Introduction

In this paper, we model CO_2 levels (measured in ppm) from 1958 to 2021 using the Mauna Loa dataset from the Scripps CO_2 program (Sampling station records, n.d.). We begin with preprocessing and visualizing the original data. The next step is to create Bayesian statistical models and initialize them in the computational tool called Stan. We proceed by predicting the CO_2 levels for the years 2021-2060 and evaluating the model with the help of pair plots, autocorrelation plots, and posterior predictive checks. The paper ends with a discussion of model applicability in real-life scenarios and potential improvements that could be done to achieve more robust inference.

Our initial dataset has two columns, which are observed parameters in this case — time and CO_2 levels. We also need to incorporate a few unobserved variables into the model, namely, parameters for time trend, parameters for seasonal fluctuations and noise.

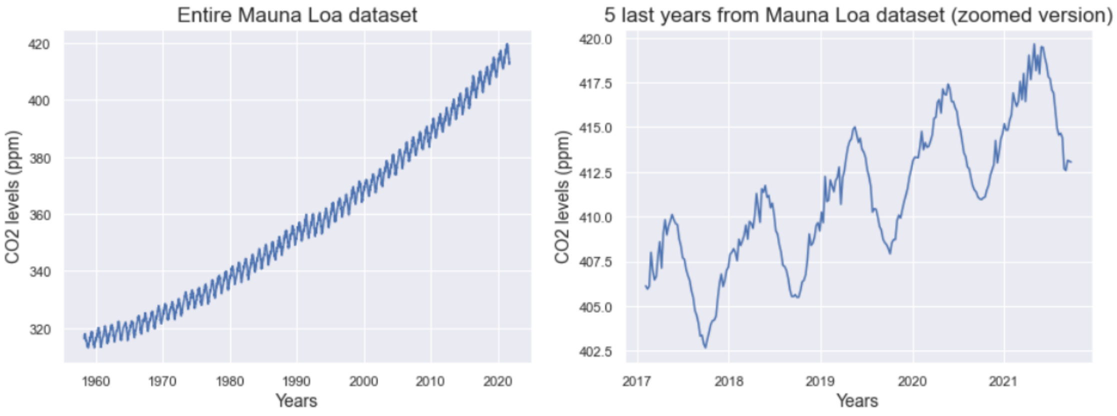


Fig. 1. CO_2 levels (ppm) from 1958 to 2021 (left figure). CO_2 levels (ppm) from 2017 to 2021 (right figure).

Looking at the figure above, we can notice a clear upward trend along with some seasonal fluctuations and noise. Measurement and technology consistency are some important assumptions to mention in this case. We believe that there was no change in either component because, otherwise, it could have contributed to the values visualized above.

Default Model

The default model can be represented by the following equation:

$$p(x|\boldsymbol{\theta}) = N\left(c_0 + c_1 t + c_2 * \cos\left(2\pi * \frac{t}{365.25} + c_3\right), c_4\right),$$

where $c_0 + c_1 t$ represents a linear trend, $c_2 * \cos\left(2\pi * \frac{t}{365.25} + c_3\right)$ shows the seasonal change and c_4 includes the noise obtained from a normal distribution with mean of 0.

Since the model parameters are unobserved, we need to identify priors over them. We do not know much about the potential range of values, which the coefficients could take. Thus, we have decided to model them using normal distributions. We have centered the first coefficient on 316 with a standard deviation of 15, where the mean equals the first measurement in the dataset. We also assume that other coefficients do not take large values since this would result in a huge decrease of CO_2 levels prediction. Thus, we think it makes sense to define broad priors such as N(0.5, 0.75) for c_1 and N(0, 2) for c_2 and c_3. Given that the standard deviation should be positive, we use a gamma distribution to determine appropriate values.

