CS146

Final Project (CO2 Level Predictions)

Abdul Qadir

Modeling Scenario

Since 1958, the Mauna Loa Observatory in Hawaii has been measuring CO2 ppm (parts per million) levels. The data available to us is in weekly intervals i.e. we get the CO2 ppm levels for every week since 1958 to the present day. Rising CO2 levels are a threat to our planet as CO2 contributes to the level of global warming. In present days, climate change is a hot topic for policy making all around the world (specifically, for countries that are more susceptible to its effects such as Somalia and other coastal countries). It is imperative to use past data on CO2 levels to try and model future CO2 levels. Accurate modeling can help us predict when CO2 levels will cross the 450 ppm mark which is considered as a high risk level for dangerous climate change.

Model Details

I made three models in total. The structure of all models was quite similar with minor differences in between to account for more accurate modeling. The observed quantities in our model are just the data on CO2 levels (and the date when the data was collected), while the unobserved quantities are model parameters described in the following pages.

Linear Model

The first model was the linear model specified in the assignment instructions, it had a total of five priors and a normal likelihood function. The priors were:

1. C_0 which was the y-intercept to our linear likelihood function. I defined the prior over a cauchy distribution with mean 315, and a standard deviation of 10. These prior parameters were determined by observing the very first CO2 ppm levels in 1958.

Personally, I do not think the choice of priors mattered much for this parameter, since even when I used a prior of mean 0 and standard deviation 1, it resulted in the same posterior results. This irrelevance of prior choice is due to the large amount of data points we have.

2. C_1 which is the coefficient multiplied with our data point (time) to account for the linearity in our trend over time. I chose a broad cauchy prior with mean 0 and a standard deviation of 1 since I have no prior knowledge on this.
3. C_2 which is the coefficient multiplied with the seasonal changes in our model. The seasonal effect is determined by a cosine function applied on yearly changes in temperature. I chose a broad cauchy prior with mean 0 and a standard deviation of 1 since I have no prior knowledge on this.
4. C_3 which is the coefficient added to our seasonal impact of time every year. I chose a broad cauchy prior with mean 0 and a standard deviation of 1 since I have no prior knowledge on this.
5. C_4 which is the noise in our measurements of $CO2$ levels. This is the standard deviation for our normal likelihood function. I chose a broad cauchy prior with mean 0 and a standard deviation of 1 since I have no prior knowledge on this.

The likelihood is of the form:

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

This model suffered from r-hat values greater than 1, which was a product of one of the parameters (c_3) having a bimodal distribution. Moreover, the autocorrelation was extremely high between the samples and we had a very low number of effective samples (see

Table 1 for more details). These conclusions were reached by observing the autocorrelation plots, and pair plots for the samples (see Appendix A for the plots). I did not bother making predictions and performing further statistical inference with this model because of the flaws mentioned above.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
c0	305.96	2.3e-3	0.14	305.68	305.87	305.96	306.06	306.24	3698	1.0
c1	4.3e-3	1.7e-7	1.1e-5	4.3e-3	4.3e-3	4.3e-3	4.3e-3	4.3e-3	3747	1.0
c2	2.72	0.09	0.15	2.44	2.6	2.71	2.84	2.99	3	1.63
c3	2.94	2.08	2.94	2.3e-4	2.3e-3	2.9	5.88	5.93	2	130.19
c4	3.85	0.03	0.06	3.73	3.8	3.85	3.9	3.97	4	1.38
lp__	-5858	25.72	36.41	-5898	-5894	-5862	-5822	-5820	2	24.97

Table 1. Stan Parameter results for all parameters in the first linear model.

Modified linear model

As the initial linear model had problems due to the parameter c_3 having a bimodal distribution, I constrained the parameter between 0 and 3 which essentially cut the distribution in half. This does not impact the results as c_3 is a parameter in our cosine function, and the results are symmetric based on the samples.

The model itself was exactly the same as the first linear model, except that c_3 was constrained within 0 and 3. The results in Table 2 indicate that our sampling method converged and we have a good posterior distribution. Moreover, there was little autocorrelation in the samples and the pair plots showed no signs of bimodal distributions (see Appendix B for the plots)

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
c0	305.97	2.6e-3	0.14	305.69	305.87	305.97	306.06	306.24	2972	1.0
c1	4.3e-3	2.1e-7	1.1e-5	4.3e-3	4.3e-3	4.3e-3	4.3e-3	4.3e-3	2718	1.0
c2	2.6	1.7e-3	0.1	2.4	2.53	2.6	2.66	2.8	3324	1.0
c3	3.2e-3	8.3e-5	3.3e-3	9.9e-5	9.2e-4	2.1e-3	4.5e-3	0.01	1540	1.0
c4	3.89	1.1e-3	0.05	3.8	3.86	3.89	3.92	3.99	1940	1.0
lp__	-5895	0.04	1.59	-5899	-5896	-5895	-5894	-5893	1473	1.0

Table 2. Stan Parameter results for all parameters in the modified linear model.

Since I had a model with a solid posterior, I compared model predictions to our observed data to see how it performs. Figure 1 shows that the model has good predictions for 2008 but then the model starts to underfit as the data starts to curve upwards, while the model predictions linearly stay below the data. Due to this flaw in the model, I did not carry on with further analyses on the model. However, Appendix C has plots for future predictions from this model if the reader is interested.

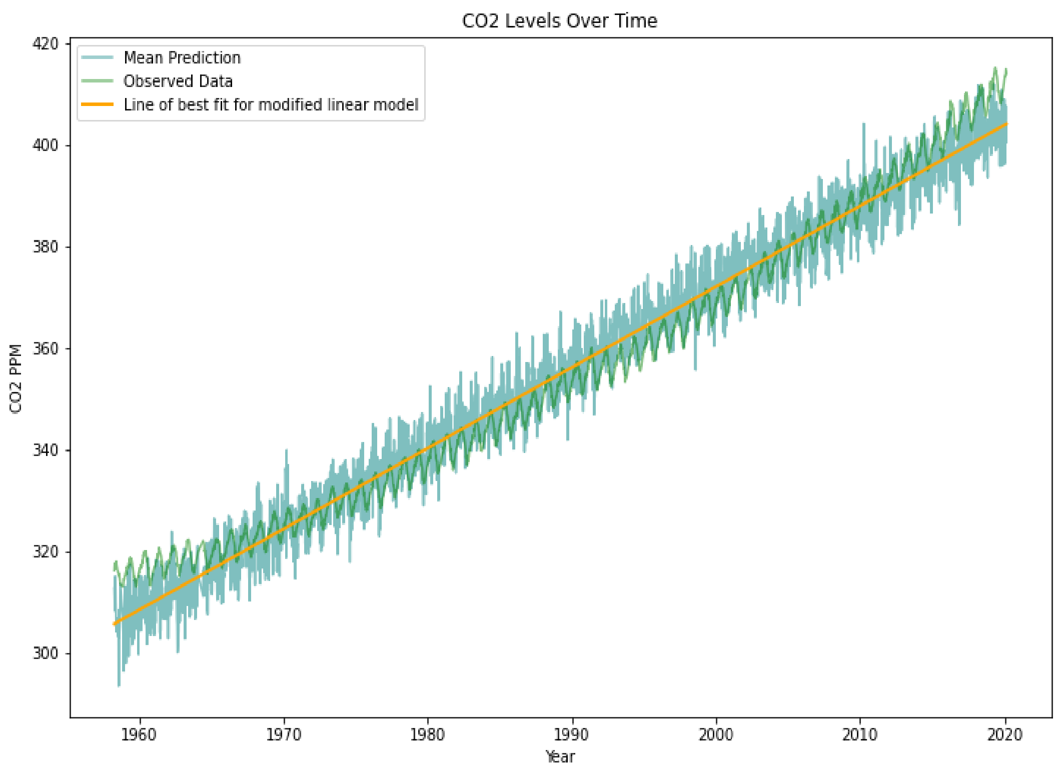


Fig 1. Modified Linear Model Results for Observed Data

Quadratic Model

Due to the flaws in the previous model, I assumed the CO2 levels follow a quadratic trend since they are curving upwards (as can be seen in Figure 1). Hence, I made the following quadratic model with six priors instead of five, and a normal likelihood function. The first five priors and their respective parameters are the same as the linear model. The new parameter is c_5 which is the coefficient for our quadratic term. The parameter is normally distributed with mean 0 and a standard deviation of 1 as I did not have any prior knowledge on it.

