Modeling and forecasting atmospheric CO_2 from 1958 into the future

Muhammad Abdurrehman Asif

Introduction

The Mauna Loa observatory has been tracking atmospheric carbon dioxide levels weekly since 1958. The data is publicly available and takes the form of CO2 in parts per million (PPM). Climate change has become an increasingly pressing issue in our lives. Understanding the impact of increasing carbon dioxide in the atmosphere on global warming is of utmost importance to policy makers and global citizens. A 'high-risk' level of 450 PPM has been identified, and it is essential that we anticipate when this dangerous zone will arrive using past data. This paper will focus on using all the existing observations to predict the expected behavior of atmospheric CO2 for the next 40 years.

The Models

Linear Model

At first, I wanted to model the dataset up until the most recent observations. The reason for doing this was to get more comfortable with the sort of data I was going to work with. Moreover, it gave direct insight as to what shape and trends to expect.

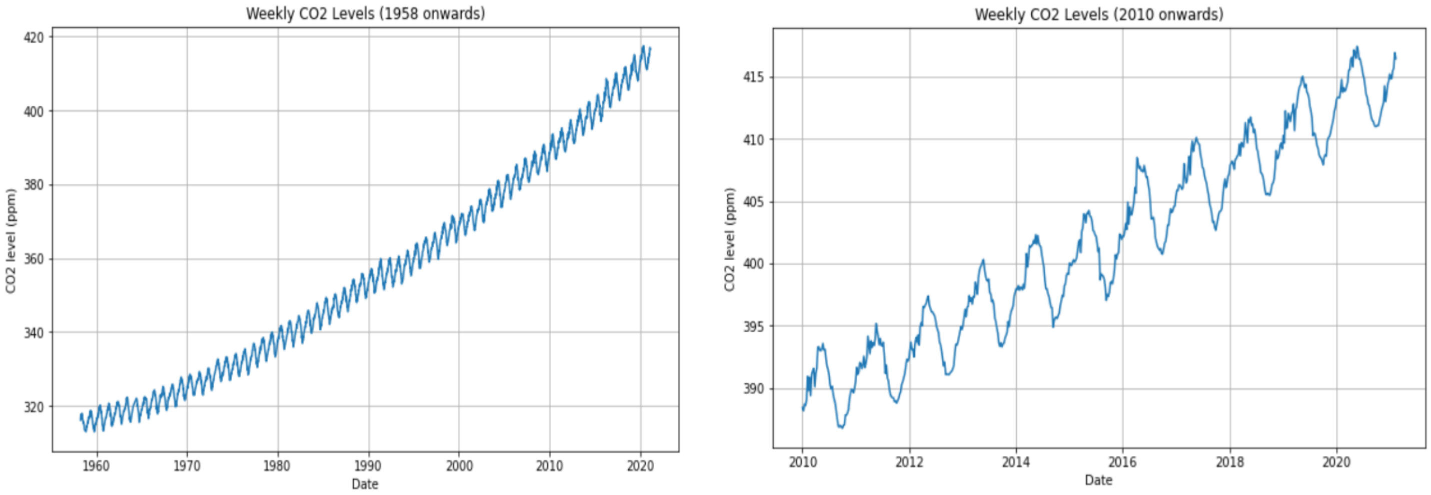


Figure 1 & 2. Figure 1 shows the entire dataset. Figure 2 shows a section of the dataset, from 2010 onwards only.

As shown above, I first plotted all the observations. There is a clear upward trend with periodic phases that resemble trigonometric functions. This was a good starting point. Digging a little deeper, there were 3 main things I needed to be clear about in my model.

1. The actual trend of the data
2. The seasonal aspect - as outlined in the task as well as the Keeling curve[1]

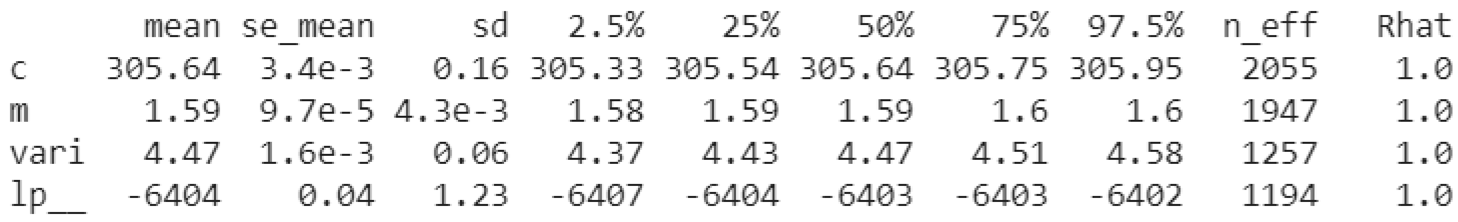
[1] The Keeling Curve is a record of past CO2 observations and future predictions. I have had some past experience with this while researching and studying the impact of CO2 on global warming.

3. The randomness or the noise in the model

The first model I implemented was a simple linear model. There was a set of data points, a visible increase in the CO2 concentration and a starting y-intercept. Using the relationship of Y = Mx + c, the priors were set. There are two parameters in this relationship. The priors would have to be over the gradient (m) and the y-intercept (c).

- Y-intercept; c: N(300,15). This was based on looking at the starting point of the data, I can see that it is roughly around the 300 PPM mark, and the annual fluctuations are about 10,15.
- Gradient; m: N(1,1.5). I did not have much prior knowledge about how to set the gradient up, but I did not want to be too overbearing so I settled for an average initialization. This gives room for the model to have greater influence.
- Variance; vari: cauchy(10, 5). Figure 2 shows a zoomed in subsection which shows fluctuations happening constantly. Since there isn't a great deal of accountancy for fluctuations and variance in this linear model, I wanted to use a broad-tailed cauchy that would allow room for this, overall adding to the accuracy of the model.

There was a solid convergence that happened and the model seemed to have done a good job according to the results extracted from Stan. There was a large number of significant samples and Rhat values were all at 1. The table below shows this. All pairplots and autocorrelation plots are in Appendix A.



```
       mean se_mean     sd   2.5%    25%    50%    75%  97.5%  n_eff   Rhat
c    305.64  3.4e-3   0.16 305.33 305.54 305.64 305.75 305.95   2055    1.0
m      1.59  9.7e-5 4.3e-3   1.58   1.59   1.59    1.6    1.6   1947    1.0
vari   4.47  1.6e-3   0.06   4.37   4.43   4.47   4.51   4.58   1257    1.0
lp__  -6404    0.04   1.23  -6407  -6404  -6403  -6403  -6402   1194    1.0
```

Table 1. Stan summary for the linear model

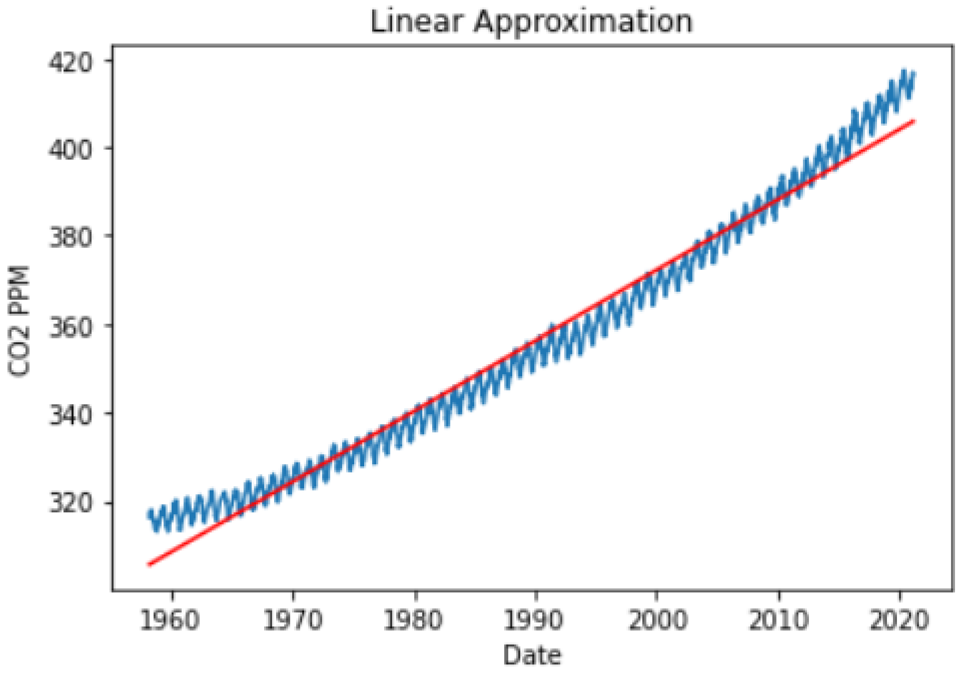


Figure 3. The linear model

As seen in figure 3, the model does have clear lapses. It does not capture the upward trend of the CO2 levels. Even though Stan noticed convergence, the convergence does not mean that the model accurately depicts what it should.