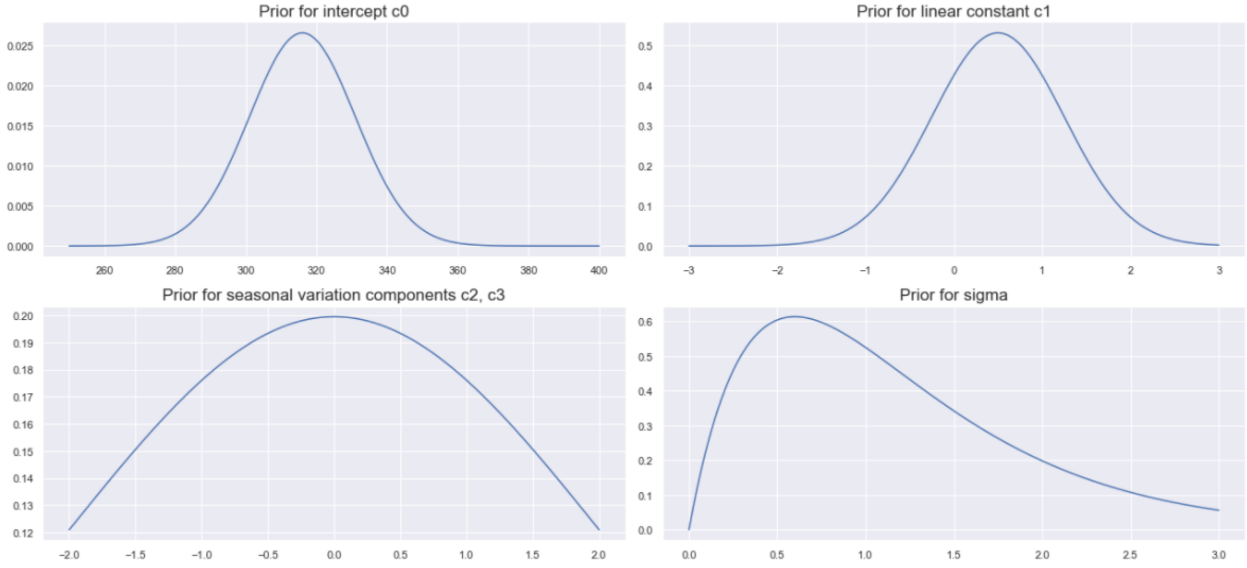


Fig. 2. Priors for model coefficients.

After initializing and compiling the default model in Stan, we should plot the values it outputs for 1958-2021. We do not incorporate seasonal variation in the results. Looking at the visualizations below, we notice that the linear trend is not a great representation of the given data. This becomes more obvious when we zoom in and look at the results for the last 14 years. The predicted values are lower than the true ones.

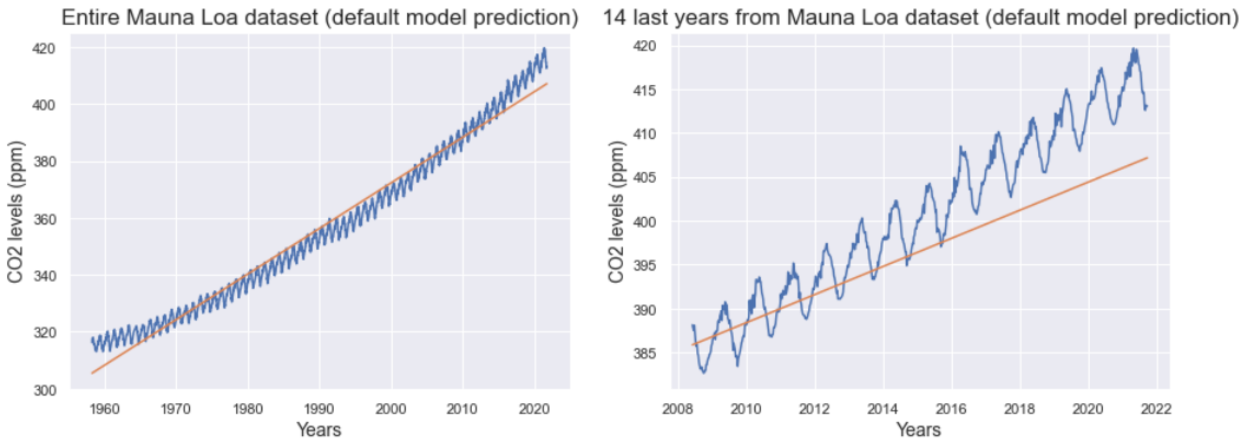


Fig. 3. Default model (full version on the left, zoomed version on the right) and the linear trend.

We can also check the prediction quality by plotting the residuals and examining root mean squared error. We notice a curve contributing to the general pattern on the graph below. This is not something we should see in an accurate representation of CO$_2$ levels.

