CS146: Final Project

CO2 Data Modeling and Prediction

Minerva Schools at KGI

CO2 Data Modeling and Prediction

Introduction

In this paper I will model the CO2 measurements data recorded since 1958 at the Muna Loa observatory and use that model to predict the future CO2 level up until 2060. I will also estimate when CO2 level will reach the alarming level of 450 ppm. A high level of CO2 is very dangerous for all the living things in earth and also for the ecological balance of the earth. If we can make a good prediction on how CO2 level can further rise in the upcoming years, we can provide more concrete evidence to the policy makers to take necessary steps to control the CO2 emission. The prediction is valid only under the assumption that the sources and sinks of CO2 doesn't change drastically in near future and CO2 increase continues to follow the same pattern (i.e. statistical model) as the previous recorded pattern (i.e. the model that we want to build here). In this paper, we compare between three different models and choose the one which gives us the lowest mean squared error when compared with the recorded data.

Data Preprocessing

I collect the data from the weekly data set from Mauna Loa Observatory (links added in appendix). The data starts from 29 March 1958. I transformed the dates into the number of year after the first observation. Figure 1 shows the full data and the close up view of the data since 2018. We can see that there is a salient overall trend of increasing the CO2 levels as the years go. This makes sense as we constantly using fossil fuels which throws a huge amount of CO2 constantly to the environment. We can also see an periodical change in CO2 emission in every year. It represents the fact that in winter more energy is consumed, thus more CO2 is emitted compared to summer, and in summer plant grows faster and consumes more CO2. From the

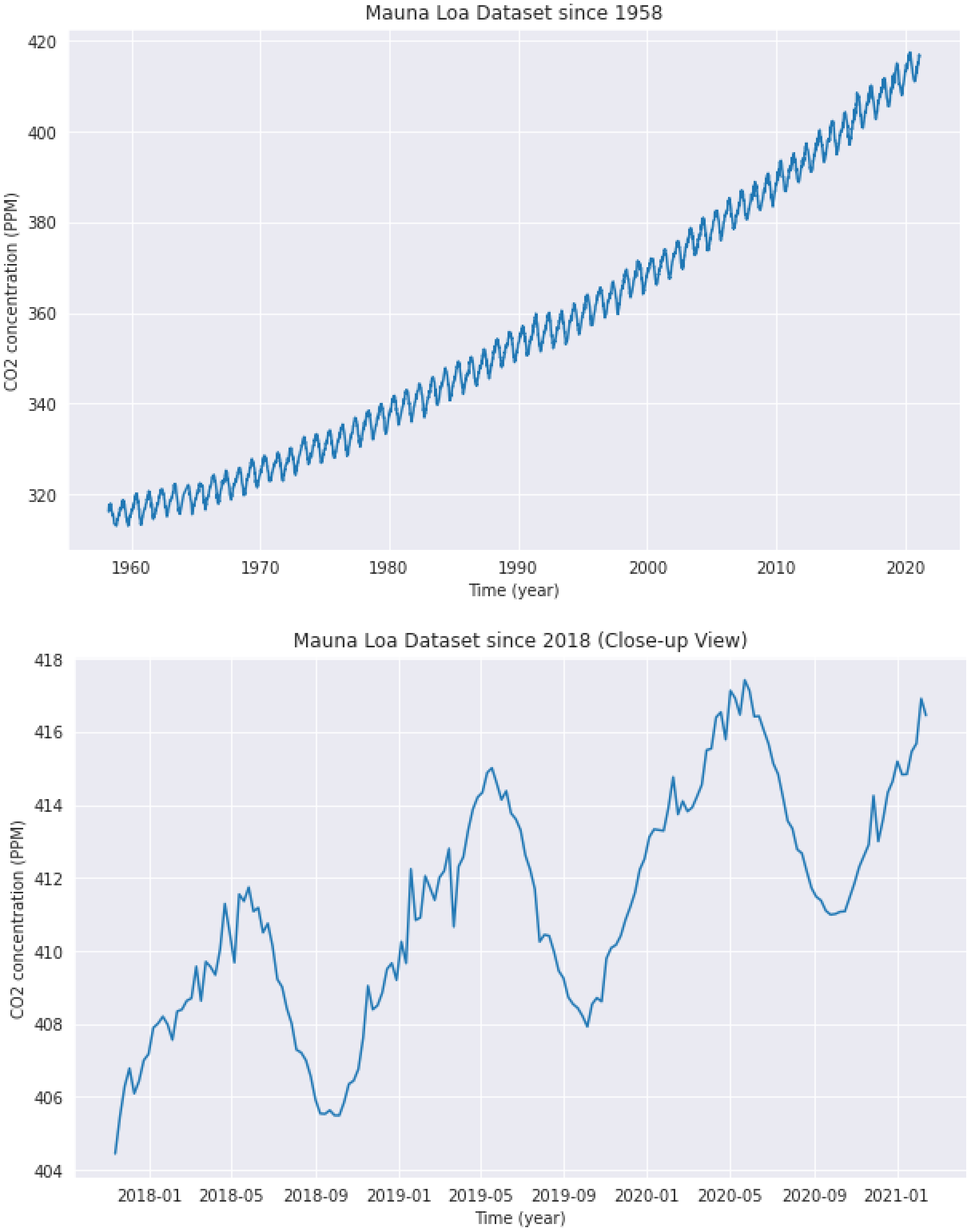


Figure 1: Overall increasing trend and sinusoidal seasonal variation in CO2 levels in atmosphere as measured in Mauna Loa Observatory since 1958.

close-up view, the seasonal variation looks like a relatively well fitted sinusoidal curve, so we can model this variation using a sine or a cosine function. The overall increasing trend can be modeled as a linear or high degree polynomial or an exponential function. Here we will compare three different model, where all of them models the seasonal variation as cosine or sine function. The first model is identical as proposed in the final project description: a linear overall trend. Other two models assume the overall trend as a quadratic and cubic function respectively.

Model 1: Linear

This model assumes a constant increase of CO2 over time. This could be plausible if human being's interference with the CO2 emission is very low and the major reason to increase CO2 is the natural dissipation. The long term trend is approximated as the following function: $c_1 + c_2 x_t$, where c1 is the intercept and c2 is the slope of the curve and the seasonal variation is approximated as a cosine function: $c_3 cos(2\pi x_t + c_4)$, where c3 is the amplitude of the sinusoidal curve and c4 is the phase shift at the initial time. I already changed the unit of time variable $x_t$as years after the first observation and the period of seasonal variation is 1 year, thus we don't need to divide $x_t$by anything inside the cosine. Along with the additional noise due to random fluctuation, the full model is the following likelihood function:

$$p(y_t \mid \theta) = N(c_1 + c_2 x_t + c_3 \cos(2\pi x_t + c_4),\ c_5^2)$$

Here $y_t$is the CO2 levels in ppm on day x_t. The x_t and $y_t$are already observed, but we need to estimate the unobserved parameters denoted as c_i. The prior of $c_i$variables are taken as follows. One thing to keep in mind that as we have more than 3000 data values, the likelihood will likely to dominant these chosen priors once we compute the posterior.

c1 (intercept): This is the y-intercept of the linear model (i.e. the CO2 levels at initial time (1958)). From the top figure of fig 1, it seems like the initial CO2 value is around 310 ppm. Depending on other parameters of the model, this value can change a little, but it is very unlikely that it will go below 270 or above 350. Thus I assumed that 95% quantile of the distribution stays in this range by taking a normal distribution with a mean of 310 ppm and an SD of 20 ppm.

c2 (slope): From figure 1, we can see that the CO2 level rises around 100-120 ppm in around 60 years. Thus the likely value of a slope in a linear mode is around 2 ppm per year. Still, to ensure enough flexibility, I took an SD of 0.5. As CO2 always increases the slope cannot be negative. So C2 was drawn from a truncated normal distribution of N(2, 0.5).

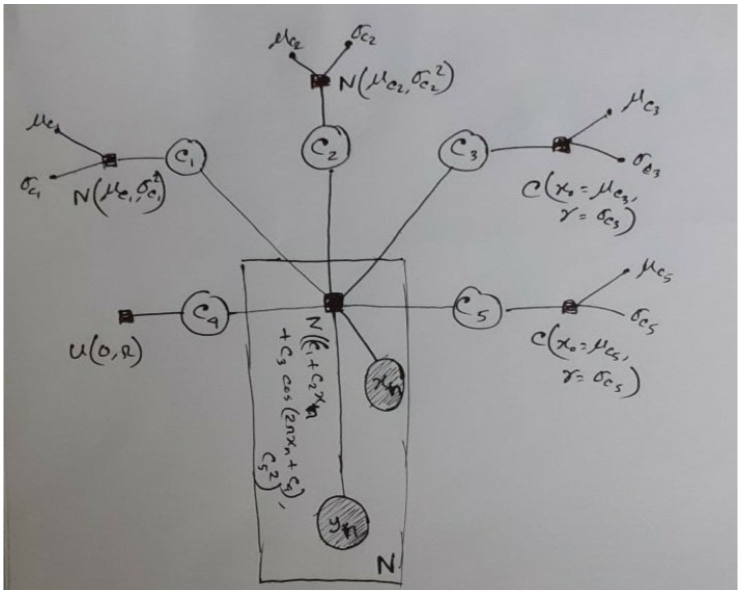
c3 (amplitude of the seasonal variation): From the bottom graph of Fig 1, the minimum to maximum range for a given period seems like around 10 ppm, thus an amplitude of 5 ppm. But this varies a lot for different periods, but it can never be less than 0. As there is a chance of having high values, c3 is drawn from a half-cauchy (i.e. truncated at 0) distribution with x0 = 5 and γ= 1.

c4 (phase shift of initial time): The possible value of the phase shift is from 0 to 2π. But both cosine and sine functions are symmetric, thus it is enough to have a value between 0 to π. As I do not have strong prior information, I drawn c4 from the uniform distribution: U(0, π).

c5 (random noise): From Fig 1, it seems like that the random noises are much smaller than the seasonal variation amplitude, c3. It is at least less than half of the amplitude, but as it is random fluctuations, there is a chance that we could get a high value. Thus I took c5 from a cauchy distribution with with x0 = 2 and γ= 1, as cauchy has high tails. But I truncated it at 0 (make it half-cauchy) as we know that the sigma cannot be negative.

