Final Project

CS146 - Fall 2019

Ahmed Abdelrahman

Introduction:

With the increase of global warming risk, it became crucial for us to estimate the carbon dioxide ratio over time and take the required steps to regulate the CO2 ratio in the air. In this paper, we are creating a statistical model that explains the CO2 ppm historical data set well and use it to forecast the CO2 ppm level in the future.

Old Model Critique:

The provided model has used a linear long trend function to predict the increase of the ppm overtime. Knowing that the CO2 ppm ratio depends on the world population, industrial revolution and the green areas concentrations, it's unlikely that the long trend has an absolute linear behavior. That's due to the big number of non-linear events that happened to each of the CO2 factors.
When we plotted the data, we have observed that CO2 ppm rate is increasing over time but not in a straight linear way. It is more like a line with some curvatures, which is more realistic.

New Model Approach:

In order to select the best overall model, we have used spilitted the problem into a long term trend and a seasonal one. Then, we have suggested a couple of possible models that can predict the behavior of the data. Then, we have compared which one fits the data the most and select it to be in the overall model.

Long-term trend:

For the trend part of the problem, we have examined four different equations to model the data set and represented their results compared to the data on the graph

Linear equation: $Normal(\ c0 + \ c1\ t,\ noise)$
Quadratic equation: $Normal(\ c0 + \ c1\ t^2,\ noise)$
Polynomial equation: $Normal(\ c0 + c1\ t + c2\ t^2,\ noise)$ (linear-quadratic)
Exponantial equation: $Normal(\ c0 + e^{(c1\ t)},\ noise)$
Logarithmic equation: $Normal(\ c0 * lg(c1 * t),\ noise)$ # This model didn't converge

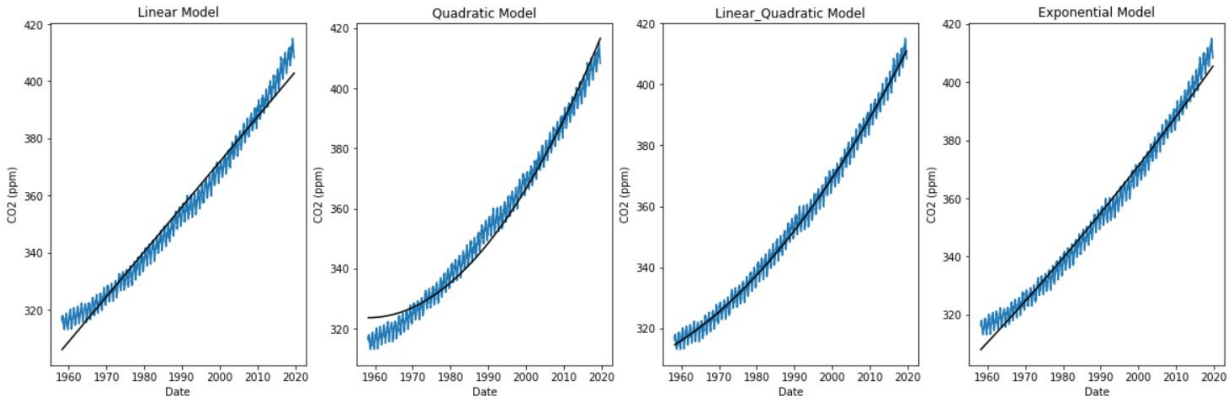

Figure1. The figure shows a comparison of various models to fit the long-term trend of the data.

The graph shows that the linear_Quadradic trend is the best fit for the data. Also, it's better than the proposed linear model in the assignment description.

Seasonal :

To get the seasonal variation model, I have subtracted the long-trend data from the original data and applied the same concept to choose the best function that can describe the seasonal data.