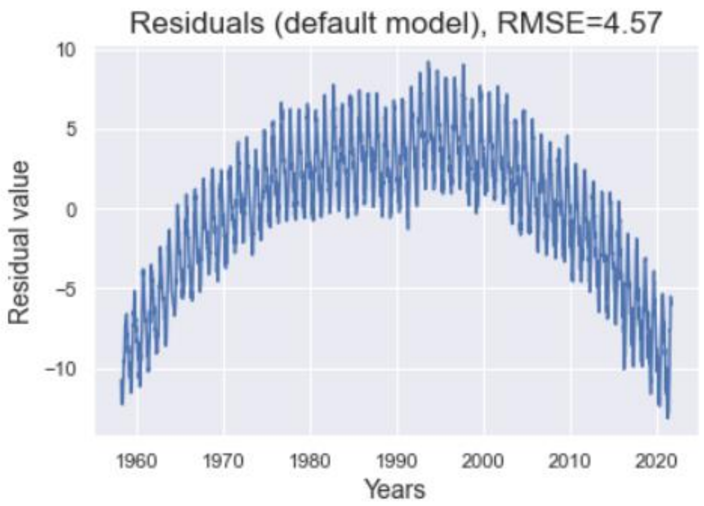


Fig. 4. Plot of residuals from the default model. The pattern is similar to a parabola, which is a sign of the model not being a great fit for the given scenario.

Suggested Model

The following equation can represent the suggested model:

$$p(x|\boldsymbol{\theta}) = N\left(c_0 + c_1 t + c_2 t^2 + c_3 * \cos\left(2\pi * \frac{t}{365.25} + c_4\right), c_5\right),$$

where $c_0 + c_1 t + c_2 t^2$ represents a quadratic trend, $c_3 * \cos\left(2\pi * \frac{t}{365.25} + c_4\right)$ shows the seasonal change and c_5 includes the noise obtained from a normal distribution with a mean of 0.

Here is a corresponding factor graph that allows understanding the relationships between the variables in the suggested model:

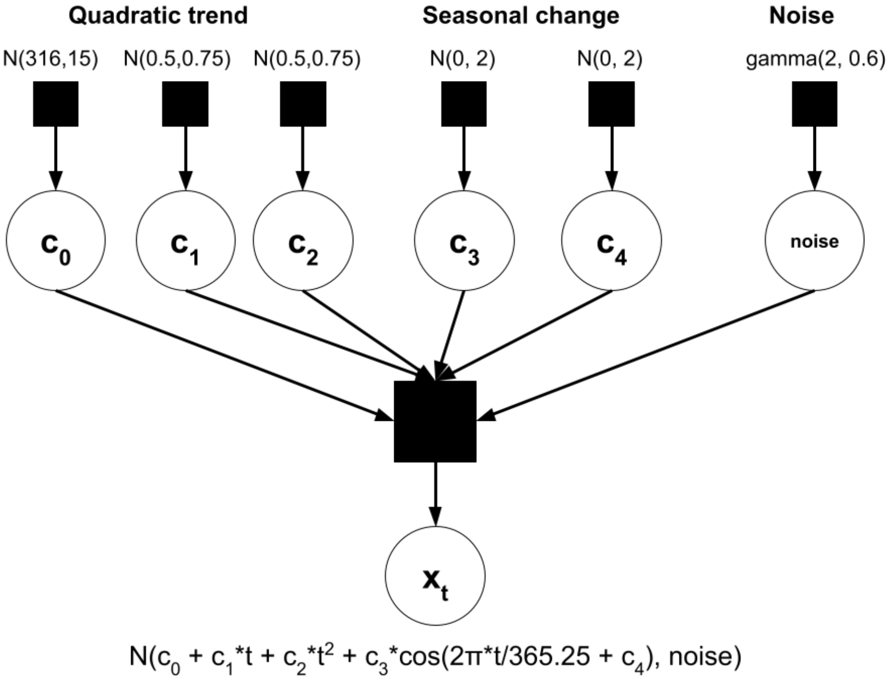


Fig. 5. Factor graph of the suggested model. Blank circles represent unobserved variables, while filled squares correspond to the distributions where the values are drawn from.

Since we cannot observe the parameters of the new model, we perform the same procedure as in the case of the default one. The only difference is that we have to define a prior for the quadratic constant c_2. We believe it should be in the same range as c_1 and, therefore, define a prior of N(0.5, 0.75) over the parameter. The rest of the priors remain unchanged.

Our model runs successfully and provides the Rhat values of 1.0 for all parameters (see Appendix A). It means that the chains mixed well, and the algorithm covered the whole distribution. Additionally, we observe high values for the number of effective samples, which hints that there are enough independent samples. Thus, we can trust the model results.

After incorporating c_2 into the Stan model, we plot the values of 1958-2021 again. We do not incorporate seasonal variation in the results. We notice from fig. 6 that the quadratic trend results in a smooth fit to the data. The zoomed-in version of the plot proves because the line of fit goes through the middle of the seasonal change in every period. We also examine the plot of the residuals (see fig. 7) to see that they do not fluctuate as much as in the case of the default model and have a lower root mean squared error.

