Statistical Model Report

Modeling and forecasting atmospheric CO_2 from 1958 into the future

---

Proposed by:

Carlos Rafael Garduño Acolt

For:

Minerva Schools at KGI, CS146: Modern Computational Statistics

San Francisco; United States - Amsterdam; Netherlands

April, 2021

Table of Contents

The scenario:

Ever since the beginning of the industrial revolution in the 18th century, levels of CO2 in the atmosphere have been steadily increasing. CO2 is a greenhouse gas and when it collects in the Earth's atmosphere it absorbs sunlight and solar radiation bouncing off the planet's surface. Greenhouse gases trap heat which would otherwise escape into space, leading to climate change, particularly global warming. Understanding the consequences of burning fossil fuels is fundamental while ensuring humanity does not endanger life on Earth, including human communities.

The Mauna Loa Observatory in Hawaii contributes to the search for such understanding by making weekly measurements of concentrations of CO2 in the atmosphere (recorded in parts per million, ppm). Such measurements have been carried since 1958 and are empirical evidence of global warming (See Figure 1). We can take advantage of the dataset by building a model that explains temperature variations through the years. Model which then can be used to forecast future measurements of CO2 levels in the atmosphere of the Earth if the current trend continues (no effective actions taken to stop global warming).

This report summarizes and presents the results of the statistical analysis carried on the data set, where CO2 measurements of over 60 years are used to predict future concentrations of the green gas over the next 40 years.

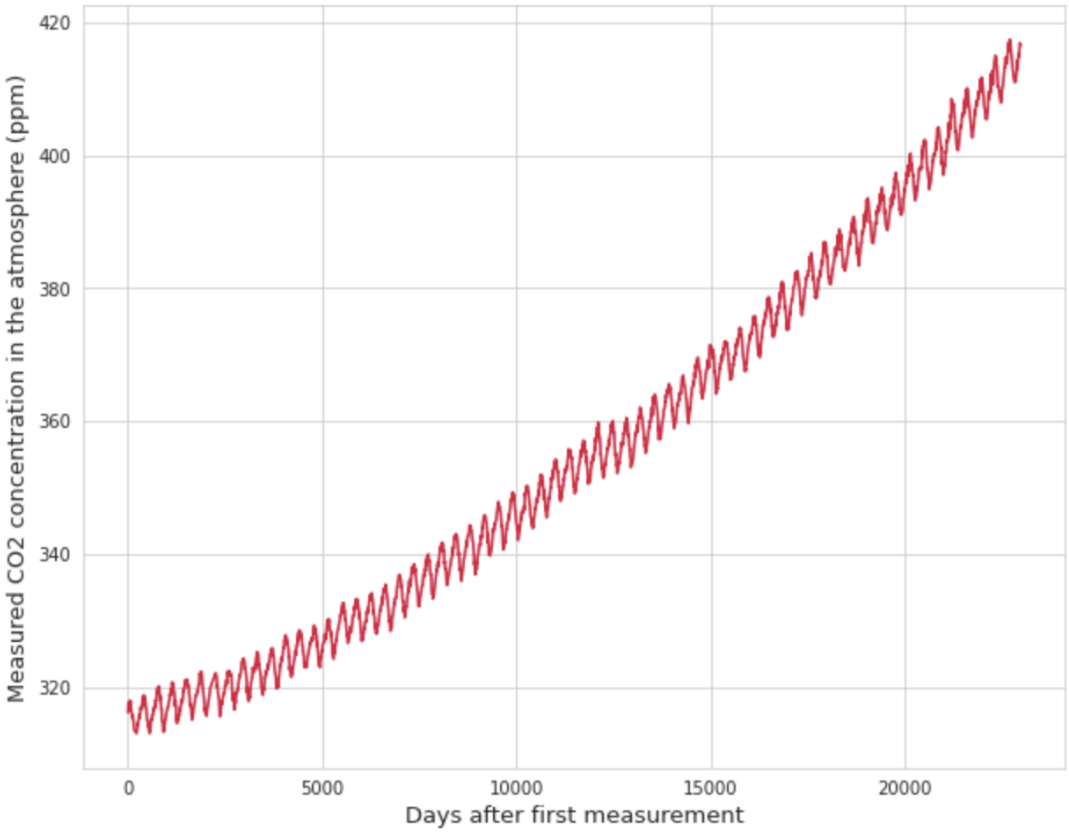


Figure 1: Weekly measurements of CO2 concentrations in the Earth's atmosphere starting in 1958 and ending in 2021.

About the statistical model:

Assumptions

We assume CO2 level measurements are accurate and representative of the actual variation through the years. The old model (example model) assumes the long-term trend grows linearly, seasonal variations are cyclical and can be modeled using a cosine function, and random fluctuations (noise) are normally distributed.

Parameters

More specifically, the likelihood function of the old model is:

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

Where:

$c_0 + c_1 t$ = Linear long-term trend

$c_2 cos(2\pi t/365.25 + c_3)$= Seasonal variation (every 365¼ days)

c_4 = Standard deviation of Gaussian portraying random fluctuations (noise)

θ= Set of unobservable parameters ($c_0,\ c_1,\ c_2, c_3,\ c_4$)

x_t= Values, CO2 ppm measurements

t= Time, number of days since measurements started in 1958

Observed and unobserved quantities

Our likelihood function is constructed of both observed and unobserved quantities.

There are 5 unobservable parameters $\theta = (c_0, c_1, c_2, c_3, c_4)$, for which we establish prior distributions.

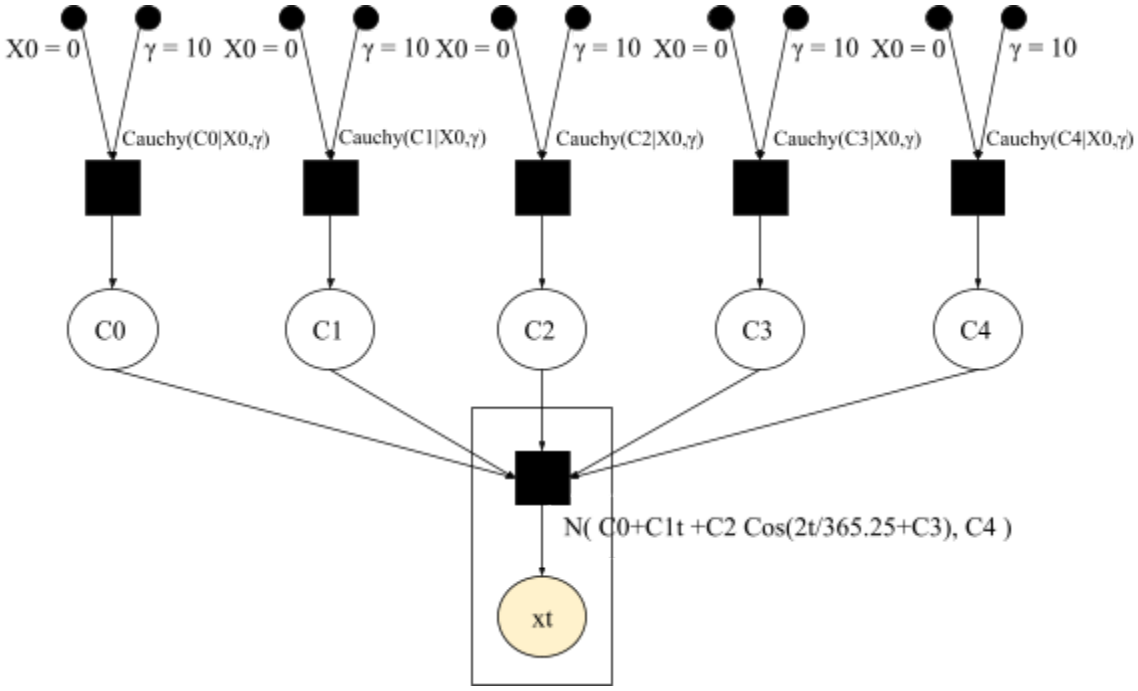
There are 2 observable quantities (the two columns in the original data file). One is the CO2 ppm weekly measurements, and the second is the time of each measurement, which we transformed from date recordings (YYYY-MM-DD) to a positive integer representing the number of days after the very first CO2 ppm measurement (starting on March 29th, 1958).

Priors over parameters

Before being able to input the model structure into STAN for carrying out inference, we must specify prior distributions for each one of the model's unobservable parameters. Priors can either encompass our a priori belief about the parameter value or they can be uninformative. For this particular case we have no context about the parameter values. We could try to infer those from the data set and the way the trend looks. Nevertheless, conditioning our prior distributions using observations from the data set (which will build the likelihood) will overfit the model. Furthermore, we are working with a large data set (3210 observations), which means that any prior will have minimal effect over the posterior values relative to the data. In other words, by using an uninformative prior, we assume very little about the parameter values, and we let the data take over while defining the posterior. An uninformative prior for a normal distribution as likelihood (as the case) is a Cauchy distribution. The uninformative Cauchy distribution is

similar to a Gaussian, but it has heavier tails, which implies its support is infinite. The prior distributions will be centered at 0, with a scale parameter of 10 as it would not be surprising to see such variation in the value of the parameters.

