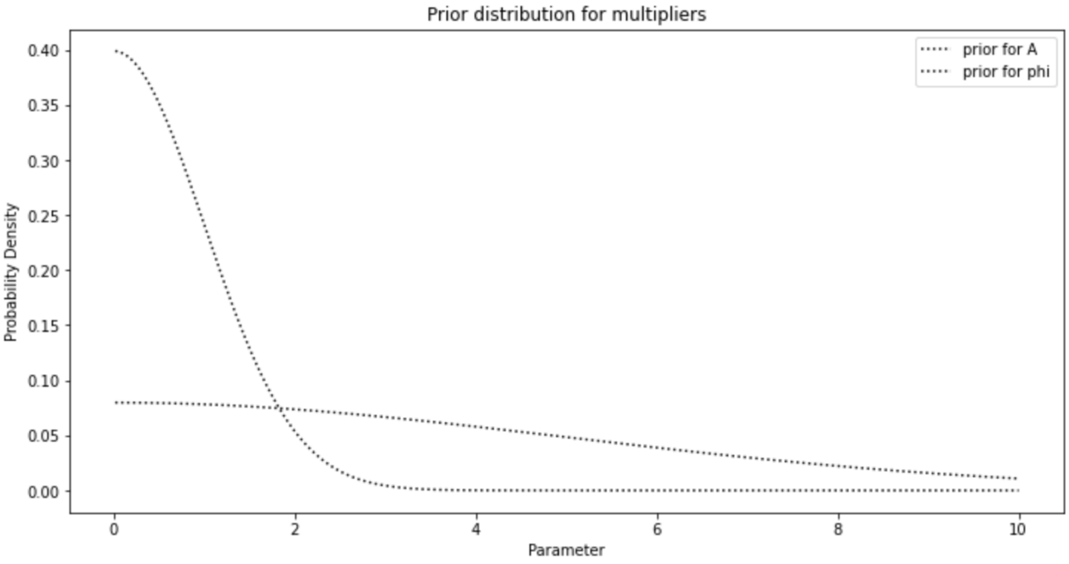


Figure 4. Prior distributions for A and phi

3. Variance for likelihood function (σ_3^2)

$$\sigma_3^2 \sim Gamma\,(1,\ 3)$$

A Gamma distribution is chosen to model the distribution for the variance because gamma distribution is a conjugate prior for normal distribution with known mean. Parameters chosen for the Gamma distribution are $\alpha = 1,\ \ \beta = 3$, governing the variance to be most likely 0.