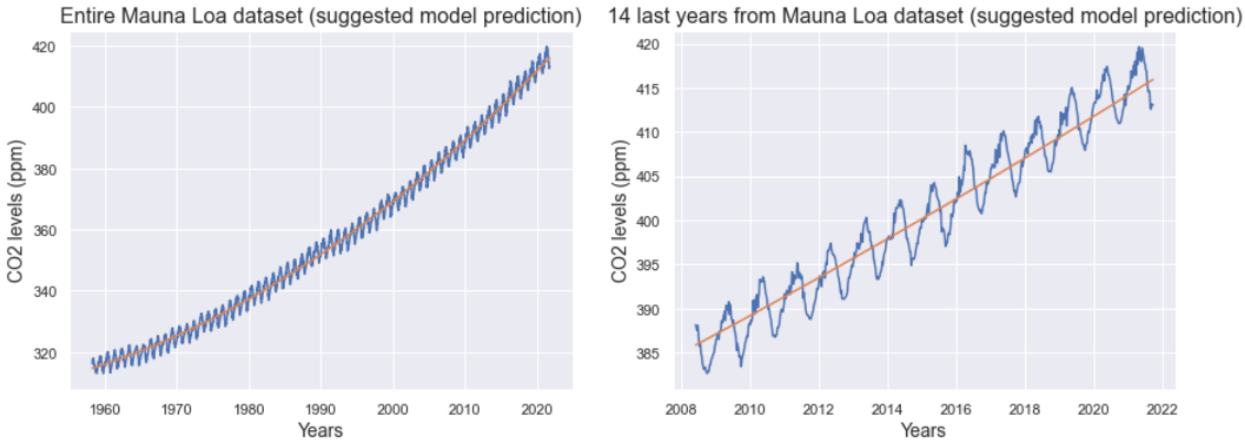


Fig. 6. Suggested model (full version on the left, zoomed version on the right) and the quadratic trend.

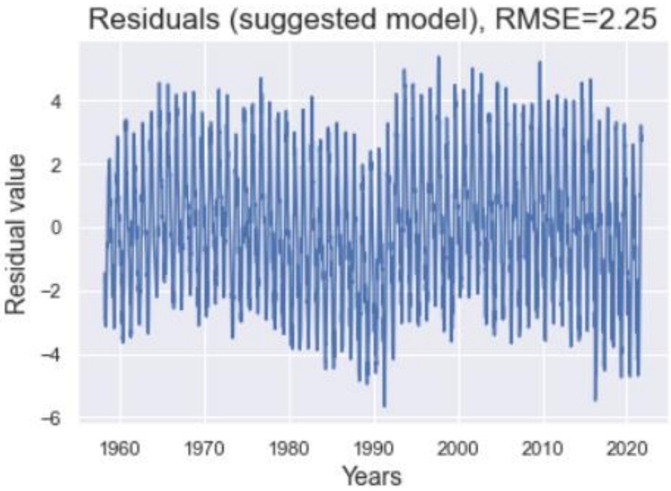


Fig. 7. Plot of residuals from the suggested model. The pattern indicates that there is only little curvature present, which makes the suggested model an appropriate fit.

Comparison and Predictions

We continue comparing the default and suggested model by inspecting the produced predictions of CO_2 levels. The difference is significant, with the default model being more optimistic about the increase of CO_2 levels in the next 40 years. However, based on the discussion above, we have identified the suggested model to be more appropriate for the given scenario. Thus, we believe that the values will reach the critical threshold of 450 ppm in 2034. By 2060, we expect CO_2 levels to equal 528 ppm. The 95% confidence interval for the year 2060 ranges from 526 to 530 ppm. It is worth highlighting that the confidence intervals for the suggested model are narrower than the ones for the default model, which demonstrates the higher level of certainty present in the suggested model. The model's assumption during the prediction phase includes the constant standard deviation. We are not sure this is an entirely accurate approach because the confidence intervals should become wider the further into the future the model is trying to predict due to the higher uncertainty related to potential changes in the biosphere or the human behavior related to the greenhouse gas emissions etc. Thus, a potential suggestion to the modeling process would be to gradually increase the uncertainty of predictions when calculating the CO_2 levels in the far future.