Factor Graph



Inference in the model:

Once we have defined the priors over parameters and the original (old model) likelihood function, we proceed to get the posterior distributions using STAN through Python.

The STAN package is a tool for automated inference. Once a Bayesian model is specified, it generates samples from the posterior distribution by implementing a Hamiltonian Monte Carlo algorithm. Although the program is meant to ease the computation of posteriors, results are not always perfect (e.g. highly correlated samples). To check for the usefulness of output results we can look at two metrics of correlation presented in the STAN results: n_eff (effective number of samples, an estimate of how many of the 4000 samples are uncorrelated or independent), and $\widehat{R}$ (Rhat, an estimate of whether the MCMC chain has converged). Reasonable values for these metrics are a few hundred for n_eff, and an Rhat value of around 1. Rhat values of more than 1.1 mean the Markov chains are not mixing well and the MCMC chain is not converging.

First model (linear long-term increase)

The STAN results for our first model were:

```
WARNING:pystan:n_eff / iter below 0.001 indicates that the effective sample size has likely been overestimated
WARNING:pystan:Rhat above 1.1 or below 0.9 indicates that the chains very likely have not mixed
Inference for Stan model: anon_model_66516250502160067cd23563ed4832f1.
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   305.64  2.4e-3   0.14 305.35 305.54 305.64 305.74 305.92   3778    1.0
c1   4.4e-3  1.7e-7 1.1e-5 4.3e-3 4.4e-3 4.4e-3 4.4e-3 4.4e-3   4036    1.0
c2     2.16    0.72   1.02   0.26   1.48   2.71   2.79   2.93      2  10.56
c3      2.3    2.19    3.1   0.41    0.5   0.55   3.98   8.05      2  20.04
c4     4.14    0.13   0.19   3.94   4.01   4.05   4.26   4.53      2   4.21
lp__  -6162  100.69 142.44  -6411  -6264  -6080  -6079  -6078      2  95.37

Samples were drawn using NUTS at Tue Apr 20 19:46:41 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).
```

As seen on the last two columns of the output, n_eff and Rhat values are outside the range of acceptable values. The Markov chains are not mixing well, not getting to convergence, and the effective number of samples is too small. Another measurement of the usefulness of the samples comes from looking at the autocorrelation and pair plots of the results (See Figure 2 and Figure 3). It is clear from Figure 2 that samples of c2, c3, and c4 are highly correlated. We would expect all 5 plots to look like the first 2 if samples were independent. From Figure 3 we see that samples of c2, c3, and c4 are bimodally distributed, and do not follow the unimodal Gaussian shape we would expect, hinting at the lack of convergence after inference.

To solve the issues and generate a larger number of effective samples, some reparametrization in the STAN model definition can work. After a few rounds of trial-and-error, changing the parameter bounds for c2, c3 was enough to generate better samples. More specifically, the second version of the model sets upper bounds for c2 and c3 at 10 and 5, respectively.

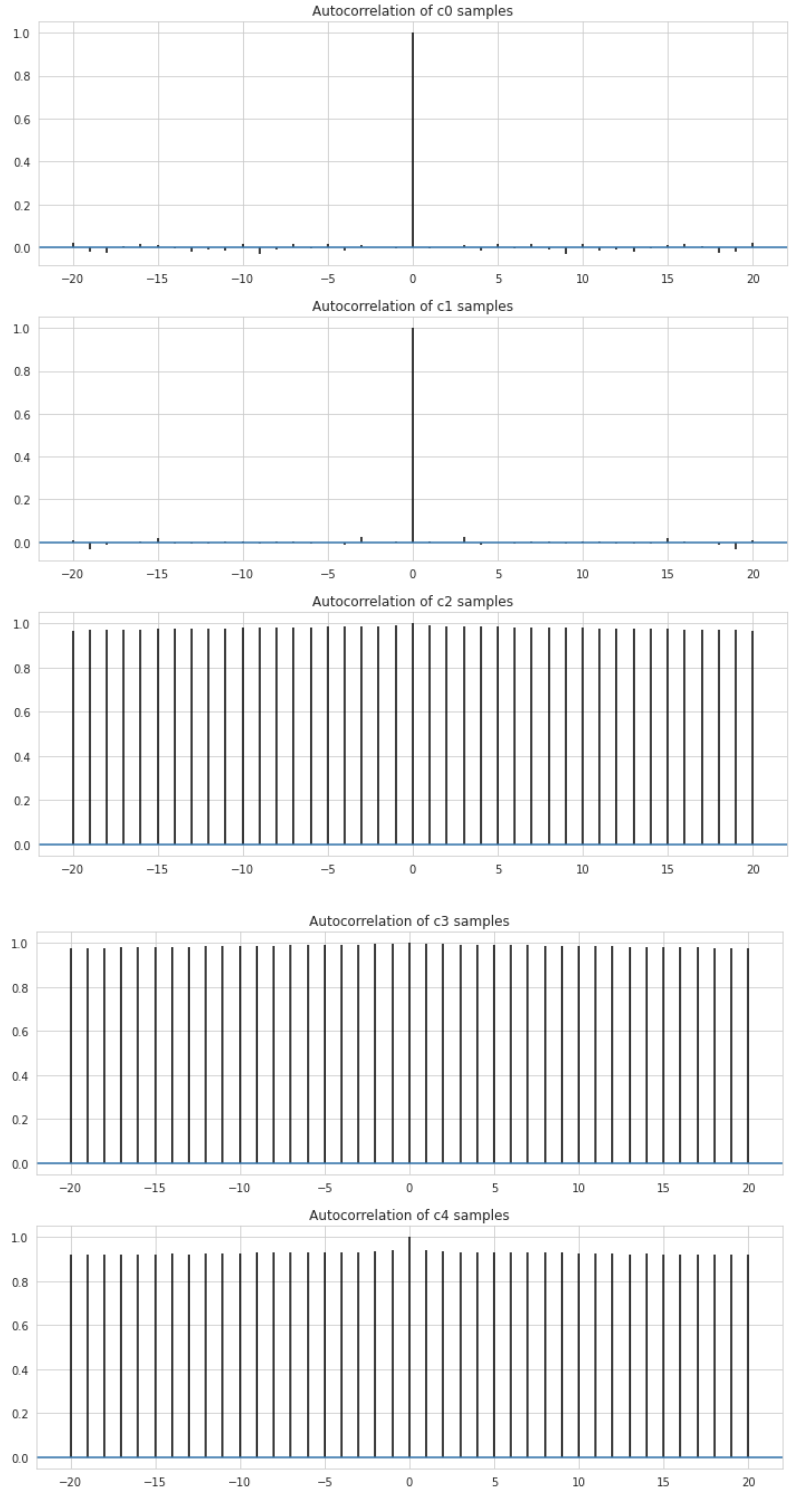


Figure 2. Autocorrelation plots of first STAN input model. Samples of unobserved parameters c2, c3, and c4 are highly correlated.