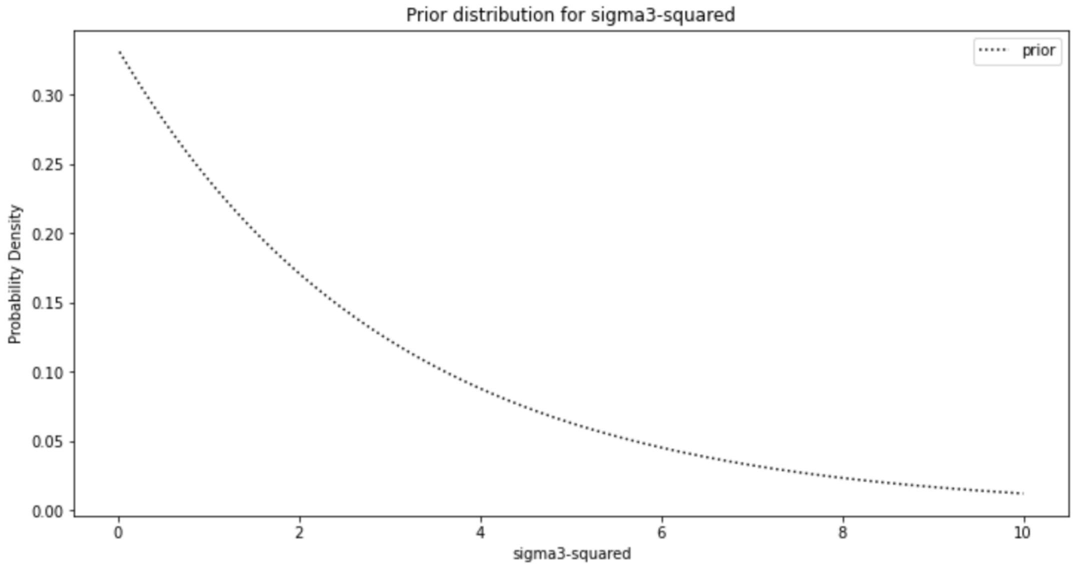


Figure 5. Prior distribution for σ_3^2

With the distributions and data specified above, the model is built and compiled using PyStan and feed in with the pre-processed data.

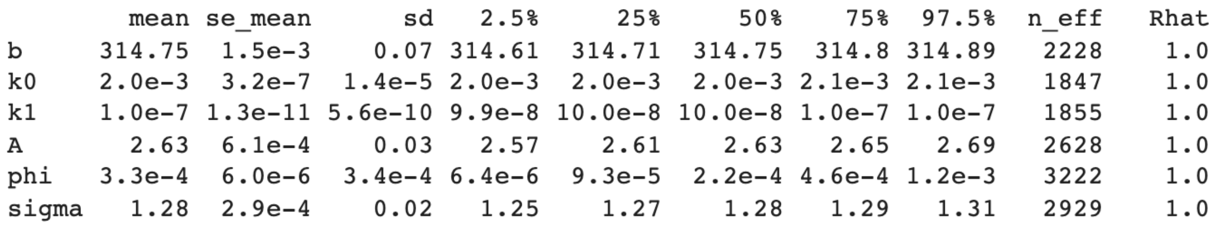
Results and Discussion

Convergence Diagnostics

Inference from the model is done by running Hamiltonian Monte Carlo using PyStan with 4 Markov chains. For each chain, there are 1000 steps for warm up to reach the parameter space and 1000 steps for the actual sampling. From the result below we can see the estimated mean and confidence intervals.

Rhat values of 1.0 for each parameter indicate that the Markov chains have converged properly because the average variance of samples within a chain is similar across chains.

The n_eff values represent the effective number of samples - the number of samples that are independent. The values are higher than the suggested couple of hundred.



```
        mean se_mean      sd   2.5%     25%     50%    75%  97.5%  n_eff   Rhat
b     314.75  1.5e-3    0.07 314.61  314.71  314.75  314.8 314.89   2228    1.0
k0    2.0e-3  3.2e-7  1.4e-5 2.0e-3  2.0e-3  2.0e-3 2.1e-3 2.1e-3   1847    1.0
k1    1.0e-7 1.3e-11 5.6e-10 9.9e-8 10.0e-8 10.0e-8 1.0e-7 1.0e-7   1855    1.0
A       2.63  6.1e-4    0.03   2.57    2.61    2.63   2.65   2.69   2628    1.0
phi   3.3e-4  6.0e-6  3.4e-4 6.4e-6  9.3e-5  2.2e-4 4.6e-4 1.2e-3   3222    1.0
sigma   1.28  2.9e-4    0.02   1.25    1.27    1.28   1.29   1.31   2929    1.0
```

Figure 6. Parameter result from sampling from the posterior distribution

Autocorrelation plots are also generated. From the figures we can see that there is little correlation between the samples for each parameter, which corresponds with the high number of effective samples.