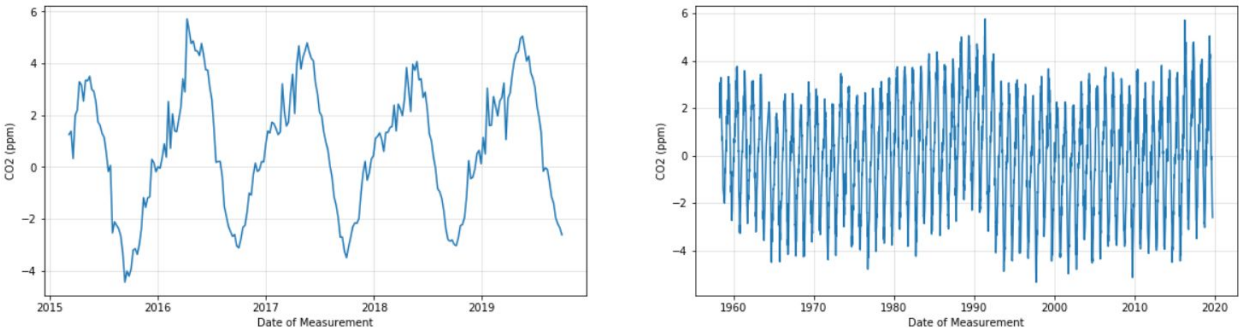

Figure2. The figure shows a clear representation of the seasonal data without the trend one.

From our analysis, we have observed that the season function has a periodic behavior that can be modeled by any trigonometric function. Thus, we have examined twp different equations to model the data set. *Note that the sine and the cosine function will have the same behavior.

Cosine equation: $Normal(\ c0 * \ cos(2\pi t/365.25 + \ c1),\ noise)$

Two_Cosines equation:

$Normal(\ c0 * \ cos(2\pi t/365.25 + \ c2) + (\ c1 * \ cos(2\pi t/365.25 + \ c2),\ noise)$

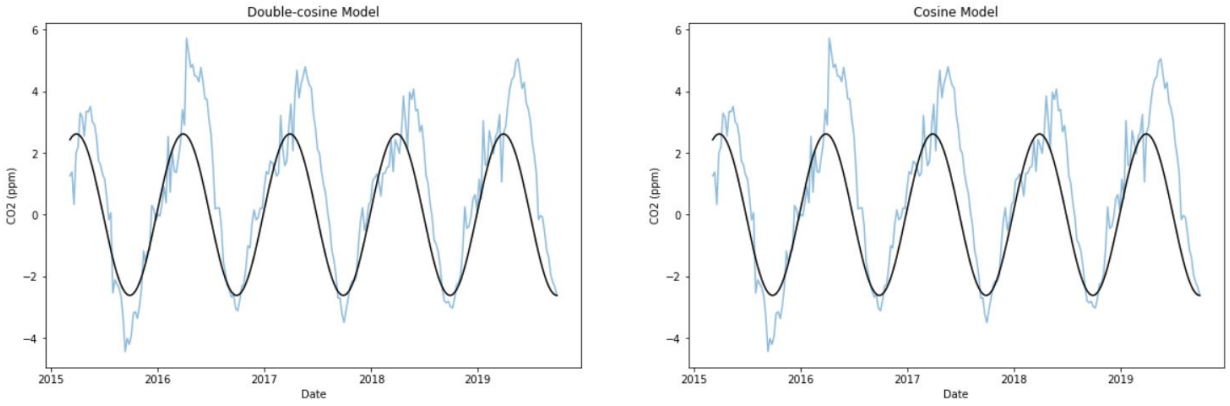

Figure3. The figure shows a comparison between the Double-cosine model and the cosine model trying to fit the seasonal data.

