Final Project

Johannes M. Halkenhaeusser

Minerva Schools at KGI

CS146 Prof. Scheffler

Spring 2019

Code is available in the appendix, the attached file, and here:

https://gist.github.com/Halkenhaeusser/303d32b17ab1168c4efc8e6bd218292a

Table of Contents

Executive Summary *2*

Motivation *3*

The Data *4*

Modeling *6*

Full linear Model **6**

Parameters of periodic functions 7

Results 8

Introducing a test statistic 10

Quadratic model **12**

Improvements and set-up 12

Results 14

Inference and discussion 17

Works Cited *19*

Appendix *20*

Linear Sampling Checks **20**

Robustness checks for Quadratic Sampling **24**

Appendix - Code *28*

Executive Summary

Climate change has impacted biodiversity, caused draughts, storms, and will change the way humans live on our planet forever unless we can prevent reaching critical changes in CO_2 concentration levels of 450 ppm. By specifying a Bayesian model of CO_2 concentrations based on 60 years of data, we are able to show that CO_2 levels are increasing and that the rate of increase is accelerating. We reduce the error in seasonal variations by fitting a double sine curve. Predicting the long-term concentrations over the next 40 years, we estimate that we will reach these critical levels between 2034-02-22 and 2032-03-16 but most likely at 2033-03-09.

Motivation

Climate change can be seen as the complex yet pressing issue of our generation. Scientific evidence has been discredited by politicians at the highest level of government (BBC News, 2018). Having an accurate understanding of how gases like CO_2 that contribute to climate change have changed in recent decades is crucial in being able to make actionable predictions and communicate them clearly, all while being backed up by concrete data. Hence, this project will analyze data from the Mauna Loa Observatory in Hawaii that has tracked CO_2 concentration in the atmosphere for over 60 years now. We will explore how long-term trends and seasonal variation adjustment can be used to accurately model concentrations of the past and predict when critical CO_2 concentrations will be reached.

Figure 1. Climate change already affects the costs of Hawaii (NOAA, 2016)..

The Data

The data set is a weekly record of CO_2 concentration measured at the Mauna Loa Observatory, Hawaii, between 1958 and 2020. Each has one entry denoted by the data. To make the data most workable, we will convert it to the number of days or years after the first recorded observation in 1958. This way there are no issues with models that do not "understand" dates. The data shows seasonal variation in the individual years and an upwards-sloping long-term trend (Fig. 2). The seasonal variation is characterized by yearly periodic changes where CO_2 levels incline over around 200 days and then decline sharply (Fig. 3).

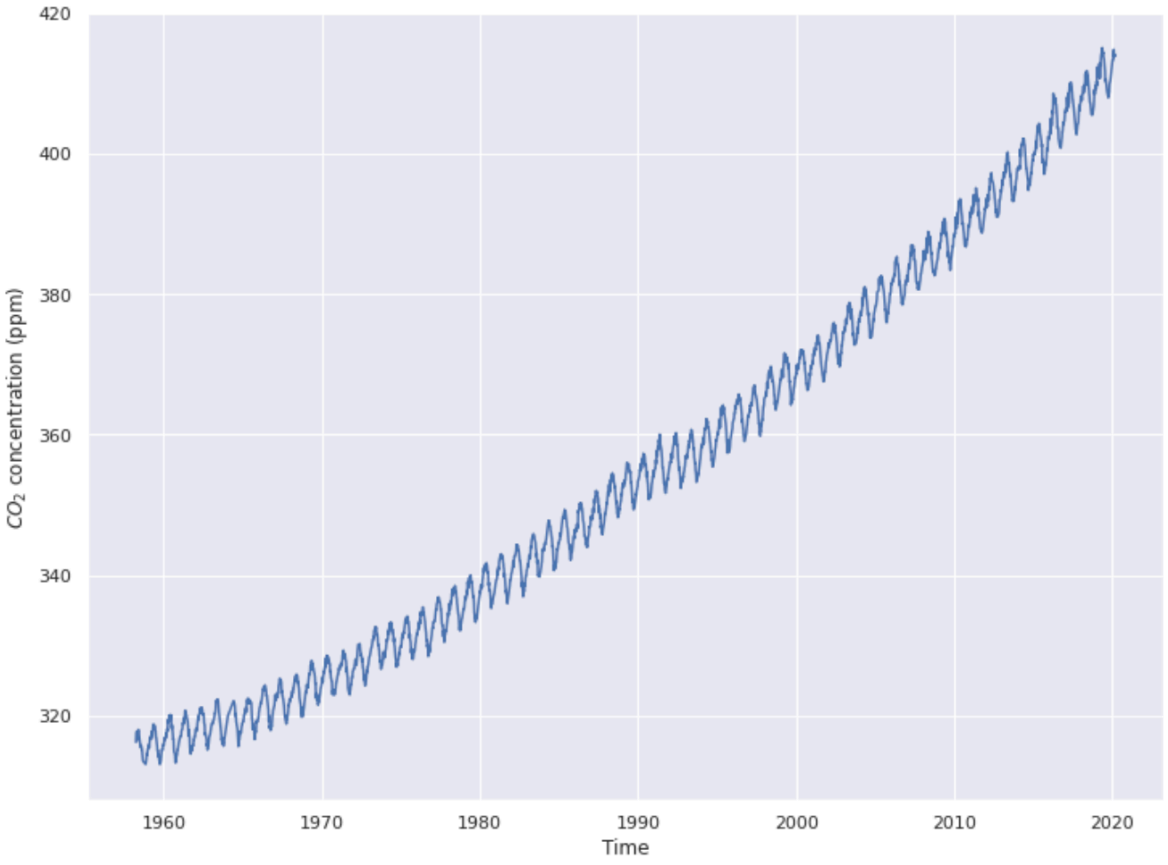


Figure 2. The change in CO_2 concentration over time.

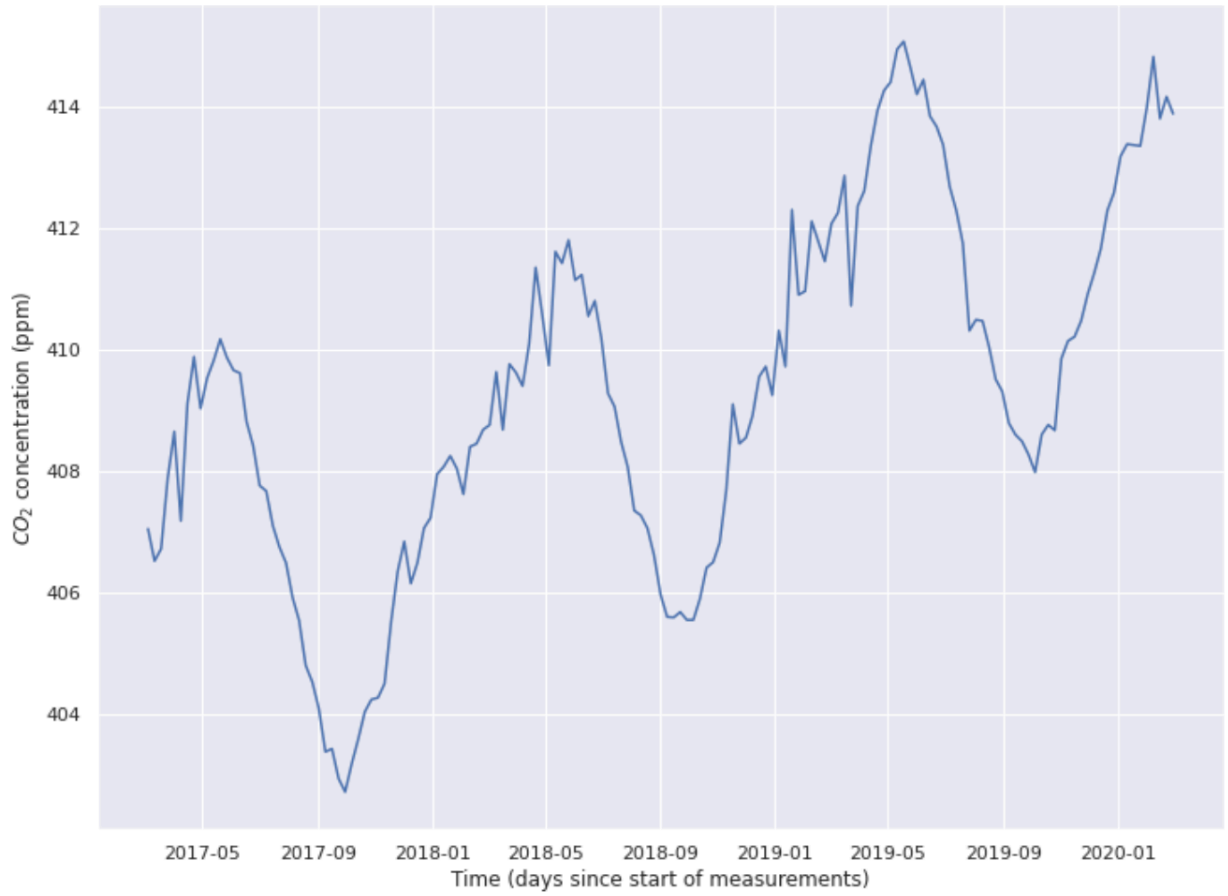


Figure 3. A roughly three-year snippet of the real data to show the tilt of the sin function.

Modeling

Full linear Model

The model proposed by the assignment description is a combination of a linear trend over time and a cosine variation across a year and normally distributed noise distributed around zero with a fixed standard distribution:

- Long-term trend: linear, $c_0 + c_1 * t$
- Seasonal variation (every 365¼ days): cosine, $c_2 \cos\left(\frac{2\pi t}{365.25} + c_3\right)$
- Noise: Gaussian with mean 0 and fixed standard deviation, c_4
- The c_i variables are all unobserved parameters of the model.

Combining these three components gives the following likelihood function

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

- where θ represents the set of all unobserved parameters. Since there are 3156 data, the full like comprises a product overall 3156 values, x_t. To complete the model, we define priors over the 5 model parameters (Tab. 1).

To convert the dates to a linear scale, they have been converted to the days after the first measurement. Therefore, the parameters represent are hence trained to show how a one-day increase changes the CO_2 concentration (c_1).

Table 1. Priors over the linear model.

Parameter	Prior	Justification
c_0	$N(310, 30)$	We can roughly estimate the y-intercept.
c_1	$cauchy(0, 0.01)$	We are unsure of the gradient. Hence, we use a heavy tailed distribution.
$c_2 = A$	$cauchy(0, 5)$	Similarly, uncertain. Using thick tailed Cauchy distribution.
$c_3 = phase$	$N(0, 1)$	We know that the phase cannot be more or less away than π. Hence, we set it around 0.
$c_4 = sigma$	$Gamma(1, 2)$	The error has to be non-negative, so using the Gamma distribution, which has positive support makes sense.