The distribution of the priors are visualized in Appendix Fig 1. With all these prior distribution, we run Stan model. The output of the stan model shows large enough effective sample size (more than 3000) and Rhat values of close to 1 suggests a convergence. I plotted the autocorrelation function (ACF) plot which shows that samples for each of the variables are independent as the all the samples in lags and leads shows a correlation close to 0. Then in the pairplot (Appendix Fig 2), almost all the parameter samples shows gaussian random distribution as expected. Only c1 (intercept) and c2 (slope) shows a negative correlation. This makes sense because if the model takes a higher intercept value, then the slope of the model must be smaller to adjust with the data. Overall we can say that it is an effective stan model.

Now we can extract the parameter samples and derived the estimated co2 levels from these samples. In Fig 2, we compared the output of the linear model with the observed data. We can see that the model is not performing very well. The overall trend is not followed well by the linear model, the 95% confidence interval is very wide and the observed data is very far away from the estimated mean of the linar model. So linear model is not a good model in this problem.


μ_{c_2}
σ_{c_2}
$N(\mu_{c_2}, \sigma_{c_2}^2)$
μ_{c_1}
σ_{c_1}
$N(\mu_{c_1}, \sigma_{c_1}^2)$
c_1
c_2
c_3
μ_{c_3}
σ_{c_3}
$C(x_0 = \mu_{c_3}, \gamma = \sigma_{c_3})$
c_4
$U(0, \Omega)$
$N(c_1 + c_2 x_n + c_3 \cos(2\pi x_n + c_4), c_5^2)$
x_n
y_n
N
c_5
μ_{c_5}
σ_{c_5}
$C(x_0 = \mu_{c_5}, \gamma = \sigma_{c_5})$

Graphical Model 1: Linear Model

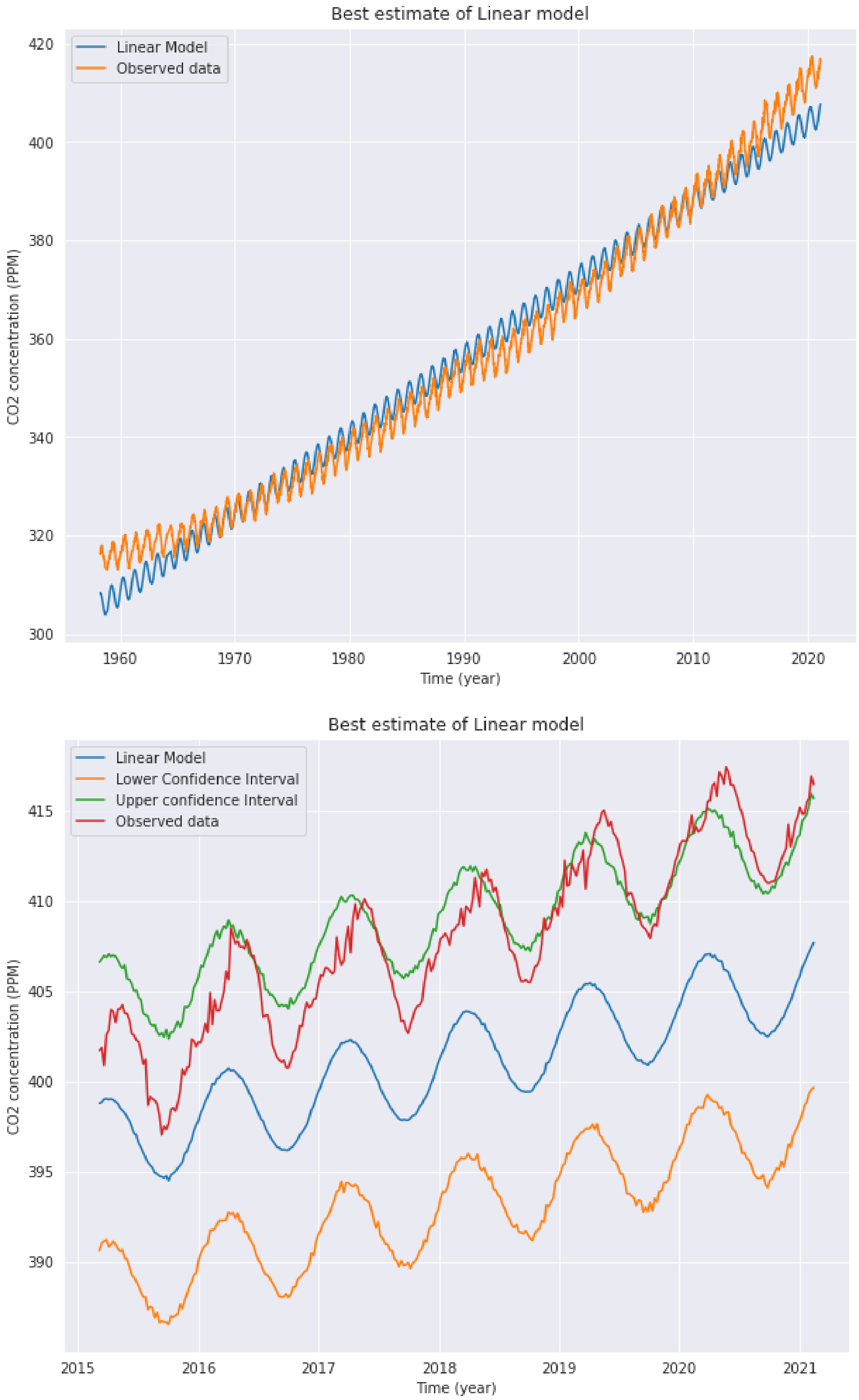


Figure 2: The best estimate of the linear model does not follow the overall trend of the observed data (top) and the 95% confidence interval is also very wide (bottom)

Model 2: Quadratic

Now we will model the overall trend using a quadratic function: $c_1 + c_2 x_t + c_3 x_t^2$. This will be reasonable if the CO2 is increasing with a constant acceleration. This can be the case as since the 18th century, we started to use fossil fuels in industrial level and the use of fossil fuels is increasing day by day, which can accelerate the CO2 rise. The likelihood of this model:

$$p(y_t \mid \theta) = N(c_1 + c_2 x_t + c_3 x_t^2 + c_4 \cos(2\pi\, x_t + c_5),\, c_6^2)$$

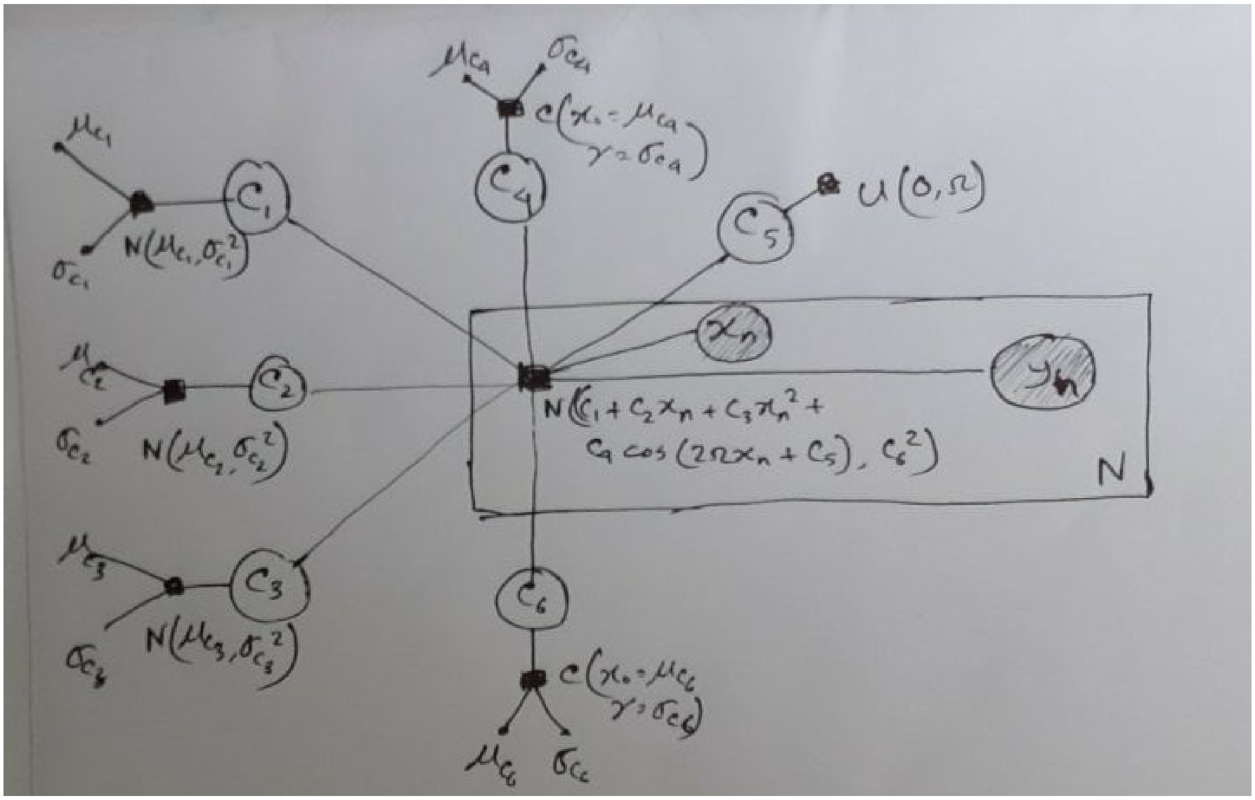
Now again x_t and y_t are observed variables and others are unobsevred. Most of the parameters are same as before and thus have the same prior distribution for similar reasoning as described above. Only c2 and c3 have a new prior distribution in this model:

c2 (slope): Last time, we have only the slope to capture the overall trend, thus we assume it to have a mean of around 2. But now both the first order and second order term collectively capture the trend. Thus the expected value of the slope should be smaller as it will capture only a portion of the trend and the remaining part will be captured by the second order term. So now c2 is drawn from a truncated normal distribution of N(1, 1).