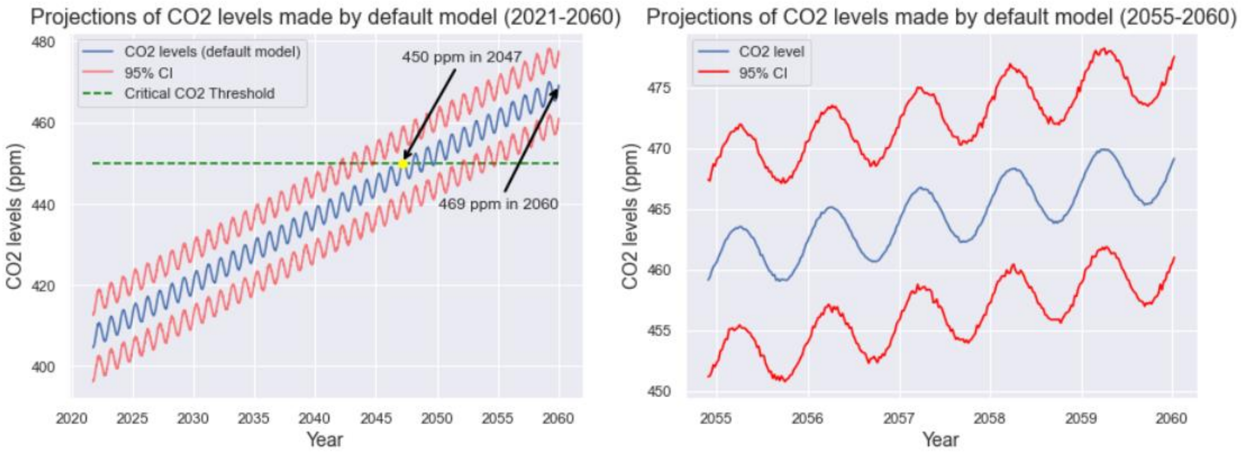


Fig. 8. Default model predictions (incl. critical threshold) and corresponding confidence intervals (full version on the left, zoomed version on the right).

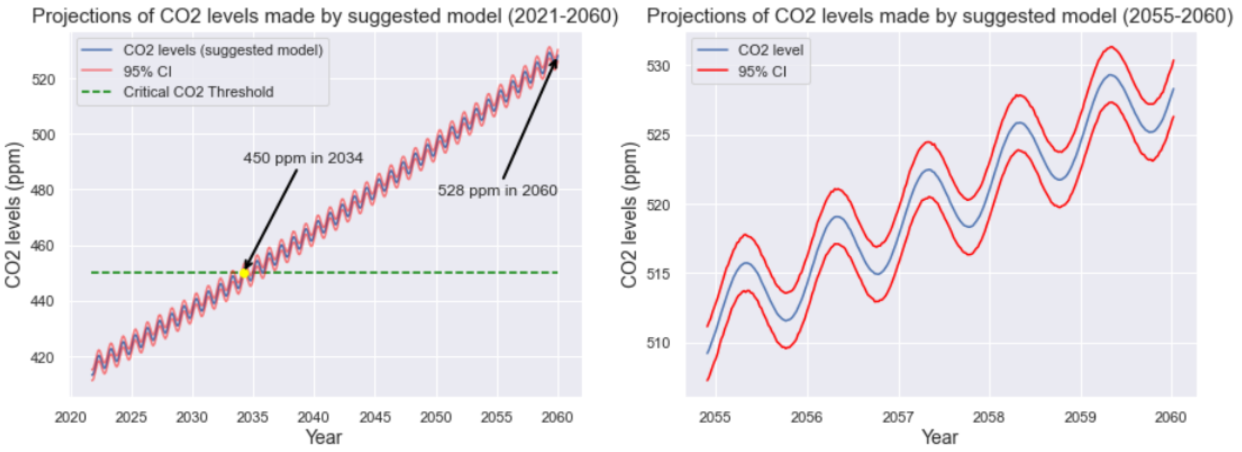


Fig. 9. Suggested model predictions (incl. critical threshold) and corresponding confidence intervals (full version on the left, zoomed version on the right).

Suggested Model Evaluation

We have performed additional checks to ensure that the suggested model makes sense in the given scenario. For instance, we have created a pair plot (see Appendix B), which shows that all parameters are normally distributed. We observe a negative correlation between c_0 and c_1, c_1 and c_2, and a positive correlation between c_0 and c_2. All parameters have a positive range except c_4 because it was not constrained to be equal or more than zero in our model. We also plot autocorrelation values for samples of each parameter. This allows us to notice that there is almost no autocorrelation between samples and proves the high level of sample independence shown in the Stan model output (see Appendix A).