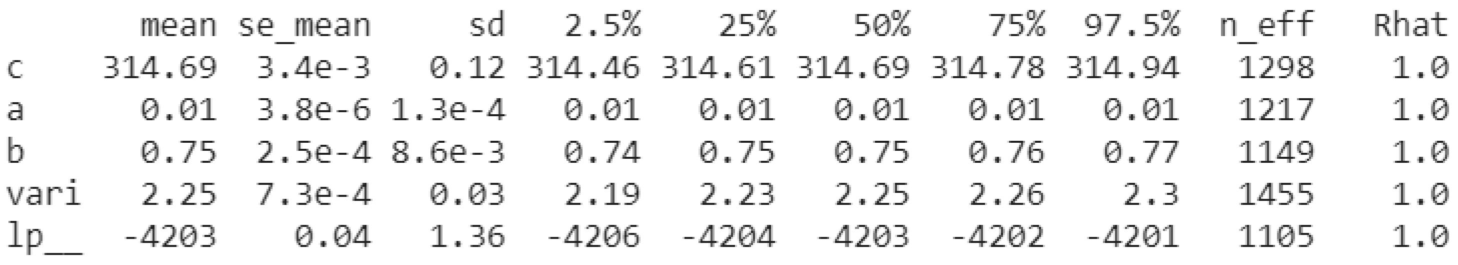
Quadratic Model

On first glance I recognized that a quadratic model would fit the data much better. The quadratic trend is usually in the form $ax^2 + bx + c$

The priors in this case would need to be altered slightly.

- Y-intercept; c: N(300, 30). Again, similar reasoning as the linear model. Slightly more variance since this is a quadratic model and values can increase in magnitude quite significantly.
- a: N(1, 1). I did not have much information about what to do with a. By reading a couple sources on the keeling curve, I was informed that quadratic mapping of CO2 requires higher order polynomials (relative to the quadratic equation) to have lower values (Stark, 2020).
- b: N(0, 5). A lack of information led me to playing around with this parameter. As long as the initialization was with lower inputs, it was relatively correct, so I went with this.
- Variance; vari: cauchy(10, 5). Same as for linear.

Although the number of effective samples took a hit, Rhat values were still at 1 showing promising results. I was more confident that quadratic trending would accurately map the dataset and after looking at the results I was very glad. The model had captured the upward trend as the years went by and was truly a much better fit than the linear one. The table below shows the Stan results, and appendix B contains the autocorrelation and pair plots.



```
        mean se_mean     sd   2.5%    25%    50%    75%  97.5%  n_eff   Rhat
c     314.69  3.4e-3   0.12 314.46 314.61 314.69 314.78 314.94   1298    1.0
a       0.01  3.8e-6 1.3e-4   0.01   0.01   0.01   0.01   0.01   1217    1.0
b       0.75  2.5e-4 8.6e-3   0.74   0.75   0.75   0.76   0.77   1149    1.0
vari    2.25  7.3e-4   0.03   2.19   2.23   2.25   2.26    2.3   1455    1.0
lp__   -4203    0.04   1.36  -4206  -4204  -4203  -4202  -4201   1105    1.0
```

Table 2. Stan summary of the quadratic model

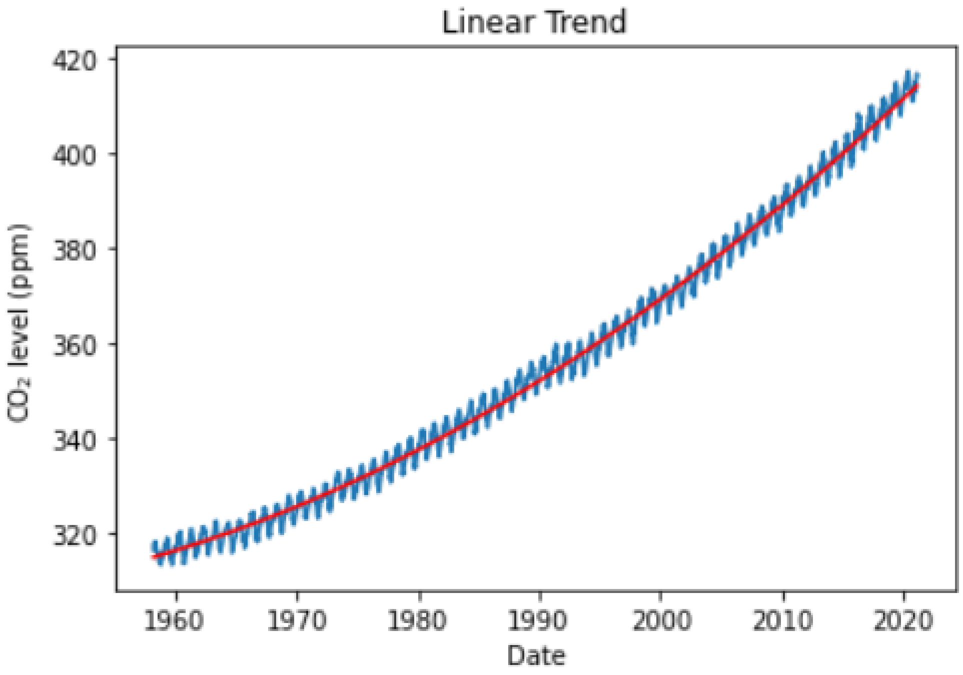


Figure 4. Quadratic approximation line against the observed data

Figure 4 illustrates the convergence of the model and the data. Now that the ‘trend’ aspect of what I wanted to model was completed, I wanted to focus on the randomness and the seasonality.

Seasonal Analysis

To begin the seasonal analysis, I chose the quadratic trend. This was clearly evident as it was a much better predictive model. Then, the upward trends from the observations needed to be removed.

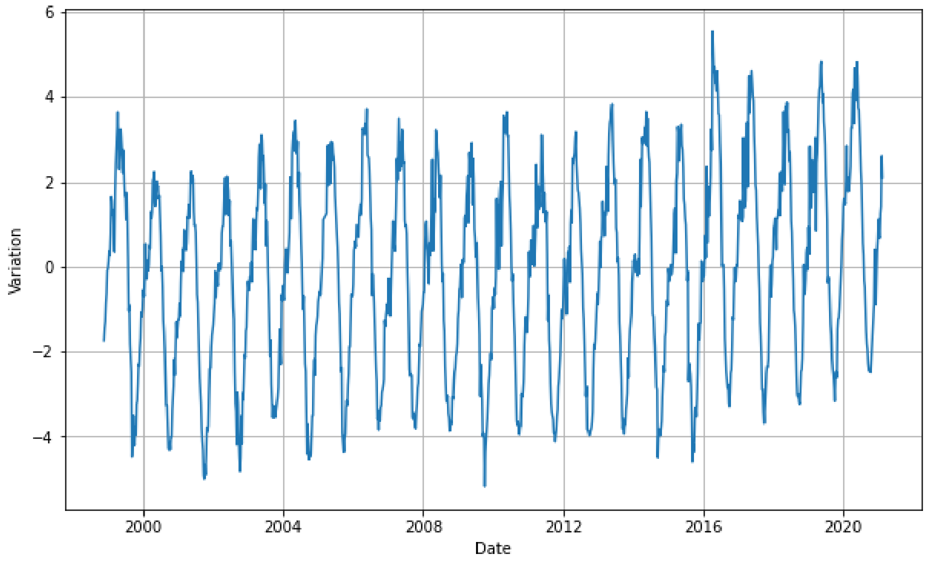


Figure 5. This graph shows the variation (or the difference between) the quadratic approximation and the observed values from 2000 - 2020)

All these differences were separated and added to a new column. The strategy was to model the seasonal data, and layer it on top of the quadratic approximation that I had done. I looked at figure 5 and noticed that there was a trigonometric pattern that could be observed. In the assignment instructions, a hint also said using a cos-based model to predict the seasonal variance.

- The equation given as the hint was: $c_2 cos((2\pi t/365.25 + c3)$.
- I tweaked it to make it: $c_2 cos((2\pi t) + \Phi)$

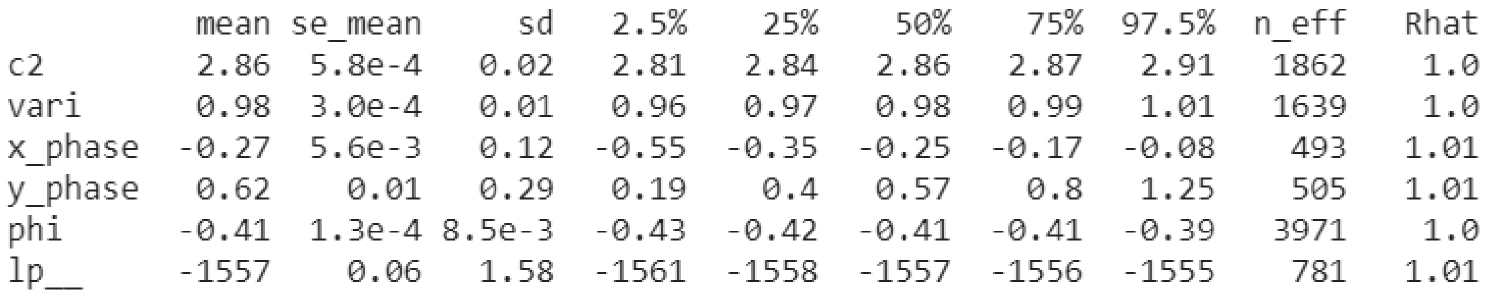
The reason for doing so was simple. Dividing $2\pi t$ by 365.25 was accounting for the seasonal variance to be spread out across the years, however, I had already normalized my date parameters accordingly. And so, my 'time' period had already been divided by years and was split into week-long intervals relative to the starting year. Secondly, $c3$ was an unknown parameter I had no knowledge about. I knew that the Keeling Curve relied on some sort of sinosidual trends, but since I was using a cosine-based model, I wanted to incorporate 'phi', a parameter dependent on the bi-phasic nature exhibited by the seasonal variance. And thus, phi became a part of my model, however, this was dependent on two transformed parameters; x-phase and y-phase, meant to mimic the peaks and troughs.