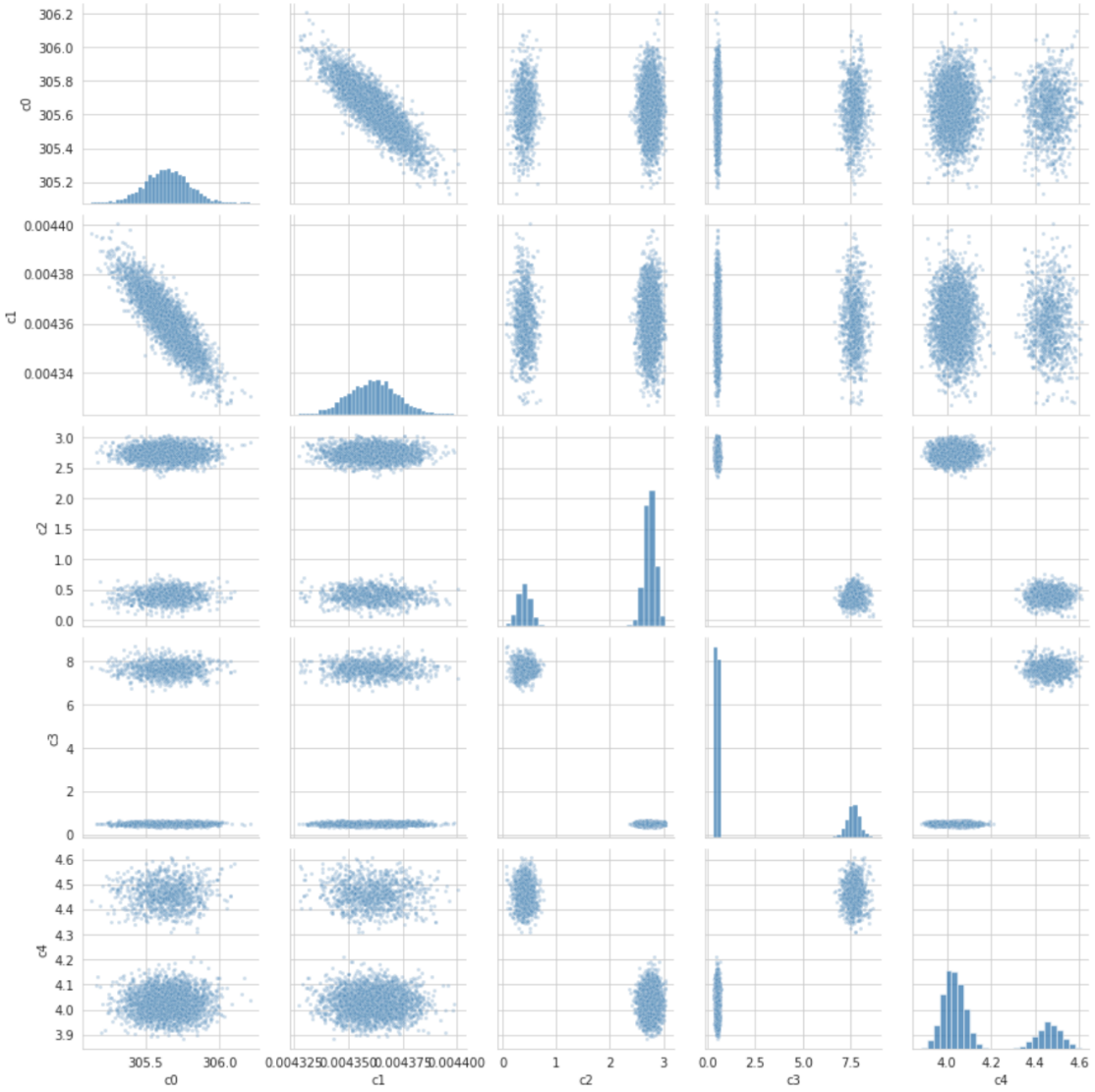


Figure 3. Pair plots of first STAN input model. Sample distributions of unobserved parameters c2, c3, and c4 follow a bimodal distribution rather than a unimodal Gaussian (as one would expect).

The STAN results for our second version of the model were:

```
Inference for Stan model: anon_model_d21b89b006e652140f18c6eb036d9124.
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   305.64  2.4e-3   0.15 305.36 305.54 305.65 305.74 305.93   3895    1.0
c1   4.4e-3  1.7e-7 1.1e-5 4.3e-3 4.4e-3 4.4e-3 4.4e-3 4.4e-3   4100    1.0
c2     2.75  2.1e-3    0.1   2.56   2.68   2.75   2.81   2.95   2303    1.0
c3     0.52  1.2e-3   0.06    0.4   0.48   0.52   0.56   0.64   2223    1.0
c4     4.03  1.0e-3   0.05   3.93   3.99   4.03   4.06   4.12   2459    1.0
lp__  -6080    0.04   1.63  -6084  -6081  -6080  -6079  -6078   1597    1.0

Samples were drawn using NUTS at Tue Apr 20 20:50:32 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).
```

This time values for n_eff are all over 1500, and Rhat values are all exactly 1. We have succeeded while building a model which after STAN's automated inference leads to a sufficiently large number of independent samples. Figures 4 and 5 present Autocorrelation and Pair plots for the second version of the model. The new Autocorrelation plots have only values close to 0, which means the samples can be considered independent or uncorrelated. The new Pair plots show all five distributions for the parameters following a Gaussian-like distribution, which we expect due to our definition of Cauchy priors.

Now that we have a first version of a model describing the changes of concentration of CO2 in the atmosphere, we can compare it to the real-world data to see how good the approximation is. Figure 6 presents a visual comparison of the modeled trend and the real-world measurements. The modeled version follows the positive change of CO2 ppm through the years quite closely. However, it is clear that it can be improved. The increase of CO2 ppm does not seem to follow a linear trend in the long-term, but it seems to be increasing exponentially.

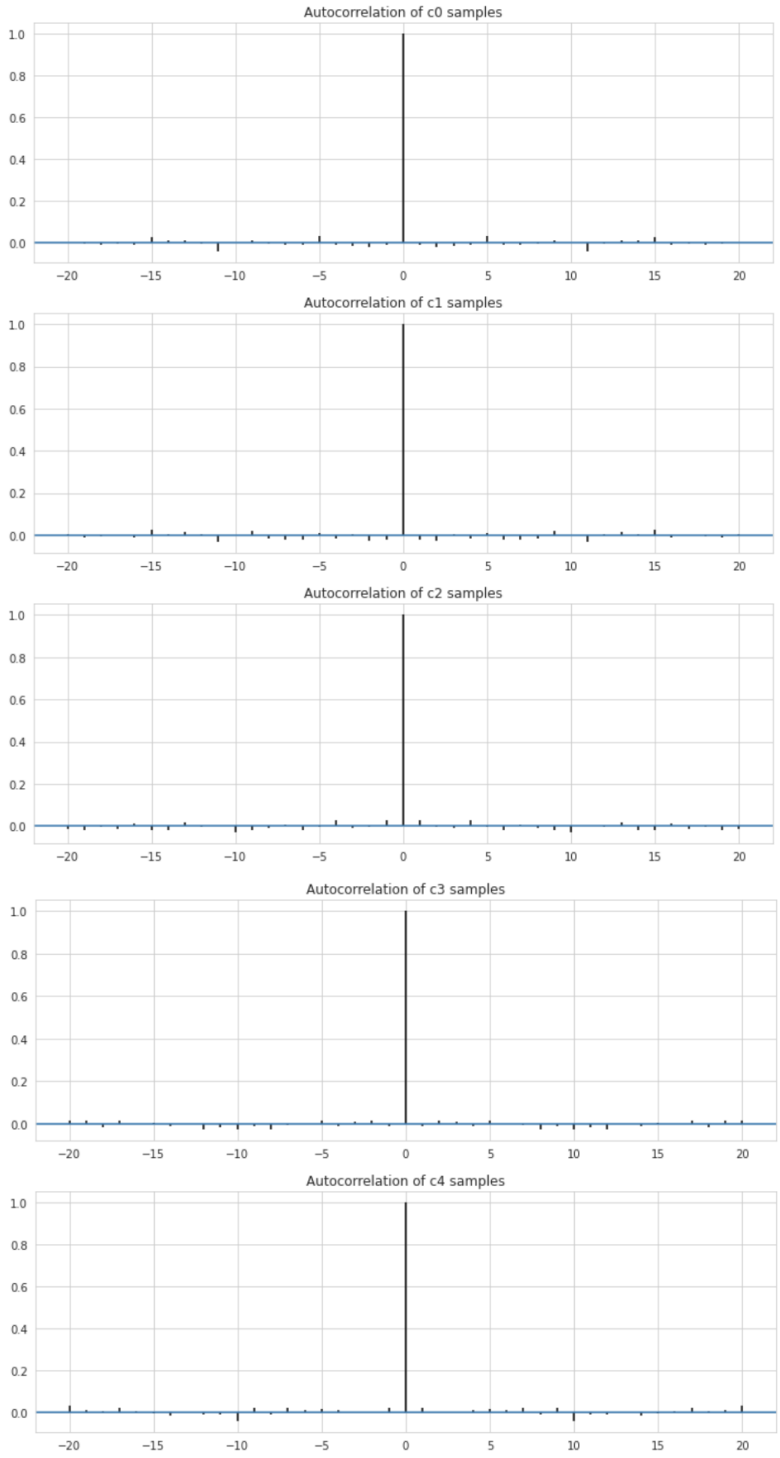


Figure 4. Autocorrelation plots of second version of STAN input model (after reparametrization to avoid correlation of samples). Samples of all unobserved parameters are sufficiently uncorrelated.

