FINAL PROJECT

CS146: Modern Computational Statistics
Minerva Schools at KGI

Khanh Nguyen
December 2020

FINAL PROJECT

1. Introduction

In this project, we will work with the CO2 dataset provided by the Mauna Loa Observatory in Hawaii. CO2 level in the atmosphere has been increasing steadily since the industrial revolution at an alarming rate, which is the main cause of global warming. We will try to use statistical models to explain the CO2 dataset, as well as generating predictions for the rise of future CO2 level until 2060. We will discuss what the results entail, address the drawbacks of such methods and how we can improve the model in the future.

2. The scenario and dataset

Data description

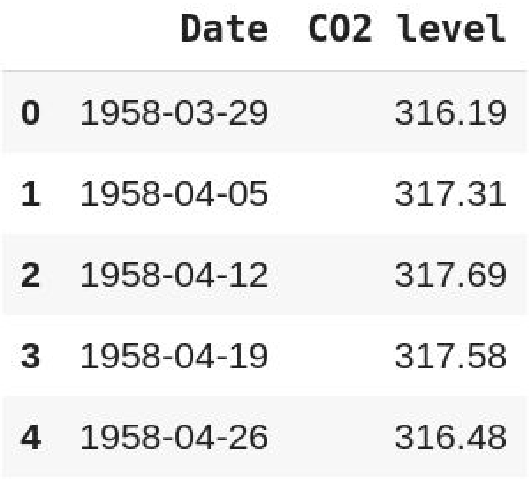
The dataset is the weekly measurement of CO2 level in the atmosphere from March 29, 1958 to November 28, 2020 recorded at the Mauna Loa Observatory in Hawaii. There are 3,199 observations in the dataset, each with the CO2 level (ppm) and the date of the measurement. The dataset can be downloaded from the Scripps CO2 program, and also attached in the Appendix A below

We are interested in explaining the relationship between each weekly CO2 level measurement by statistical method. Therefore, in this project we will employ different statistical models to find the best fitted model. In order to do so, we assume that the weekly CO2 levels are related statistically.

Once we find the best fitted model, we will use this model to predict the future trend of the CO2. The higher the CO2 level is, the more dangerous it is to the environment. Therefore, we will try to predict how high the CO2 level is by 2060, which is 40 years from now. In addition, CO2 level of over 450 ppm is considered dangerous for climate change, so we will also estimate when we will reach that threshold in the future.

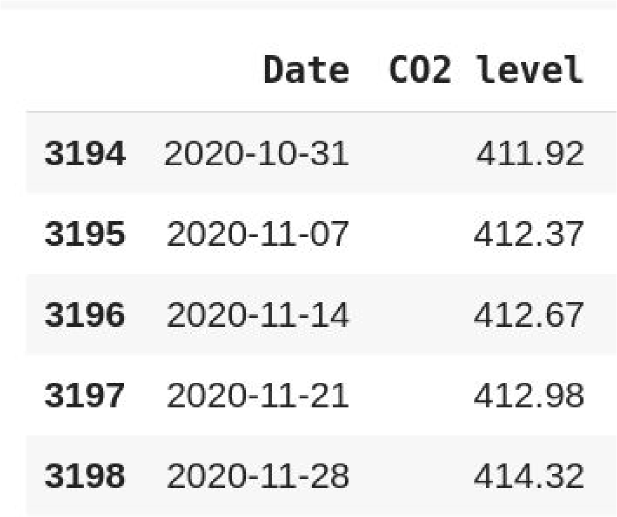
Data exploration

There are 3,199 observations in the dataset, each with the CO2 level (ppm) and the date of the measurement. A sneak peek of the first 5 rows and last 5 rows of the dataset is as below:



	Date	CO2 level
0	1958-03-29	316.19
1	1958-04-05	317.31
2	1958-04-12	317.69
3	1958-04-19	317.58
4	1958-04-26	316.48

Figure 1: *The first 5 observations of the CO2 dataset*



	Date	CO2 level
3194	2020-10-31	411.92
3195	2020-11-07	412.37
3196	2020-11-14	412.67
3197	2020-11-21	412.98
3198	2020-11-28	414.32

Figure 2: *The last 5 observations of the CO2 dataset*

I also plot the trend of the CO2 level as below:

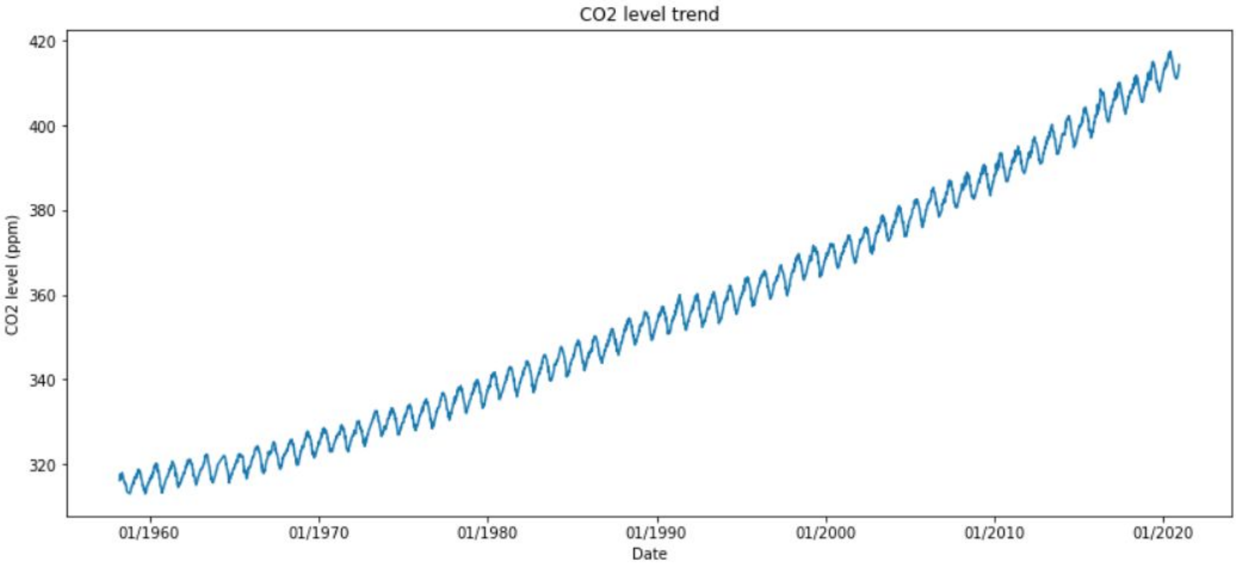


Figure 3: *CO2 level trend from March 29, 1958 to November 28, 2020*

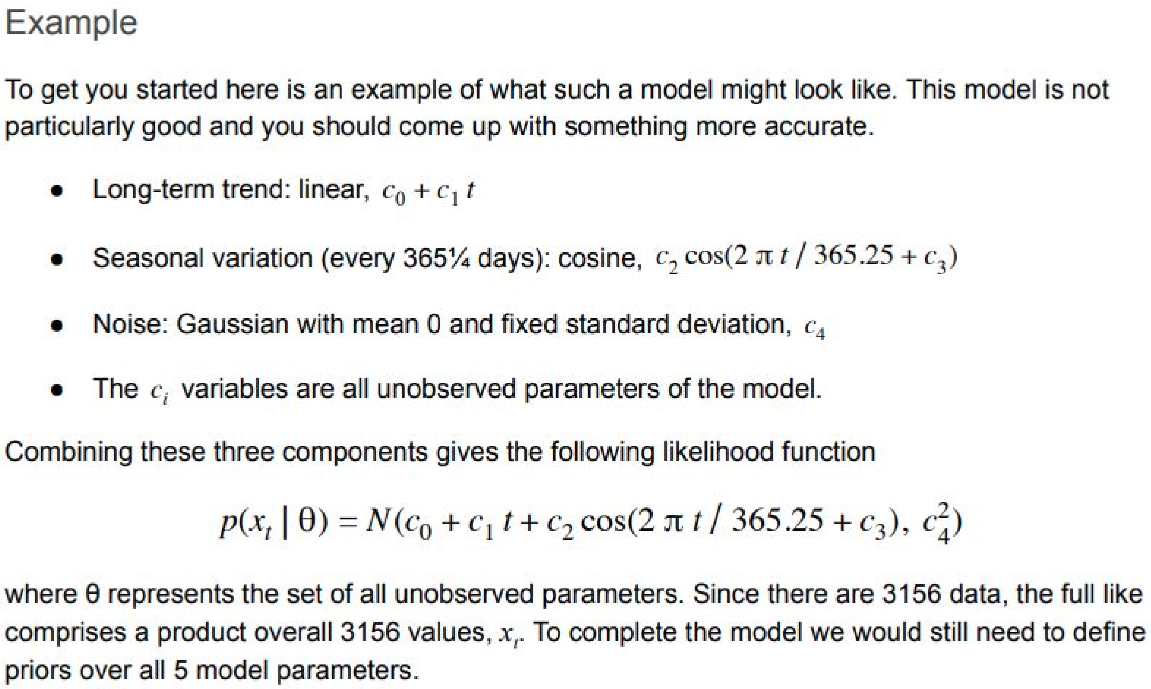
<u>Data pre-processing</u>

The date in the dataset, as shown in Figure 1 and 2, is in string format with "YYYY-MM-DD". This was a trouble when I tried to plot Figure 3 because matplotlib doesn't recognize this as a date. For the modeling step in part 3 (explained more below), this format is not helpful for our equation because it is not meaningful for the equation to make predictions. Therefore, I changed the date to 'Days interval" with days interval from the first observation, March 29, 1958. For example, the next observation has a date of April 5, 1958 and the days interval is 7, and for April 12, 1958 the days interval is 14 days.

We also need to pre-process the CO2 level because it is currently on a different scale with our parameters in the models in part 3. The parameters are mostly drawn from a standard Cauchy distribution with $\mu = 0$ and $\sigma = 1$, therefore the prediction we are making for the CO2 level should be around this range too for the model to converge faster. We can do this by normalizing each CO2 level according to this equation: $\frac{CO2 - CO2_{min}}{CO2_{max} - CO2_{min}}$, so each of them would be in range (0, 1) after the normalization. After we finish approximating the parameters, we can denormalize the CO2 level again to get predictions.

3. The model

In this project, we will use 3 statistical models to fit the data. The first model is given in the prompt with a linear long-term trend, seasonal variation according to a cosine function, and noise according to a Gaussian distribution. The second and third models are similar to the first model, but with quadratic and linear-quadratic equations, respectively, for the long-term trend.



Example

To get you started here is an example of what such a model might look like. This model is not particularly good and you should come up with something more accurate.

- Long-term trend: linear, $c_0 + c_1 t$
- Seasonal variation (every 365¼ days): cosine, $c_2 \cos(2 \pi t / 365.25 + c_3)$
- 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

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