- X-phase: N(0, 0.5). I kept it broad and simple, trying to let the data influence the posterior as much as possible.
- Y-phase: N(0, 0.5). Same as above.
- c2 : inv_gamma(6, 1). The maximum peaks of the changes were +- 6, and so I wanted this parameter to reflect this more than the variance parameter. I did not have much information on this parameter so I wanted

to experiment and I knew that the posterior was a normal distribution so I thought using an inverse gamma with high alpha value and low beta value would result in thin, long peaks and troughs as shown in figure 5.

- Variance; vari: cauchy(0, 2). Since parameter c2 was trying to project most of the data, I decreased the influence of the variance. I still opted for a cauchy with broad tails but a smaller peak so that it did not influence the 'amplitude' of the posterior.

The results were very strong again. There were a solid number of effective samples that were generated. Rhat convergence also occurred, with only x-phase and y-phase being off by a hundredth.



```
          mean se_mean      sd   2.5%    25%    50%    75%  97.5%  n_eff   Rhat
c2        2.86  5.8e-4    0.02   2.81   2.84   2.86   2.87   2.91   1862    1.0
vari      0.98  3.0e-4    0.01   0.96   0.97   0.98   0.99   1.01   1639    1.0
x_phase  -0.27  5.6e-3    0.12  -0.55  -0.35  -0.25  -0.17  -0.08    493   1.01
y_phase   0.62    0.01    0.29   0.19    0.4   0.57    0.8   1.25    505   1.01
phi      -0.41  1.3e-4  8.5e-3  -0.43  -0.42  -0.41  -0.41  -0.39   3971    1.0
lp__     -1557    0.06    1.58  -1561  -1558  -1557  -1556  -1555    781   1.01
```

Table 3. Stan summary for seasonal variance model.

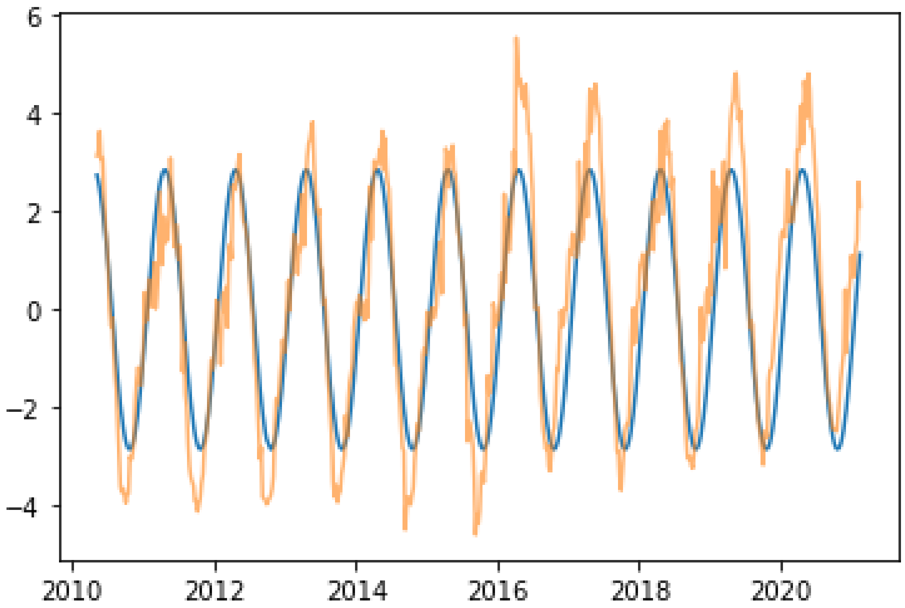


Figure 6. The seasonal variance predicted plotted on the observed.

As figure 6 shows, the posteriors were not only predicting the trend, but the phases as well. If I was really critical of the model, I would say there is still room for improvement as the peaks can be higher, but I was happy with what I had currently and was confident that these slight differences would be accounted for in the 95% confidence interval.

Final Model

The final model became simple now. There were three key aspects I outlined at the beginning. The trend, the seasonal variance and the random noise. The trend was captured by the quadratic nature, the seasonal variance by the model I created and the random noise would be captured by the parameters in the model once I combined both initial models. The Φparameter, the coefficients of $ax^2 + bx$, and the variance would all contribute to this randomness. I tried dedicating a separate variable to account for the noise, however, due to the nature of my model, this was resulting in overfitting. Thus, I chose to forgo it. The final equation I came up with, then was:

$$p(x_t | \theta) = N(ax^2 + bx + c + c_2 cos((2\pi t) + \Phi), vari)$$ [2]

The directed graph is shown below.

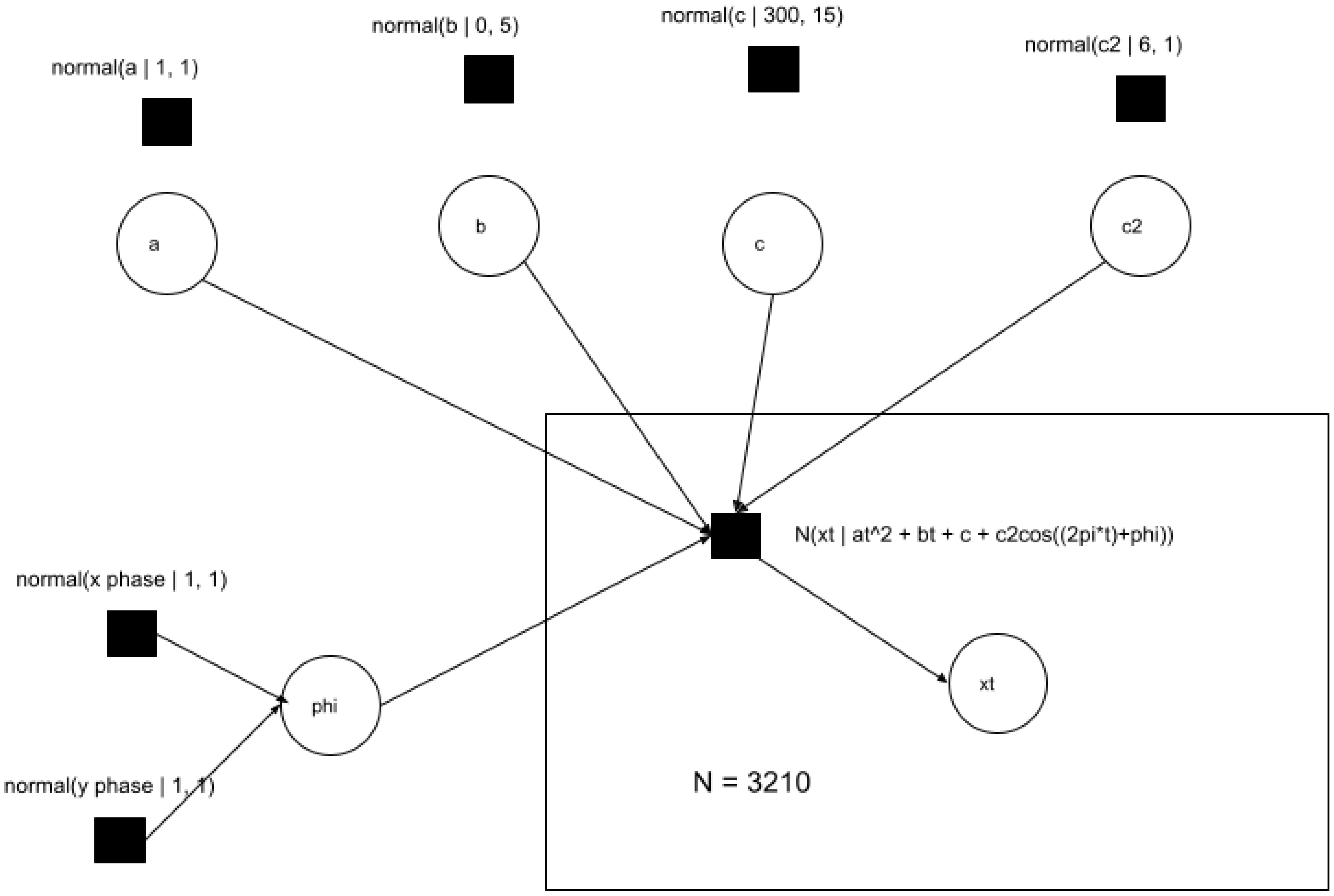
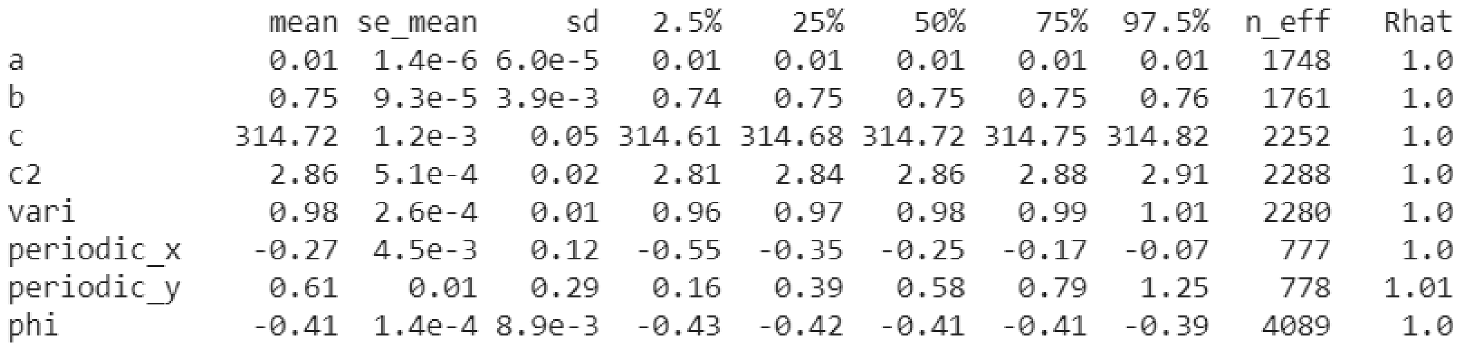


Figure 7. The factor graph of our final model. N=3210 as that is the total length of the data for which we iterate.