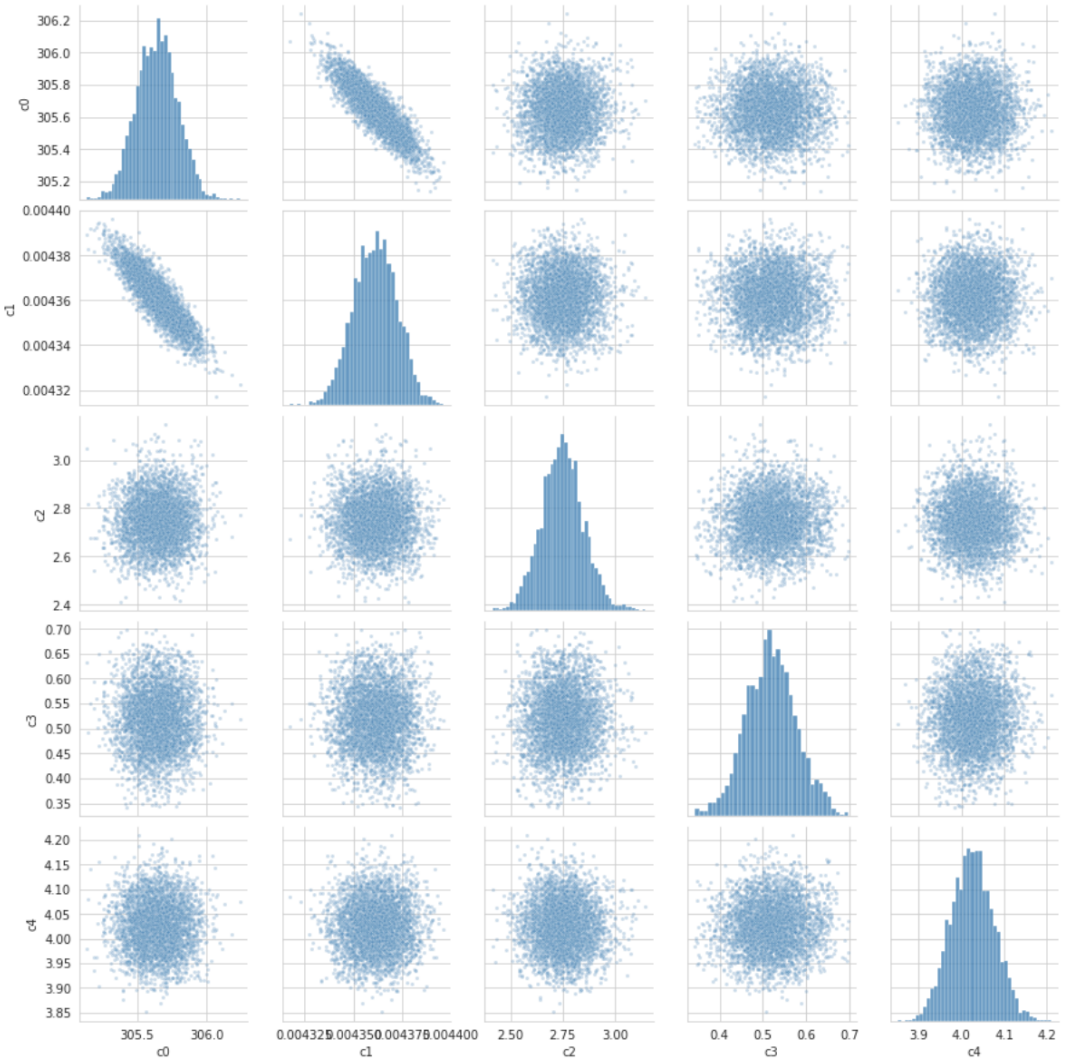


Figure 5. Pair plots of second version of STAN input model. Sample distributions of all unobserved parameters follow a Gaussian-like distribution. Note c0 and c1 present negative correlation between their values.

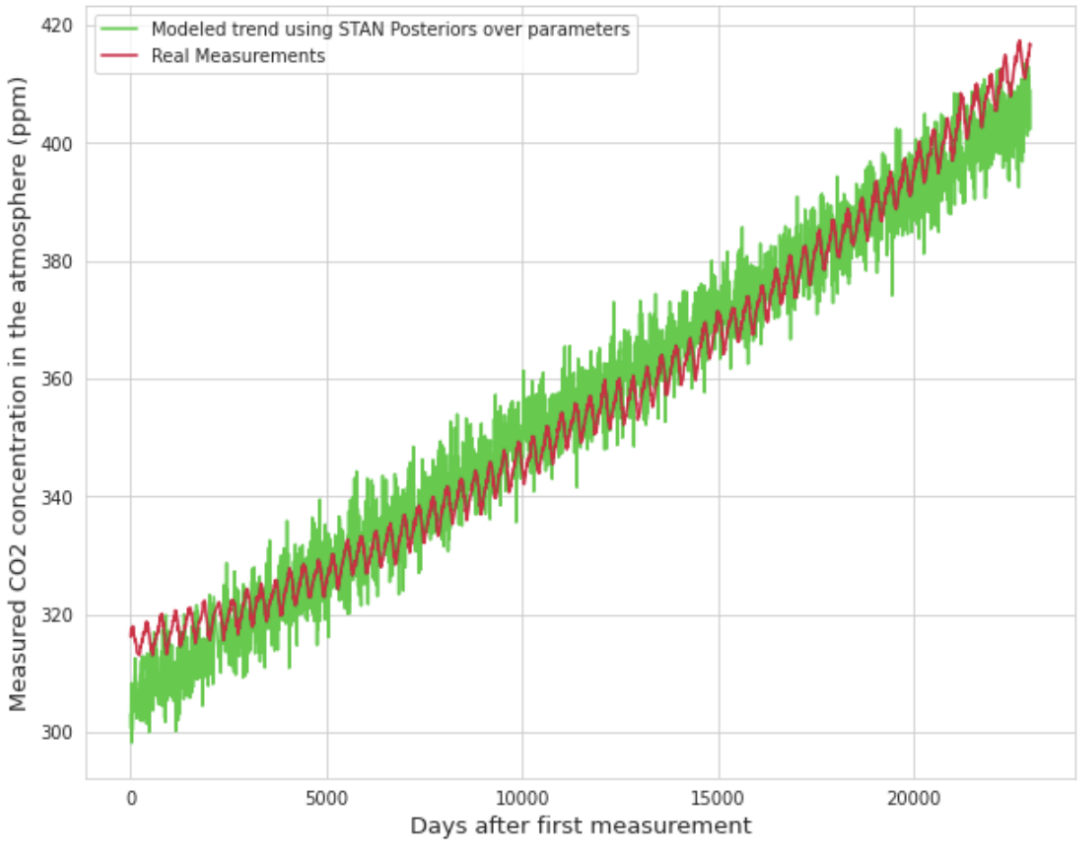


Figure 6. Real-world changes of CO2 concentration in the atmosphere through the years (in red) compared to modeled estimates using STAN posteriors (in green). The model approximates long-term variations as a linear increase with respect to time.

Second model (quadratic long-term increase)

Given the previous results, we can build on top of the previous model hoping to improve our results. For the next iteration we will increase the degree of the factors defining the long-term behavior of the data. For the previous model this was approximated with the term $c_0 + c_1 t$ in the likelihood function. Now, we try $c_0 + c_1 t^2$ since the red line (representing actual measurements)

seems to be an upward curve, similar to what we see from a quadratic function. The, the new likelihood is defined as:

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

All other parameters will stay the same. We run the new model through STAN, results are the following:

```
Inference for Stan model: anon_model_c04feeaea8c5c9642d86769234d0fdde.
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   323.78  1.4e-3     0.1  323.6 323.72 323.78 323.85 323.97   4900    1.0
c1   1.8e-7 5.8e-12 4.0e-10 1.8e-7 1.8e-7 1.8e-7 1.8e-7 1.8e-7   4850    1.0
c2      2.8  1.8e-3    0.09   2.62   2.74    2.8   2.86   2.98   2351    1.0
c3     0.54  9.8e-4    0.05   0.44    0.5   0.54   0.57   0.63   2444    1.0
c4     3.52  8.8e-4    0.04   3.43   3.48   3.52   3.55    3.6   2562    1.0
lp__  -5656    0.04    1.65  -5660  -5657  -5655  -5655  -5654   1561    1.0

Samples were drawn using NUTS at Tue Apr 20 22:49:30 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).
```

N_eff and Rhat values for all parameters are reasonable. Autocorrelation and Pair plots for the second model can be found in the zipped Notebook. Figure 7 compares the approximation to the actual measurements.

The second model seems to improve while estimating the first and last values, while the approximations for those measurements in the middle are not as closely followed.

Third model (exponential long-term increase)

To improve our estimates we can redefine the model one more time. This time we will model the long-term increase as exponential (exp = 1.5) rather than quadratic or linear. Where the part of the likelihood defining the long-term trend is set to $c_0 + c_1 t^{1.5}$. The new likelihood is defined as:

$$p(x_t|\theta) = N(c_0 + c_1 t^{1.5} + c_2 cos(2\pi t/365.25 + c_3),\ c_4^{\ 2})$$