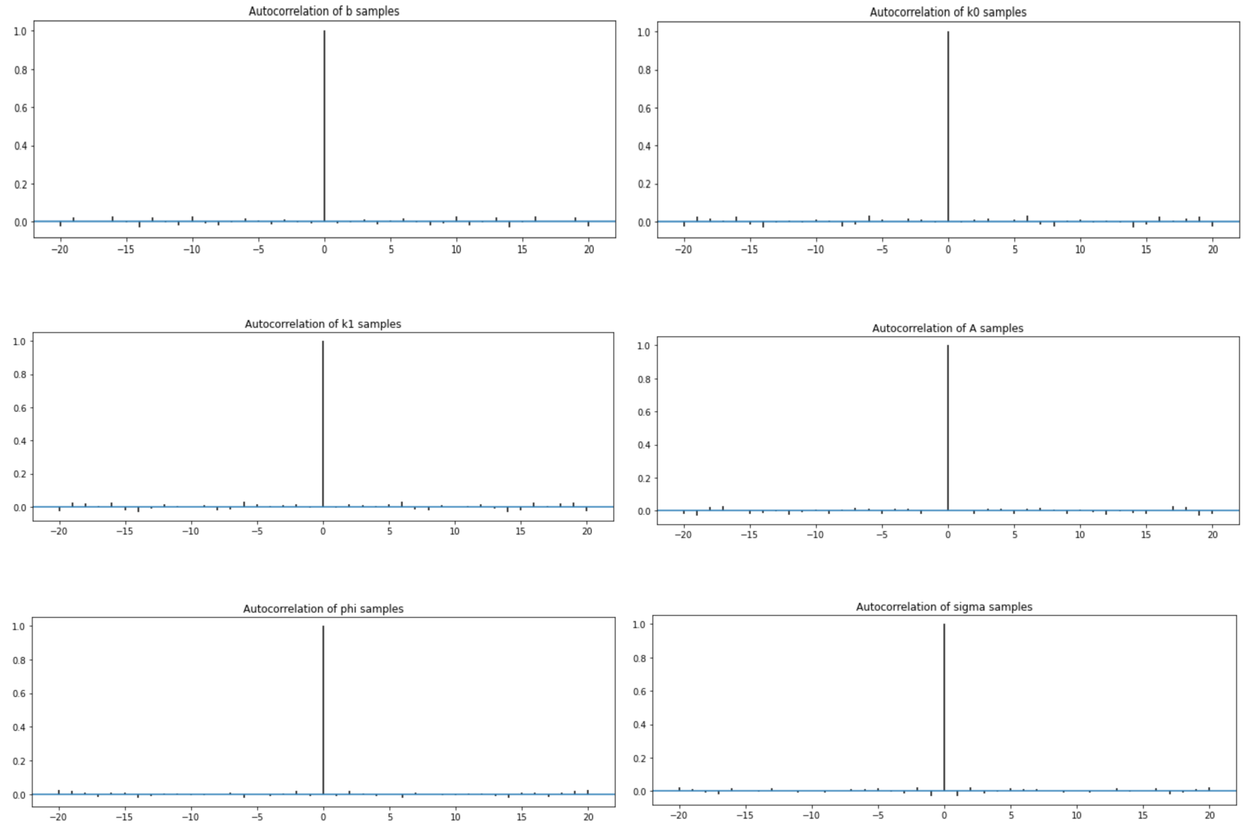


Figure 7. Autocorrelation plots from the parameter samples from the posterior distribution

From the pair plots below we can see that the distribution of samples for each parameter is unimodal, indicating the chains are mixed. Although there’s negative correlation between b and k0, k0 and k1, and positive correlation between k1 and k2, it’s not problematic.

The results above - high independence, low correlation, properly converged model and unimodal distribution of samples - mean that the Stan MCMC sampler is working properly.