[2] #**probability:** Condensed the posterior function using an iterative process where I mapped individual elements of the dataset. Some level of expected propagation and simplification is done as well which builds the model in a solid way.

- Y-intercept; c: N(300, 15). Same Reasoning as for the quadratic model. Changed the beta parameter as fluctuations will be accounted for by seasonal variance as well
- a: N(1, 1). Taken from the quadratic model.
- b: N(0, 5). Taken from the quadratic model.
- X-phase: N(0, 0.5). From the seasonal variance model.
- Y-phase: N(0, 0.5). Same as above.
- c2 : inv_gamma(6, 1).
- Variance; vari: cauchy(0, 2). This is taken from the seasonal variance model as well. This is because the variance on the quadratic model itself was too strong and if we chose that it would have a double-magnitude effect.[3]

The stan results converged and there was a strong set of effective samples. The table below summarizes these results.



```
              mean se_mean     sd   2.5%    25%    50%    75%  97.5%  n_eff   Rhat
a             0.01  1.4e-6 6.0e-5   0.01   0.01   0.01   0.01   0.01   1748    1.0
b             0.75  9.3e-5 3.9e-3   0.74   0.75   0.75   0.75   0.76   1761    1.0
c           314.72  1.2e-3   0.05 314.61 314.68 314.72 314.75 314.82   2252    1.0
c2            2.86  5.1e-4   0.02   2.81   2.84   2.86   2.88   2.91   2288    1.0
vari          0.98  2.6e-4   0.01   0.96   0.97   0.98   0.99   1.01   2280    1.0
periodic_x   -0.27  4.5e-3   0.12  -0.55  -0.35  -0.25  -0.17  -0.07    777    1.0
periodic_y    0.61    0.01   0.29   0.16   0.39   0.58   0.79   1.25    778   1.01
phi          -0.41  1.4e-4 8.9e-3  -0.43  -0.42  -0.41  -0.41  -0.39   4089    1.0
```

Table 4. Stan summary for our final model.

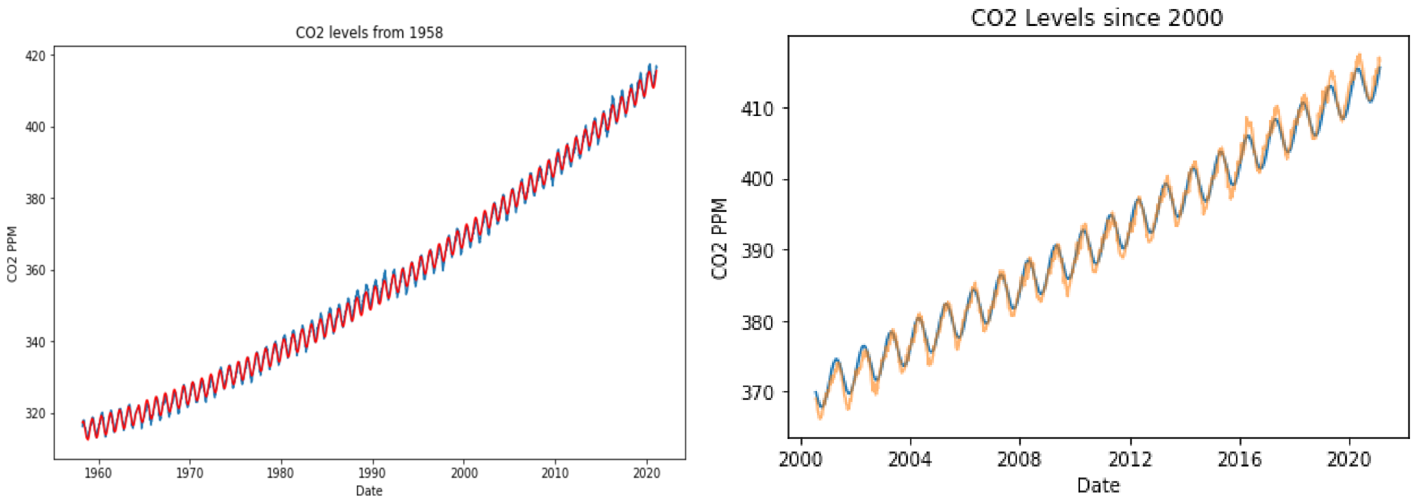


Figure 8 & 9. Both these figures show our prediction mapped onto the real results. Figure 8 worried me a little bit as I could clearly see differences in the observed and the predicted results, however, upon closer inspection and

[3] **#variables:** Identified the relevant variables to set as parameters and priors in each model. Condensed all of these and joined the quadratic and seasonal variance variables together to create the final predictive model.

zooming into a subsection I saw that the predicted results did a great job and were capturing the trends and variances of the dataset.

Next, I used the predictive values to create the confidence intervals. The upper bound and lower bound were taken and an array of the next 40 years was created so I could plot my predictions against it. The results are below.

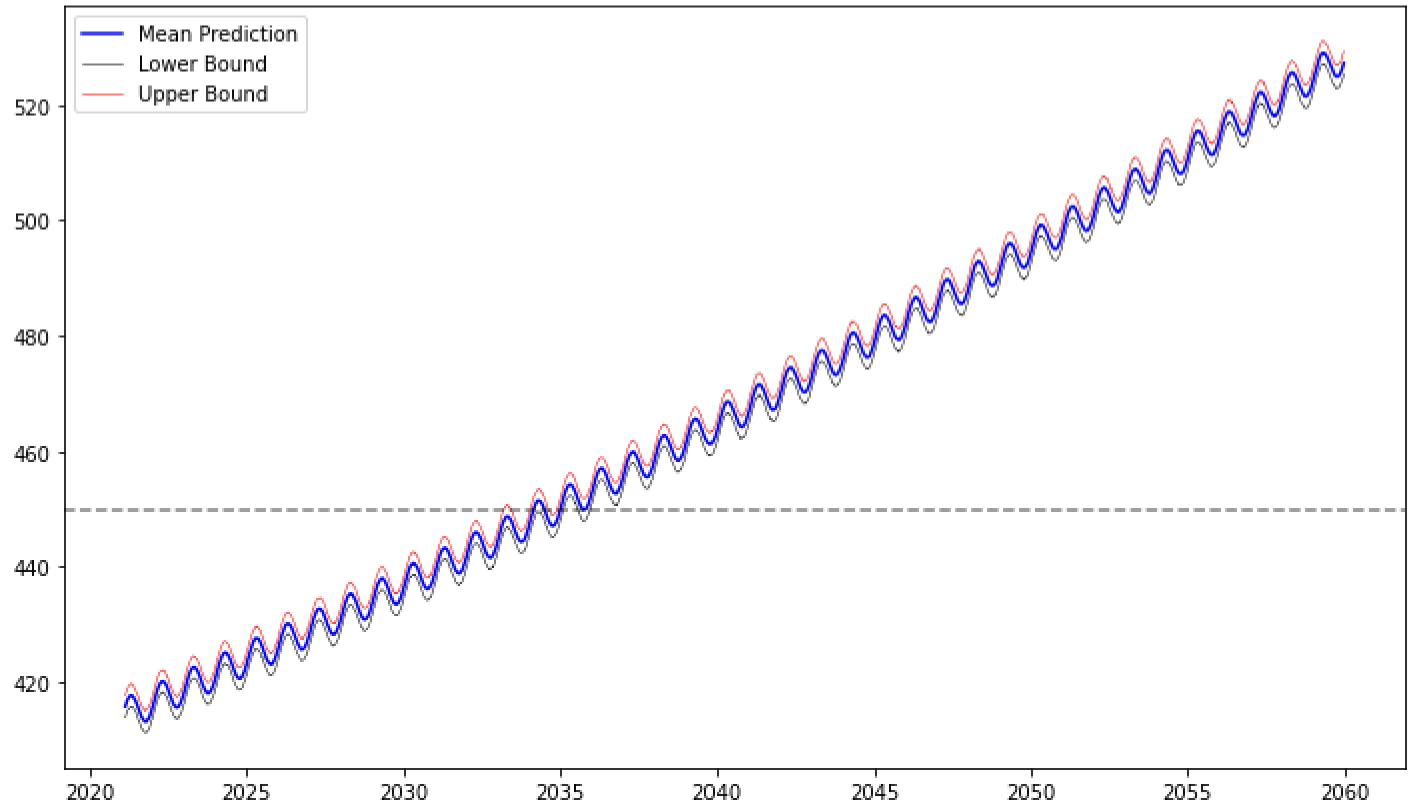


Figure 10. The figure shows the mean and 95% confidence interval of our predictions.