The likelihood is of the form:

$$p(x_t|\theta) \ = \ N(c_0 + c_1 t \ + c_5 t^2 + c_2 cos(2\pi t/365.25 \ + \ c_3), c_4)$$

Figure 2 below shows the factor graph for the model.

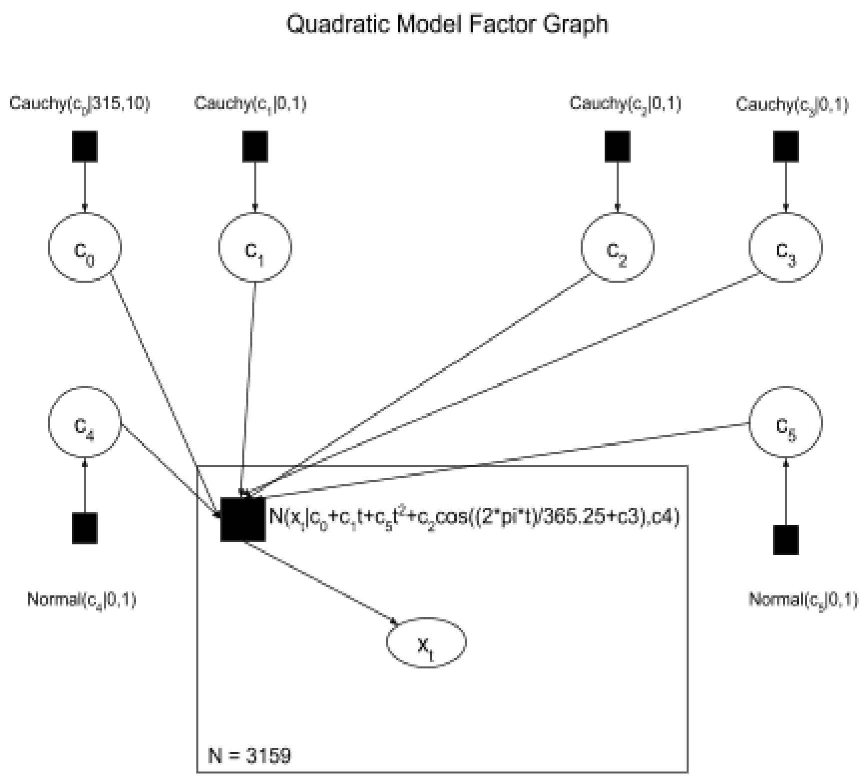


Figure 2. Factor graph for the final quadratic model.

The model did not have any sampling or autocorrelation deficiencies. The rhat values are 1 and there is a high number of effective samples and the results can be seen in Table 3. Pairplots and autocorrelation plots for this model can be seen in Appendix D.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
c0	314.6	1.9e-3	0.07	314.46	314.55	314.6	314.65	314.74	1404	1.0
c1	2.1e-3	4.3e-7	1.4e-5	2.1e-3	2.1e-3	2.1e-3	2.1e-3	2.1e-3	1079	1.0
c2	2.62	6.1e-4	0.03	2.55	2.6	2.62	2.64	2.68	2801	1.0
c3	3.5e-4	6.8e-6	3.4e-4	1.1e-5	1.1e-4	2.5e-4	5.0e-4	1.2e-3	2460	1.0
c4	1.28	3.2e-4	0.02	1.25	1.27	1.28	1.29	1.31	2595	1.0
c5	9.8e-8	1.8e-11	6.0e-10	9.6e-8	9.7e-8	9.8e-8	9.8e-8	9.9e-8	1075	1.0
lp__	-2385	0.05	1.75	-2389	-2386	-2384	-2383	-2382	1350	1.0

Table 3. Stan Parameter results for all parameters in the quadratic model.

The model fit the observed data much more accurately than the previous models as shown in Figure 3 below; hence, I used this model for further predictions and statistical analyses.

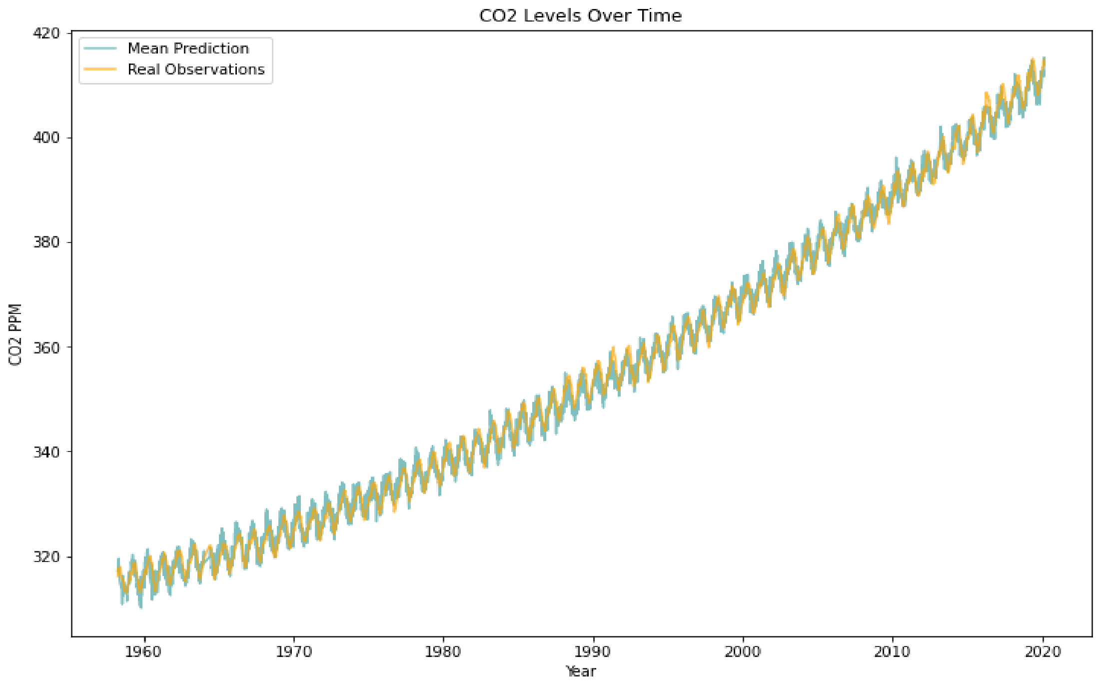


Figure 3. Plot for model predictions and observed data upto the present day.

Figure 4 shows the model's predictions from 2020 to 2060, and it predicts that we will cross the 450 ppm CO2 levels around the year 2034

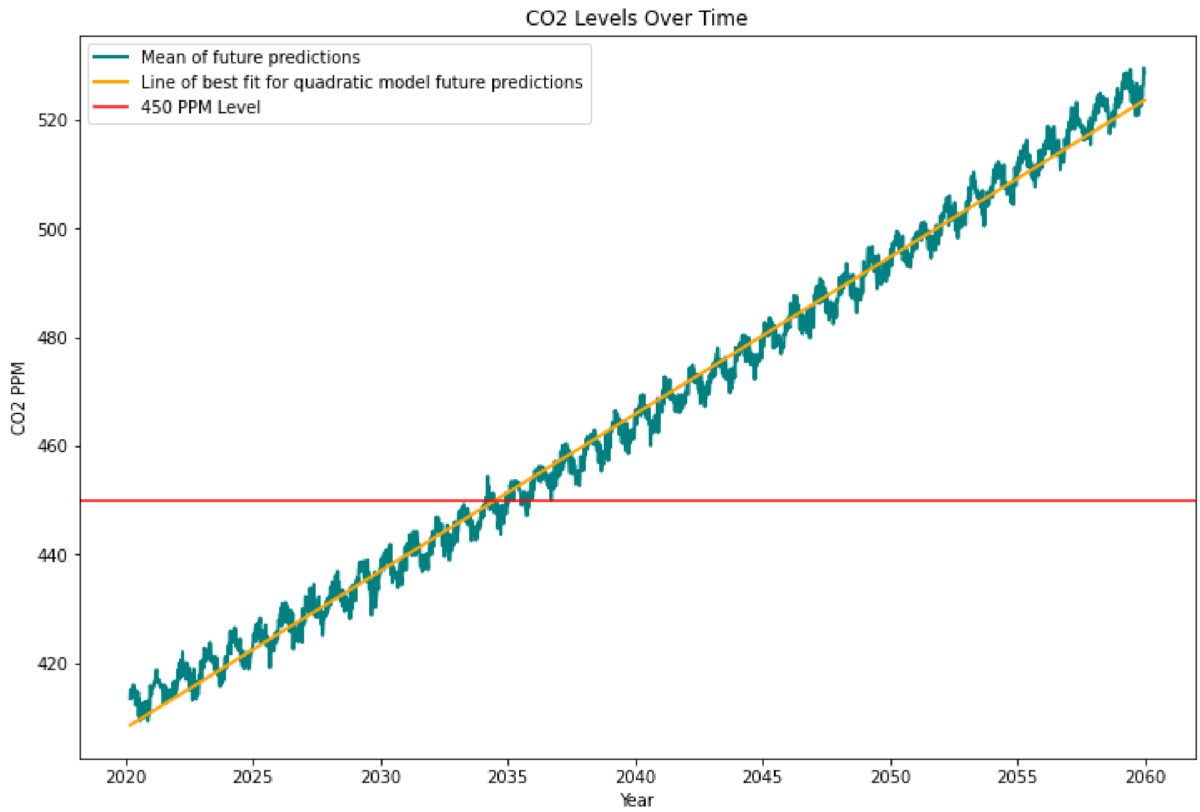


Figure 4. Plot for future model predictions until 2060.