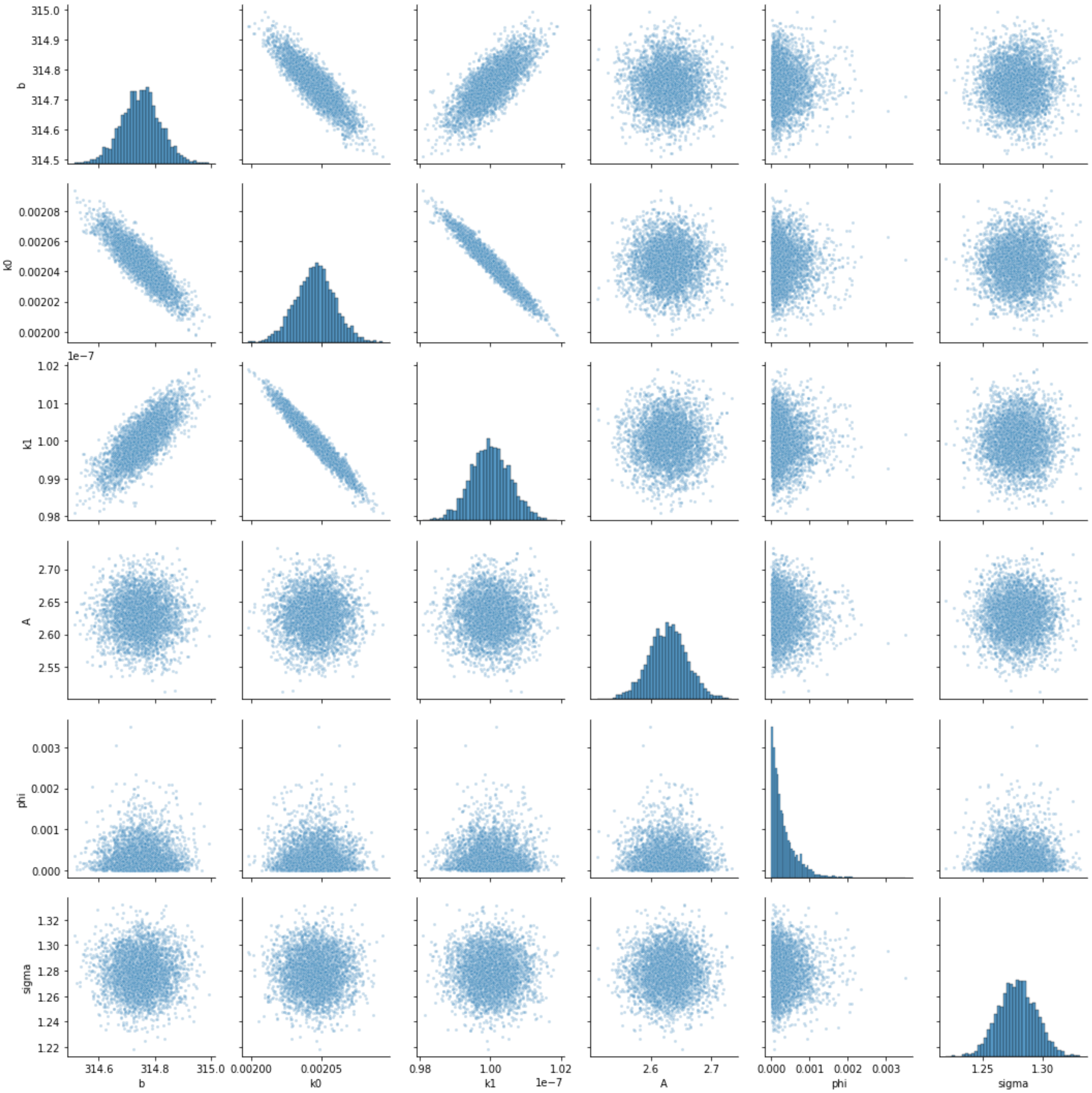


Figure 8. Pair plots for the parameter samples from the posterior distribution

Predicting with the model

Using the estimated parameters, results of CO_2 level is generated by the model.

Zooming in on the past, the confidence intervals are tight, meaning that the model has high certainty and describes the data well. Mean squared error is used to measure the training accuracy decide how well the model fits the training data. The current model has a MSE minimized to 1.63.