As the figure shows, reaching 450 PPM is not that far. Although it is slightly hard to tell, 450 PPM will be reached by around 2034. I zoomed into this time period to capture the exact predicted dates.

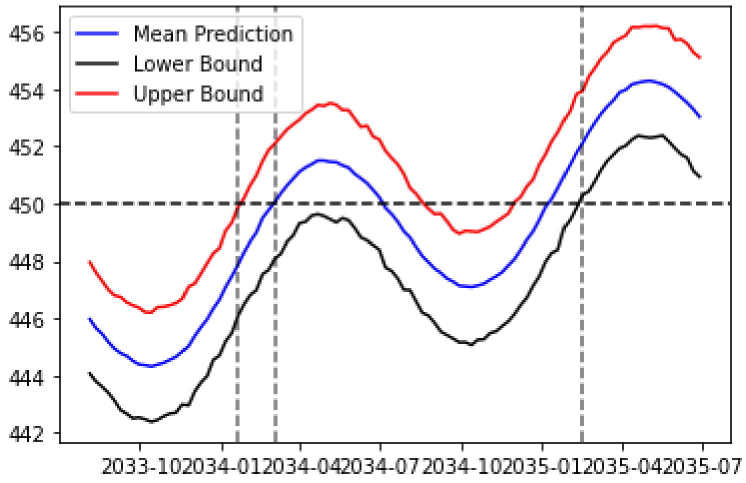


Figure 11. 2033 - 2035 predicted data and confidence intervals.

```
“450 ppm is predicted to be reached by: 2034-03-03 00:00:00 with a 95% confidence
interval between 2034-01-20 00:00:00 and 2035-02-16 00:00:00
In 2060, the expected PPM is: 527.2653144396608 With a 95% confidence interval between
525.3367163250434 and 529.3513183367469”[4]
```

Test Statistics

Three test statistics were calculated based on the final model. These included mean and standard deviation. There were other statistics that I could have used, but these two felt the most relevant. I wanted the prediction and observed means and standard deviations to be similar to validate my findings. That would further emphasize that the modeling I had done had been fairly accurate. There was still room for the confidence intervals or different percentiles to be different, as long as mean and standard deviation were supportive.

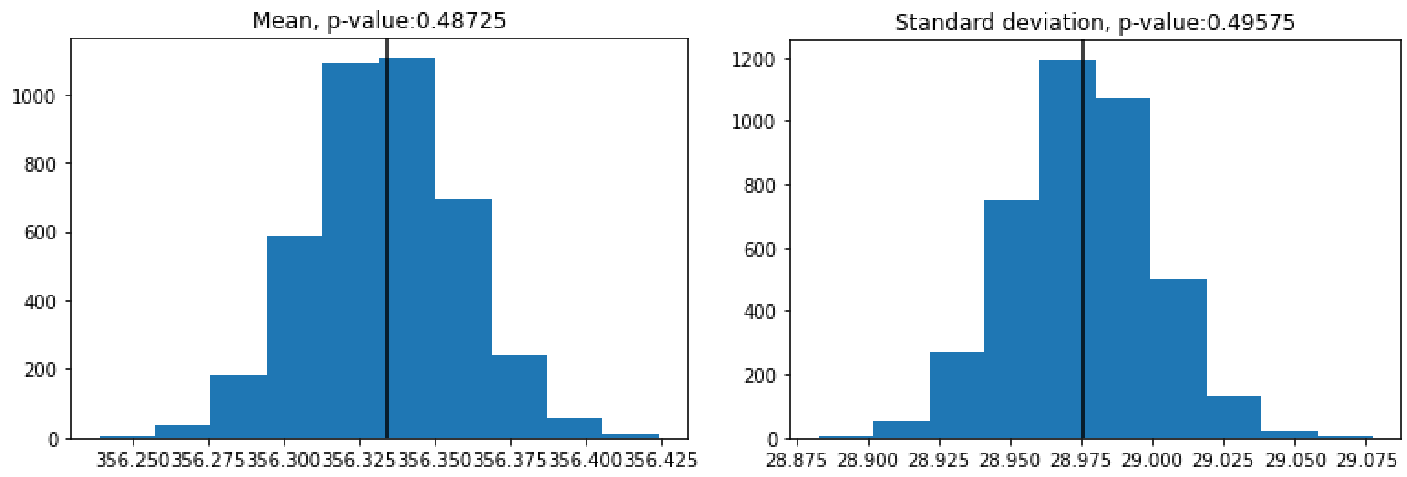


Figure 12 & 13. Mean and standard deviation test statistics.

As shown above, both the mean and standard deviations have p-values of around 0.5 supporting that our predictions closely resemble the actual observations.

Real World Implications & Critique

As the illustration shows, 450 PPM is expected to be reached by March of 2034. It can be as early as January of that year or in the best case scenario, as late as February of the following year. But it is approaching really fast. Another daunting finding is that in 2060, the PPM is predicted to be as high as 527 PPM ± 2. With the ‘danger zone’ being at 450, this means that the CO2 levels in the atmosphere will be catastrophically high. This can trigger global warming events such as polar ice caps melting, sea levels rising and unpredictable and destructive calamities such as

[4] **#confidenceintervals:** Accurately calculated, graphed and analyzed the confidence intervals of the posterior predictions. Gave exact date and times for upper and lower bounds and what this means for CO2 levels.

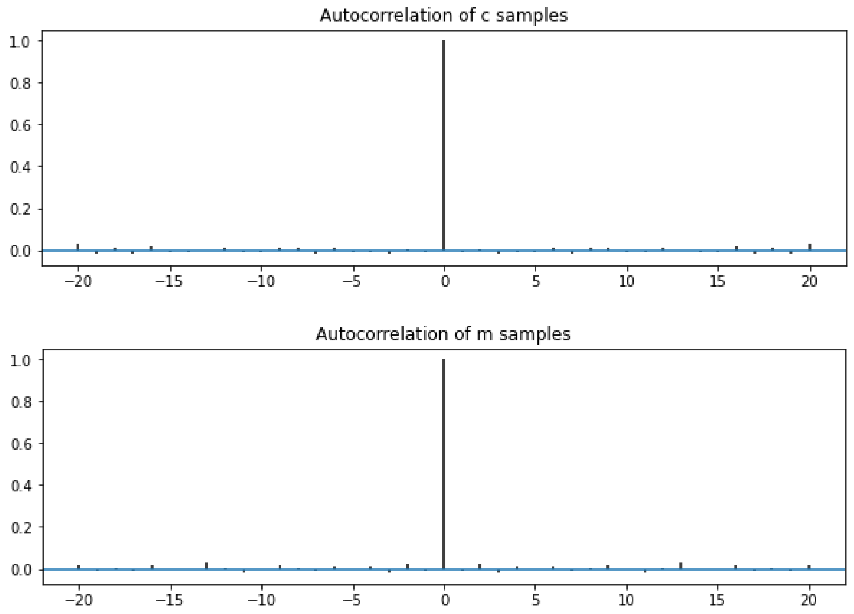
hurricanes, tornadoes and storms. It is very important to note that these are *approximations* and ultimately, the real life behavior of CO2 may or may not mimic that from the model. However, based on a number of factors, there is considerable evidence to suggest that global CO2 levels are heading towards a risky level. This means that world leaders need to come together and take collective action through strict policies and measures which ensure the safety and well being of this planet and its inhabitants.

The model does a seemingly good job, however, there are still areas for improvement. CO2 emissions vary from country to country. Perhaps modeling the impact of the 7 largest contributors to CO2 emissions can tell us the story of who is responsible and how responsible are they. This can help drive change and foster policy development in these countries faster. All in all, it is imperative to understand that whether this model has strong predictive capability or not, global CO2 levels are rising at an alarming rate, and postponing decision-making is wasting valuable time that decision-makers in the future may not have.