In order to get an estimate of what CO2 levels we can expect in the year 2060, I plotted the 95% confidence interval of CO2 levels as well as the mean predicted levels in Figure 5 below. The model predicts that we can expect an average of 525 ± 2 ppm of CO2 in 2060

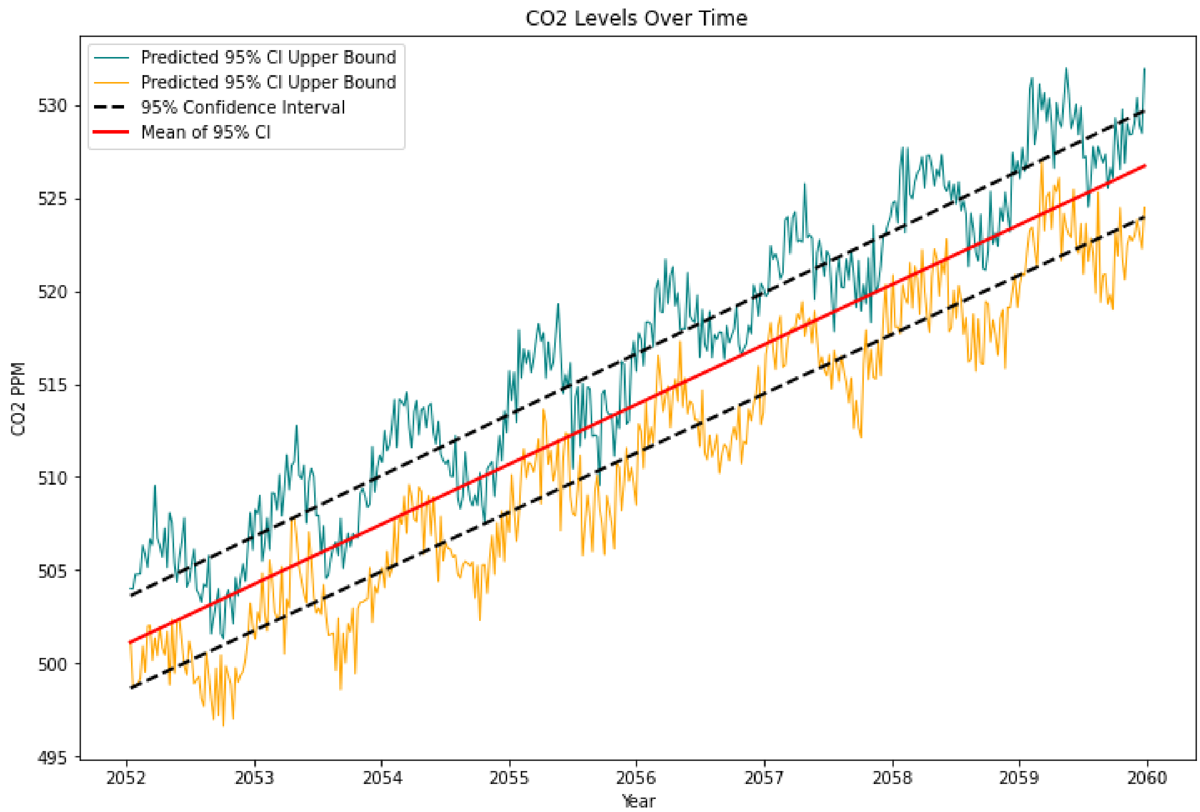
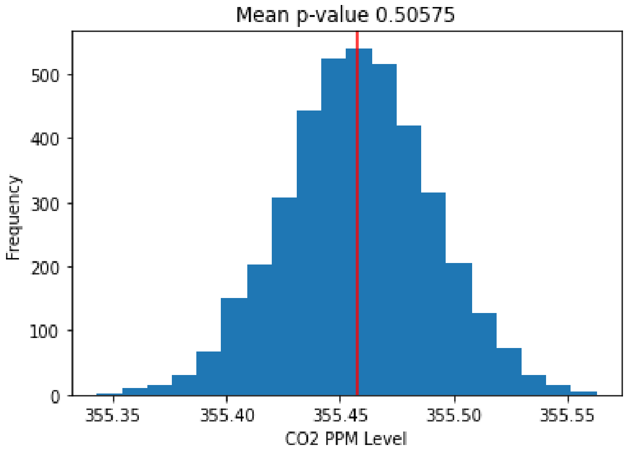
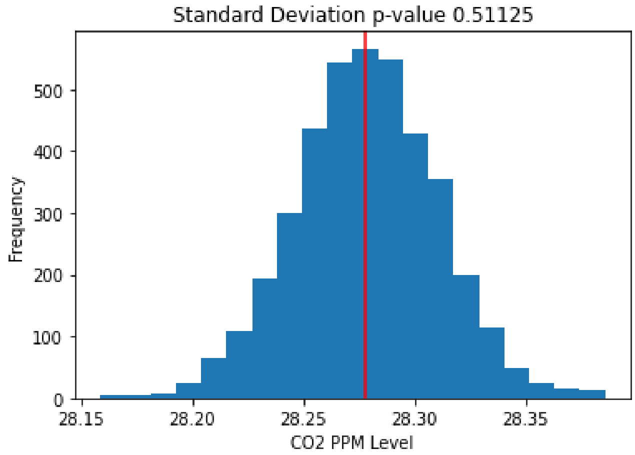


Figure 5. Plot for future model prediction from 2052 to 2060.

Lastly, since this model gave the best results out of all three models. I calculated test statistics on data generated by the model, and compared the results with the true observed data mean to calculate p-values. The two test statistics I used were mean and standard deviation. The results can be seen in Figure 6 below and the test statistics on the generated data are quite similar to the observed data with a p-value of around 0.5.