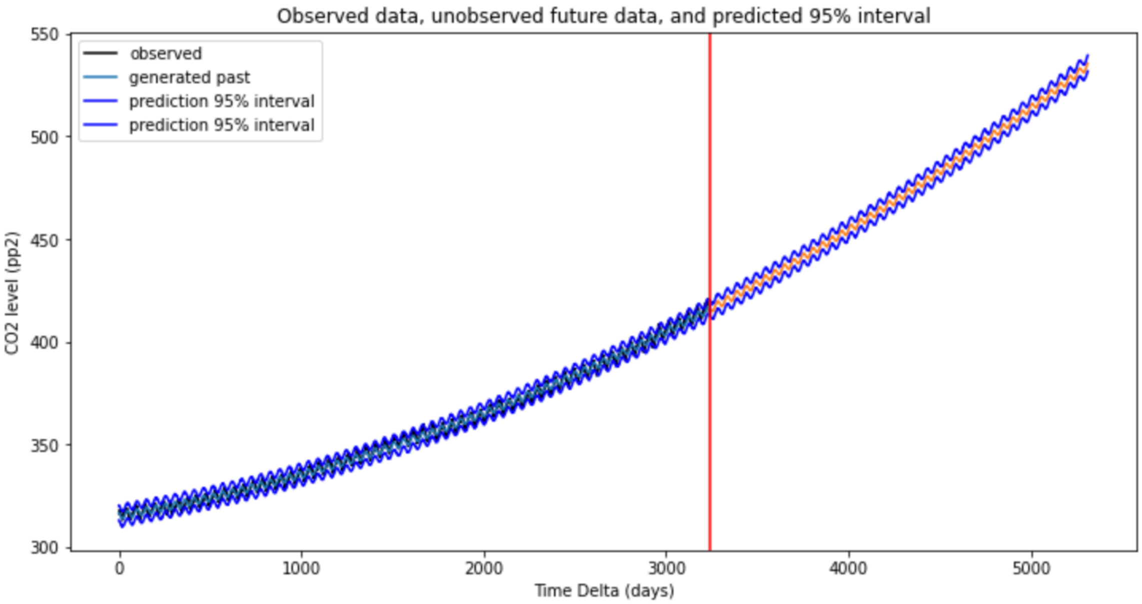


Figure 8. Generated data for the past and future

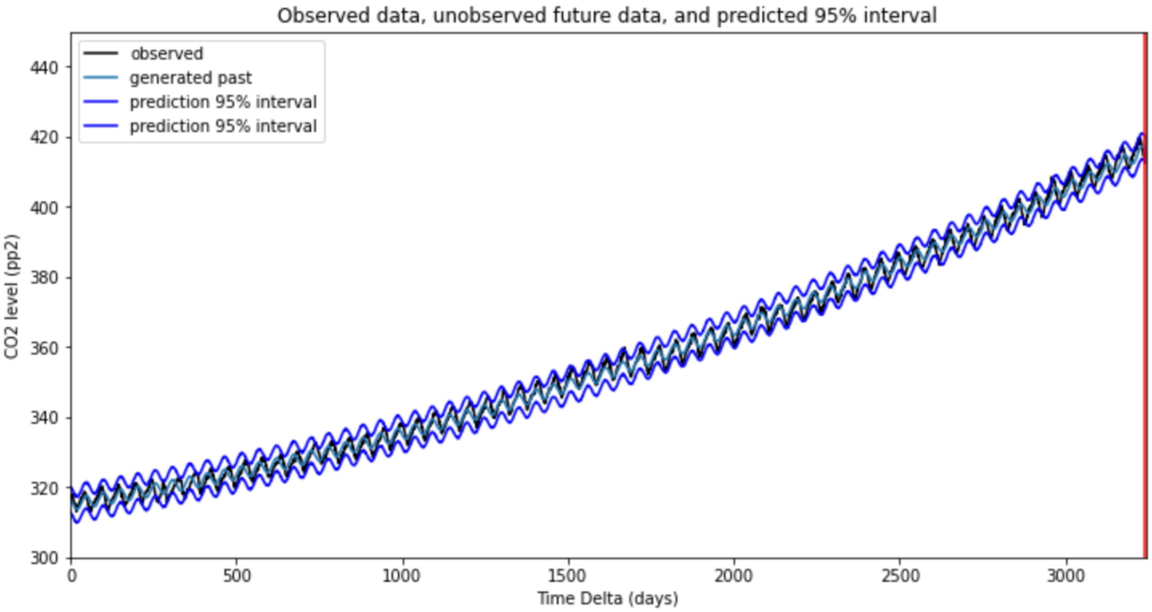


Figure 9. Generated data for the past to visualize fitness of the model

Results

The estimate for atmospheric CO_2 levels projected until the start of 2060 is 535.50 ppm (at Jan. 3rd, 2060), with a 95% confidence interval of [531.66, 539.41], meaning that we are 95% certain that the predicted CO_2 level would fall within the interval. Compared on a yearly basis, it has increased by 2.78 ppm since the same time in 2059, and 28.98% increase since the start of 2021 (415.19 ppm -> 535.50 ppm). The implication is that the CO_2 level is increasing at a high rate and actions needs to be taken or it would be out of control in 2060.