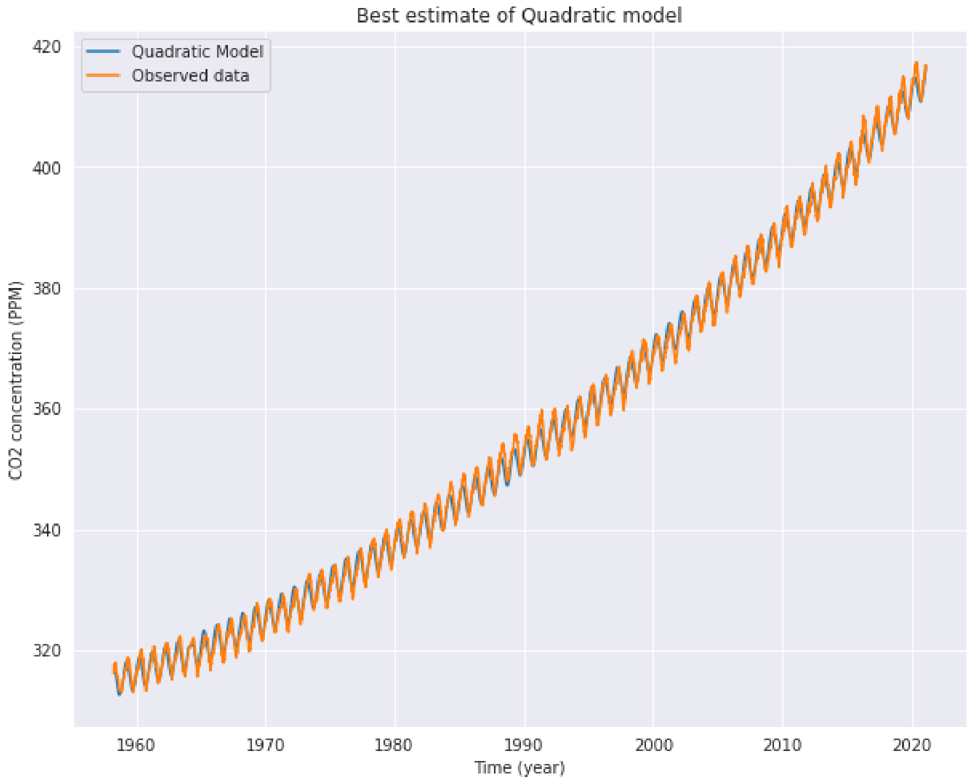
c3 (coefficient of the quadratic term): As we dont have any prior knowledge over it except that it cannot be very large. As an uninformative prior c3 is drawn from a truncated normal distribution of N(0.5, 0.5). These prior distributions are also added in Appendix Fig 3.

Now after running it through stan, we can again see that it converged as Rhat values are 1 and the effective sample sizes are also large enough (more than 1500) for all the parameters. The ACF plot shows that even the immediate lags and leads samples have a correlation of close to 0, suggesting independent samples for each of the parameters. The pair plot (Appendix Fig 4) again shows random distribution for most of the parameters. Again c1 (intercept) and c2 (slope) have a negative correlation (as we discussed above). For the same reason, c2 (slope) and c3 (coefficient of quadratic term) also have negative correlation, because if we have a lower slope, then there are more portion of the trend need to captured by the 2nd order term, thus its coefficient increases. So still we do not have much problem with the stan output.

Now we can extract the parameter samples and derived the estimated CO2 levels from these samples. In Fig 3, we compared the output of the quadratic model with the observed data. We can see that the model captures both the overall trend and seasonal variance very well. The 95% confidence interval is also comparatively narrow and the observed data mostly lies between the intervals. So quadratic model seems to be a plausible model in this problem.


μ_{c_4}
σ_{c_4}
$C(x_0 = \mu_{c_4}, \gamma = \sigma_{c_4})$
C_4
μ_{c_1}
C_1
$N(\mu_{c_1}, \sigma_{c_1}^2)$
σ_{c_1}
C_5
$U(0, \Omega)$
x_n
y_n
μ_{c_2}
C_2
σ_{c_2}
$N(\mu_{c_2}, \sigma_{c_2}^2)$
$N(C_1 + C_2 x_n + C_3 x_n^2 + C_4 \cos(2\Omega x_n + C_5), C_6^2)$
N
μ_{c_3}
C_3
σ_{c_3}
$N(\mu_{c_3}, \sigma_{c_3}^2)$
C_6
$C(x_0 = \mu_{c_6}, \gamma = \sigma_{c_6})$
μ_{c_6}
σ_{c_6}

Graphical Model 2: Quadratic Model


Best estimate of Quadratic model
Quadratic Model
Observed data
CO2 concentration (PPM)
420
400
380
360
340
320
1960
1970
1980
1990
2000
2010
2020
Time (year)

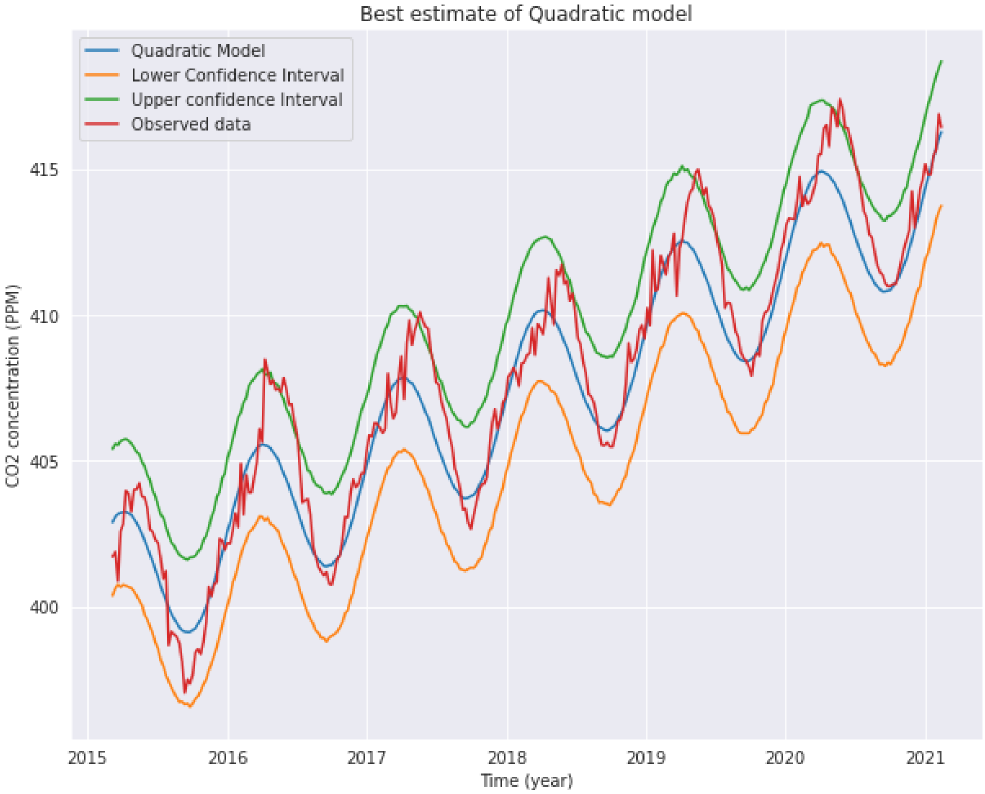

Best estimate of Quadratic model
Quadratic Model
Lower Confidence Interval
Upper confidence Interval
Observed data
CO2 concentration (PPM)
415
410
405
400
2015
2016
2017
2018
2019
2020
2021
Time (year)

Figure 3: The quadratic model follows the observed data very well, have a narrower confidence interval and observed data lies into the interval

Model 3: Cubic

Lastly we make a cubic model where the overall trend is captured in a cubic function: $c_1 + c_2x_t + c_3x_t^2 + c_4x_t^3$. This is the case when the acceleration of the CO2 rise is also increasing. This can happen if we donot control the industrial emission of CO2, rather continue to build upon that. The overall likelihood model is the following:

$$p(y_t \mid \theta) = N(c_1 + c_2x_t + c_3x_t^2 + c_4x_t^3 + c_5 \cos(2\pi\, x_t + c_6),\ c_7^2)$$

All the prior distributions are same as the last model. The new parameter c3 (coefficient of the quadratic term) is also drawn from an uninformative truncated normal distribution of N(0.5,0.5) (Appendix Fig 5).