Because both models were fitting the data in similar ways, we have calculated the RMSE for both graphs. As it's shown here both models have very close RMSE, so we are going to choose the Cosine Model for simplicity and as it has a slightly smaller number. Although both graphs are not giving optimal answers under 1, we can accept them as it will not totally change the results of the full model.

```
The Root Mean Square Error for Cosine Model is: 1.280795325126505
The Root Mean Square Error for Two_Cosine Model is: 1.280798021286471
```

Full Model:

$$Normal(\ c0 + \ c1\ t\ + c2\ t^2 + c3 * cos(2\ \pi t + c4),\ noise)$$

c0 is modeled from a normal distribution of $normal(313,\ 15)$ as I have observed that the Co2 level starts from 313 ppm.

c0,c1,c2,c3,c4 all are modeled from a Cauchy distribution of Cauchy (0,1) where the mean is at zero because CO2 doesn't dramatically change over time. Here, we choose Cauchy distribution as it's a broader distribution that can be suitable for our case with no strong prior knowledge other than being a small one.

Noise is modeled from an inverse gamma distribution with hyperparameters 3 and 2 to limit the probability mass of the inverse gamma between 0 and 1.

Factorial Graph of the Full Model:

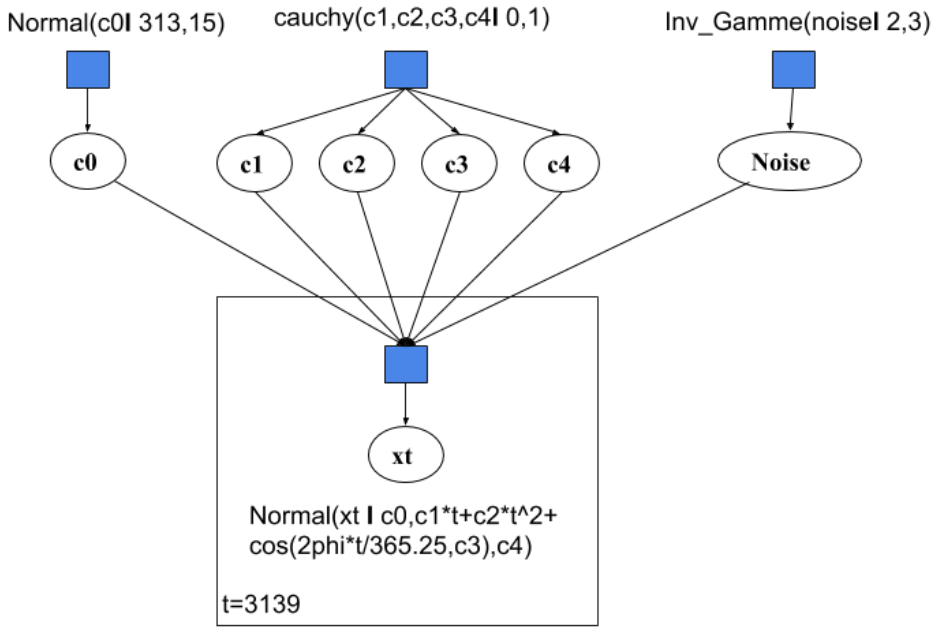

Figure4. The figure shows the factorial graph of the full model.

Model Analysis:

In our model, Markov chains converged well as we are having R-hat values that are really close to 1, which indicates that MCMC chains are mixing well. Also, the number of effective samples is more than 1000 which is pretty good for an effective sample. From the correlation graph, we can see that parameters are not correlated which means data is independent and there is no multimodally.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
c0	314.59	2.1e-3	0.07	314.44	314.54	314.59	314.64	314.72	1138	1.01
c1	0.77	1.8e-4	5.4e-3	0.76	0.76	0.77	0.77	0.78	927	1.01
c2	0.01	2.7e-6	8.4e-5	0.01	0.01	0.01	0.01	0.01	955	1.01
c3	2.62	4.7e-4	0.03	2.55	2.6	2.62	2.64	2.68	4601	1.0
c4	3.5e-4	1.2e-5	3.5e-4	1.1e-5	1.0e-4	2.5e-4	4.8e-4	1.3e-3	816	1.01
noise	1.28	2.8e-4	0.02	1.25	1.27	1.28	1.29	1.32	3583	1.0
lp__	-2371	0.05	1.77	-2376	-2372	-2371	-2370	-2369	1158	1.01