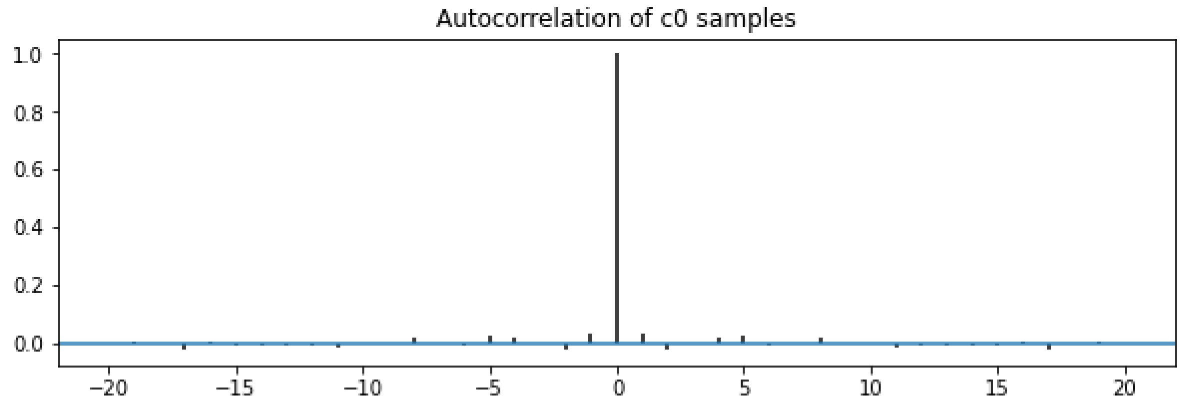


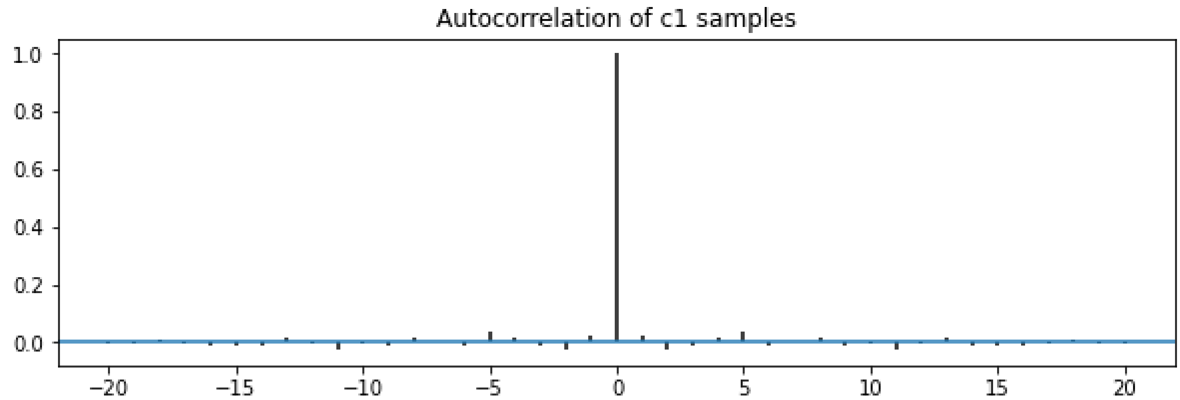
Practical Implications

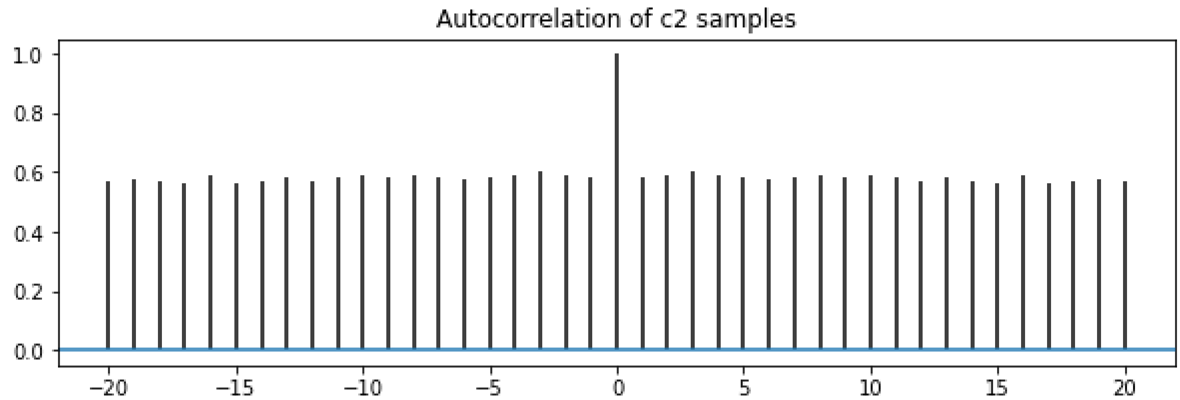
The model predicts CO_2 levels on a global scale, which can be used to gauge when we can expect the majority of the world to be impacted by rising CO_2 emissions. However, CO_2 levels vary on a country by country basis, which is why it would helpful to use the model on country based data (although, we might need a different model as the trend might be different, and this is an obvious flaw of the model for practical policy making); however, this model can still be used to guide policy makers on when they should be expected to bring about change in reducing CO_2 emissions.

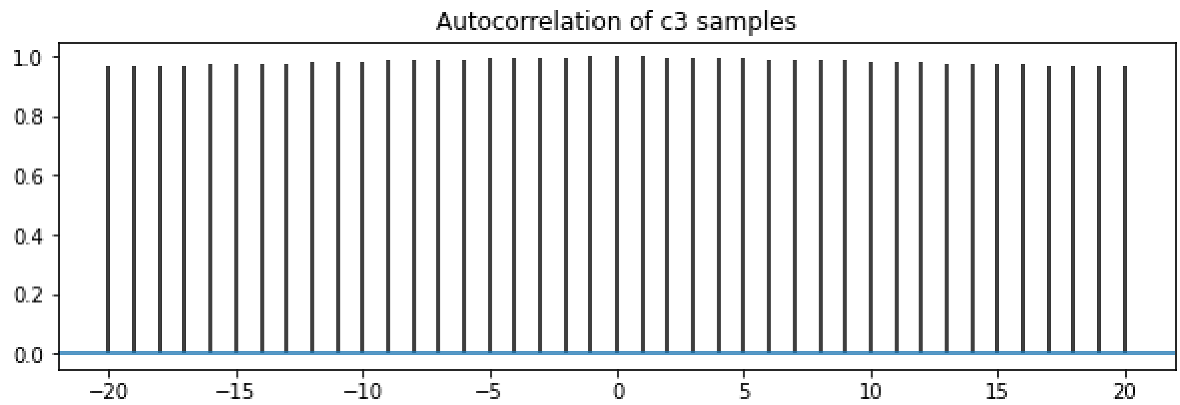
Appendix A

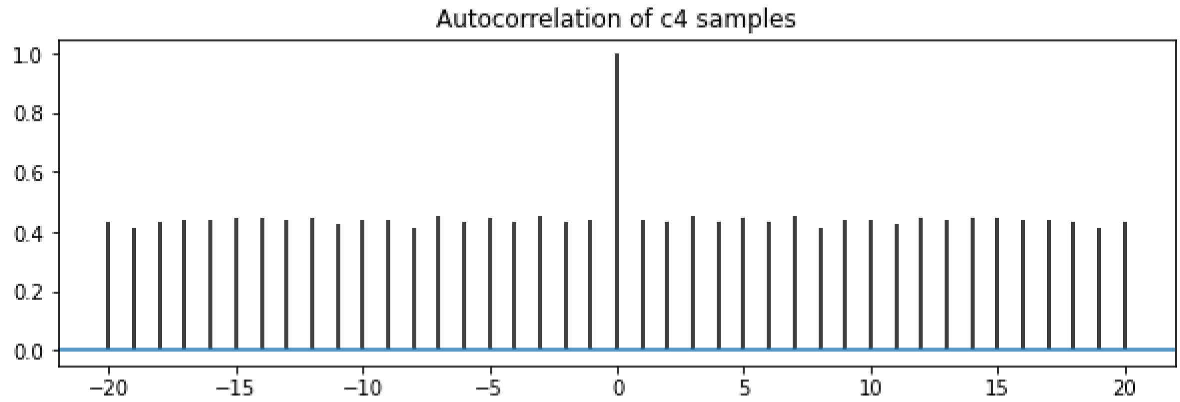
Default Linear Model Autocorrelation plots and Pair plots