Parameters of periodic functions

When cosine or sine functions are specified, the phase and amplitude can balance each other out. If the phase (start of a period) is shifted by π (when the period is 2π), then the amplitude of a sine or cosine functions can be multiplied by -1 to get the exact same curve as before. Because this is possible, stan can get confused on how to estimate the parameter because, as it samples from different chains, it will end settling on either of the two pairs of phase and amplitude. Stan samples using a Hamiltonian Monte Carlo method, where each new sample is found by taking a step away from the last accepted sample. Stan using four such chains, which each initialize the chain at a random point. If the distribution of samples from the different chains do not overlap, then they "have not mixed". Because there are different ways to get practically the same cosine functions, the chains that stan samples from do not converge. To achieve convergence, we can restrict one of the parameters to enforce only one of the pairs to be sampled. Hence, all chains should eventually converge (R_hat of or close to 1).

Results

The linear trend does not accurately capture the long-term development of the CO_2-concentration (Fig. 4). It hence seems to underestimate future concentrations. The error drives the main proportion of residuals of the mean trend in the linear trend line, not actually fitting our data (Fig. 5). However, the smaller variation in residuals also leads us to examine the error in seasonal variation, which seems to be explained by a tilt in the seasonal variations (Fig. 6). Further, we see that the amplitude seems to be underestimated, as some of the variations in the long-term noise that is not explained by the linear trend.

As shown in the pair plots in the appendix, the chains have mixed well, and the autocorrelation plots show that the samples are almost independent of the previous sample, and we can assume independent sampling.

Table 2. The estimated parameters for the linear model.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff
c_0	305.97	3.1e-3	0.14	305.69	305.87	305.97	306.06	306.24	1946
c_1	4.3e-3	2.2e-7	1.1e-5	4.3e-3	4.3e-3	4.3e-3	4.3e-3	4.3e-3	2406
A	1.98	4.1e-4	0.02	1.95	1.98	1.99	2.0	2.0	1346
phase	4.3e-3	1.5e-4	4.3e-3	9.2e-5	1.3e-3	3.0e-3	5.9e-3	0.02	848
sigma	3.92	1.3e-3	0.05	3.82	3.88	3.92	3.95	4.02	1564

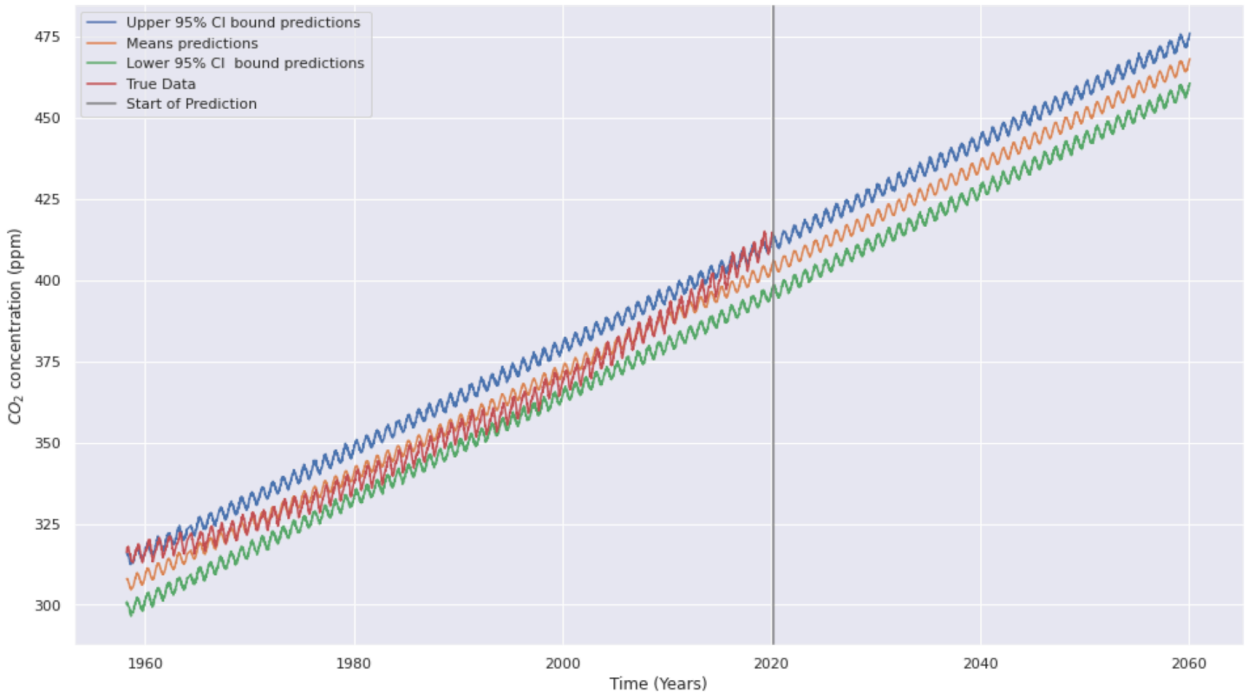


Figure 4. The linear model with 40 years of predictions. The linear trend does not fit the data smoothly.

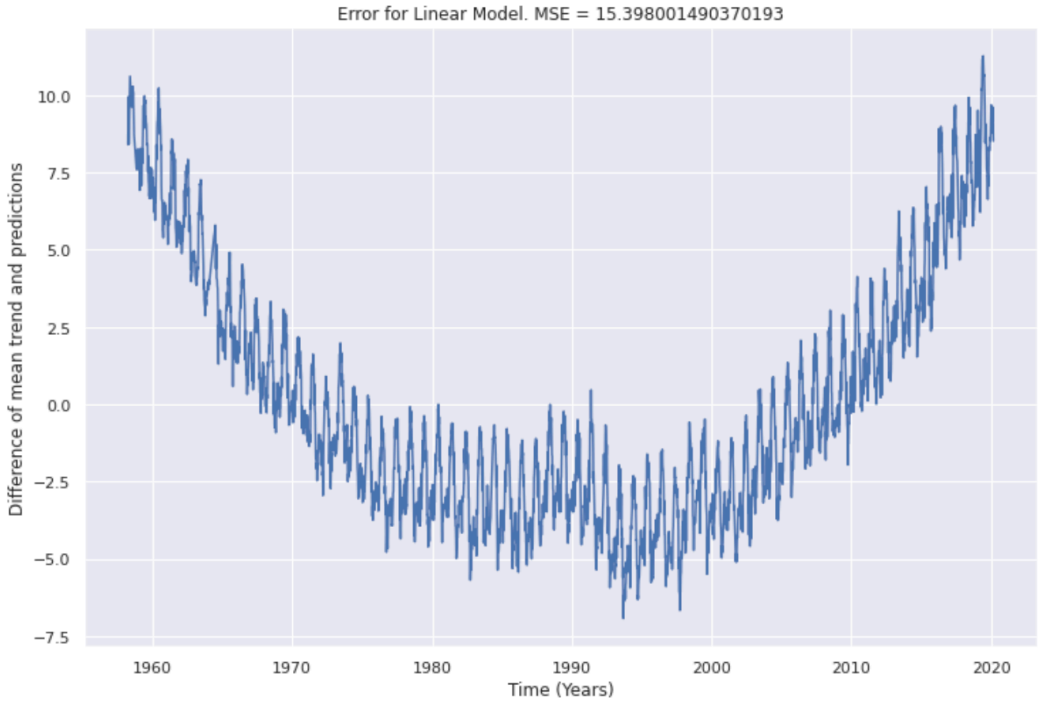


Figure 5. Residuals of with the mean predictions. Calculated by subtracting the data with the mean predictions. The main component of the error is driven by the wrong linear trend, while there seems to be misalignment in the seasonal variations also.

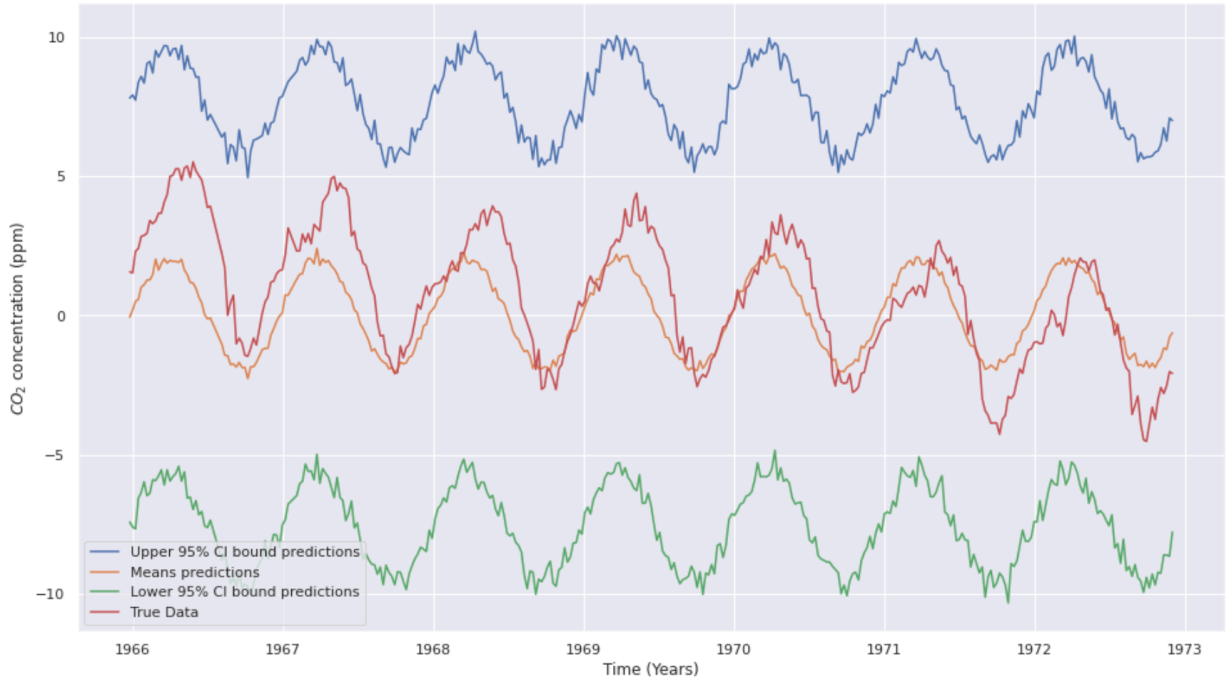


Figure 6. Removing the linear trend from predictions and data and zooming in between 1966 and 1973. Even with the linear trend component removed, we observe a trend in the real data and a tilt in the seasonal variation. The daily error seems to be correct.

Introducing a test statistic

Because the tilt is one of the most subtle features of the data, we will test if what we have produced somewhat resembles the tilt. For this, we will take the time difference between the occurrence of the minimum and maximum concentration of CO_2. Fig. 2 shows how the maximum of a year occurs only shortly before the minimum of that year. This seems to be different in our predicted data, where the maximum is as far away from the minimum on either side (Fig. 6).