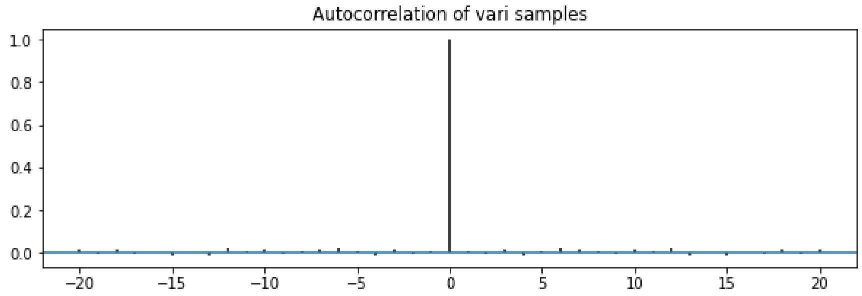
References

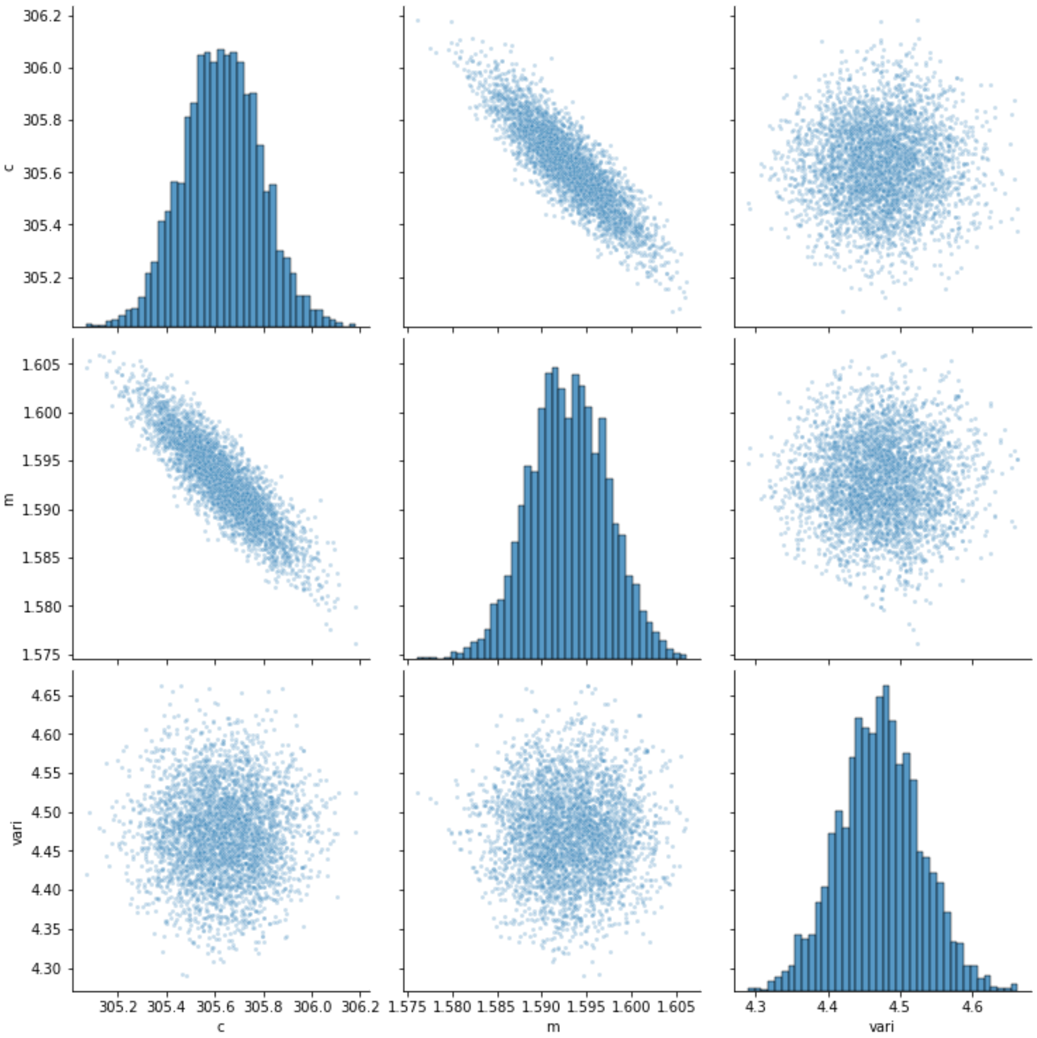
Stark., Antonio. (June 2020). Using Data Science to Understand Climate Change: Atmospheric CO2 Levels (Keeling Curve) — Model Fitting and Time Series Analysis. Retrieved from https://towardsdatascience.com/timeseries-data-science-curve-fitting-pandas-numpy-scipy-b0cd938ecb5

Appendix A (Autocorrelation & Pair plots for Linear Model)