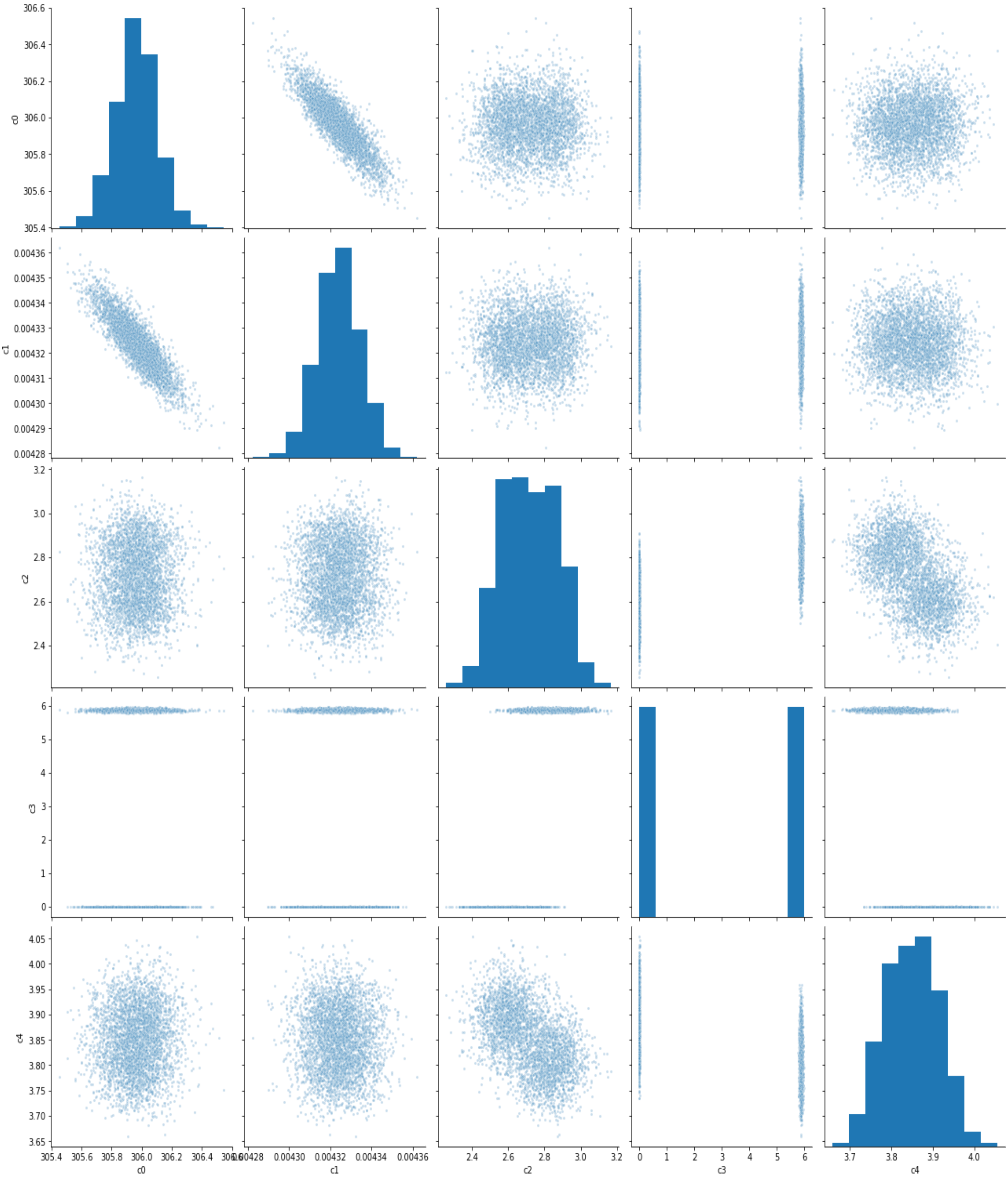







Autocorrelation of c3 samples
1.0
0.8
0.6
0.4
0.2
0.0
−20
−15
−10
−5
0
5
10
15
20

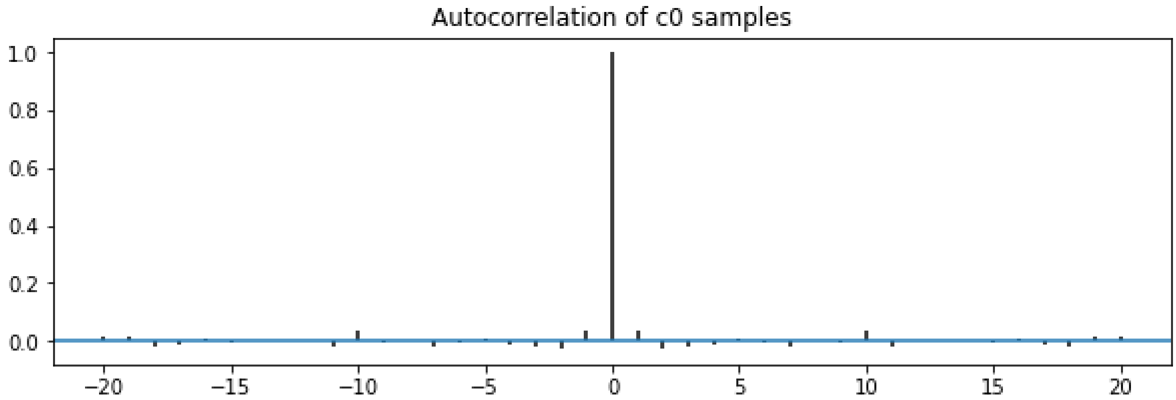

Autocorrelation of c4 samples
1.0
0.8
0.6
0.4
0.2
0.0
−20
−15
−10
−5
0
5
10
15
20