We finished the model evaluation by performing posterior predictive checks using the test statistics such as mean and standard deviation (see fig. 10). We generated posterior data samples from the suggested model and compared the test statistics of the real data to the distribution under our samples from the posterior. Obtained p-values strengthen our opinion that the model does not have many deficiencies because both p-values are in the range of [0.05 0.95].

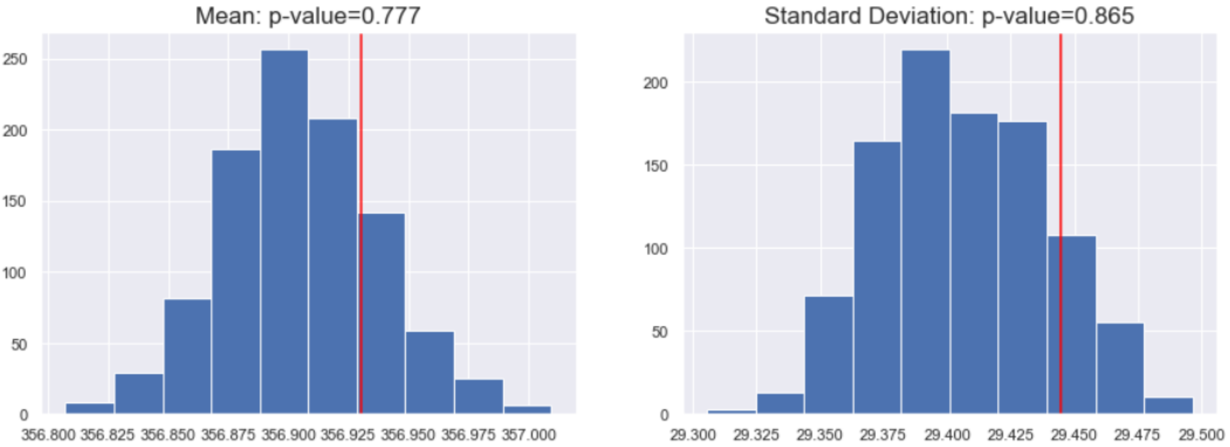


Fig. 10. Posterior predictive checks for mean (left) and standard deviation (right). The corresponding p-values support our discussion about the fit of suggested model in the given scenario.

Conclusion

Based on the analysis above, the suggested model provides reasonable and trustworthy estimates that the scientific community can use. Model output is concerning because the predicted CO2 level for 2060 is 1.5 times bigger than that measured in 1958. If humanity does not act up immediately and put significantly more effort into preserving the planet and slowing down climate change, the negative consequences are inevitable.

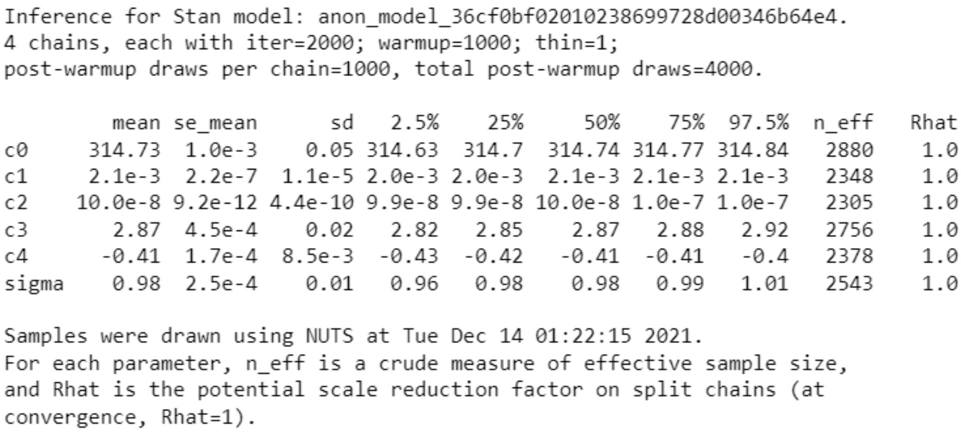
We also believe there is always room for improvement in terms of coming up with the most suitable model. As mentioned above, the quadratic trend fits the existing data fairly well. However, the model only considered data from the past 63 years, whereas CO2 levels started changing long before. Thus, we do not reject the chances of another trend being a better fit once more data is available.

References

Sampling station records. MLO Station Data | Scripps CO2 Program. (n.d.).