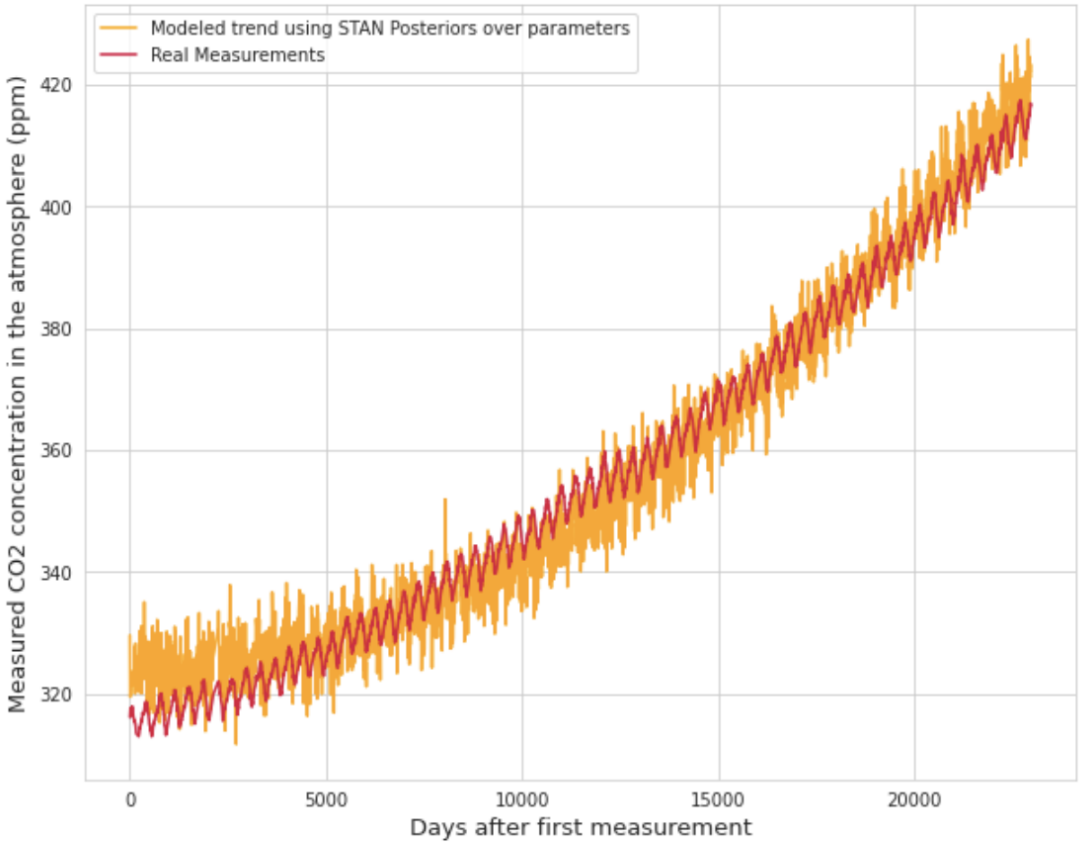


Figure 7. Real-world changes of CO2 concentration in the atmosphere through the years (in red) compared to modeled estimates using STAN posteriors (in orange). The model approximates long-term variations as a quadratic increase with respect to time.

All other parameters will stay the same. We run the new model through STAN, results are the following:

```
Inference for Stan model: anon_model_3979a6ad7609b961c0a67ccb3591cb1b.
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   317.14  5.5e-4   0.04 317.07 317.12 317.14 317.17 317.21   4272    1.0
c1   2.8e-5 3.1e-10 2.1e-8 2.8e-5 2.8e-5 2.8e-5 2.8e-5 2.8e-5   4339    1.0
c2     2.77  7.0e-4   0.03   2.71   2.75   2.77   2.79   2.83   1914    1.0
c3     0.53  3.8e-4   0.02   0.49   0.52   0.53   0.54   0.56   1987    1.0
c4     1.22  3.2e-4   0.01   1.19   1.21   1.22   1.23   1.25   2160    1.0
```

Once again, values for n_eff and Rhat are reasonable for all parameters. Autocorrelation and Pair plots for the second model can be found in the zipped Notebook. Figure 8 compares the approximation to the actual measurements.

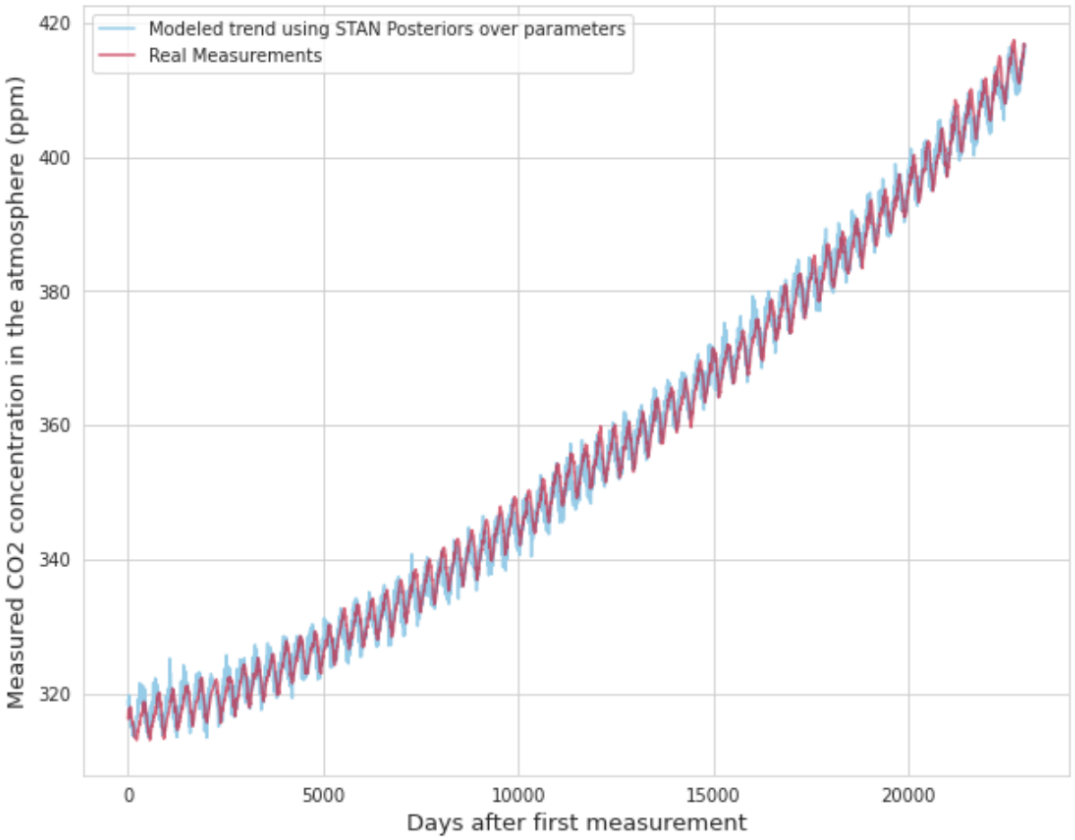


Figure 8. Real-world changes of CO2 concentration in the atmosphere through the years (in red) compared to modeled estimates using STAN posteriors (in blue). The model approximates long-term variations as an exponential increase (exp=1.5) with respect to time.

The new approximation is a great improvement from the previous two attempts. The estimates closely follow the increase of the actual measurements through time. Figure 9 compares estimates from the three different models (linear, quadratic, and exponential long-term increase) with the actual data set.