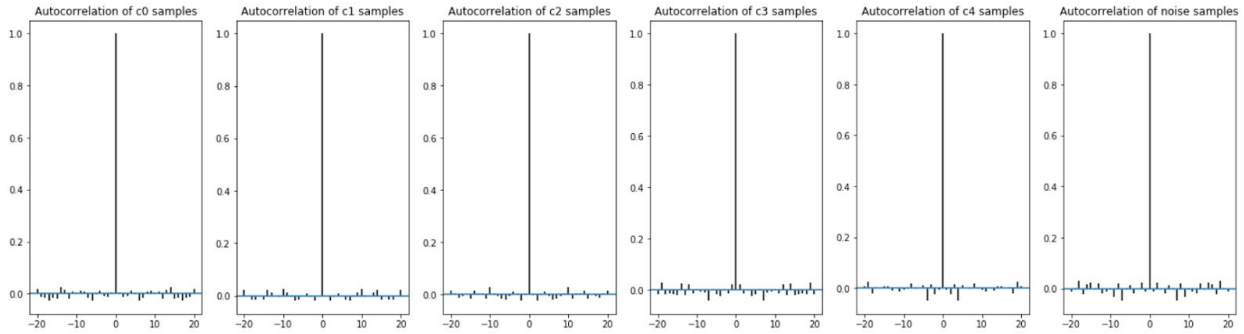

Figure5. The figures are showing Stan's results from computing the full model, and the correlation graph between the parameters.

Moreover, we have plotted a pair plot of the posteriors' overall parameters, which showed that all chains converged with no multimodality (Figure is shown in the Appendix).

Model Performance:

In order to see how our model fits the real data, we have plotted both of them and choose years as a representation the time passed after 1958-03-29 till now.

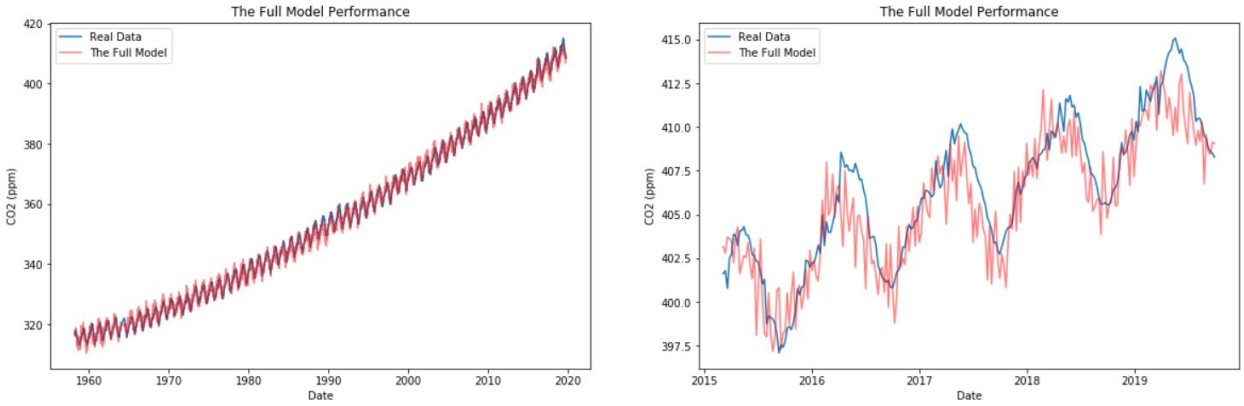

Figure6. The figure shows the full model fitting the historical data from 1958.