Retrieved from https://scrippsco2.ucsd.edu/data/atmospheric_co2/mlo.html

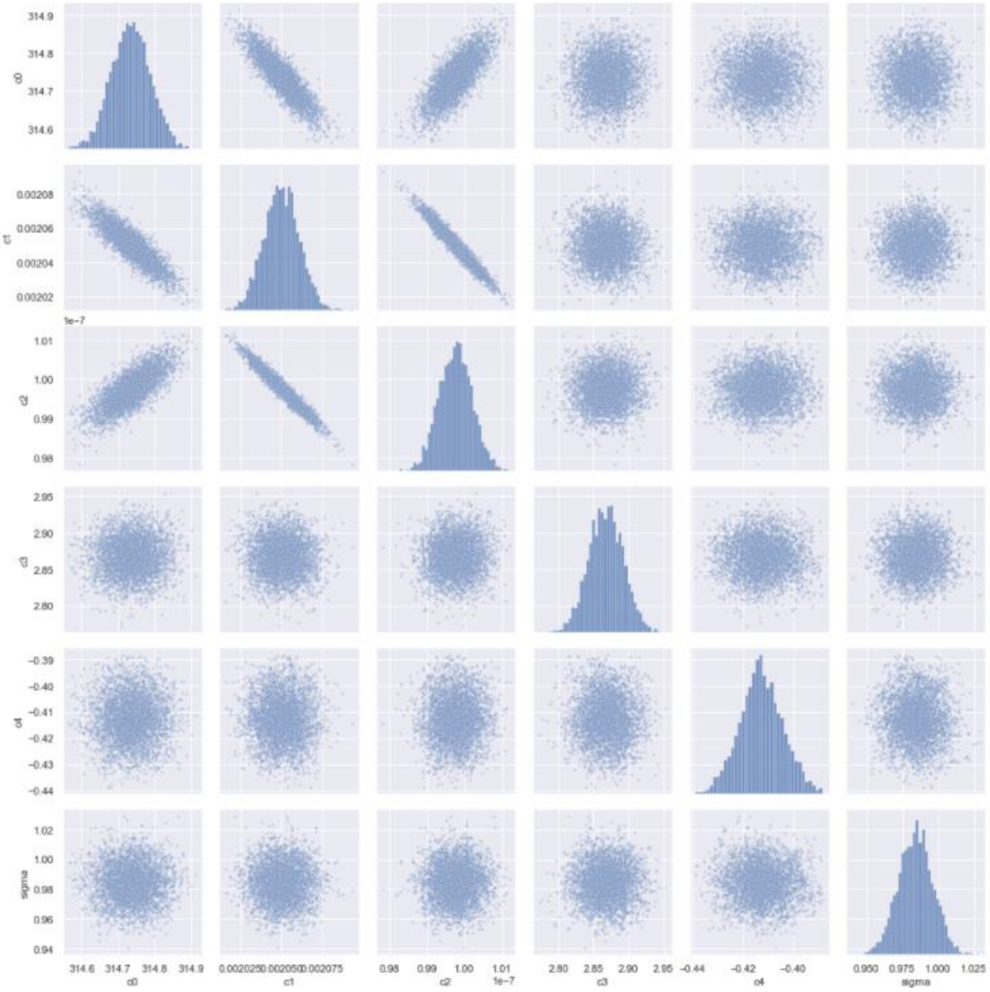
Appendix A: Sampling Results from the Suggested Model