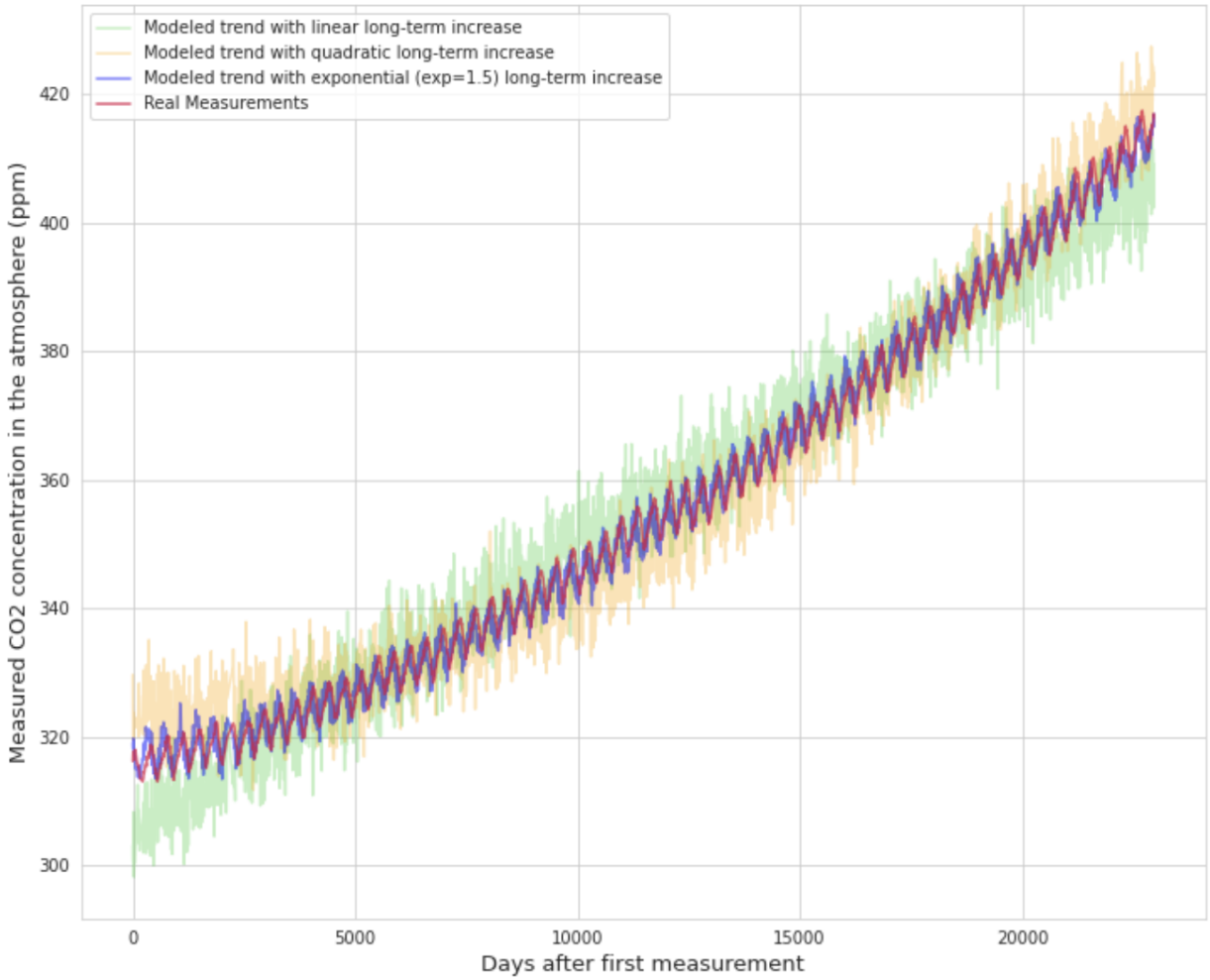


Figure 9. Real-world changes of CO2 concentration in the atmosphere through the years (in red) compared to modeled estimates using STAN posteriors. The models approximate long-term variations as a linear, quadratic, and exponential increase (exp=1.5) with respect to time. The third model (in blue) is the one that follows the actual measurements the closest.

Posterior predictive checks

Although it is clear that our third attempt at the model is the best one, we can support this argument quantitatively by applying posterior predictive checks. For this technique, we compute a test statistic from the real data. Then, we take samples from our posterior distribution and create a distribution over such test statistic. After that we compare the test statistic value of the real data set with the distribution of test statistics from sampling the posterior. We ask the question "How unusual is the test statistic of the real data, given the test statistic of the replicated samples?" The more unusual the test statistic, the less likely it is that our model design is accurate.

We can quantify how unusual the test statistic of the real dataset is with a p-value (different form the widely-known p-value). The p-value must be quite extreme before worrying about it. Values between 0.05 and 0.95 support the idea that the model is accurate.

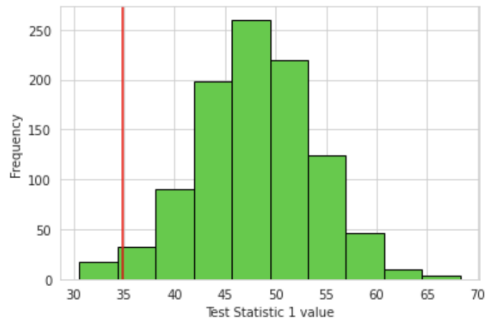
While using the mean as a test statistic is common, given the shape of the trend we expect not to get great insight from it. All of our three models are quite symmetric around the center, so the mean values will converge (extrema will 'cancel each other out,' leading to similar values.

Since the biggest differences among models are located among the first and last hundreds of days after the first measurement, we define two test statistics as:

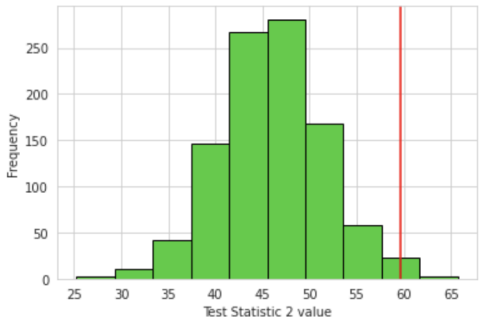
1. Absolute change in CO2 concentrations in the first 1000 data points
2. Absolute change in CO2 concentrations in the last 1000 data points

The results of the posterior predictive checks are the following (red line signals where the value of the test statistic for the real data set lies along each distribution):

For the first model (linear long-term increment):

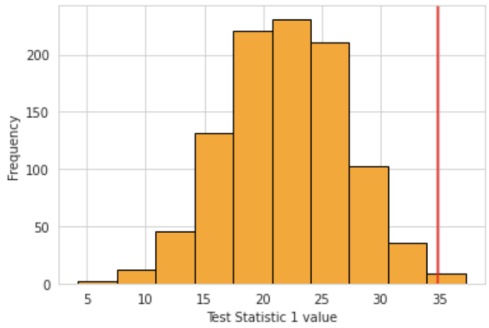


P-value = 0.982

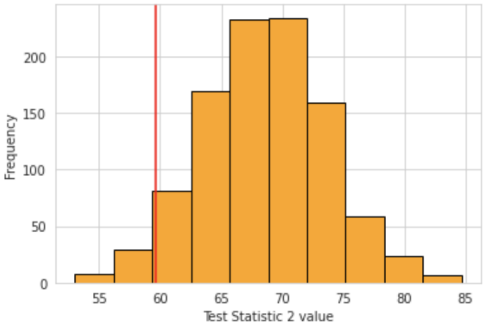


P-value = 0.01

For the second model (quadratic long-term increment):

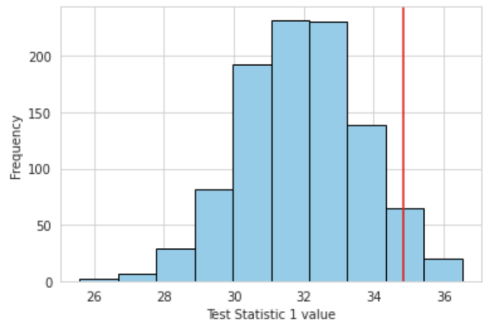


P-value = 0.005

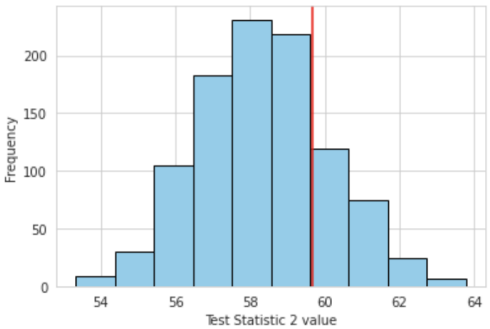


P-value = 0.958

For the third model (exponential long-term increment):



P-value = 0.044



P-value = 0.218

The presented results agree with the argument we made before about the third model being the best while modeling the increase of CO2 concentrations. All four p-values for the first two models fall outside the range we established as desired ($p>0.05$, $p<0.95$). In regards to the third model results, the second p-value is well inside the desired range. The first p-value for the third model is barely outside of the desirable range (0.044). Still, it represents the best p-value obtained through the first test.

Results:

Now that we have a defined model, we move into using the existing posteriors to forecast what atmospheric CO2 levels will be for the next 40 years (as measured from Mauna Loa). We use STAN to generate predictions for what the weekly measurements of CO2 should look like until 2060 assuming the current trend stays the same. Figure 10 shows the predicted trend.