Hence, we calculate the mean number of weeks between the minimum and maximum concentration that occur in an interval of 52 weeks (i.e., a year) in the real data and compare it to the same time differences in the predicted data (Fig. 7). This looks at the long build-up time before the sharp decline. During calculations, we had to exclude two years from the beginning of the data set because their minimum occurred before the maximum and hence had a large negative number (see appendix for which data was excluded). If we had included them and, in that interval, we had seen that the difference was actually smaller

than that of the generated data, the p-value would be reduced by 0.048. In the case of the linear trend line, even if factoring this uncertainty, the real data has a significantly longer pre-maximum period than the generated data at the 5% significance level.

This makes sense since we are basically probing for half the period and some variation due to the error term. We have fixed the period in our model, so it has little chance to pass the t-test here. However, instead of changing the period, which we are pretty certain is a year, we will modify the specification of the model to introduce a tilt. By changing the period we will not be able to mirror the asymmetry between maxima and minima.

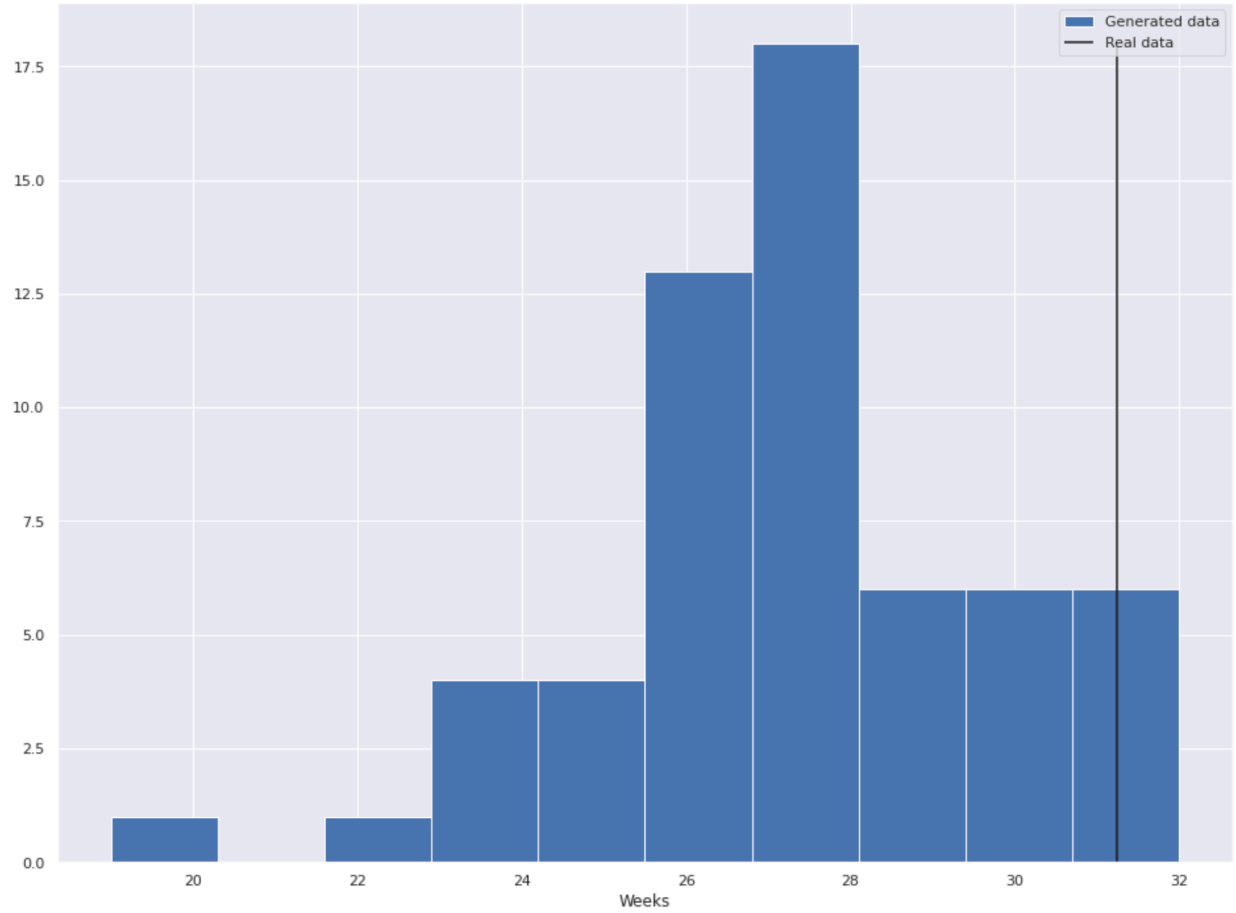


Figure 7. Difference between highest and lowest concentration in a one-year timespan. Mean for the real data: 31.237 weeks. P-value: 1.0 (could be as low as 0.952).

Quadratic model

Improvements and set-up

The proposed model has a linear upwards trend. However, it seems like the trend is non-linear (either exponential or with an added quadratic term) instead. Hence, we will change the long-term trend to be a quadratic function. While this might seem like a minor change, it will have a huge effect specifically on our future predictions. The long-term trend part of the model will be:

$$c_0 + c_1 * t + c_2 * t^2$$

Further, when we examine the yearly change closely, we notice that the curves are not a clean periodic function but that they are tilted towards the right. Because a tilted cosine function (or sine function) is a sawtooth function, it can be implemented using a combination of multiple different cosine functions. Similar to the sawtooth function we know from Fourier series (Conrad, 2017):

$$sawtooth(x) = \sum_{i=1}^{\infty} \frac{\mathrm{a_i}}{\mathrm{i}} \sin\left(\frac{x}{i}\right)$$

Because Fig. 2 and Fig. 4 show only a slight tilt and not a complete sawtooth-like behavior, we can go less extreme and just add one other cosine function to the model. Additionally, we will add noise to the model again. Hence, we have our extended model to:

$$N\left(c_0 + c_1 * t + c_2 t^2 + c_3 \cos\left(\frac{2\pi t}{365.25} + c_4\right) + c_5 \cos\left(\frac{2\pi t}{365.25} + c_4\right), c_6\right)$$

The phase (how much the sine curve is shifted), is as explained above a periodic variable. That means should the phase be the same as the period then then curve is shifted by so much that it is the exact same as before. To help the sampling algorithm, we can instead sample two values that lie around 0 – $\mathrm{phase_x}$, $\mathrm{phase_y}$. When they are passed to the Arctan function, it gives the angular position (in radians) from where the point ($\mathrm{phase_x}$, $\mathrm{phase_y}$) lies from the positive x-axis. Hence, this value will be between [π, -π] and we shall restrict the parameter to that value also (Wolfram, 2020).

The entire model is summarized in the factor graph in Fig. 8. The priors over the individual parameters are shown next to the factors that connect to each of the parameters.

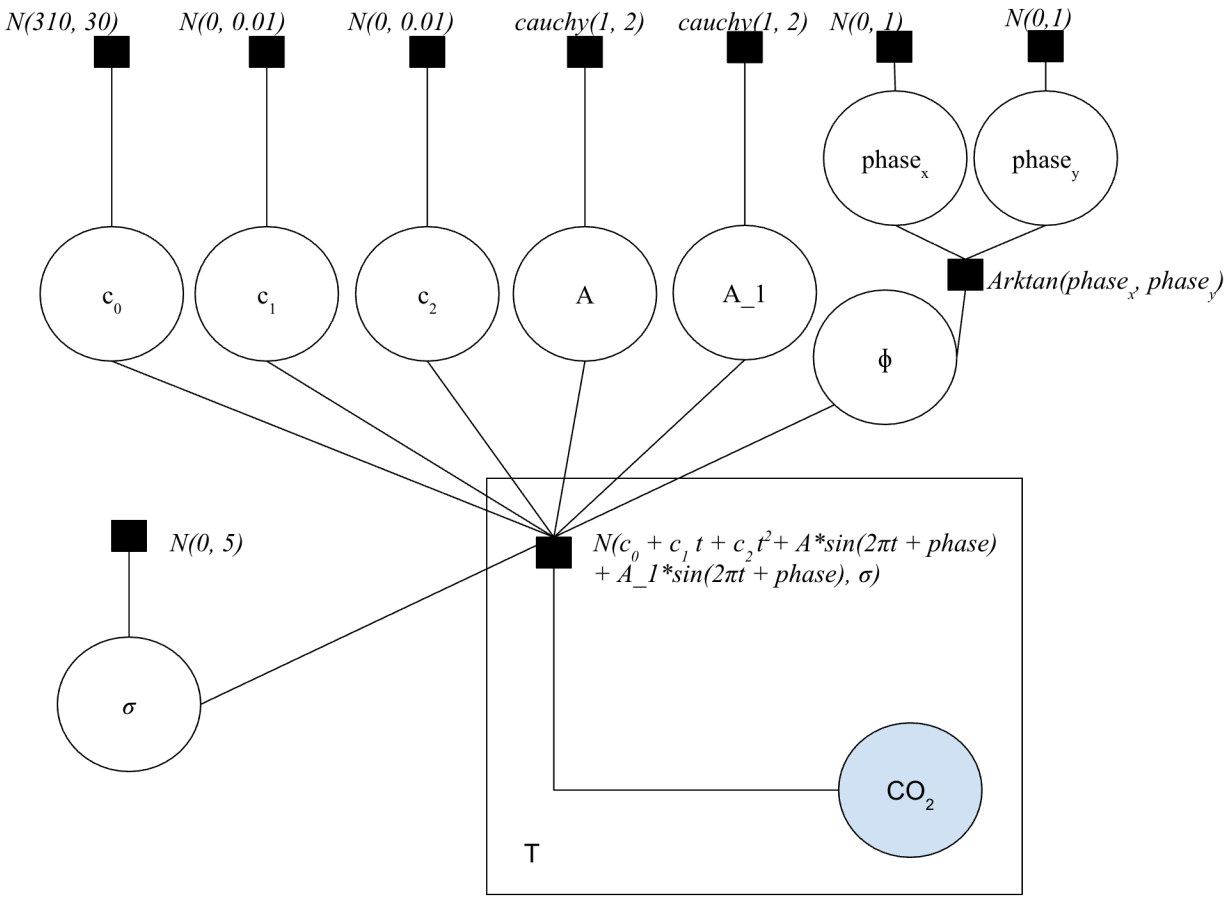


Figure 8. Factor graph of the quadratic, double-sine model.