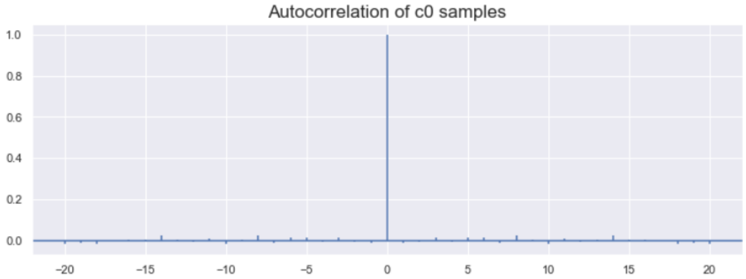
```
Inference for Stan model: anon_model_36cf0bf02010238699728d00346b64e4.
4 chains, each with iter=2000; warmup=1000; thin=1;
post-warmup draws per chain=1000, total post-warmup draws=4000.

         mean se_mean      sd   2.5%    25%     50%    75%  97.5%  n_eff   Rhat
c0     314.73  1.0e-3    0.05 314.63  314.7  314.74 314.77 314.84   2880    1.0
c1     2.1e-3  2.2e-7  1.1e-5 2.0e-3 2.0e-3  2.1e-3 2.1e-3 2.1e-3   2348    1.0
c2    10.0e-8 9.2e-12 4.4e-10 9.9e-8 9.9e-8 10.0e-8 1.0e-7 1.0e-7   2305    1.0
c3       2.87  4.5e-4    0.02   2.82   2.85    2.87   2.88   2.92   2756    1.0
c4      -0.41  1.7e-4  8.5e-3  -0.43  -0.42   -0.41  -0.41   -0.4   2378    1.0
sigma    0.98  2.5e-4    0.01   0.96   0.98    0.98   0.99   1.01   2543    1.0

Samples were drawn using NUTS at Tue Dec 14 01:22:15 2021.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

Appendix B: Model Assessment

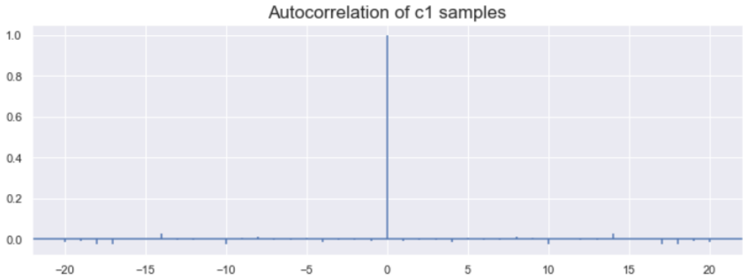
Pair plot



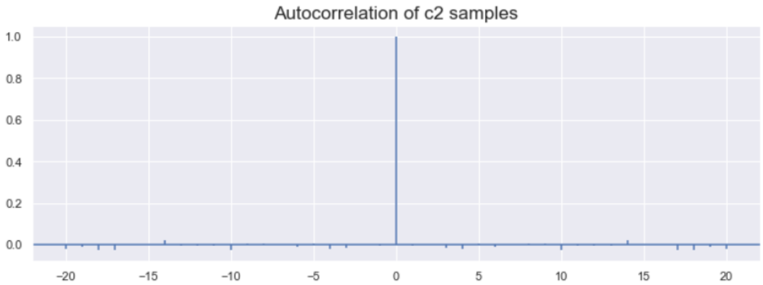
Autocorrelation plots



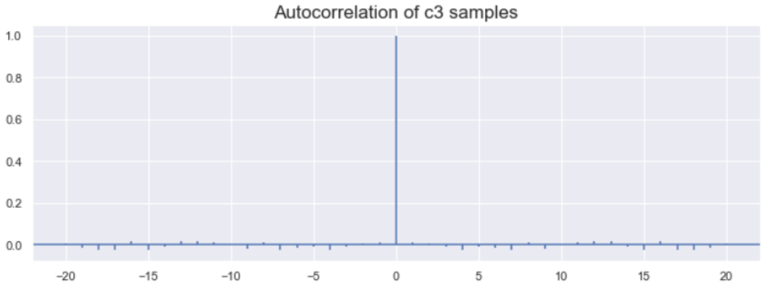
Autocorrelation of c0 samples



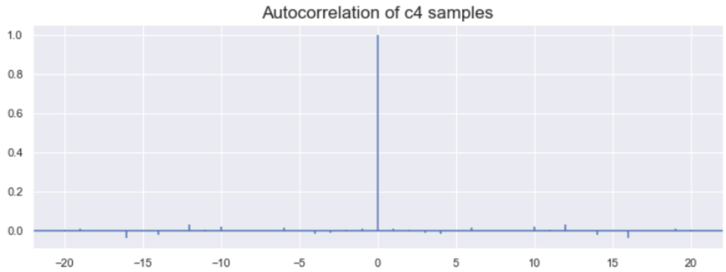
Autocorrelation of c1 samples



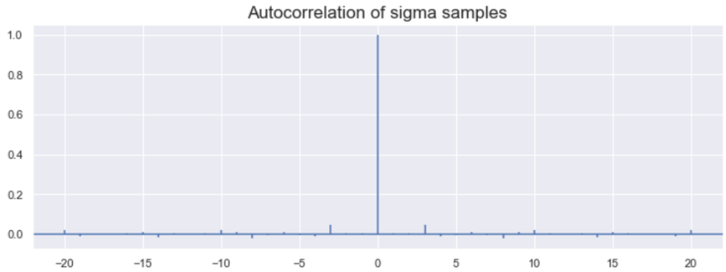
Autocorrelation of c2 samples



Autocorrelation of c3 samples



Autocorrelation of c4 samples



Autocorrelation of sigma samples