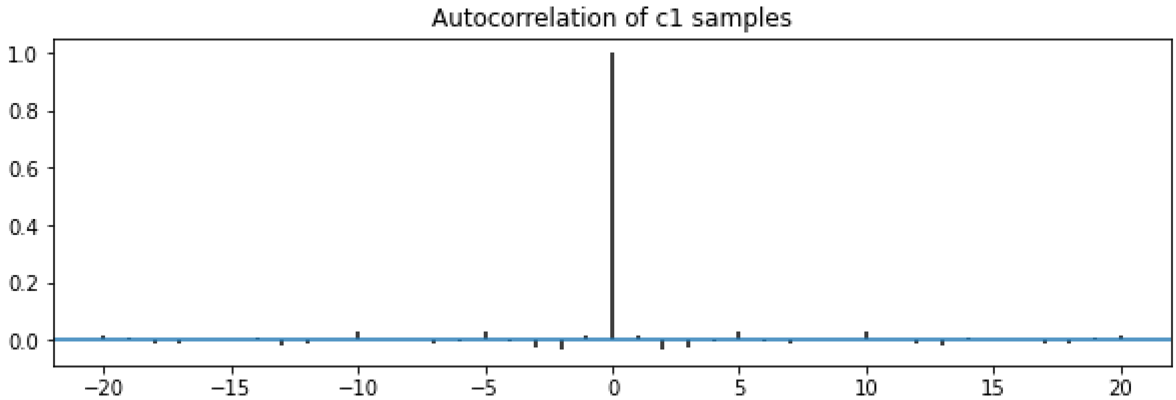


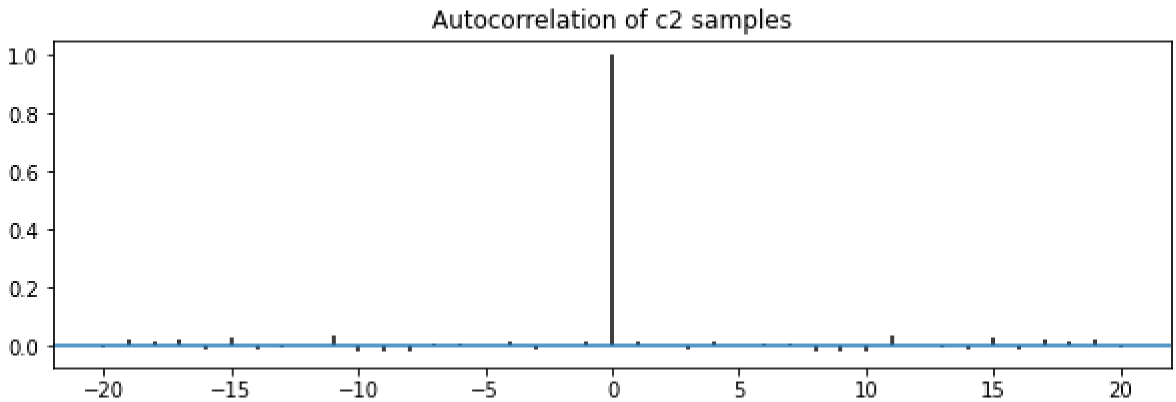
c0
c1
c2
c3
c4

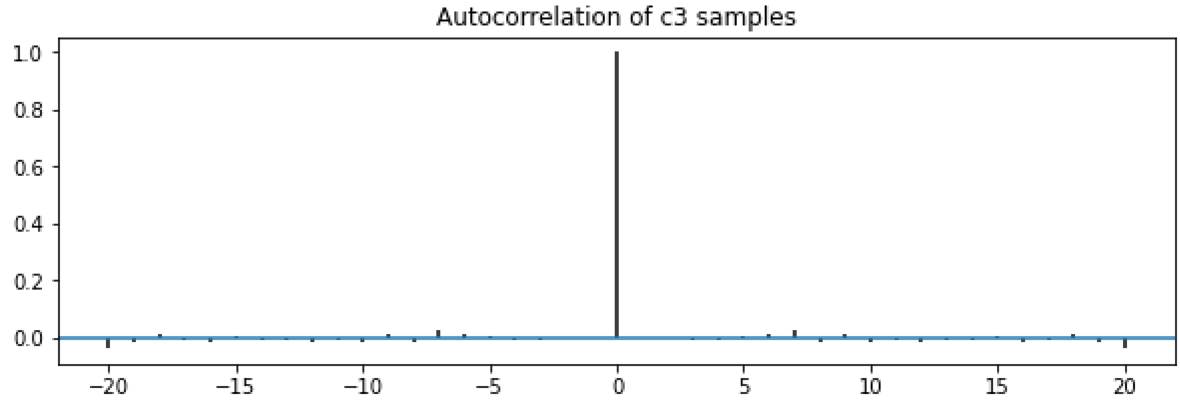
Appendix B

Modified Linear Model Autocorrelation plots and Pair plots

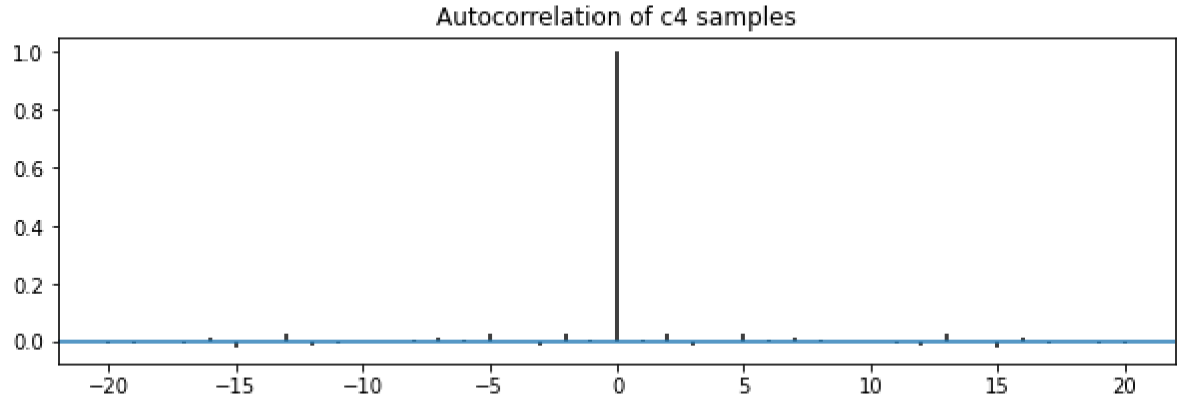




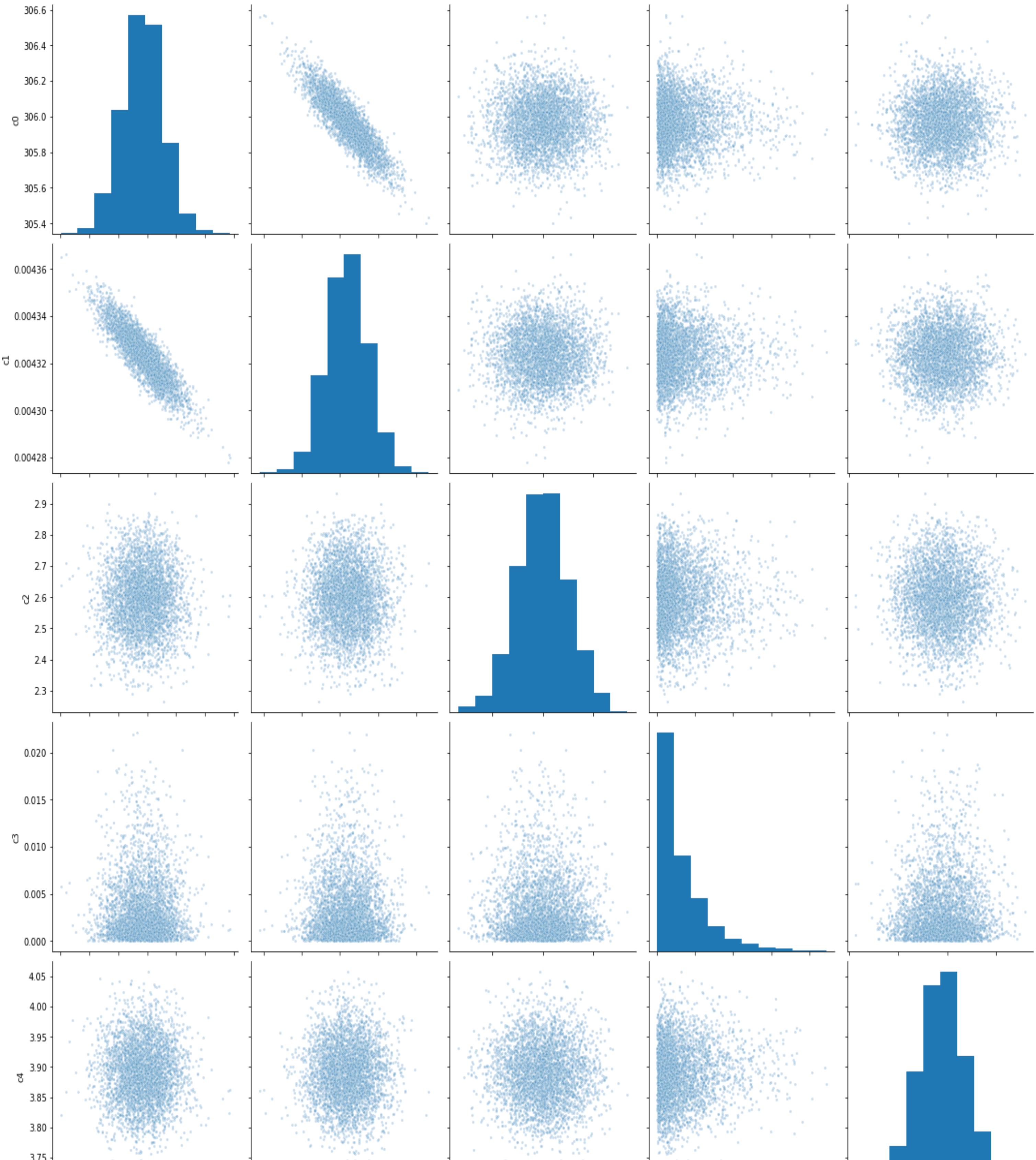




Autocorrelation of c3 samples
1.0
0.8
0.6
0.4
0.2
0.0
−20
−15
−10
−5
0
5
10
15
20