Taken CO_2 levels of 450 ppm as high risk for dangerous climate change, from the generated quantities of the model, by Dec. 25th, 2032, we would first reach this level. As shown in the graph, by June, 2023, it would drop beneath the limit because of seasonal fluctuation, but by Oct, 2033, all estimated CO_2 levels thereafter would be higher than this limit. The implication is that if we don't take any action, from Dec, 2022, we would be in high risk for dangerous climate change.

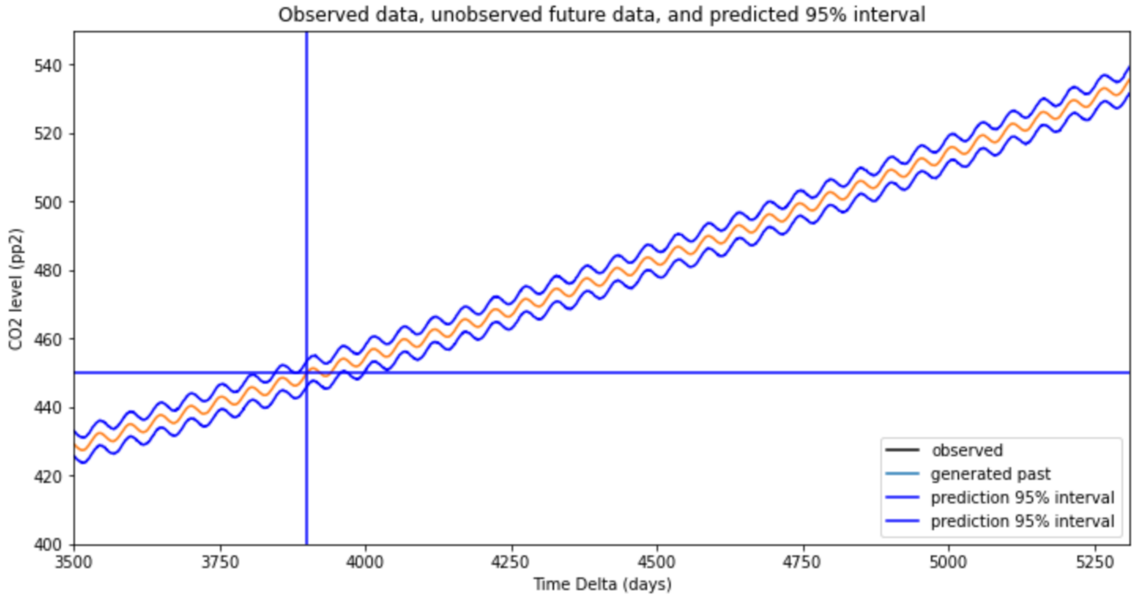


Figure 10. Prediction on the time reaching 450ppm (facing dangerous climate change)

Critique of the Model

As shown in the figure, there are still remaining uncertainties in the result. The reasons might be sudden increase of decrease in CO_2 emission caused by sudden climate change or policy decisions.