Results

Table 3. Sampling results of the model. The effective sample size and Rhat values show that the Hamiltonian Monte Carlo chains have converged.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
c_0	316.27	1.9e-3	0.08	316.12	316.21	316.27	316.32	316.42	1617	1.01
c_1	0.63	1.5e-4	5.7e-3	0.61	0.62	0.63	0.63	0.64	1517	1.0
c_2	0.02	2.2e-6	8.7e-5	0.01	0.02	0.02	0.02	0.02	1557	1.0
A	1.43	9.2e-3	0.32	0.88	1.16	1.42	1.71	1.97	1234	1.0
A_1	1.43	9.2e-3	0.32	0.88	1.14	1.43	1.7	1.97	1247	1.0
phi	1.16	1.5e-4	0.01	1.14	1.15	1.16	1.17	1.18	4347	1.0
sigma	1.17	5.5e-4	0.02	1.13	1.15	1.17	1.18	1.21	1338	1.01

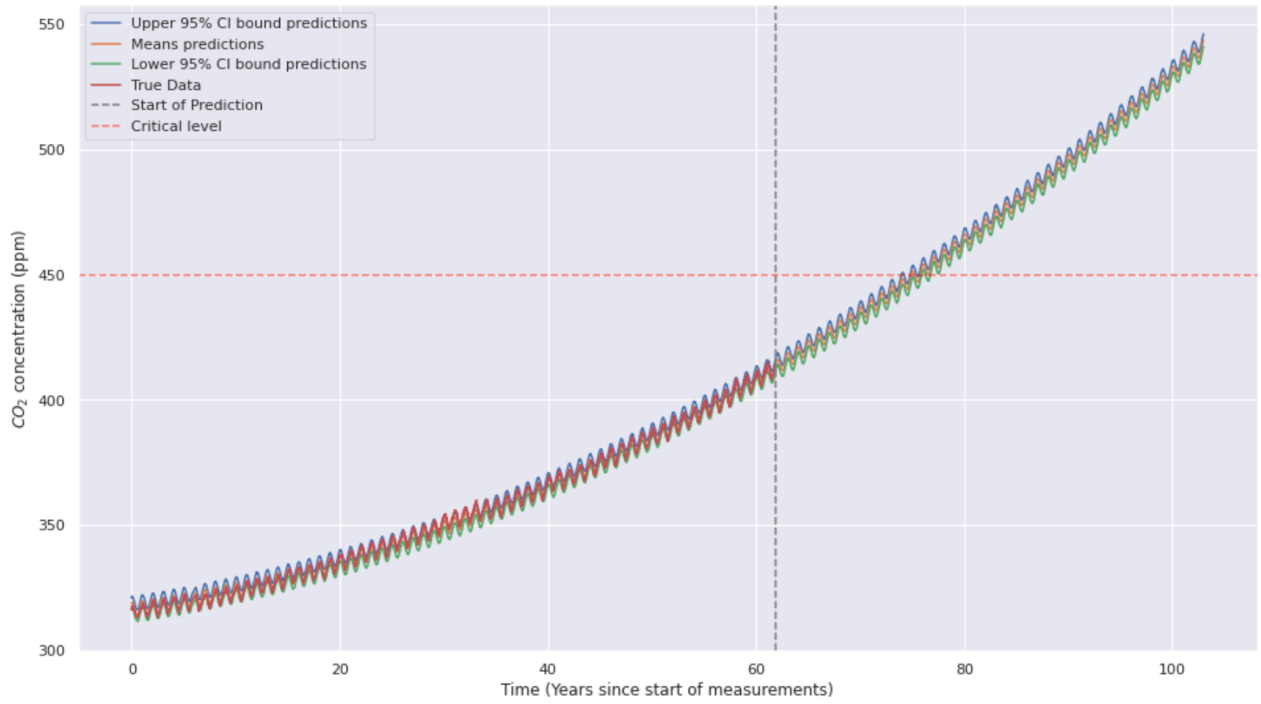


Figure 9. Replication and prediction of the CO2 concentration from the past 60 and next 40 years. The red line indicates, the point at which CO2 levels reach a critical point.

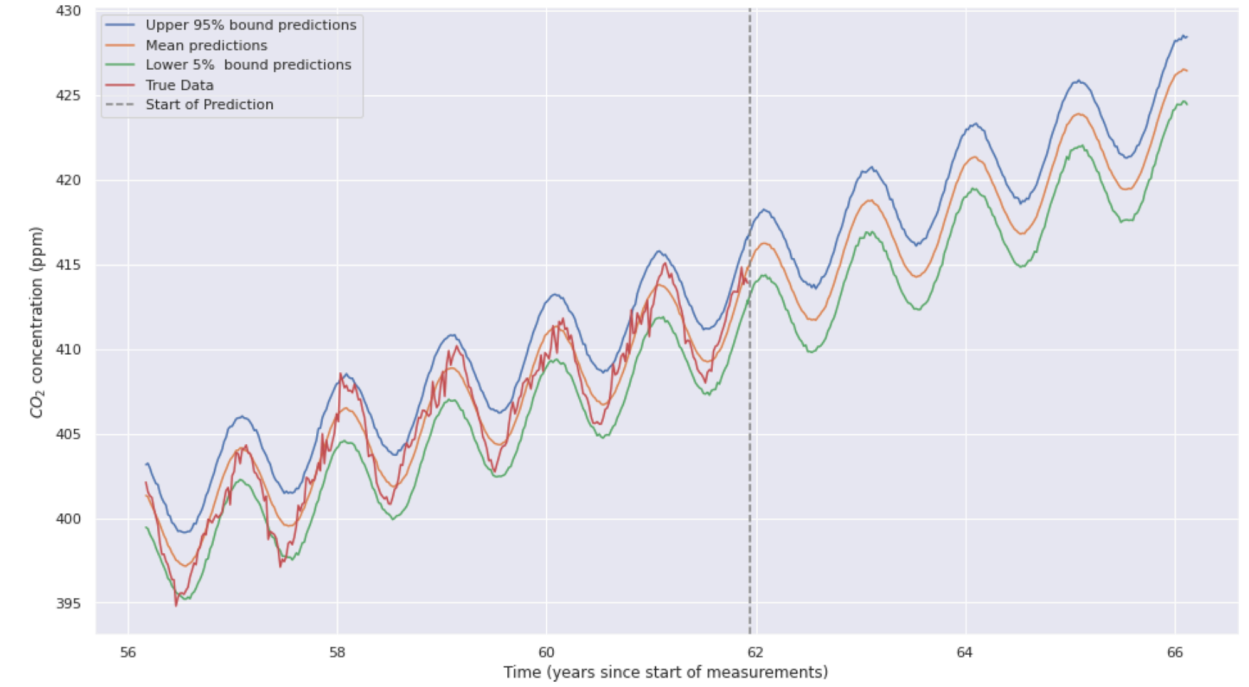


Figure 10. Replications and predictions from the last few years and the first years, zoomed in. The data lies within the 90

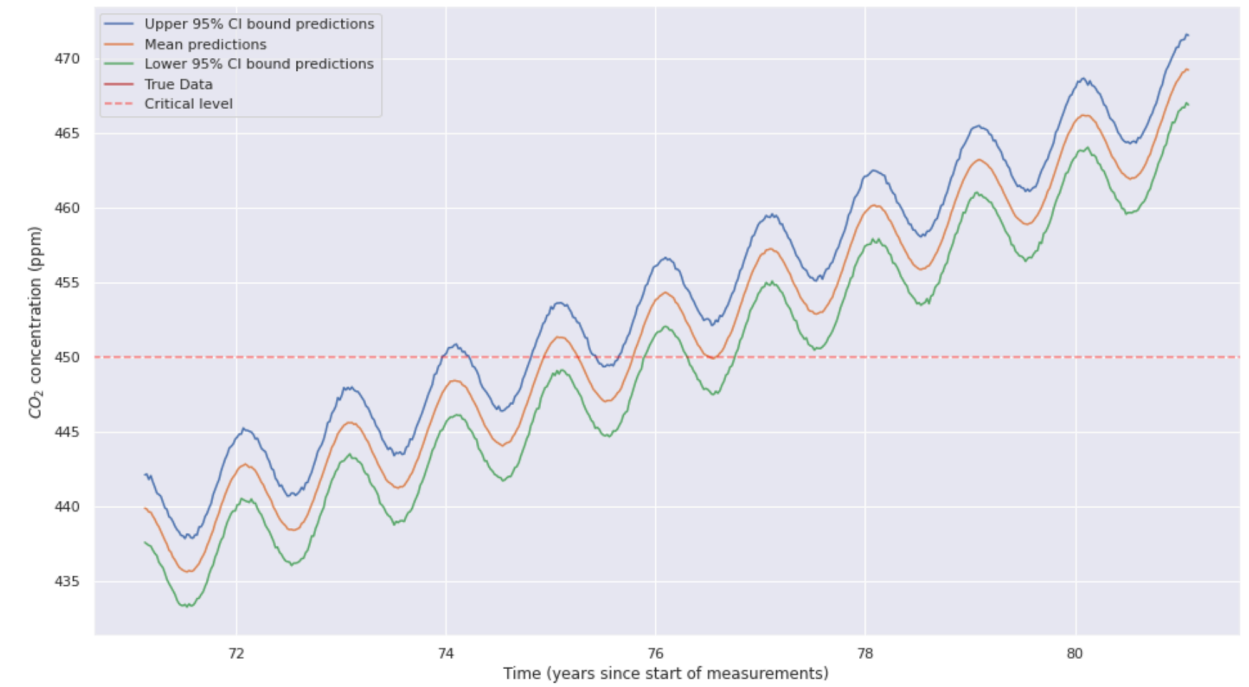


Figure 11. Close up of when the predicted concentrations will reach a level of 450 ppm.

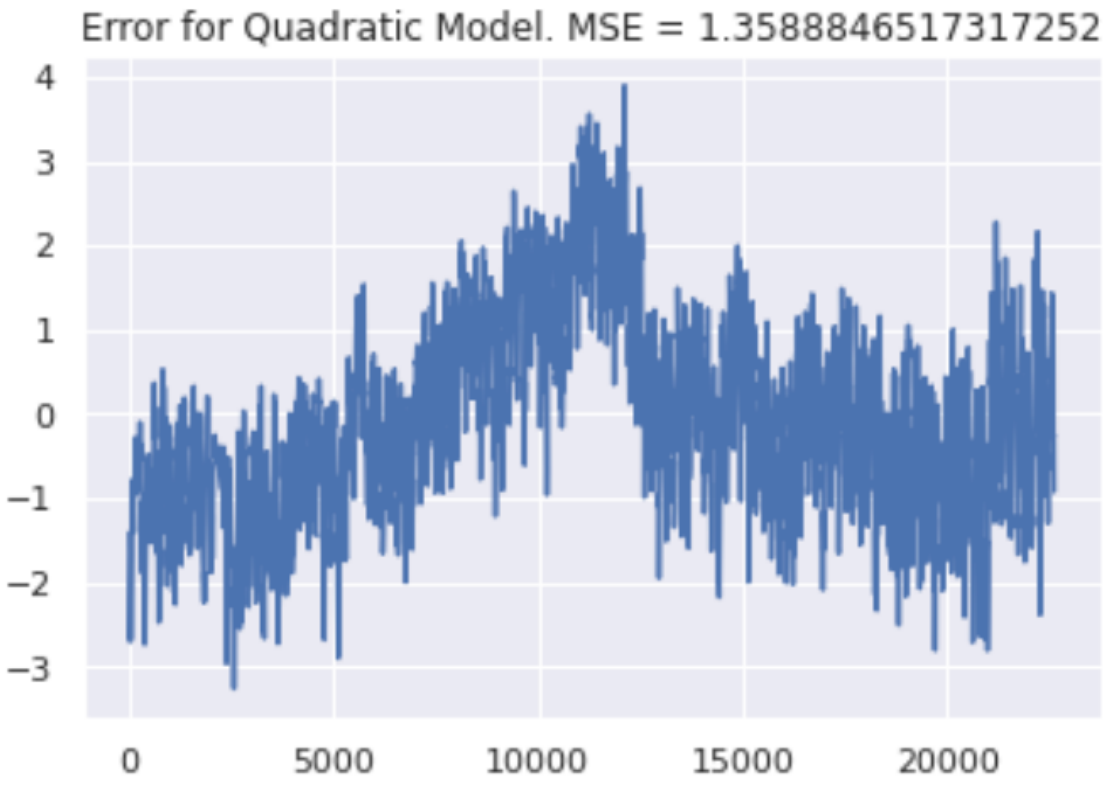


Figure 12. Error distributed over time. See appendix for histogram of error.