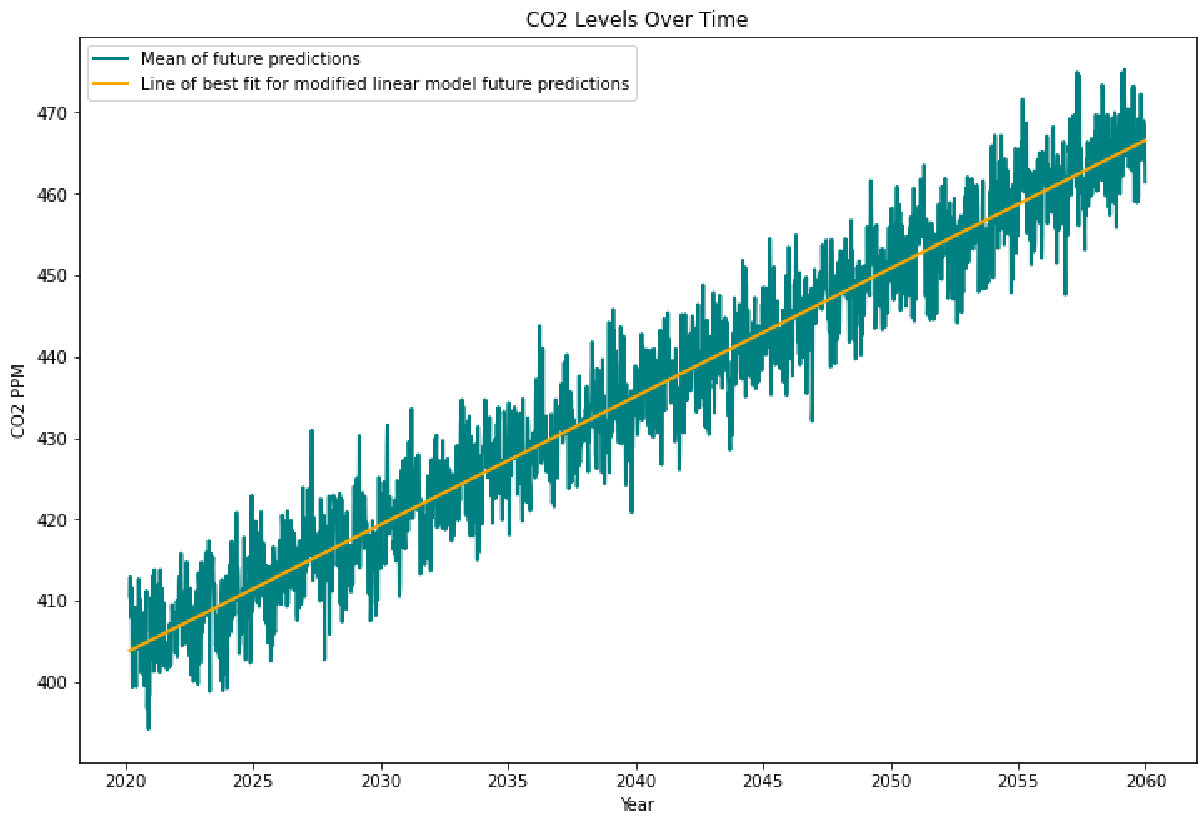


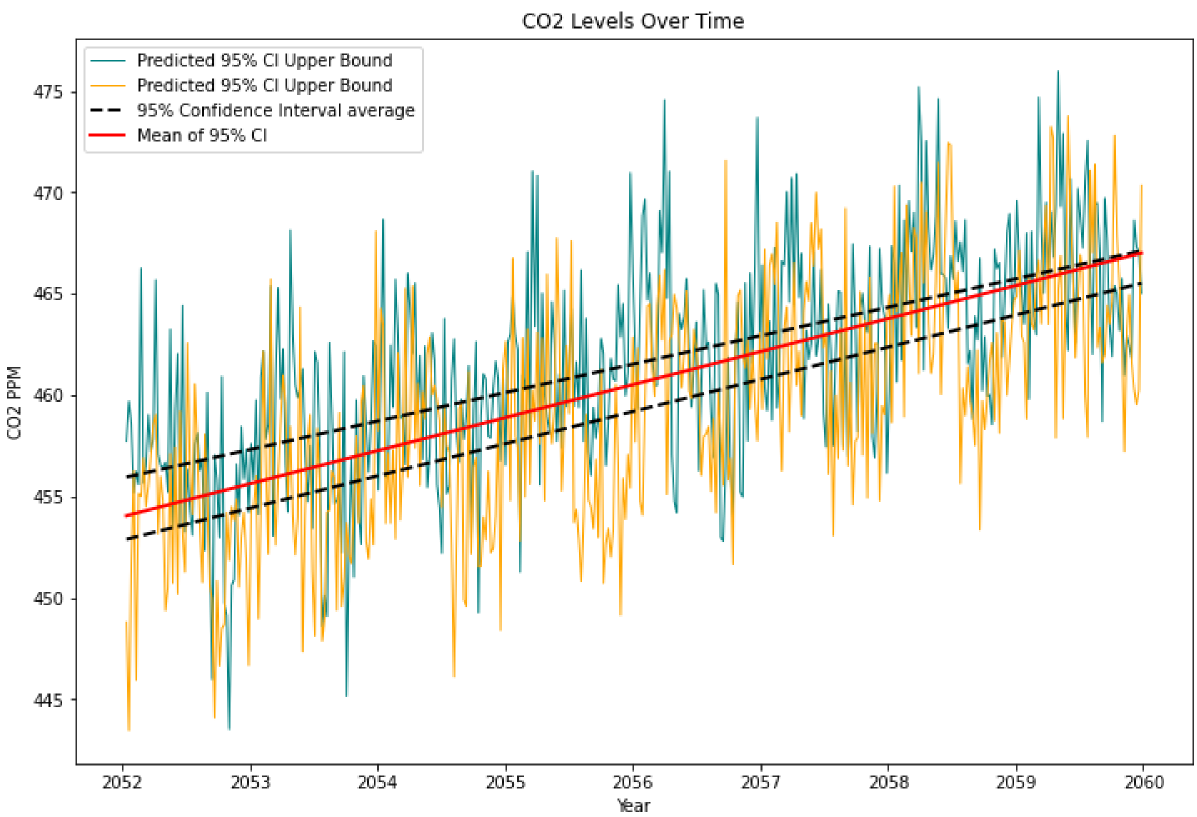
Autocorrelation of c4 samples
1.0
0.8
0.6
0.4
0.2
0.0
−20
−15
−10
−5
0
5
10
15
20


c0
c1
c2
c3
c4
c0
c1
c2
c3
c4

Appendix C

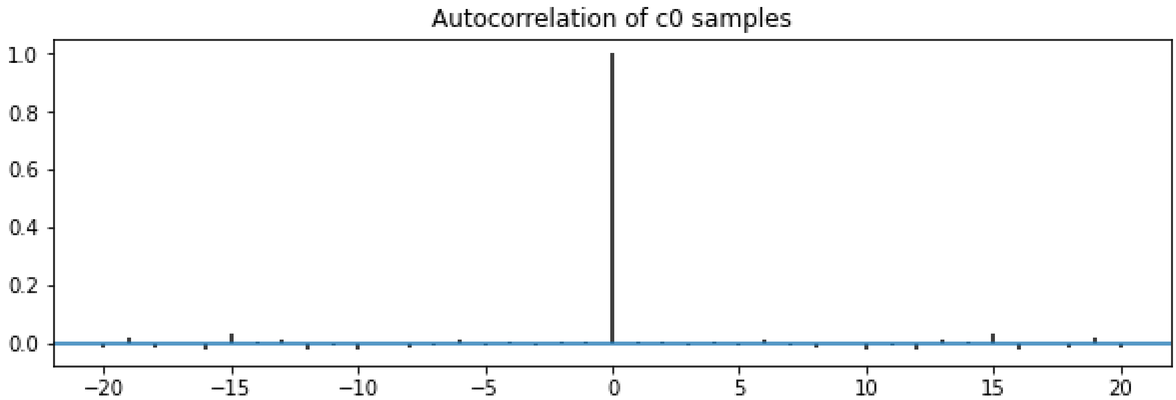
Future Predictions from Modified Linear Model