The model prediction for the atmospheric CO_2 level for the next 40 years:

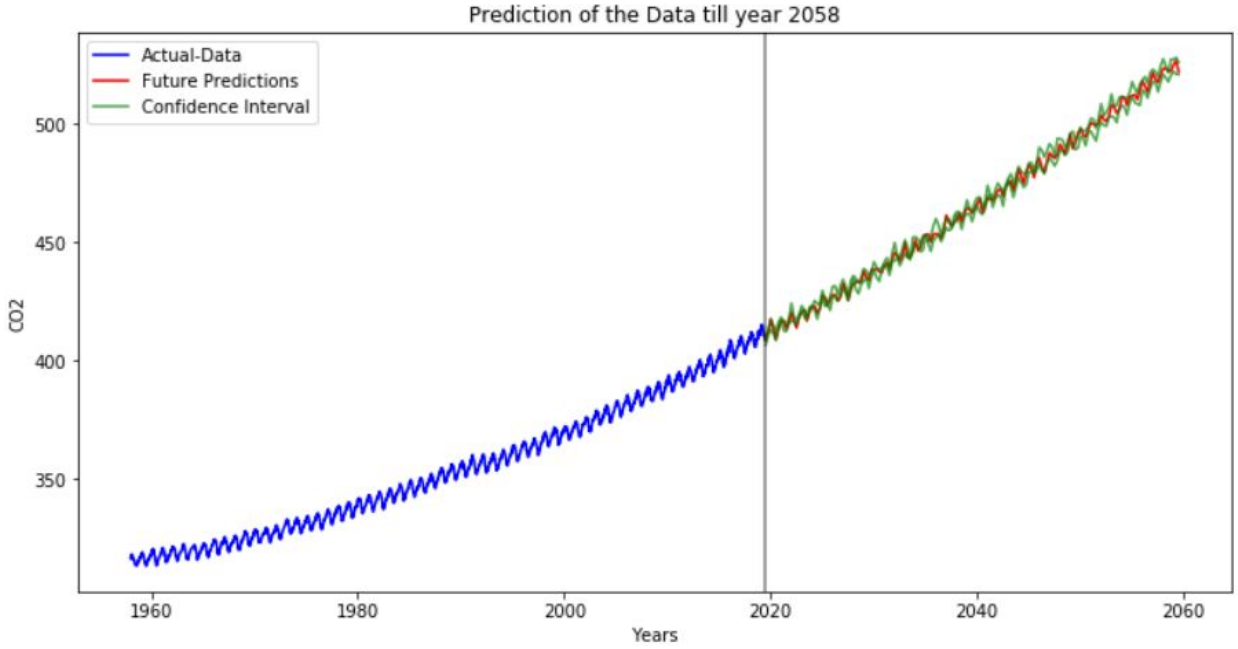

Figure6. The figure shows the full model prediction of the CO2 level until the year 2058.

From our predictive model, we are expecting that CO2 ppm value will reach 522 ppm with a confidence interval of [520.7,525]. Also, we have estimated that there is a strong probability that the CO2 level will reach a dangerous level of 450ppm by 2035. To estimate a strong probability, we have used the upper bound of our prediction of [2033,2035] confidence interval.

Model Shortcomings:

In order to critique my model, we will test how successful the model is able to fit the actual data. Earlier we have plotted the model against the real data and it showed a good fit for the data and reasonable predictions. To go into more detail, we have explored the statistical differences between our model and the data. We have examined the mean, standard deviation, the minimum and the maximum of the data. As we see in the graph, the model was able to predict the mean and the std of the real data with p_Value ranging between 0.05 and 0.95. The model has only failed to predict the minimum value of the real data and we are assuming that this is due to having limited prior knowledge of the real data. In order to fix this problem, we are suggesting improving the long term trend model by starting with a more accurate point which represents the minimum of the real data. Moreover, we have noticed a high RMSE value of 1.3 between the seasonal data and our cosine model. Preferably, we will want a value that is less than 1 and the smaller the better. To account for that, we can examine further combinations of periodic functions and see how that will fit the data.