Inference and discussion

The R_hat values have shown that the sampling has converged and that we can interpret the model (see appendix for pair plot and autocorrelation plots). The coefficients for c_1 and c_2 tell us that in one year the CO_2-concentrations increase with a rate of

$$ppm'(t) = c_0 + 2 * c_1 t$$

$$ppm'(t) = 0.63 * t + 0.06t$$

However, the model does not know this rate with certainty, as the samples of the parameters are not certain but distributed around a value. Hence, we can have a minimal and maximum development of the CO_2 concentrations. When we plot the result, then there are, for example, 100 estimates for the concentration. The interval from the 2.5th lowest and the 97th lowest ordered prediction covers 95% of the variations. However, many more estimates are close to the mean estimate than on the lower and upper end of the spectrum. Hence, the mean is the most likely trend, while with 95% certainty, we are in the range between the 2.5th lowest and 2.5th highest values.

As shown in Fig. 11, the CO_2 concentrations will reach the critical level of 450 ppm sometime between 2034-02-22 and 2032-03-16 but most likely at 2033-03-09. If we were able to decrease the long-term rise in CO_2 levels by half, we could prolong reaching the critical level until approximately 2059-04-25.

As we predict further and further into the future, the small difference in the gradient from the upper and lower bound of the confidence interval will lead to the bounds diverging a bit every year. However, because the model does not rely on a time-series structure (where each value depends on the next), we will not see a wide divergence. The model does not make the next step depend on the outcome of the previous but is more deterministic, with each new point being evaluated on its own. It means we could also solve for 450ppm if we knew the exact parameters (or took the mean like above). It, however, does not reflect the underlying mechanism that generated the data, since CO_2 levels do not magically increase over time but do

so due to (among other reasons) from the increased usage of fossil fuels, which would hence be a better independent variable. The difference illustrates why this model does not tell us anything about causation.

A particular weakness of the model is that the tilt in the sin curve is not matched yet, even though we have tried to introduce the double sine (see t-test histogram in appendix). This could be tackled by adding additional sine or cosine functions (Conrad, 2017). This would allow us to model the seasonal variation even better, which could help with the specific policy-relevant to yearly concentrations such as in the agriculture space. The model is able to capture some other relevant features of the input-data which is visually depicted in Fig. 9 and Fig. 10. Using test statistic plots for regular test statistics such as mean or standard deviation would make little sense as the mean of the data set is quite irrelevant to know. Similarly, the standard deviation of CO_2 concentrations is not of particular interest due to the fluctuations and long-term increases in the data.

Further, the model also misses some of the variation that is shown in the long-term trend of the data. Fig. 2 shows how in the 90's there was a slight kink in the increase of the concentration. It can be expected that due to events such as the covid-19 crisis, the long-term generalizability of the model may be limited, as the only input factor we have is the time and no other confounding variables that would be added in more complex climate predictions. However, with the residuals being distributed normally around zero, we can assert that there is no systematic confounder that is driving the error rate (see Appendix).

The generalizability of the model is also limited geographically as the CO_2 concentration may be different in other parts of the world and hence have a different impact.

However, with all these drawbacks, the model still teaches us that the long-term trend is upwards and increasingly so over time. This may be one of the most important lessons from the model.[1]

[1] #cs146-professionalism: I present my work in a professional, clearly structured, and easy to understand manner, leading the reader of this report through the decisions I have made and clearly communicating what the eventual results of my model entail.

Works Cited

BBC News. (2018, 11 26). *Trump on climate change report: 'I don't believe it'*. Retrieved from BBC News: https://www.bbc.com/news/world-us-canada-46351940

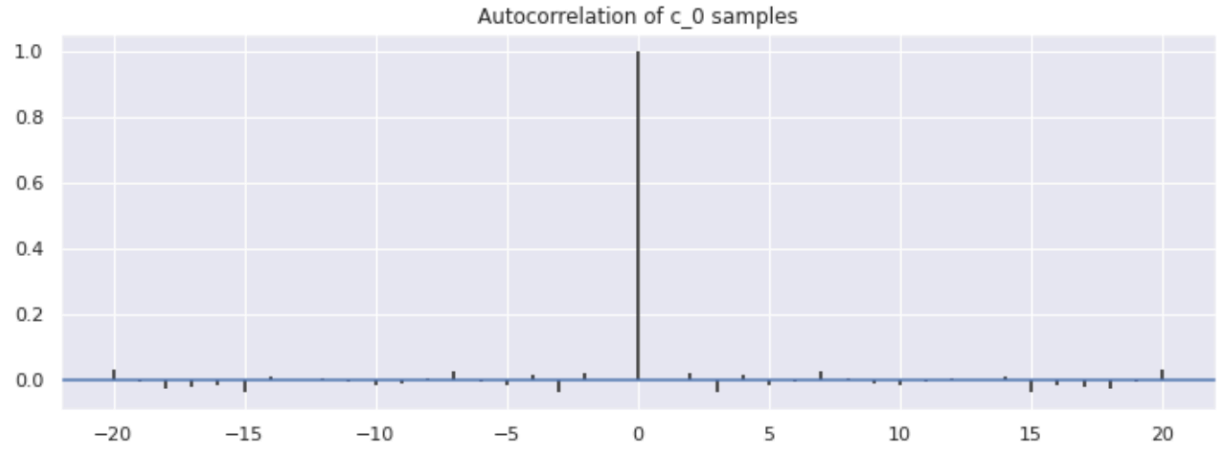
Conrad, K. (2017, 05 17). *A Sawtooth Wave*. Retrieved from kconrad.math.uconn.edu/: https://kconrad.math.uconn.edu/math1132s10/sawtooth.html

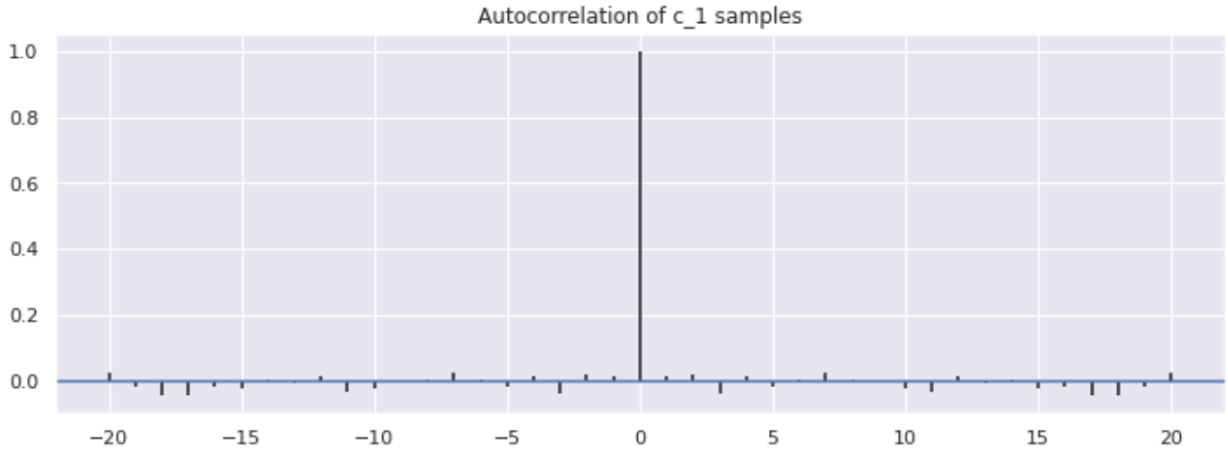
NOAA. (2016, 05 31). *Community Resilience: Is Hawai‘i ready for the impacts from climate change?* Retrieved from Sea Grant: https://seagrant.noaa.gov/News/Article/ArtMID/1660/ArticleID/559/Community-Resilience-Is-Hawai%E2%80%98i-ready-for-the-impacts-from-climate-change

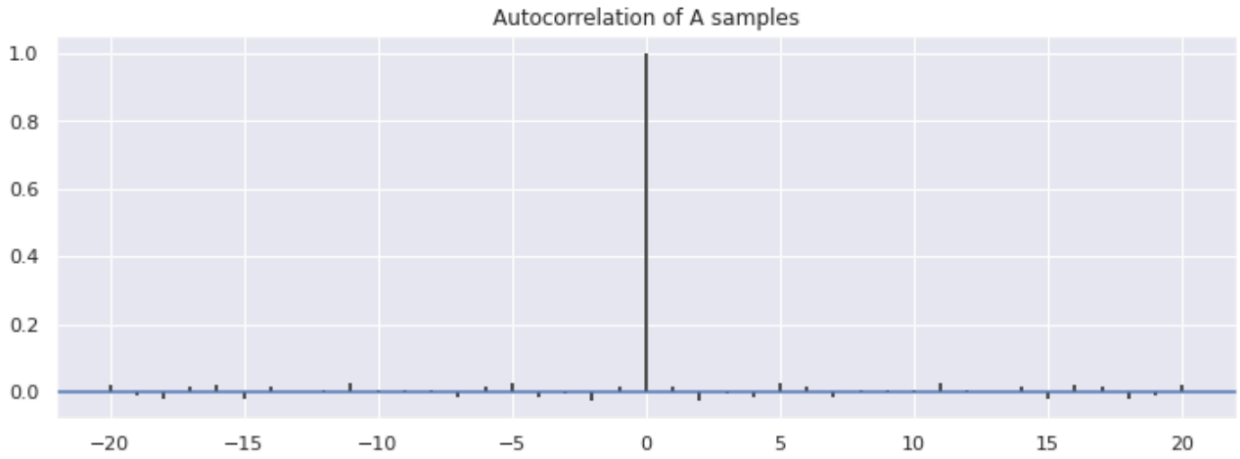
Wolfram. (2020). *ArcTan*. Retrieved from Wolfram Language & System Documentation Center: https://reference.wolfram.com/language/ref/ArcTan.html