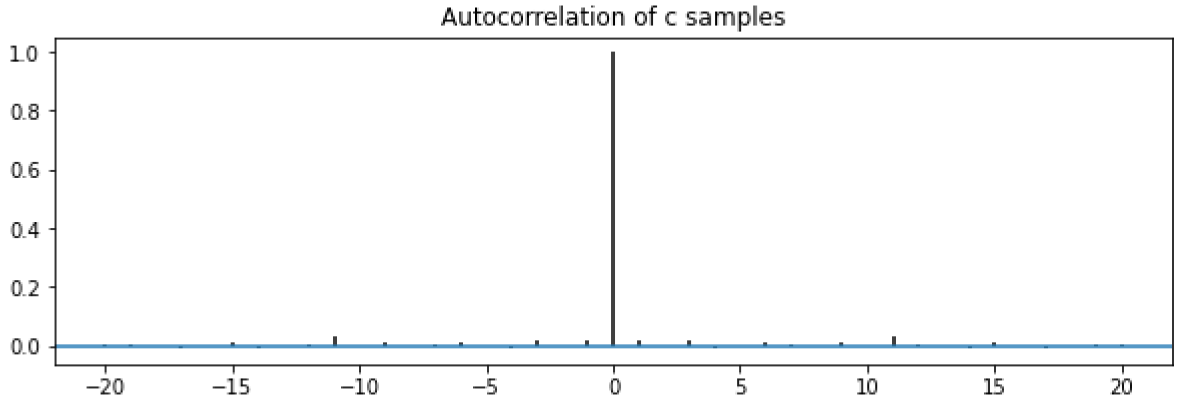

Autocorrelation of c samples

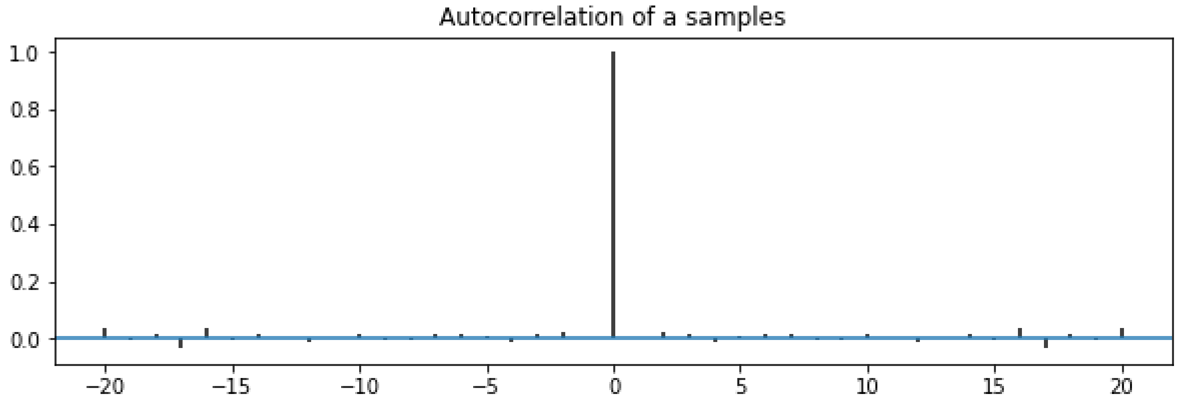
Autocorrelation of m samples

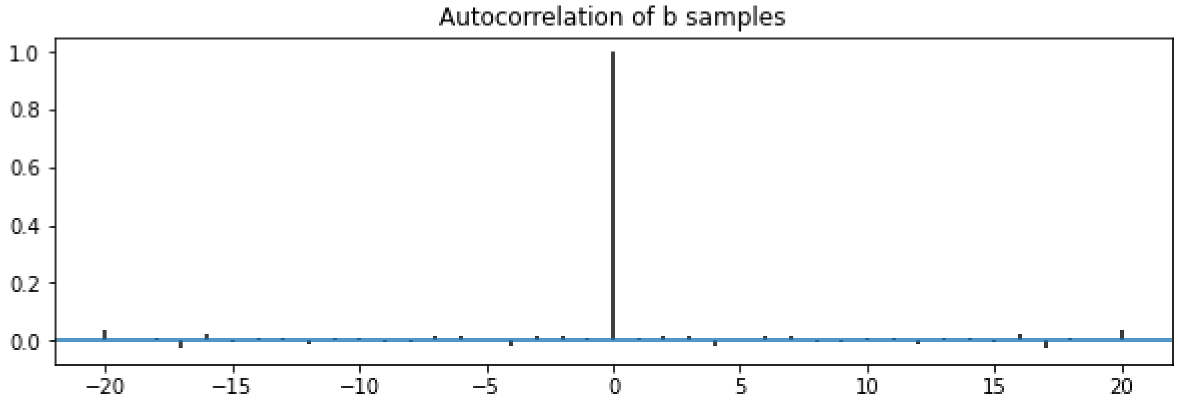


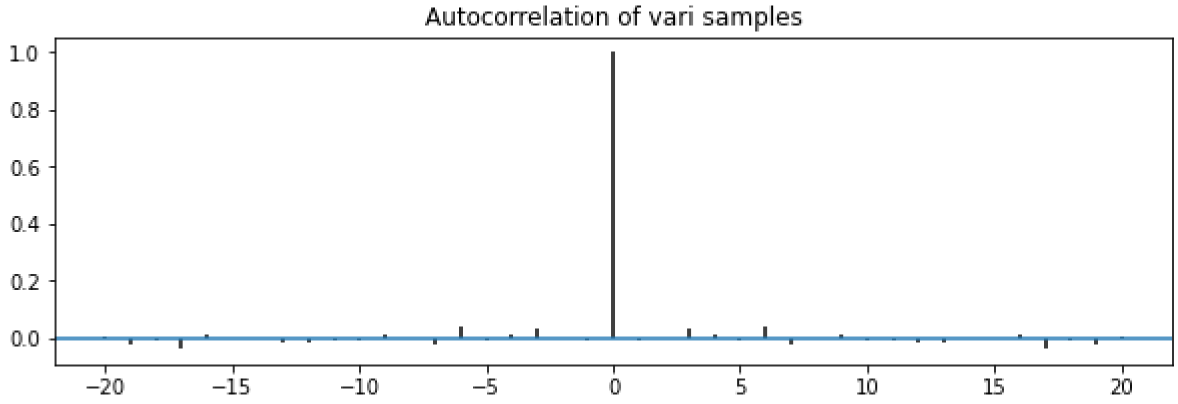


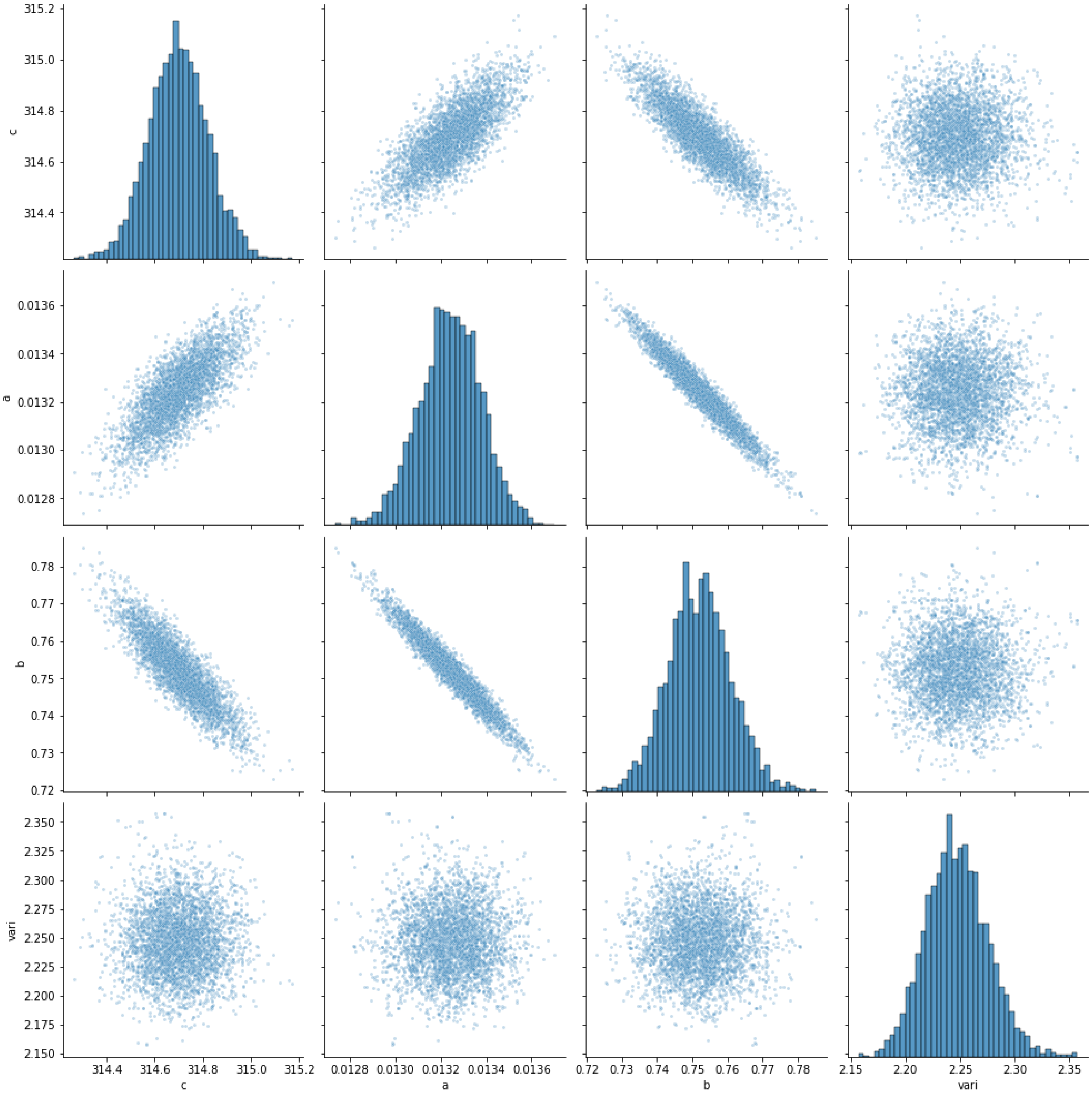
Appendix B (Quadratic Model)



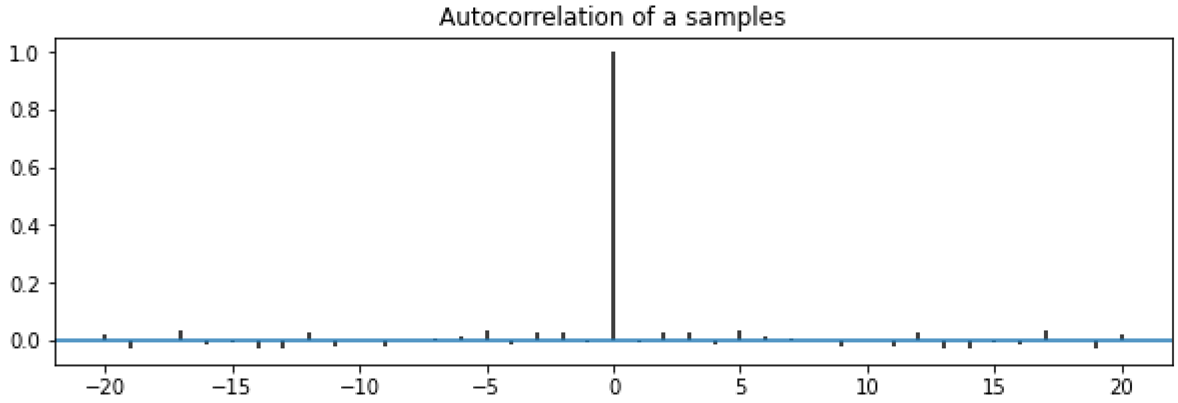

Autocorrelation of a samples
1.0
0.8
0.6
0.4
0.2
0.0
−20
−15
−10
−5
0
5
10
15
20

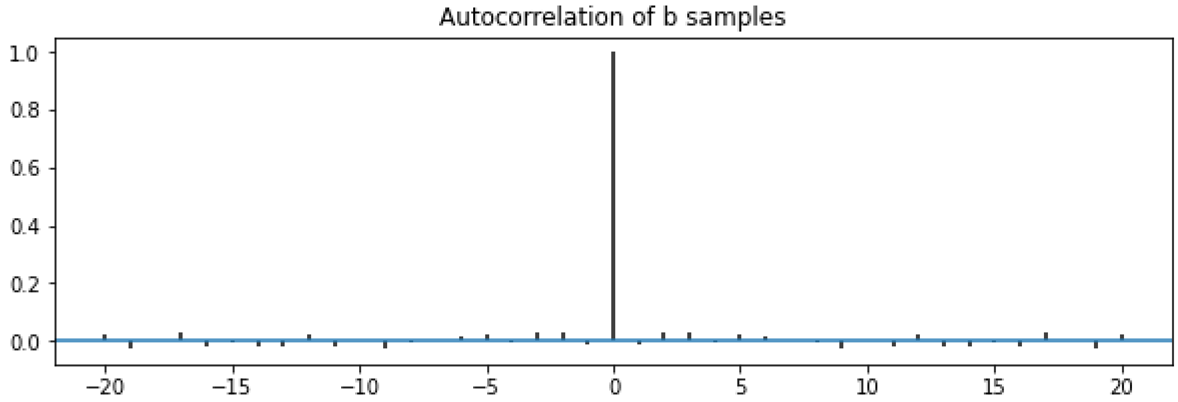

Autocorrelation of b samples
1.0
0.8
0.6
0.4
0.2
0.0
−20
−15
−10
−5
0
5
10
15
20


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

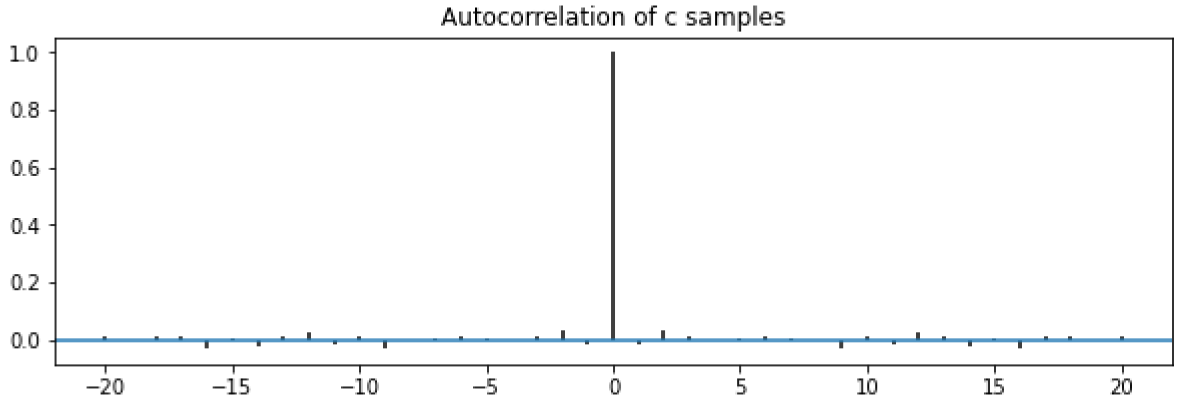


Appendix C (Final Model)