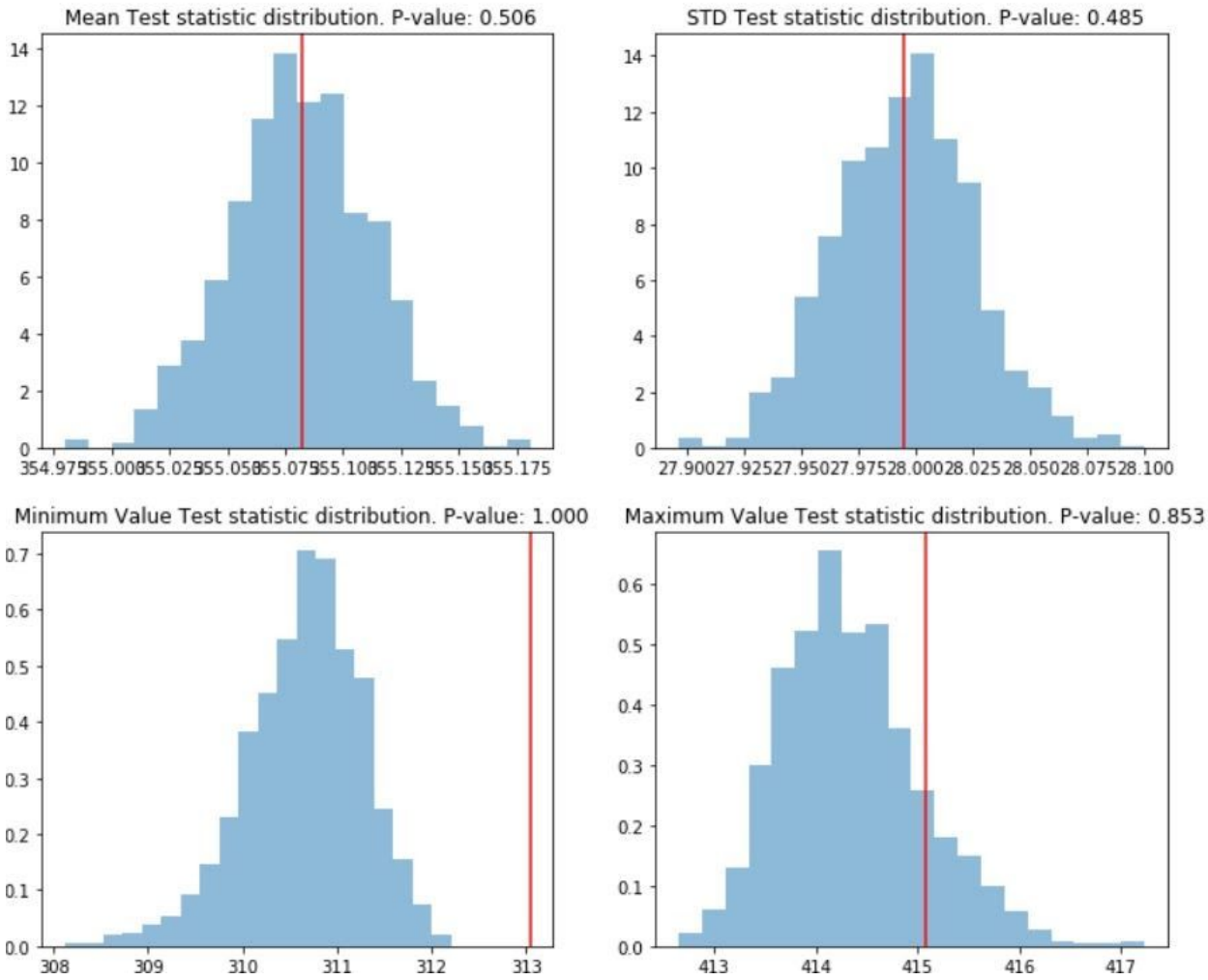

Figure7. The figure shows posterior predictive checks for different test statistics.

With the assumption that our model holds a certain level of truth, the model is suggesting the world is moving to a dangerous end. Our model is predicting that we will dangerous level of CO2 (450ppm) by 2035 and that will be even more dangerous by the end of 2058.