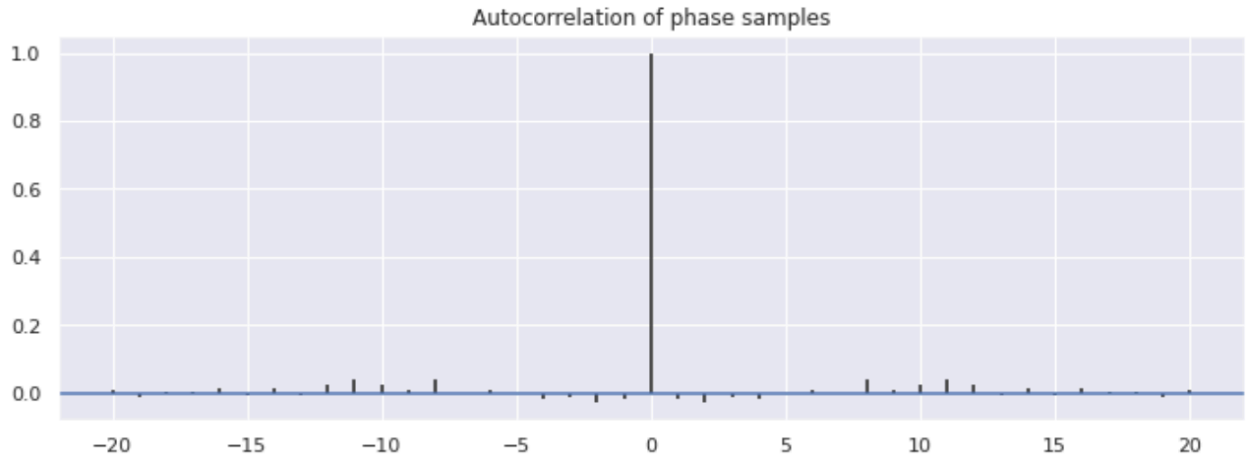
Appendix

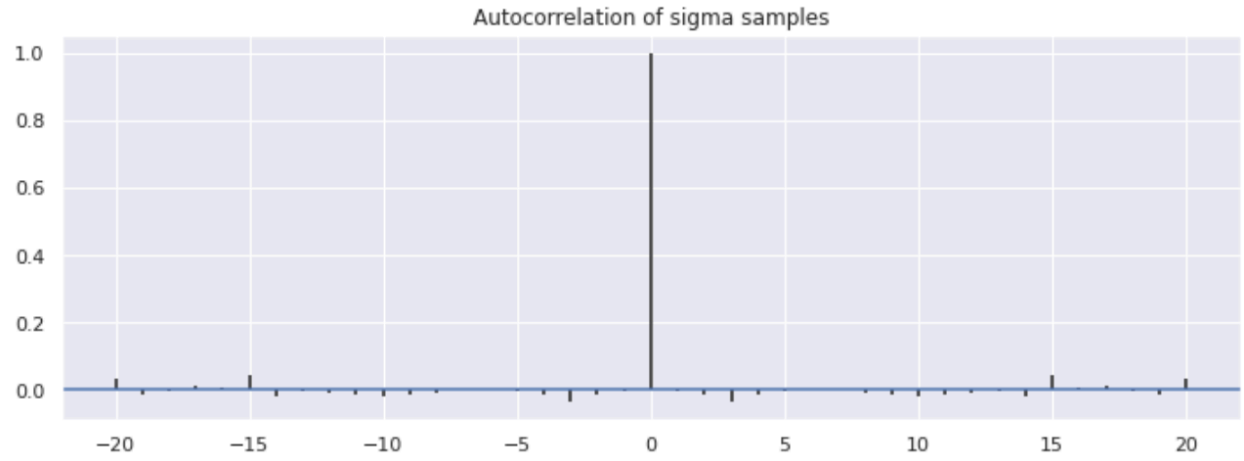
Linear Sampling Checks








Autocorrelation of phase samples
1.0
0.8
0.6
0.4
0.2
0.0
−20
−15
−10
−5
0
5
10
15
20


Autocorrelation of sigma samples
1.0
0.8
0.6
0.4
0.2
0.0
−20
−15
−10
−5
0
5
10
15
20

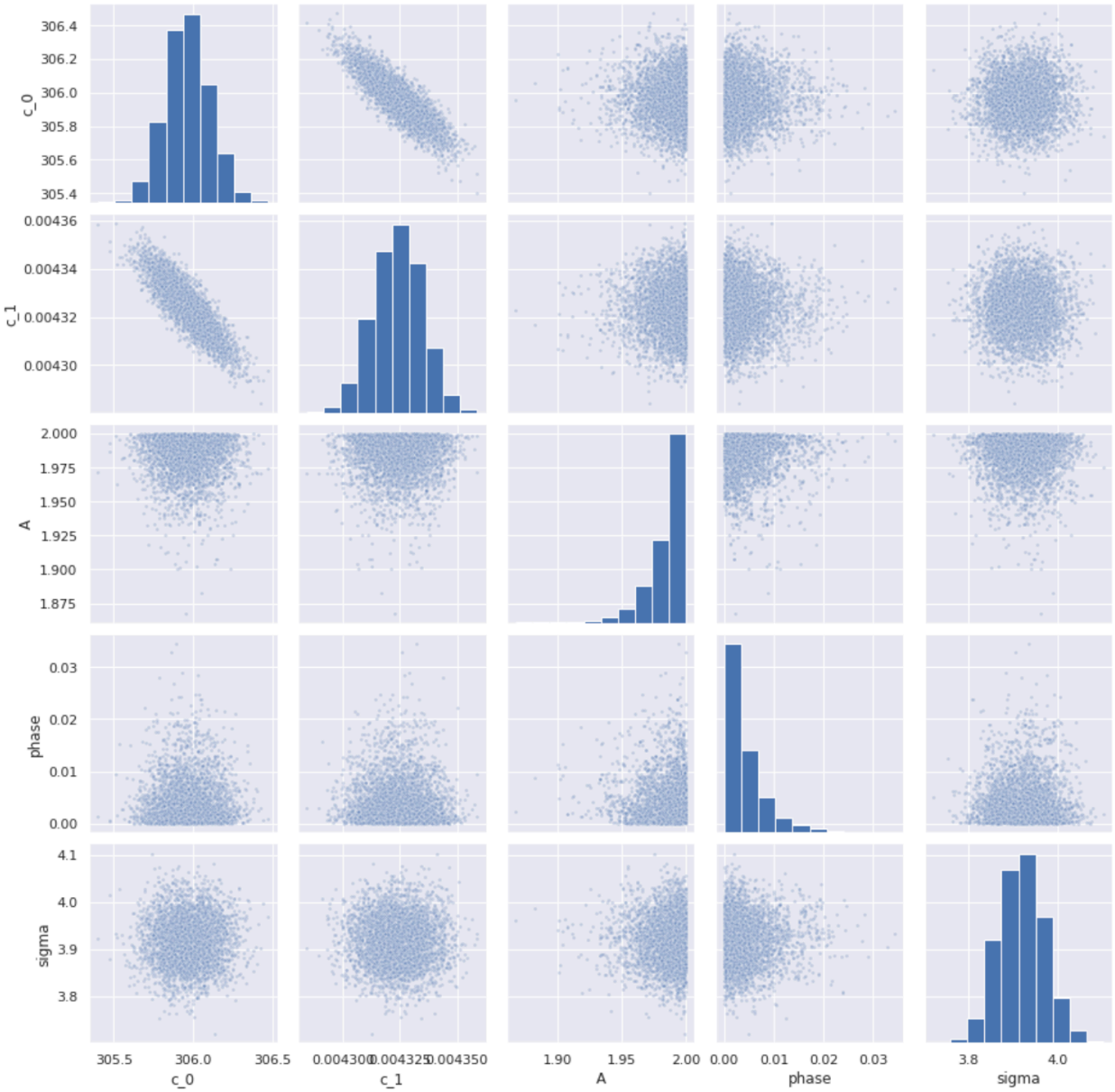


Figure 13. Pair plots of the linear samples.

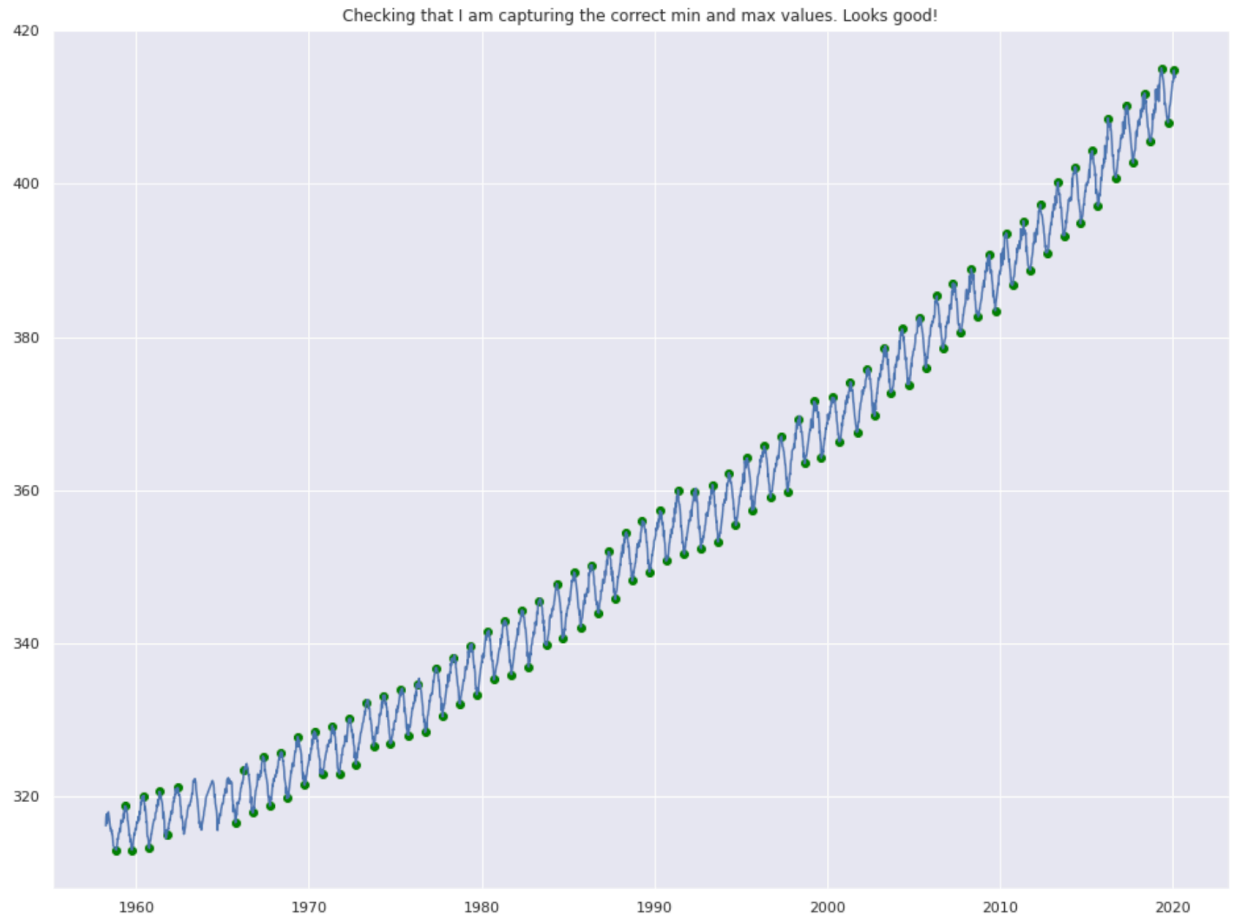


Figure 14. Green dots indicate where that this was a data point used for the Δtime calculation between the minimum and maximum point. There was a gap during the 60s.