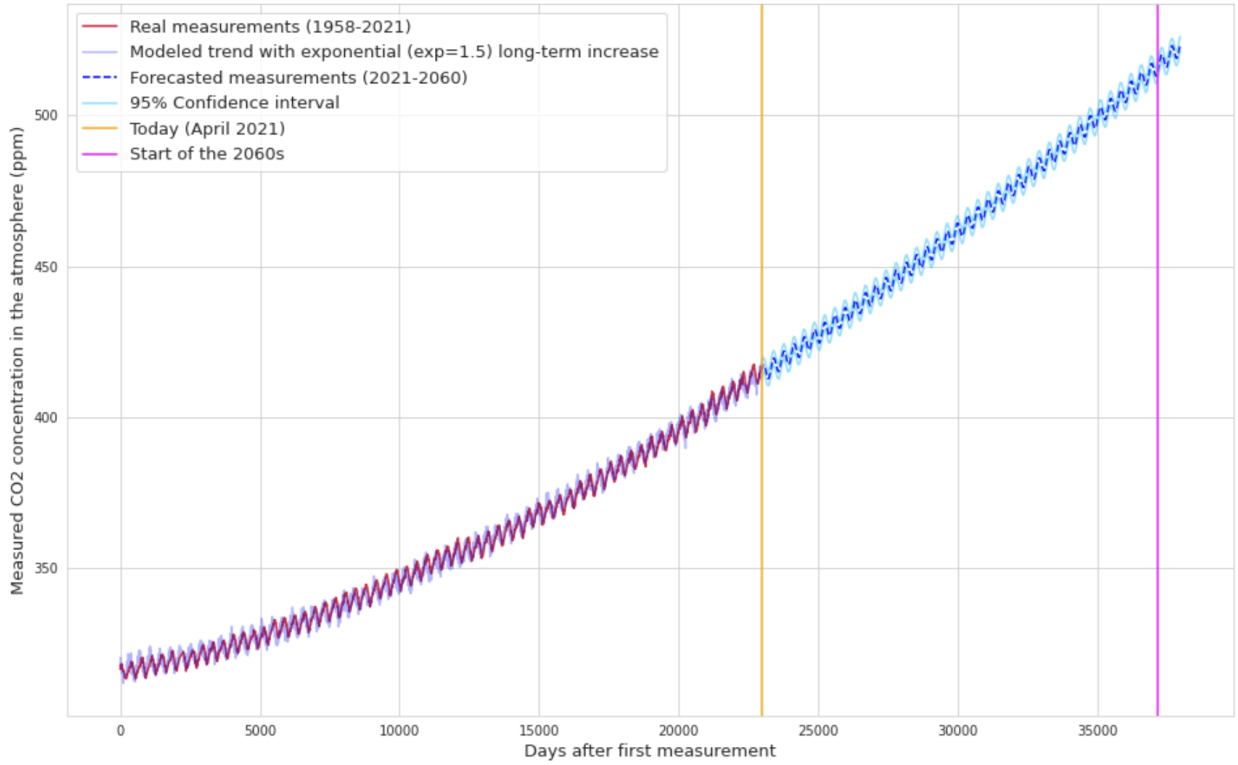


Figure 10. Results from modeling atmospheric CO2 through the years (1958-2021), and forecast of CO2 measurements over the next 40 years. The plot presents the actual measurements in red, the modeled past and future measurements in blue (with 95% confidence intervals for future (predicted) values. Vertical lines mark current time (last and most recent measurement taken into account) and date (in terms of days after first measurement) of the beginning of 2060.

As seen in the figure, we have computed s forecast of how concentrations of atmospheric CO2 should behave assuming the current trend remains the same over the next 40 years. Figure 11 presents only those forecasted values and the 95% confidence interval around them.

Estimate and 95% confidence interval for atmospheric CO2 levels projected until the start of 2060

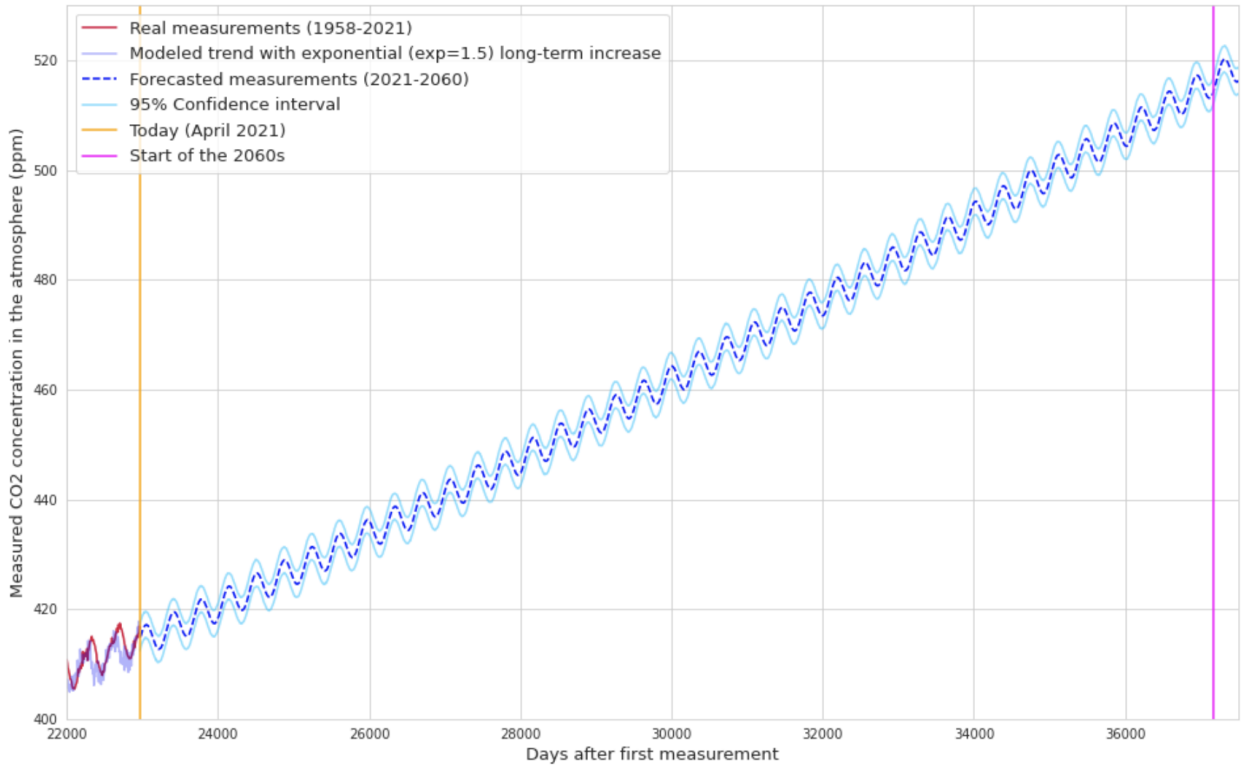


Figure 11. Measurement predictions for atmospheric CO2 concentrations over the next 40 years (2021-2060). Averages are shown with dashed blue lines, 95% Confidence intervals in light blue.

Inferring from Mauna Loa Observatory's CO2 measurements from the last decades, we expect CO2 levels to keep rapidly increasing. Our model predicts atmospheric CO2 levels to go up by around 100ppm over the next 40 years, a 25% overall increase from current levels. Figure 12 presents a closer look at the predicted trend right at the beginning of 2060.

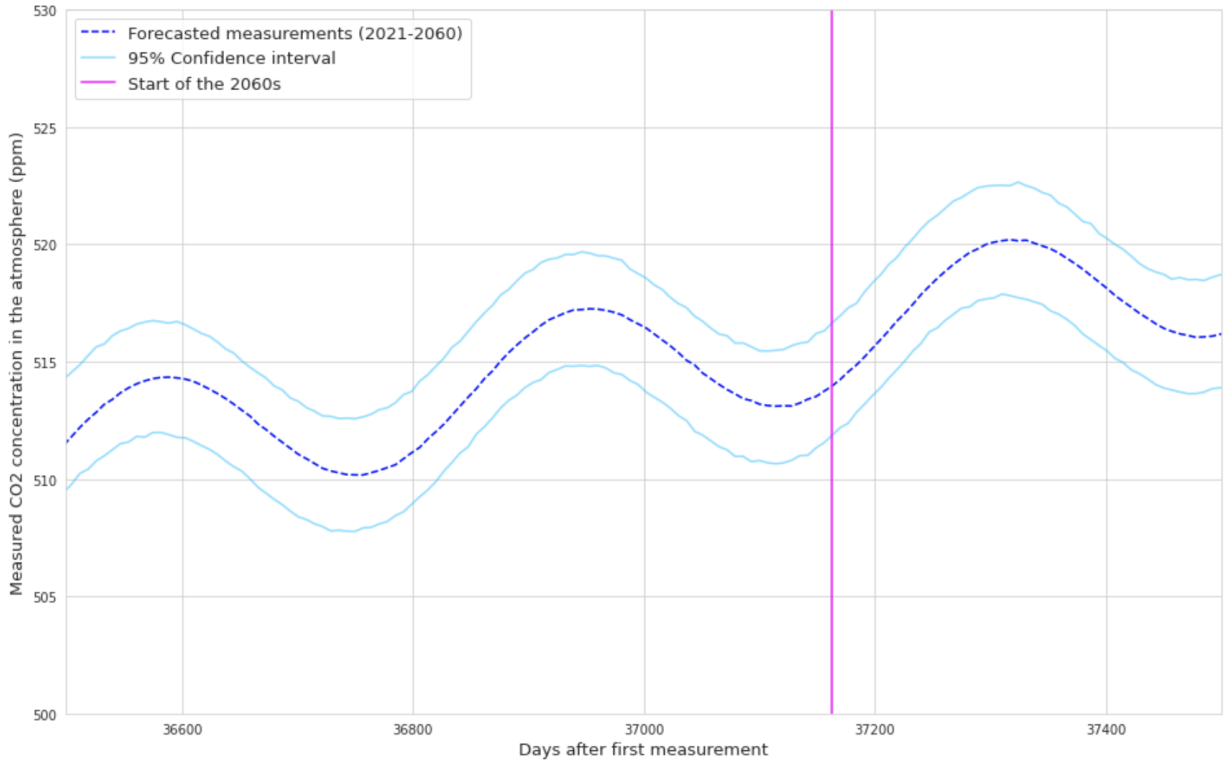


Figure 12. Predicted atmospheric CO2 concentrations at the beginning of the 2060s (timestamp shown in pink). Assuming the current trend stays the same, we predict atmospheric CO2 concentrations should rise to around 514ppm by the beginning of the year 2060.

By the year 2060 we expect Mauna Loa Observatory's atmospheric CO2 measurements to rise to a dangerous high of around 514ppm. Computed confidence intervals are quite narrow, meaning that according to our model we can be pretty certain we will rich those high risk levels if the increasing trend stays the same.