Zooming in to each seasonal variation, we can see that the actual data has a seasonal variation skewed to the left while the current model is symmetrical (using cos). Therefore, there's a potential mismatch between the model and real data and inaccurate predictions. It's manifested by MSE not close to 0 and there's still room for improvement. A candidate solution is to add another periodic function accompany the existing one to model the skewness.

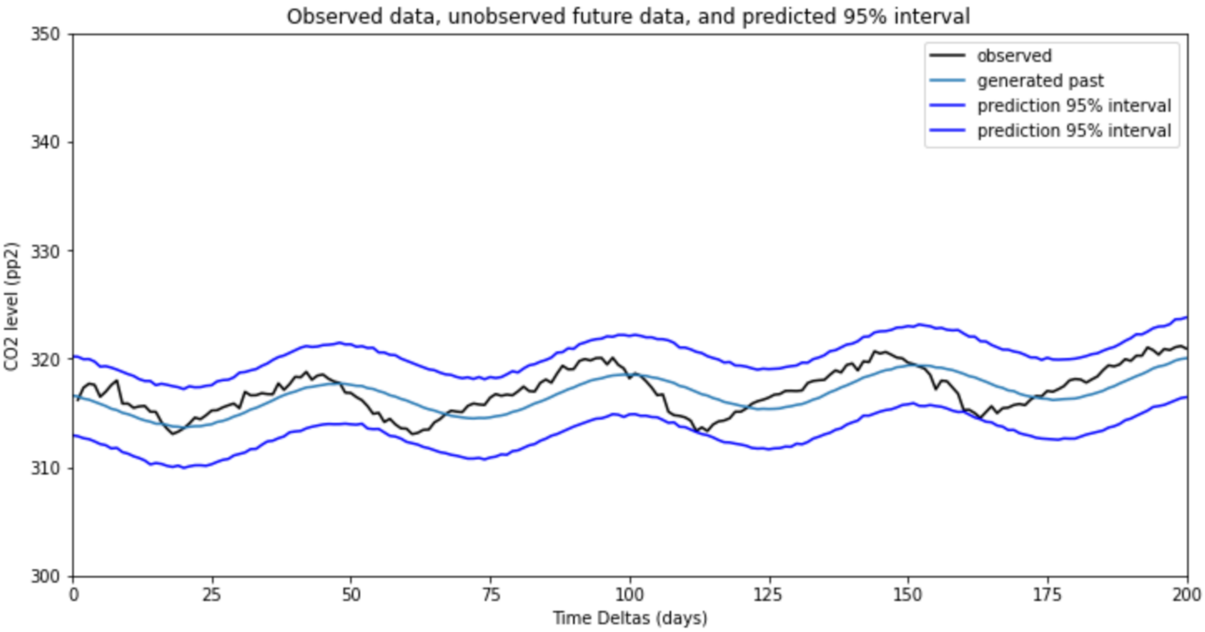


Figure 11. Comparison of actual data and generated past data using the model

Bibliography

C. D. Keeling, S. C. Piper, R. B. Bacastow, M. Wahlen, T. P. Whorf, M. Heimann, and H. A. Meijer, Exchanges of atmospheric CO2 and 13CO2 with the terrestrial biosphere and oceans from 1978 to 2000. I. Global aspects, SIO Reference Series, No. 01-06, Scripps Institution of Oceanography, San Diego, 88 pages, 2001.