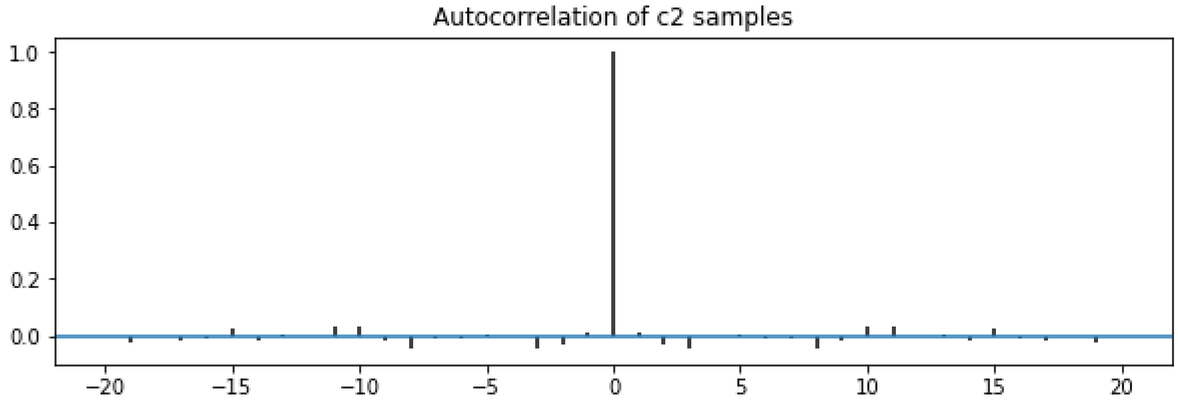




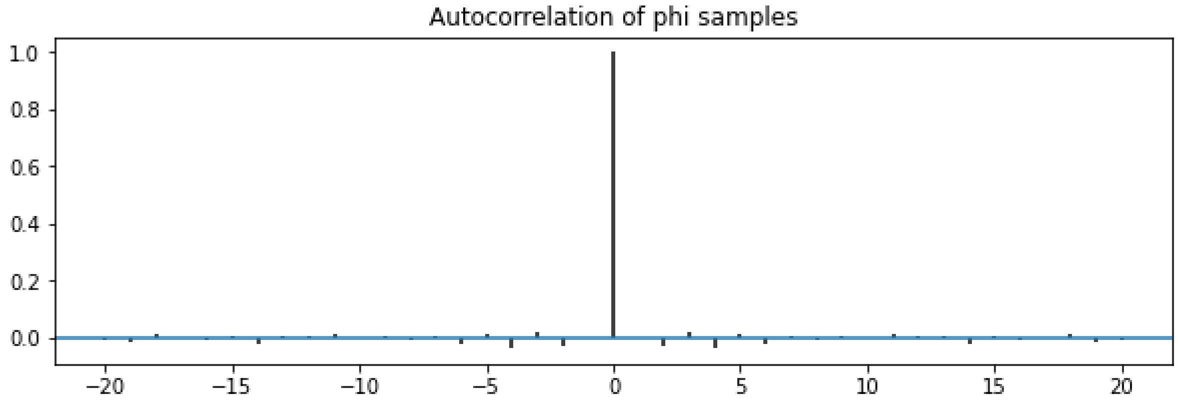
Autocorrelation of b samples



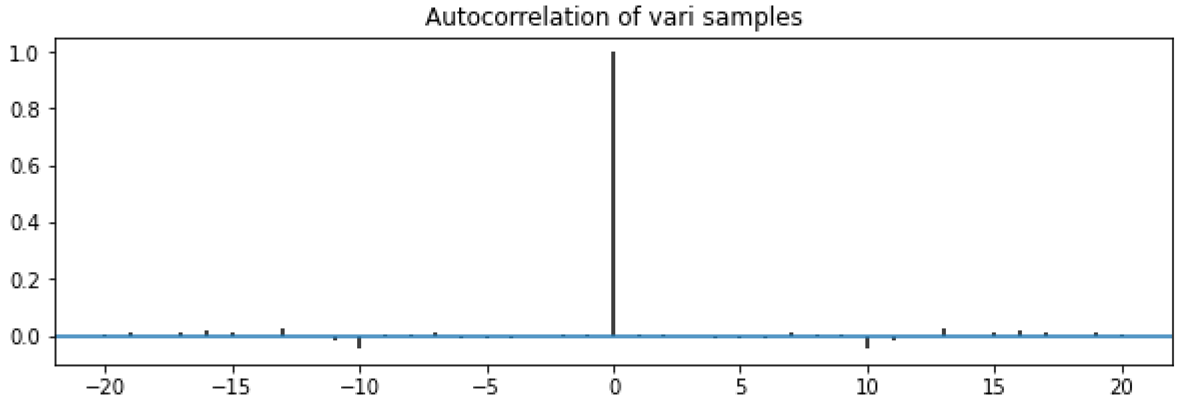
Autocorrelation of c samples



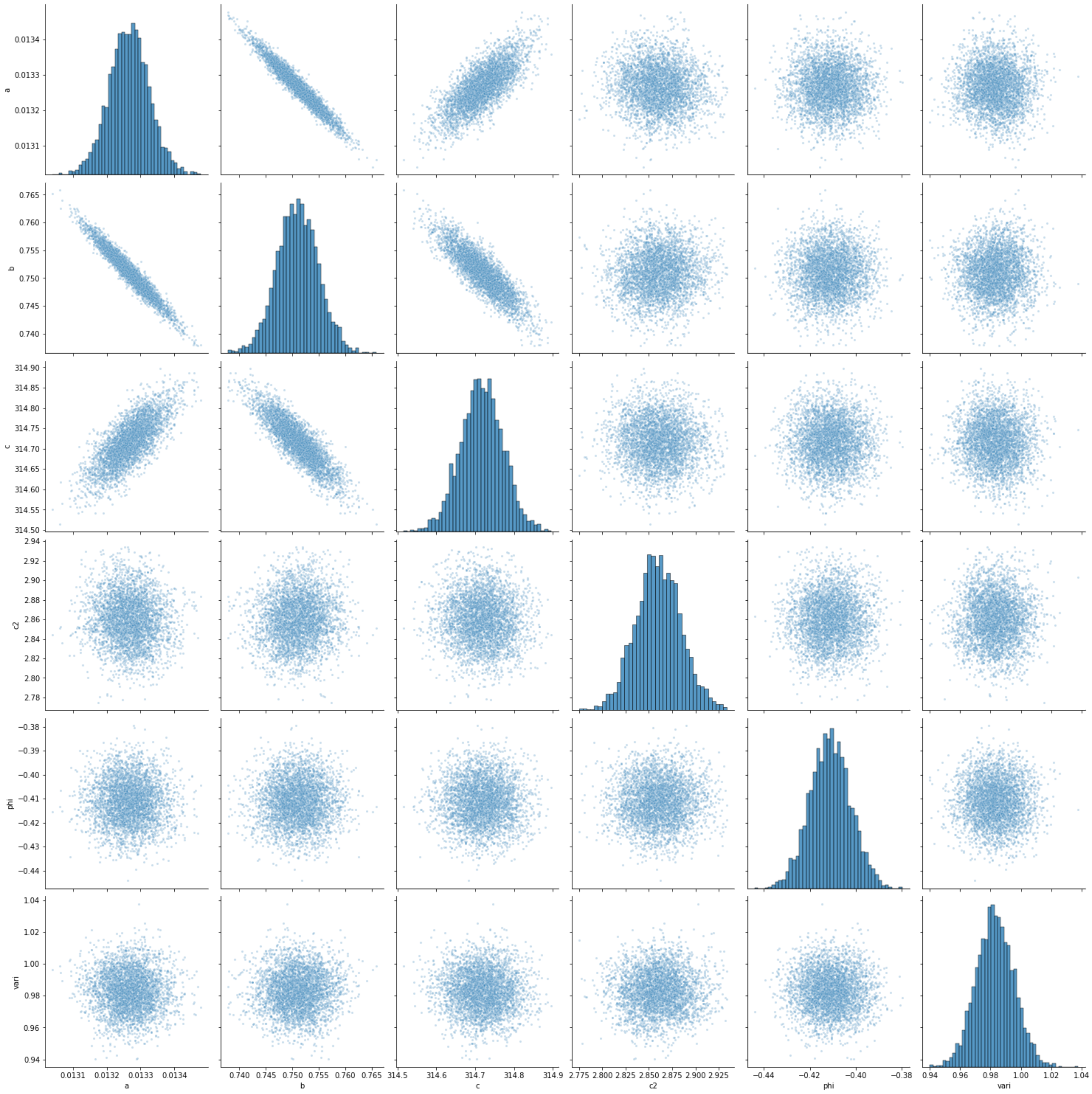
Autocorrelation of c2 samples



Autocorrelation of phi samples



Autocorrelation of vari samples


a
b
c
c2
phi
vari
0.0131
0.0132
0.0133
0.0134
0.740
0.745
0.750
0.755
0.760
0.765
314.50
314.55
314.60
314.65
314.70
314.75
314.80
314.85
314.90
2.78
2.80
2.82
2.84
2.86
2.88
2.90
2.92
2.94
−0.44
−0.43
−0.42
−0.41
−0.40
−0.39
−0.38
0.94
0.96
0.98
1.00
1.02
1.04
314.5
314.6
314.7
314.8
314.9
2.775
2.800
2.825
2.850
2.875
2.900
2.925