Again the stan output seems reasonable as we have Rhat of 1 (suggesting convergence) and effective sample sizes of more than 1000. The ACF plots are similar as before, showing no correlation and thus independent samples. Pair plot shows similar behaviour as before: random distribution for most pair of samples, but correlation for the pairs of the intercept and the coefficient of the linear model, which was expected.

Fig 4 compared the observed data with the cubic model estimate with 95% confidence interval. Though the overall trend is captured quite well in this model, the seasonal variance is not very well captured. The 95% confidence interval is relatively narrow, but most of the time the observed data goes outside of this interval, suggesting that there needs to more improvement.

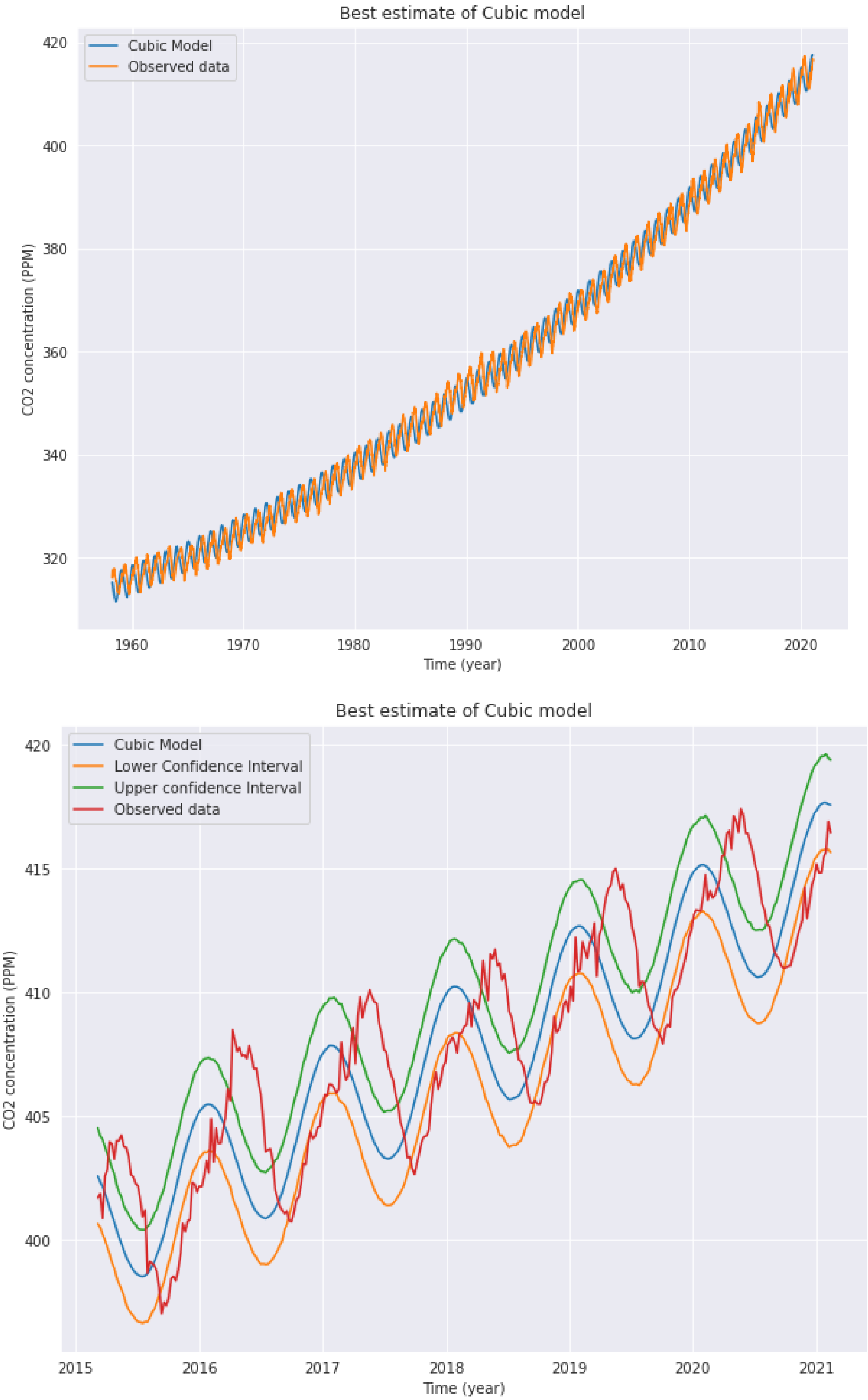


Figure 3: The cubic model follows the observed data somewhat well and have a narrower confidence interval but the observed data doesnot fit into the interval

Model Comparison and Selection

From the visual analysis, we can see that quadratic model works best among these three. But to compare the model quantitatively, we will compute the mean squared error (MSE) for each of the model compared to the observed data. To capture the variability of the model, we will not just take the MSE between the mean prediction of the model and the observed data. Instead, for each model, we will derived the average MSE of each of the 4000 sample prediction from the model compare to the observed data.

The average MSE for linear, quadratic and cubic model is respectively 33.19, 3.24 and 10.09. As the MSE is minimum for the quadratic model, we will use it for further inference and prediction.

Inference

Before prediction, we will look again into the output of quadratic model and try to compare the posterior distribution of the parameters (Fig 5) with their prior distribution (Appendix Fig 3). As their prior and posterior range was very different, we could not draw them together in single plots. Comparing both picture, we can see that the standard deviation becomes much narrower for all the parameters. The biggest difference happened for c5 (phase shift), where we didnot have any knowledge, thus took a large uniform prior, it now shows a very narrow exponential distribution with a mean very close to 0. The mean for c4 (magnitude) and c6 (random noise) shifted to the left and the prior cauchy distribution now results in a narrow normal posterior distribution. For c1 the posterior mean is close to our assumed prior mean, and the c2 (slope) and c3 (coefficient of quadratic term) decreases than the assumed uninformative mean, though the SD becomes much more narrower for all.

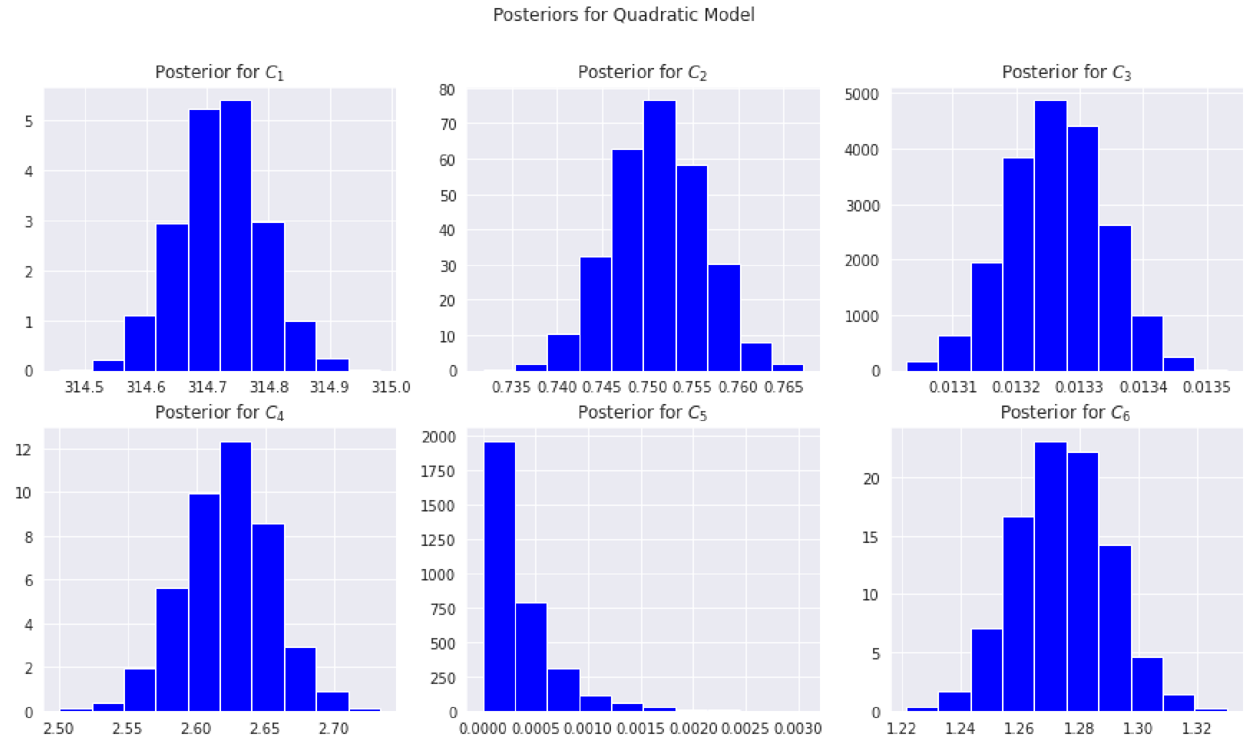


Figure 5: The posterior distribution of the parameters in the quadratic model shows narrow normal distribution, while the phase shift shows an exponential distribution.

Prediction:

Using the quadratic model, we predicted the future CO2 with 95% confidence interval level if it continues to increase similarly (Fig 6). We can see that as a quadratic model, the rise becomes more steeper as time grows. Where it takes around 60 years to rise 100 ppm from 1960-2020, it will likely to rise another 100 ppm (to reach around 520 ppm) in the next 40 years. The curve will continue to be more steeper if this trend continues, which could be a terrible news for ourselves. The seasonal variance will continue to exist with similar period (1 year) and amplitude (around 2.6 ppm with a confidence interval of [2.56-2.68 ppm]) according to the model.