Predictions on high risk CO2 levels over the next 40 years

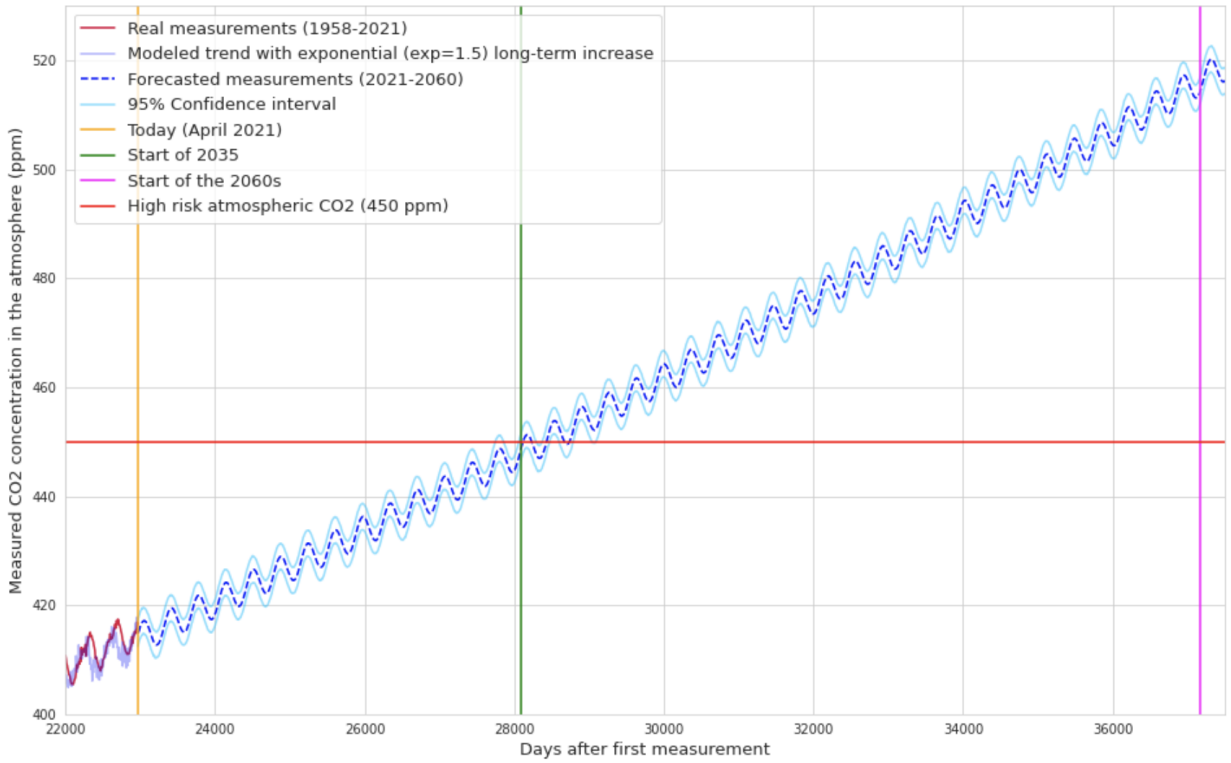


Figure 13. Measurement predictions for atmospheric CO2 concentrations over the next 40 years (2021-2060). Averages are shown with dashed blue lines, 95% Confidence intervals in light blue. High risk atmospheric CO2 levels are marked by the horizontal red line. Vertical green line shows the timestamp when we predict we should measure high risk CO2 levels (beginning of 2035).

An atmospheric CO2 level of 450ppm is considered high risk for dangerous climate change which would certainly compromise the safety of most life on Earth. According to our model, such dangerous levels should be reached by the beginning of 2035 if no further actions are taken to stop CO2 concentrations from increasing so rapidly. Figure 13 shows predictions for the next 40 years, marking with the red horizontal line the 450ppm mark and with the vertical green line the point in time when high risk CO2 levels should be reached. Figure 14 shows a closer look to the timestamp when we predict high risk CO2 levels should be reached.

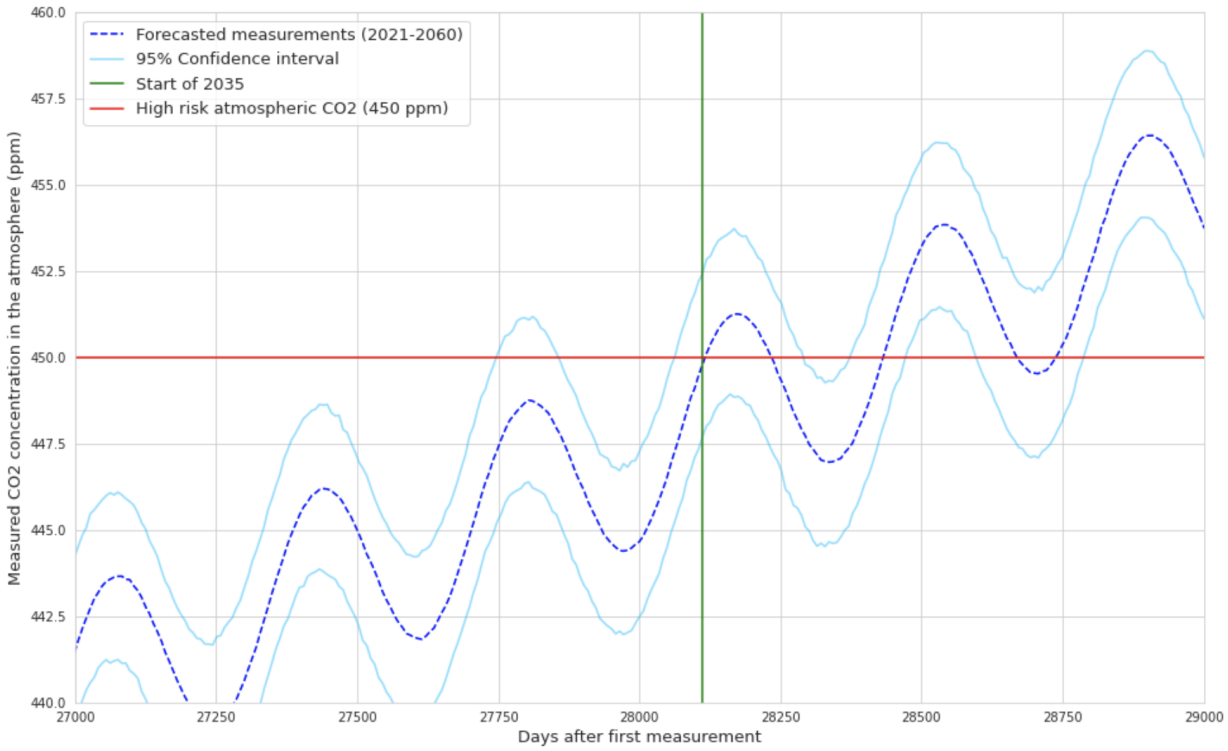


Figure 14. Predicted timestamp of when atmospheric CO2 measurements should reach high risk levels. Specifically, we expect the 450 ppm high risk mark to be reached in the first months of 2035.

We have forecasted average values for CO2 measurements, as well as 95% confidence intervals for those. The bounds of our 95% confidence intervals (shown in light blue in Figure 14) represent the uncertainty of our results. The reader should interpret the range between the lower and upper bounds as the overall prediction for each timestamp.

For instance, we expect to reach a CO2 concentration of 450ppm in the first few months of 2035 if the current trend stays the same. We are not 100% confident on that given that our model is only an approximation of the real world. Therefore, we have computed the lower and upper bound of a 95% confidence interval to be [511ppm, 516ppm]. Therefore, we are 95% certain that average CO2 levels will be inside of that range by the start of 2060.

Our model's results are alarming. We predict that high risk atmospheric CO2 levels should be reached only 14 years from now. Furthermore, this is only if the current trend stays the same. It could be the case that human settlements could increase their CO2 emissions, accelerating the dangerous increase of the green gas in the Earth's atmosphere, and making those levels reachable before 2035.

Our statistical model's results support the argument that an environmental catastrophe is approaching rapidly. Governments, companies and human societies must be aware of the dangerous consequences of climate change in the short, medium, and long terms. Worldwide intensive efforts to cut down on greenhouse gas emissions must be pursued immediately.

Furthermore, our model's results are only one piece of evidence of the climate crisis. Several research teams around the world are actively constructing and improving on highly-sophisticated models for CO2 levels forecasting, and most of them agree that action is necessary.

Shortcomings of the model:

Although the estimated values for CO2 measurements in the range 1958-2021 match the real dataset pretty well, there is clearly room for improvement. First of all, we have only estimated the values in the likelihood function with trial and error and based on a pre-existing model (given in the assignment prompt), which means we do not completely understand why the present trend has such an increase. In other words, the analyzed model is only a rough approximation to what happens in nature. To make the model more accurate, we could start by researching the variables which have affected CO2 concentrations in the atmosphere for the last centuries and then devise a likelihood function (similar to what is described for this case) which is a better representation of nature. Nevertheless, there are most probably many confounding variables and unknown factors that affect CO2 concentrations that this approach could be a highly difficult task and would require extensive understanding of the environmental sciences.