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 would still need to define priors over all 5 model parameters.

Example equation from the prompt

From the example equation, we can see that the c_i parameters are unobserved parameters of the model that we need to sample. The observed parameters is the seasonal variation with 2pi and 365.25 days for a cycle.

Explaining the parameters:

Linear model:

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

- x_t : the CO2 level we are trying to predict
- θ : the set of all unobserved parameters
- c_0 and c_1 : unobserved parameters of linear long-term trend of the CO2 level
- c_2 and c_3 : unobserved parameters of the cosine seasonal variation that the CO2 level follows
- c_4 : the fixed standard deviation of the Gaussian distribution that the noise of the model is drawn from. The Gaussian distribution has a mean of 0

<u>The priors for the parameters:</u>

By default, unless mentioned otherwise, the prior for parameters is a standard Cauchy distribution with μ = 0 and σ = 1[1]. We are using this distribution because it has thicker tails than the standard normal distribution, while it's also generic enough to capture the model that we don't have a lot of information about.

- c_0 and c_1 : unobserved parameters of linear long-term trend of the CO2 level. c_0 is the constant that would be valuable at t = 0, so c_0 should be non-negative to

[1] #distributions: Here I explain the rationale behind using Cauchy distribution as priors for the parameters. I also explain where I don't use the standard Cauchy distribution parameters but customize it in the linear-quadratic model below.

represent the existence of the CO2. c_1 is the linear trend of the CO2, which should be non-negative too given the rise of the CO2 represented in the data exploration step.

- c_2 and c_3: unobserved parameters of the cosine seasonal variation that the CO2 level follows. The CO2 level follows a periodic pattern, with c_2 as the amplitude of the periodic function and c_3 as the phase of the periodic function.
- c_4: the fixed standard deviation of the Gaussian distribution that the noise of the model is drawn from. The Gaussian distribution has μ = 0 and σ = 1, and the noise is to represent the uncertainty around the mathematical function that we have.