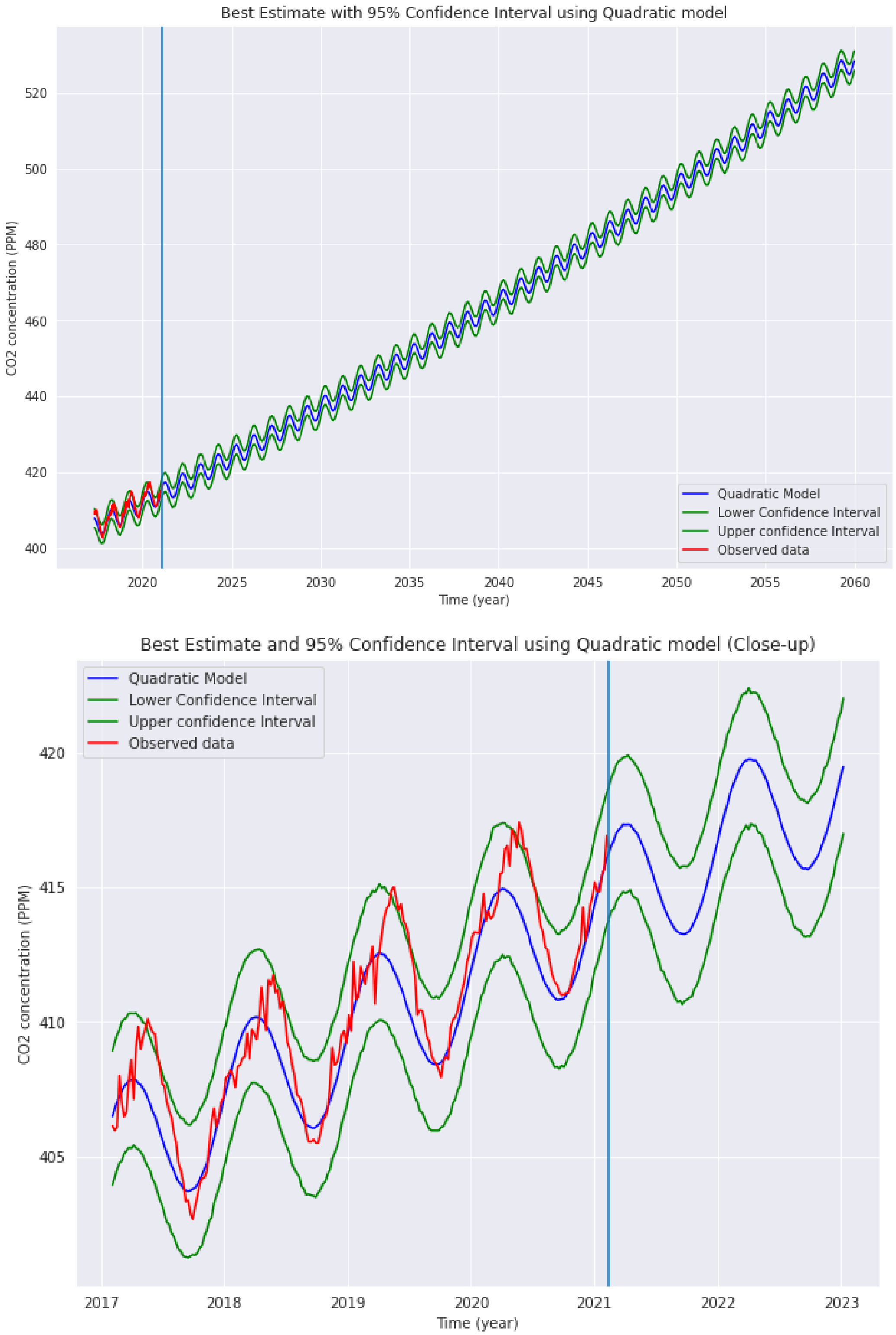


Figure 6: The prediction for the CO2 levels up until 2060 with 95% confidence intervals.

Reaching 450 ppm

According to our model, we can estimate that the mean estimate will likely to reach 450 ppm at around Feb 2034 (Fig 7). But using the 95% confidence interval of our model, we can say that there are high probability that CO2 can reach 450 in anytime between Feb 2033 to Feb 2035. In 2033, the upper limit of our confidence interval reaches to 450 and at 2035, the lower limit touches 450 ppm. Also as we can see due to seasonal variation, after reaching to 450 ppm, the CO2 will again go down for some time. But after 2036, none of the 95% confidence interval goes under 450 ppm, suggesting that there is very small chance that CO2 will go down under 450 ppm anytime after 2036.

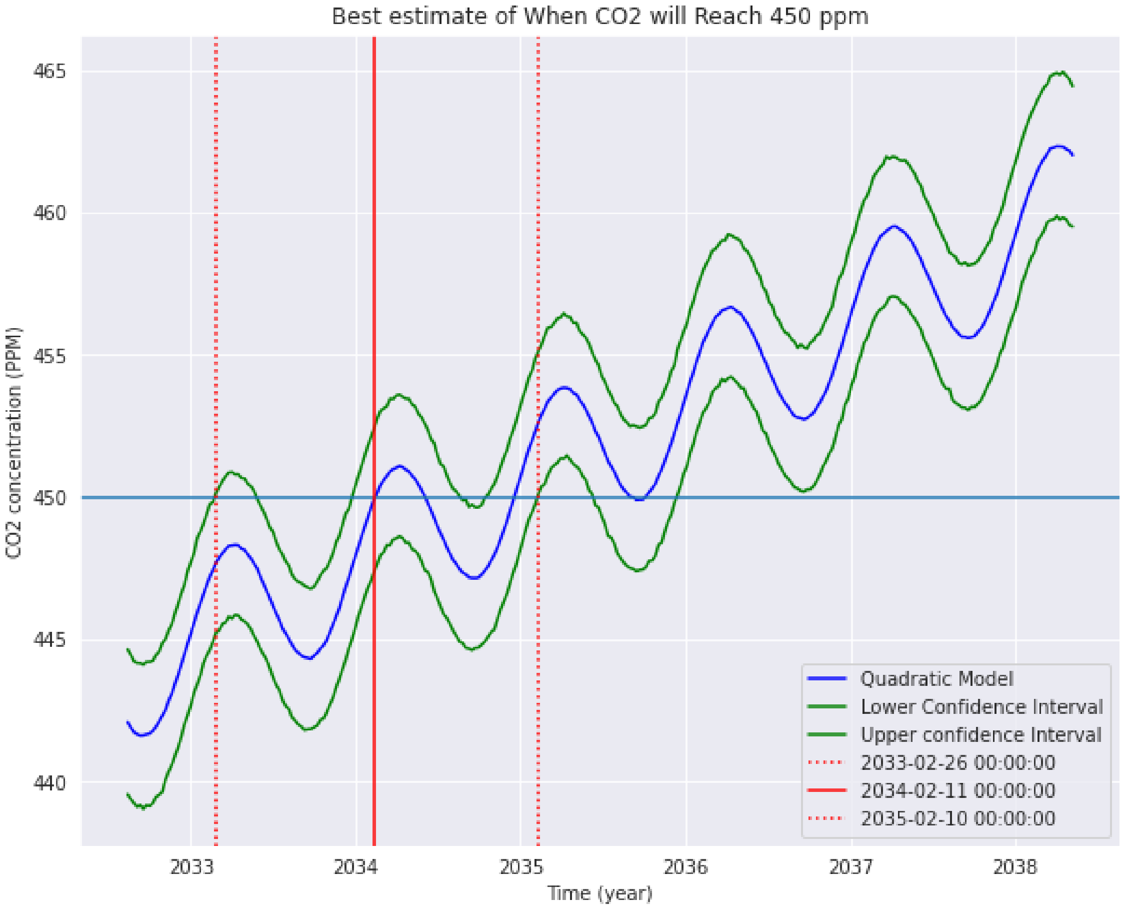


Figure 7: Best Estimate of CO2 reaching 450 ppm is on the year 2034 where the 95% interval reaches in between 2033 to 2035

Posterior Predictive Check:

We checked the mean and standard deviation of the data with that of the sample outputs of our model and the test statistics shows a p-value close to 0.5, which suggests that the model is compatible with the data (Fig 8). But then we also checked two special test statistics:

1. The average difference of CO2 values between two dates which are 52 datapoints (each datapoint is one week away, thus one year approx.) away.
2. The fraction of time, the CO2 values increases after 52 weeks (same day, next year).