Robustness checks for Quadratic Sampling

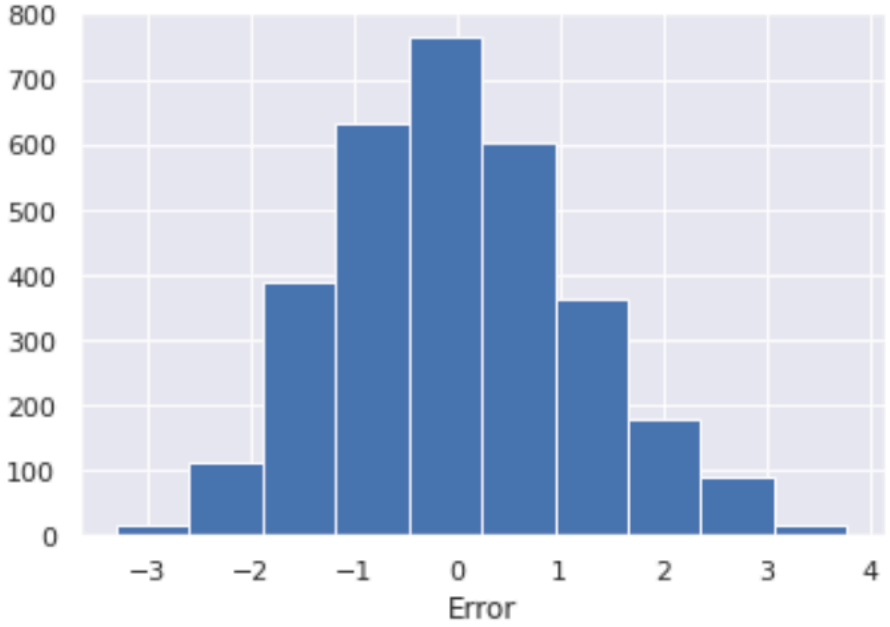
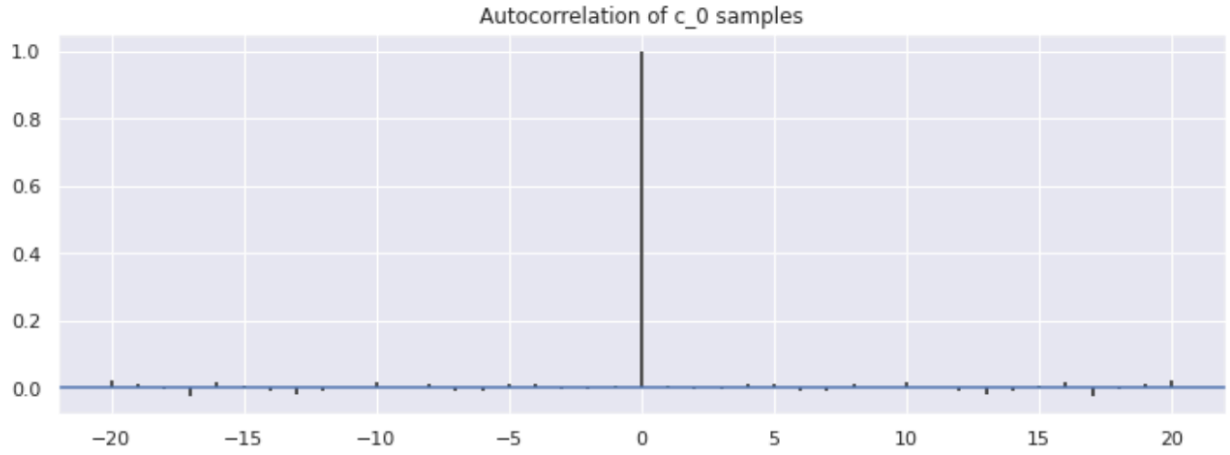
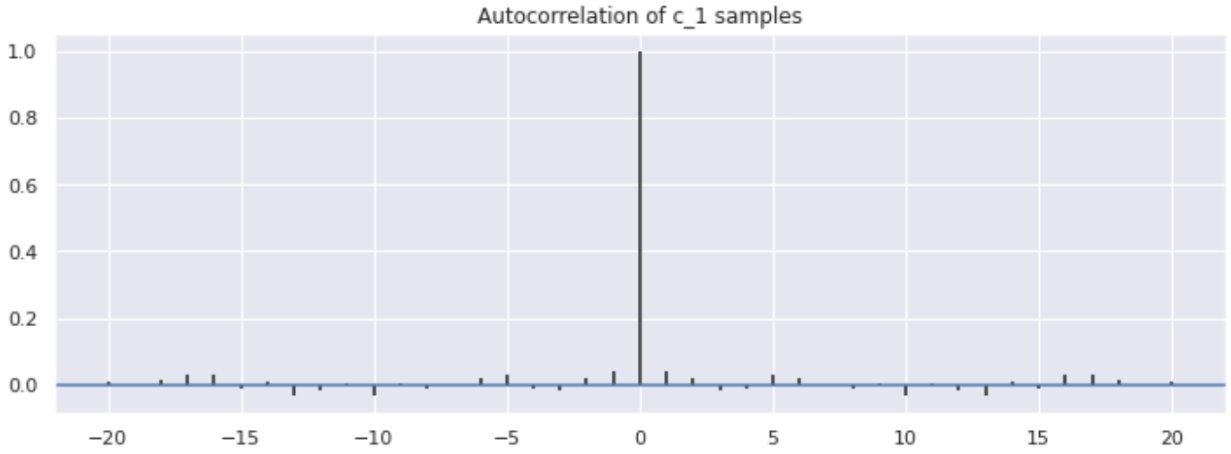
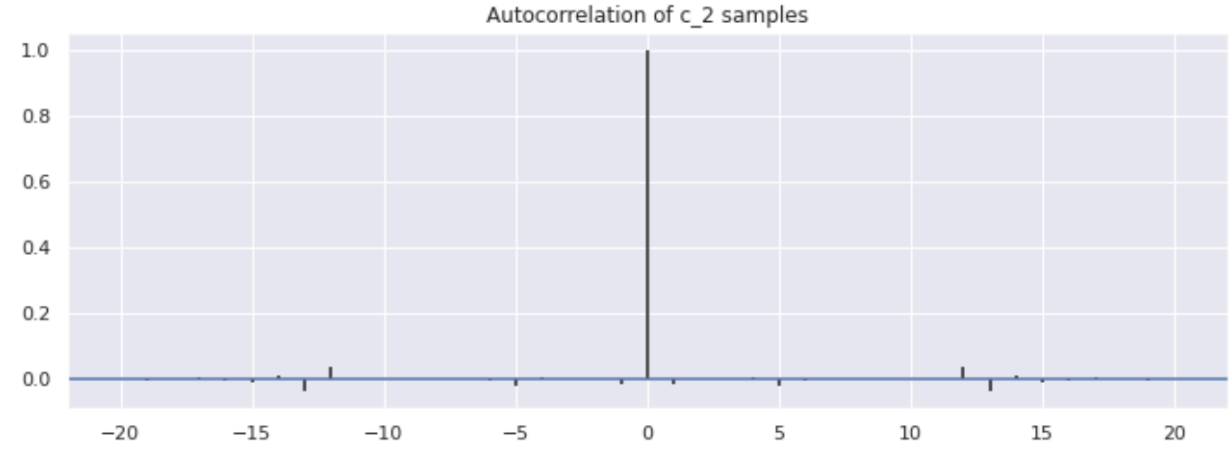
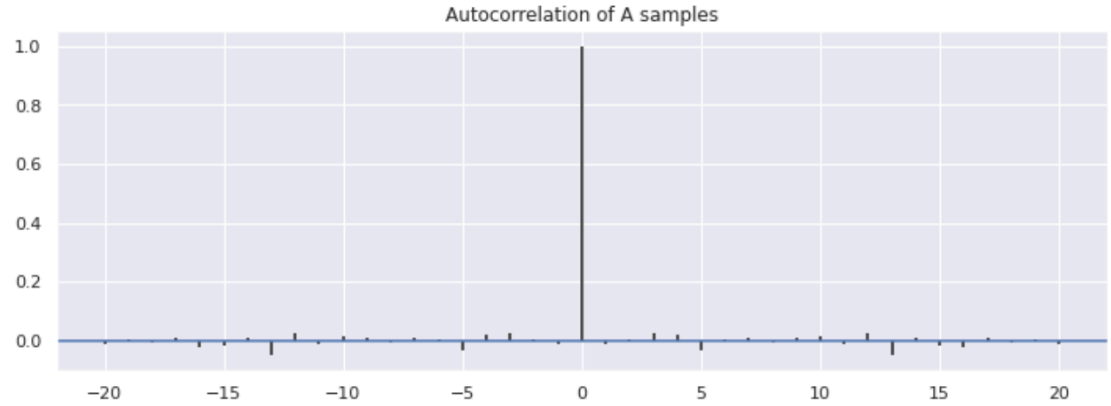


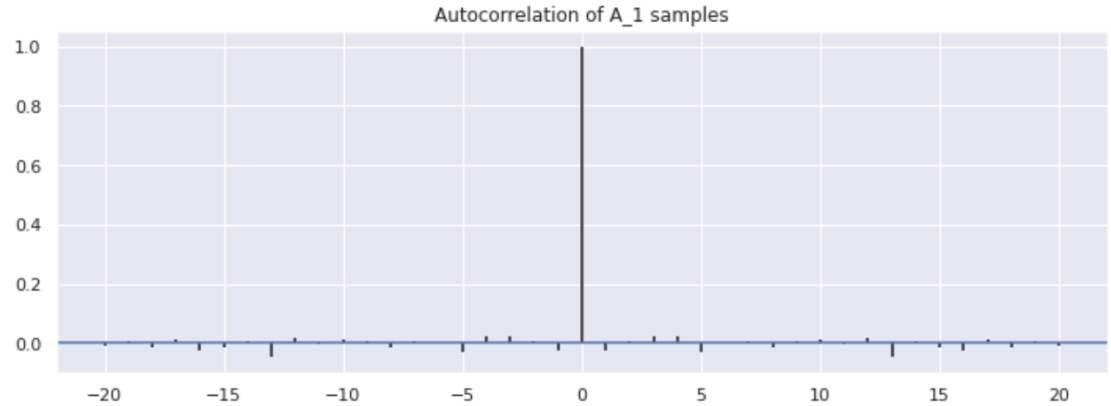
Figure 15. Errors are normally distributed. This means there is no systematic variation in the residuals.