Appendix:

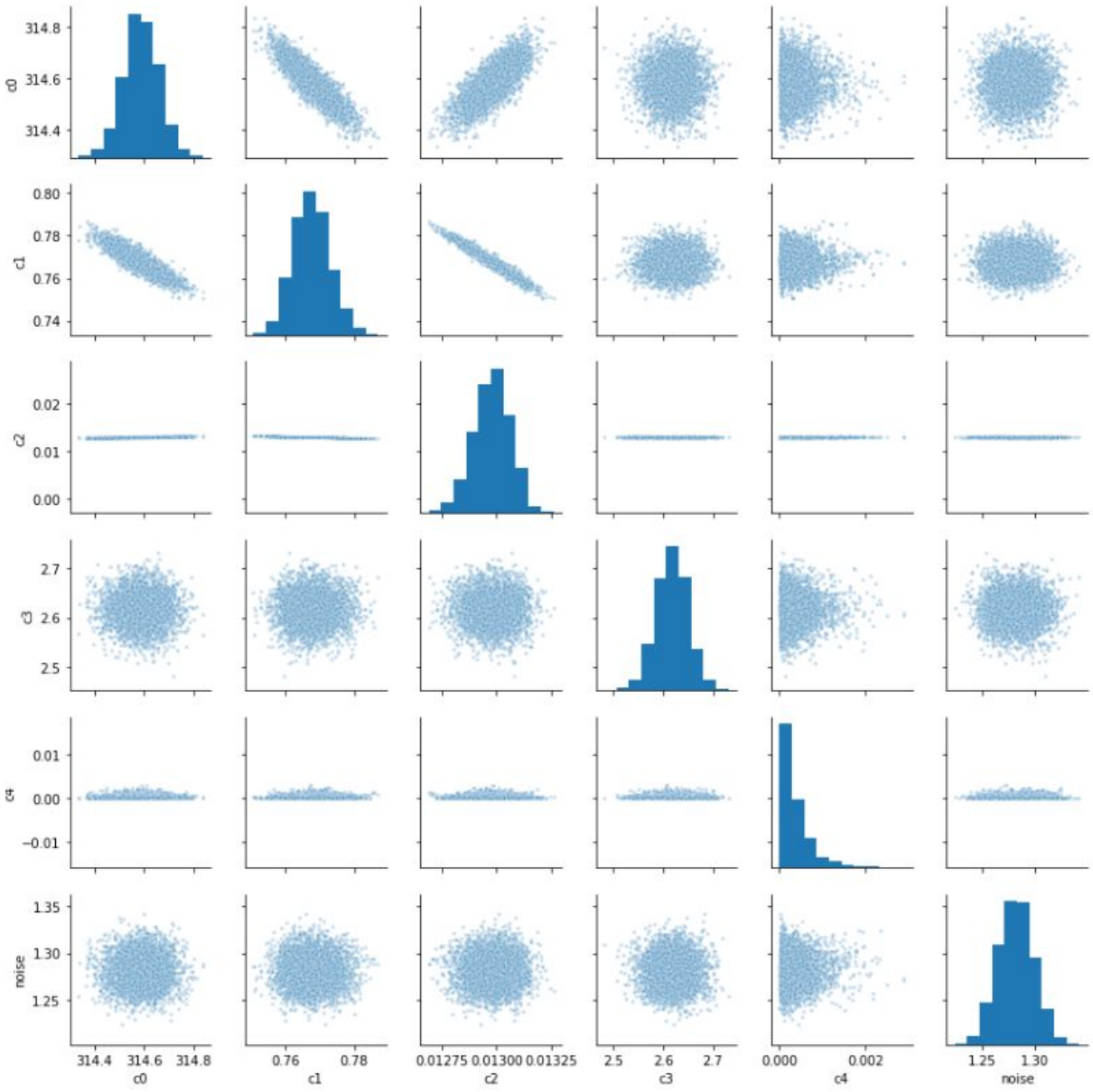

Figure8. The figure shows a pair plot of the full model prediction, which shows that the model will converge.