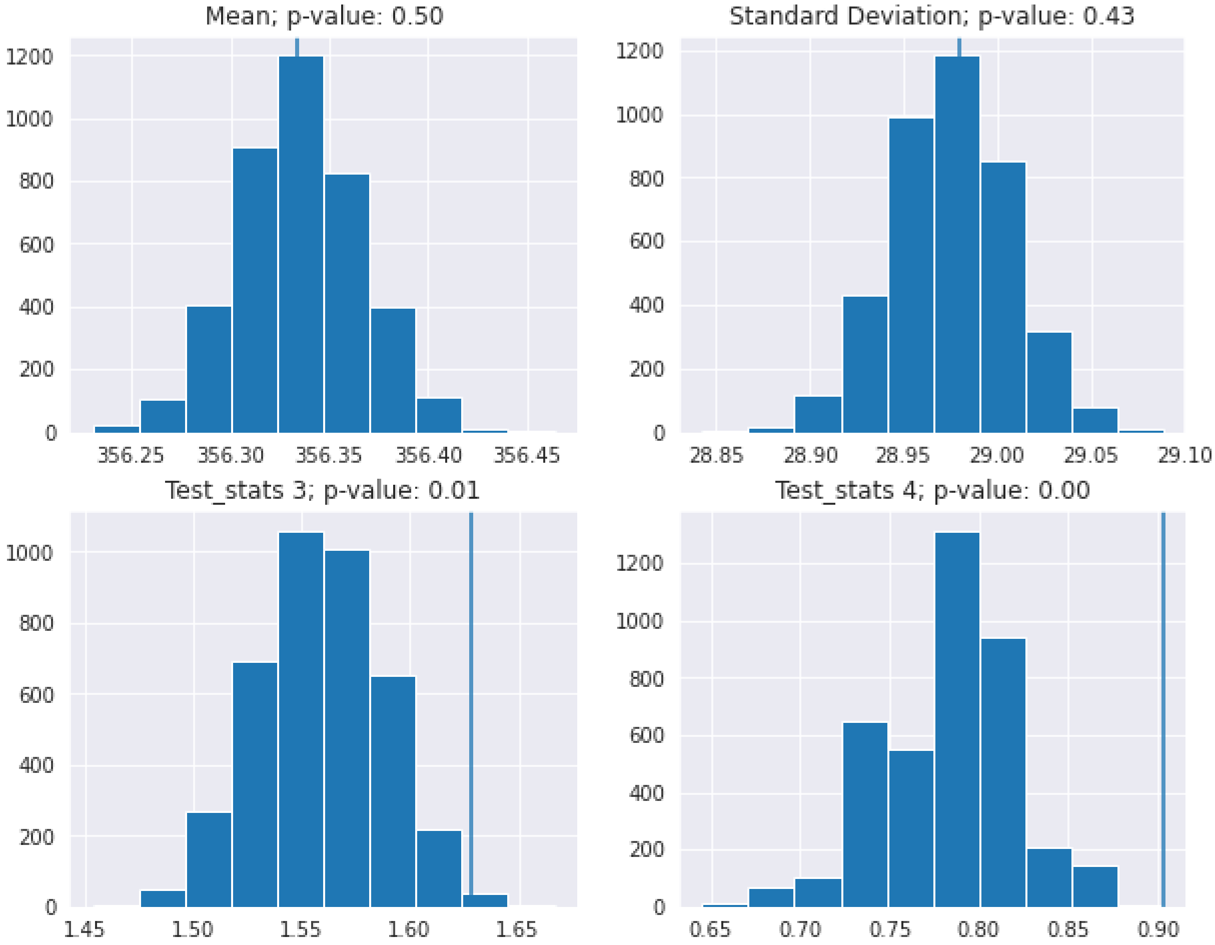


Figure 8: Posterior Predictive Check shows good p value for two statistics but bad p-value for other two test statistics.

For these two statistics, our model works very poorly with a p-value close to 0, which suggests that there are more room to improve the model and possibly there exists some other model which can represent the situation better.

Appendix

Dataset Link:

https://scrippsco2.ucsd.edu/assets/data/atmospheric/stations/in_situ_co2/weekly/weekly_in_situ_co2_mlo.csv

Code Notebook Link:

https://colab.research.google.com/gist/mahmud-nobe/60460ba89113c781dcff936096ef95a2/cs146_final_project.ipynb

Additional Figures:

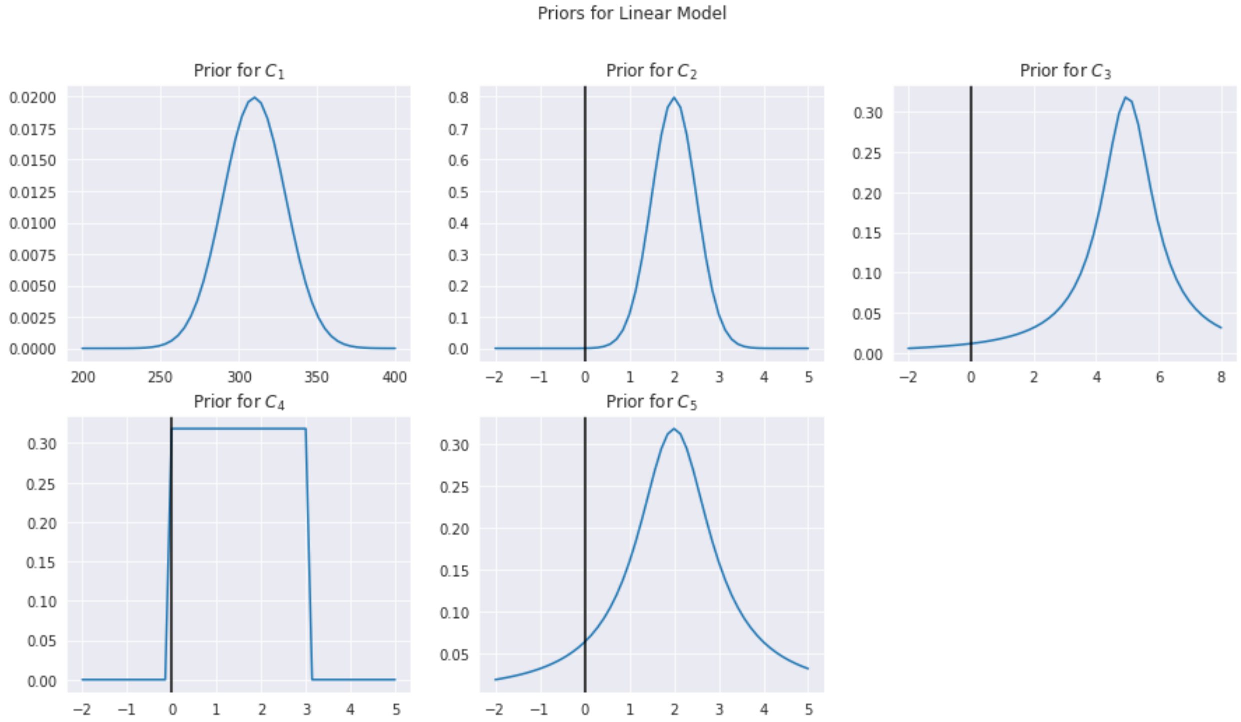


Figure 1: Prior distribution for linear model

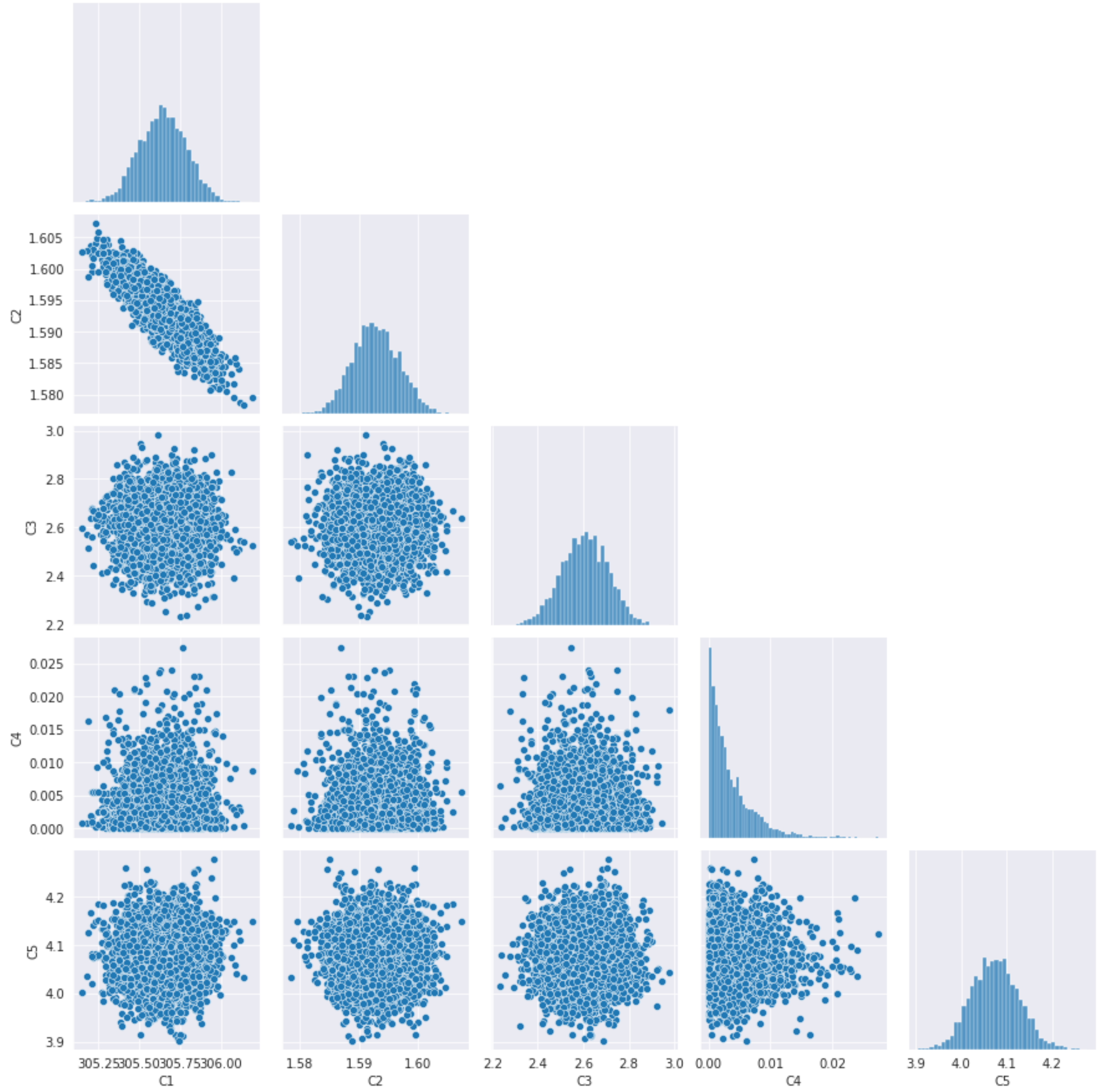


Figure 2: Pair Plot for Linear Model

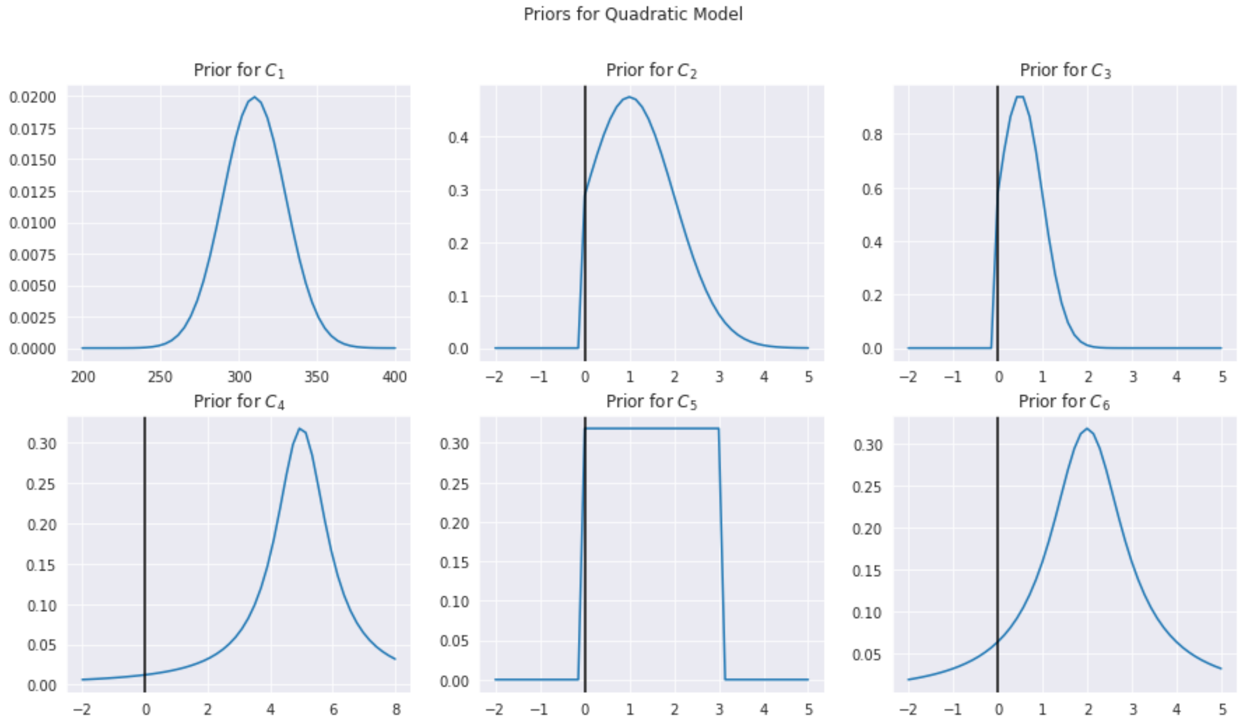


Figure 3: Prior distribution for quadratic model

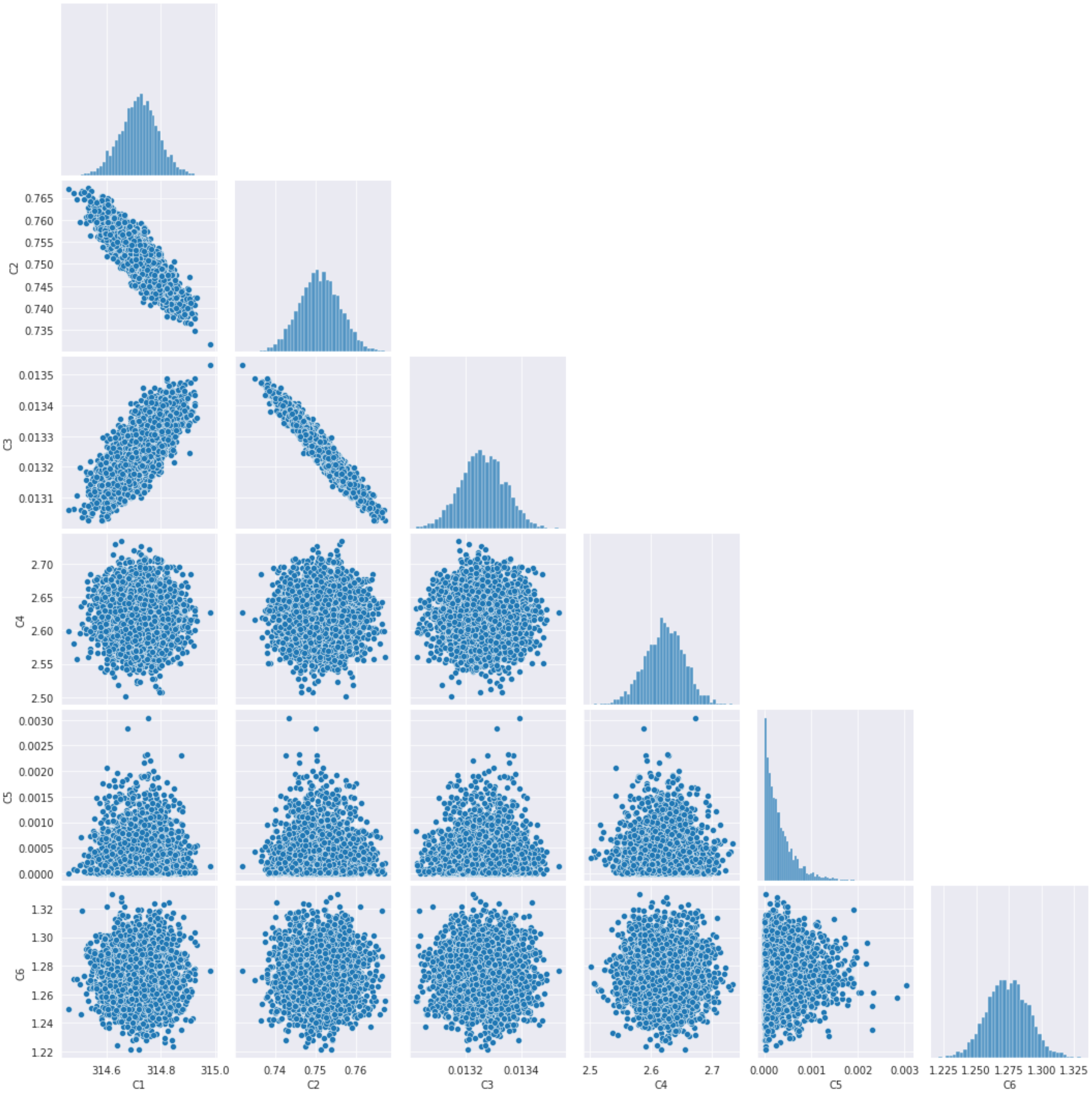