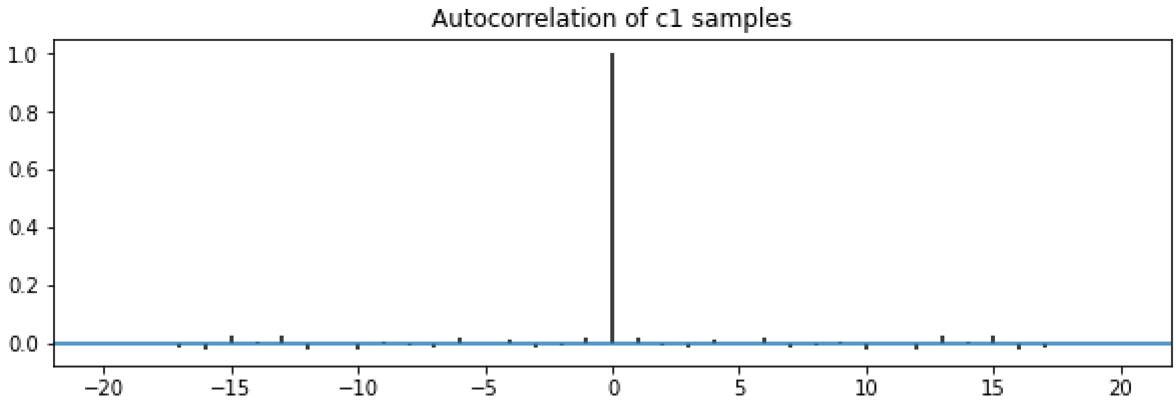


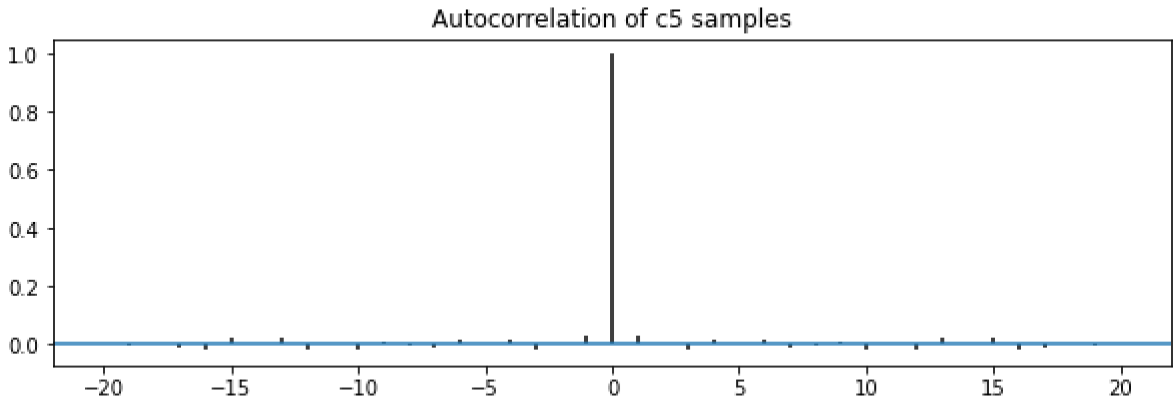


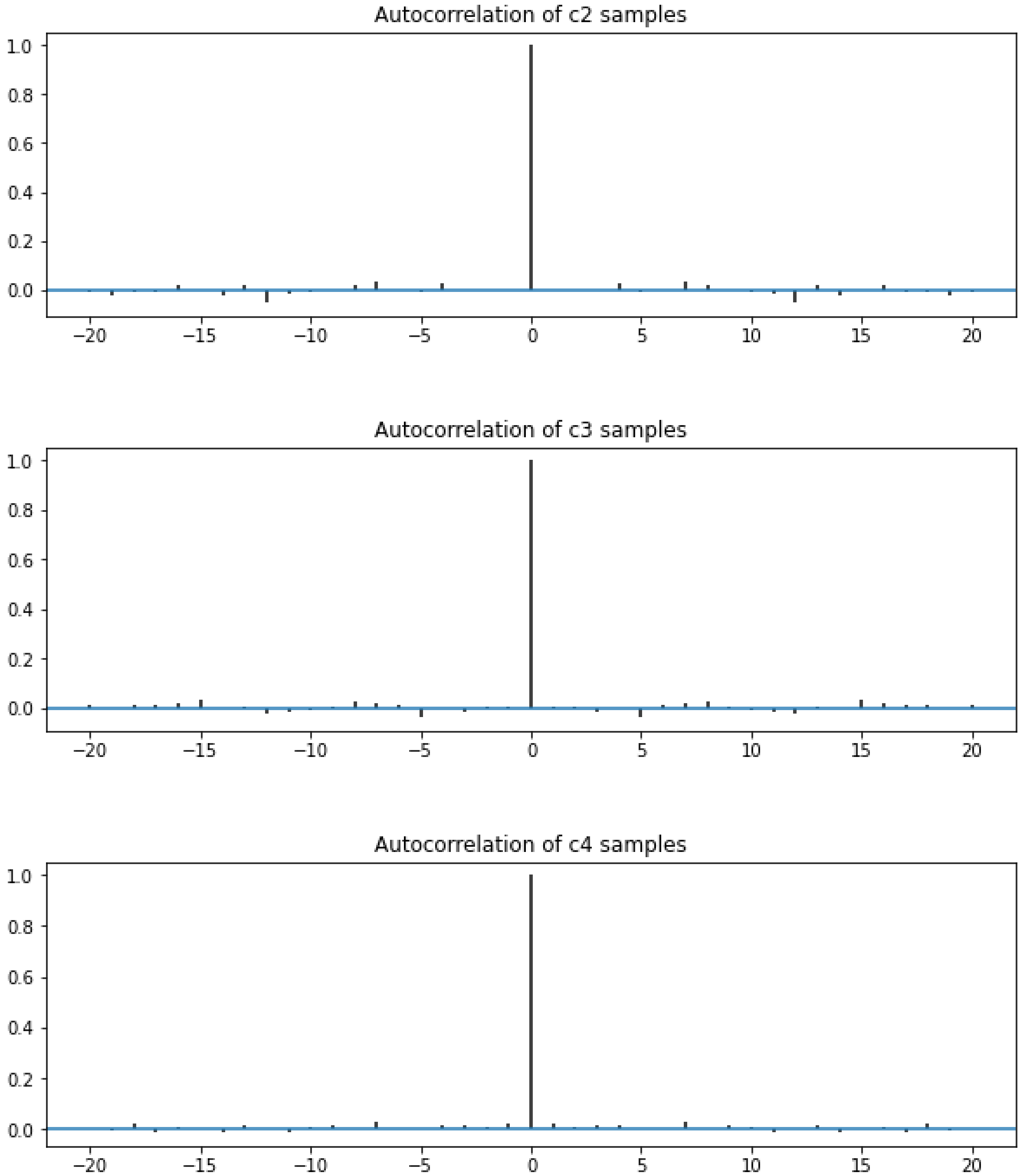
Appendix D

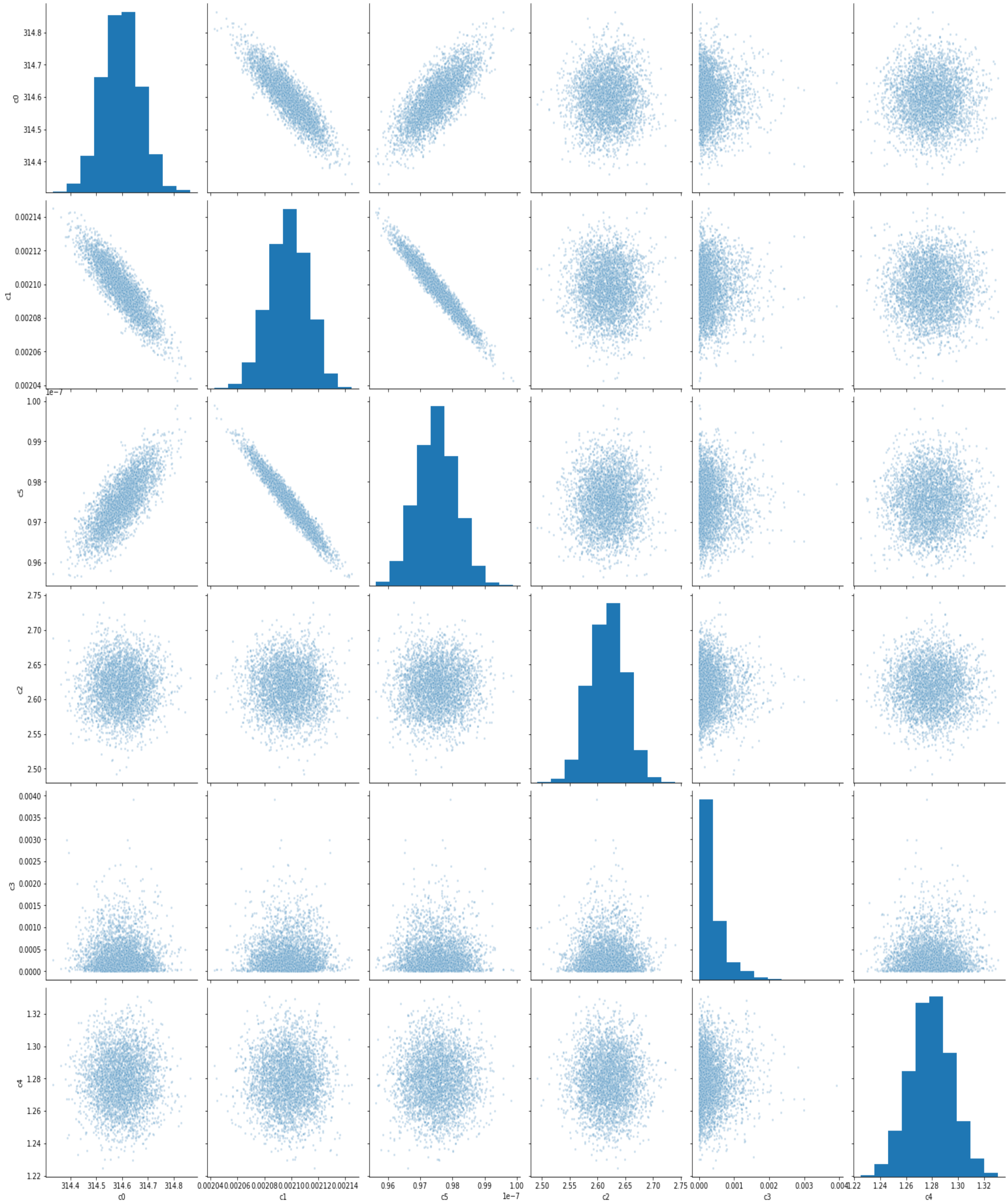
Quadratic Model Autocorrelation plots and Pair plots








Autocorrelation of c2 samples
Autocorrelation of c3 samples
Autocorrelation of c4 samples
1.0
0.8
0.6
0.4
0.2
0.0
−20
−15
−10
−5
0
5
10
15
20


c0
c1
c5
c2
c3
c4
1e-7