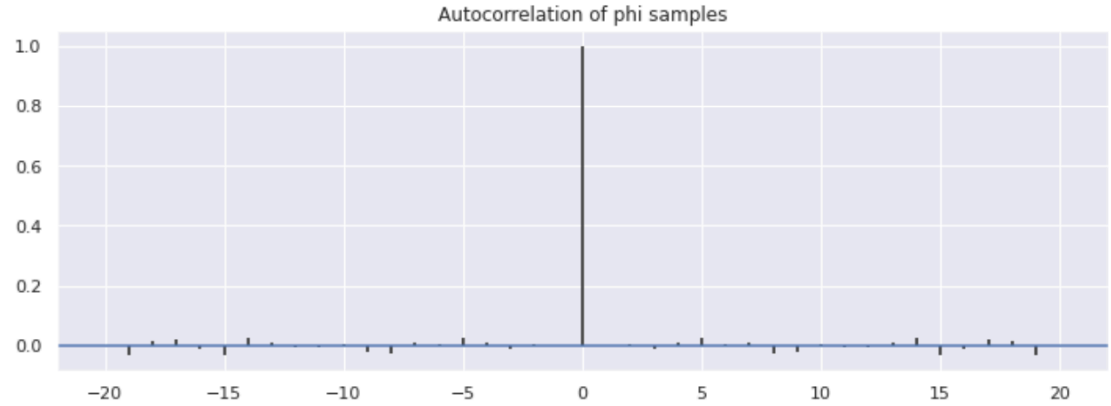


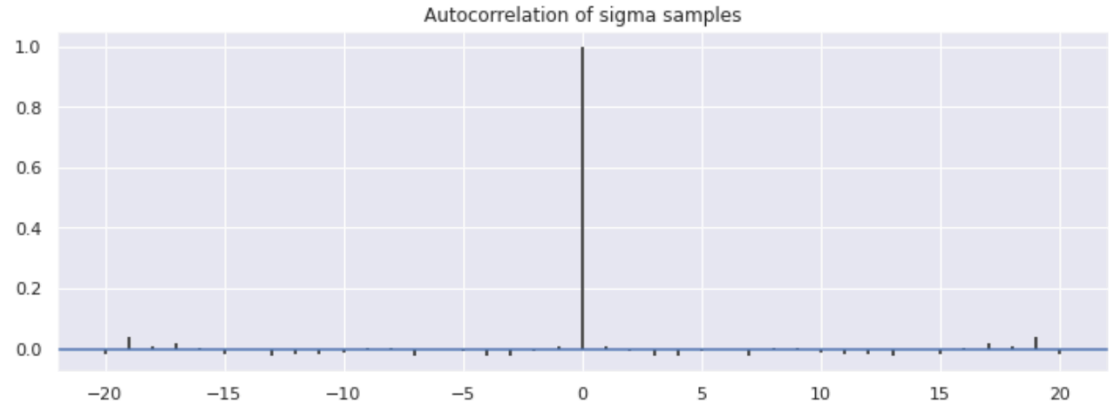

Autocorrelation of c_1 samples
1.0
0.8
0.6
0.4
0.2
0.0
−20
−15
−10
−5
0
5
10
15
20


Autocorrelation of c_2 samples
1.0
0.8
0.6
0.4
0.2
0.0
−20
−15
−10
−5
0
5
10
15
20


Autocorrelation of A samples
1.0
0.8
0.6
0.4
0.2
0.0
−20
−15
−10
−5
0
5
10
15
20


Autocorrelation of A_1 samples
1.0
0.8
0.6
0.4
0.2
0.0
−20
−15
−10
−5
0
5
10
15
20


Autocorrelation of phi samples
1.0
0.8
0.6
0.4
0.2
0.0
−20
−15
−10
−5
0
5
10
15
20


Autocorrelation of sigma samples
1.0
0.8
0.6
0.4
0.2
0.0
−20
−15
−10
−5
0
5
10
15
20

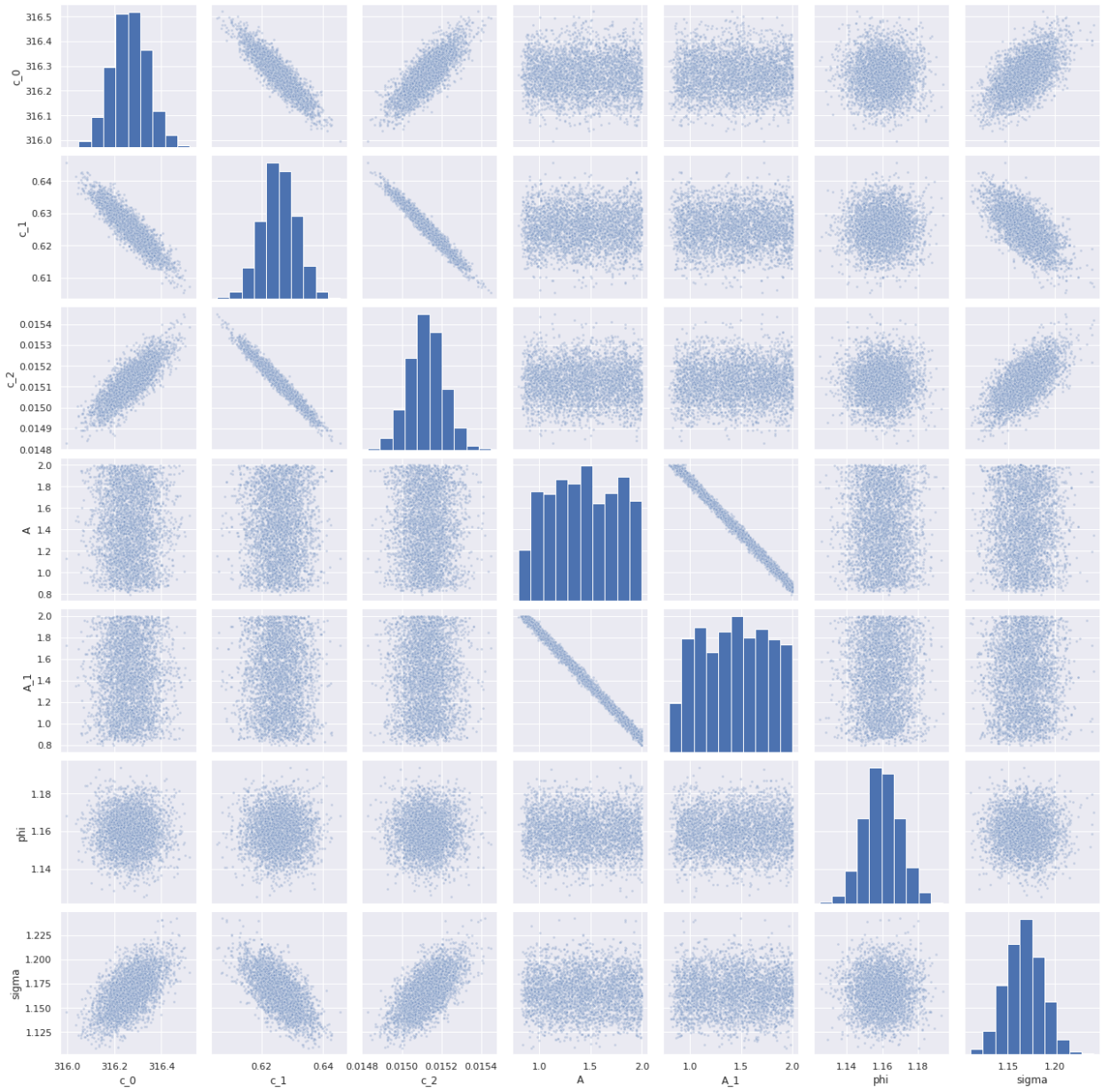


Figure 16. Pair plots for the quadratic parameters.

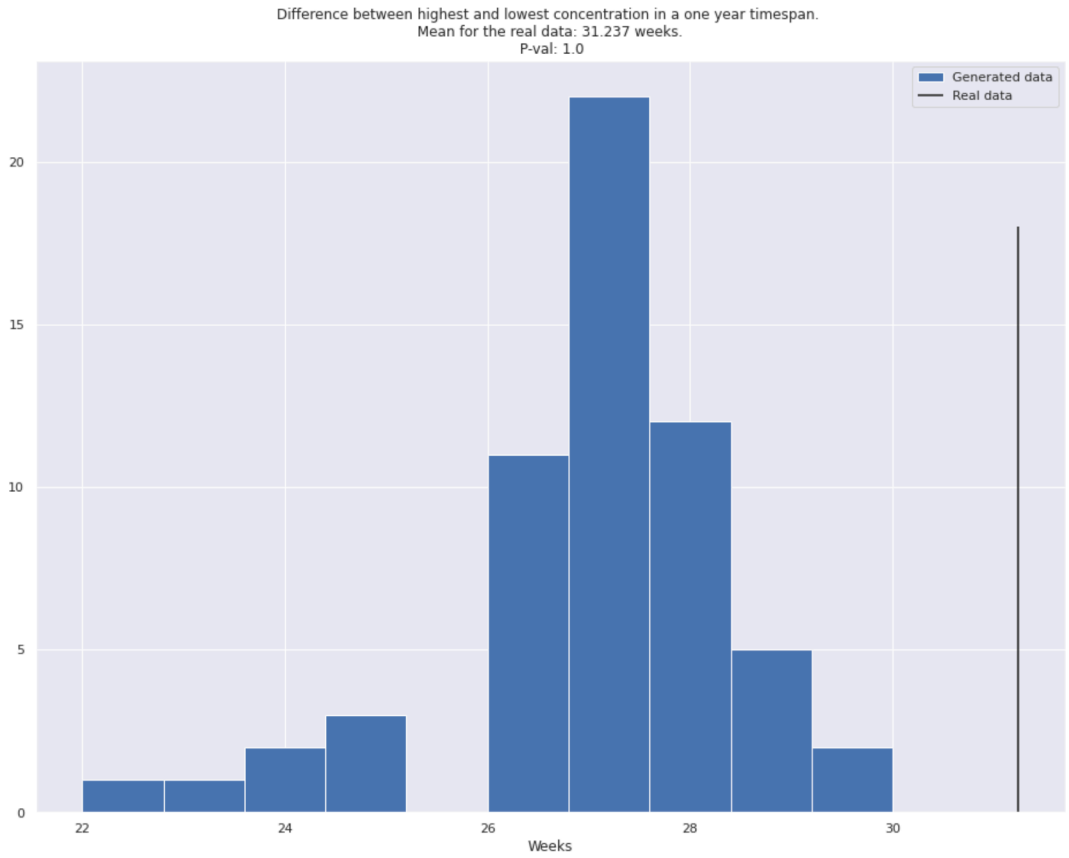


Figure 17. Testing for the tilt. There is still a significant difference in the generated samples and the tilt. It seems we would have to add additional sine or cosine functions into the model.

Calculating how what would happen if we halved the long-term trend:

$$450 = 0.01t^2 + 0.315t + 316$$

$$t = 101.075\ years\ after\ first\ observation$$

(we exclude the negative root).