Figure 4: Pair Plot for Quadratic Model

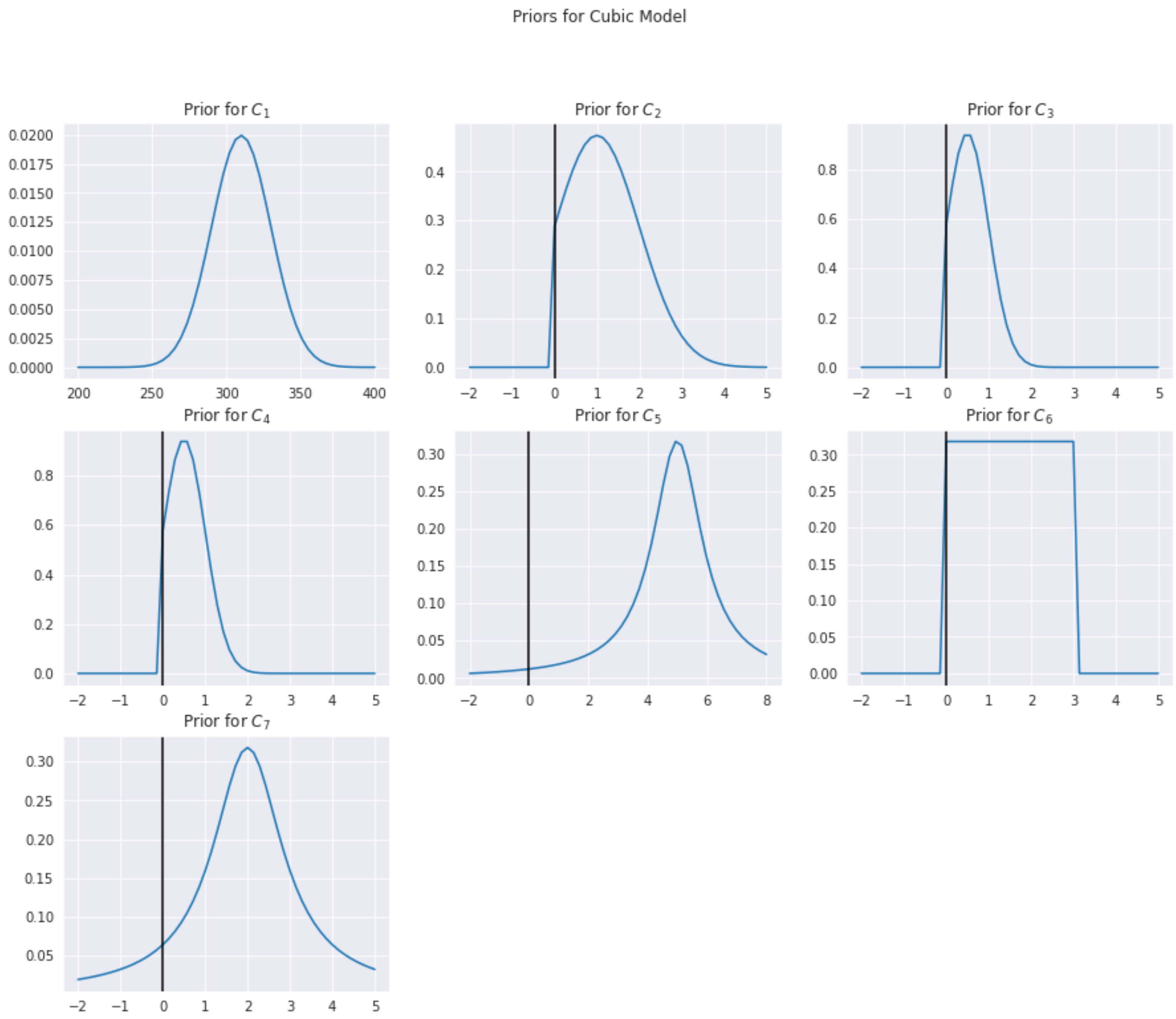


Figure 5: Prior distribution for cubic model

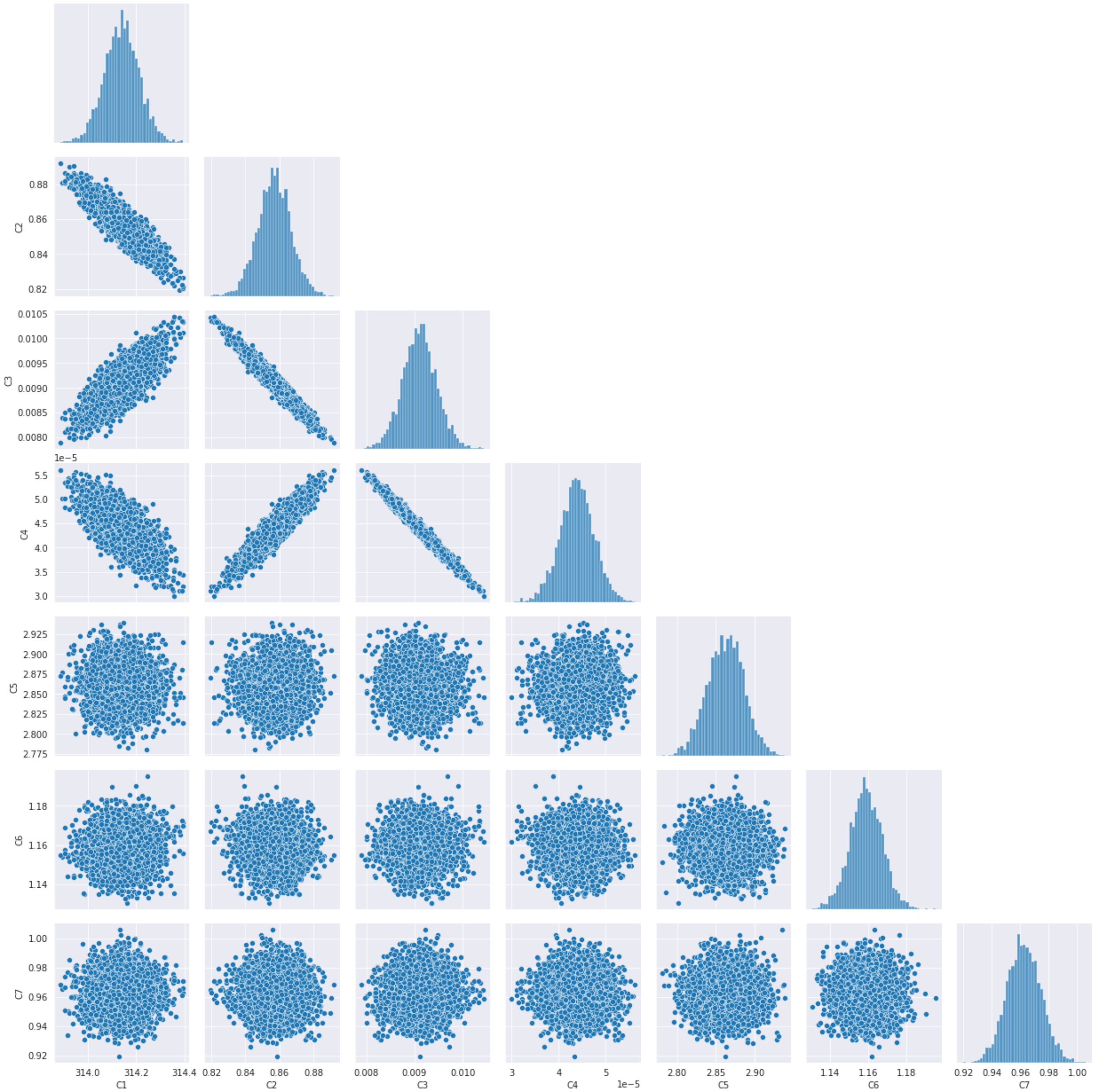


Figure 6: Pair Plot for Cubic Model

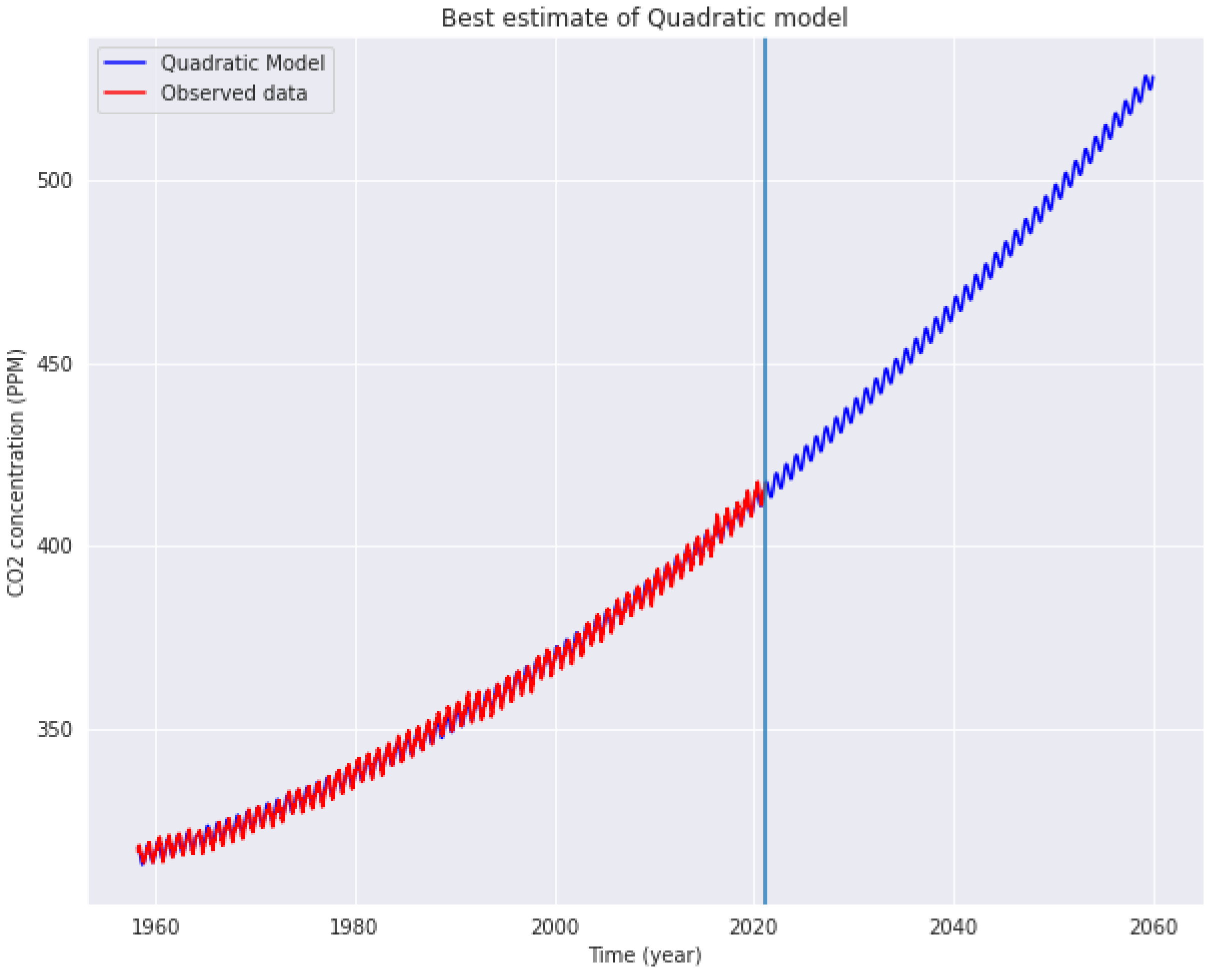


Figure 7: Observed data and Prediction of the future