Factor graph for the model:

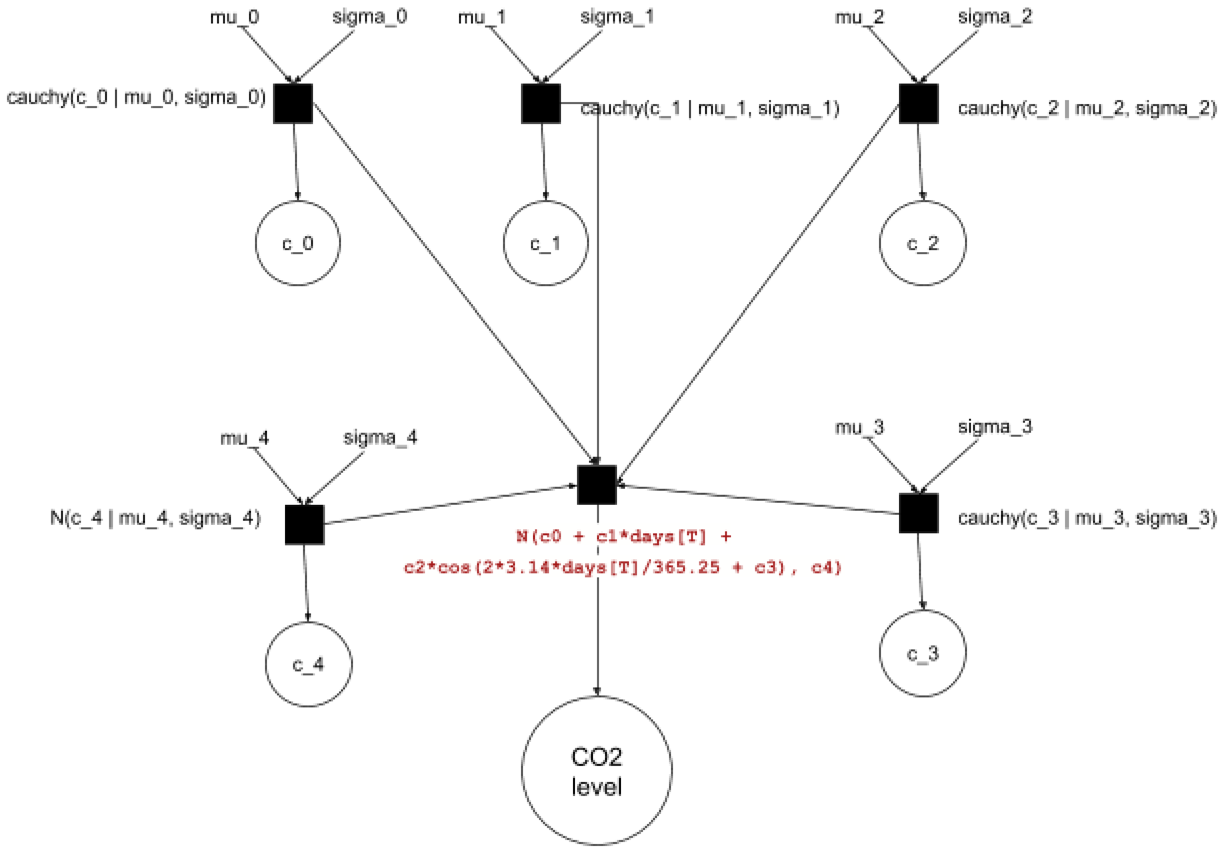

mu_0
sigma_0
mu_1
sigma_1
mu_2
sigma_2
cauchy(c_0 | mu_0, sigma_0)
cauchy(c_1 | mu_1, sigma_1)
cauchy(c_2 | mu_2, sigma_2)
c_0
c_1
c_2
mu_4
sigma_4
mu_3
sigma_3
N(c_4 | mu_4, sigma_4)
cauchy(c_3 | mu_3, sigma_3)
N(c0 + c1*days[T] +
c2*cos(2*3.14*days[T]/365.25 + c3), c4)
c_4
c_3
CO2
level

Figure 4: *Factor graph of the linear model*

From the factor graph, we can see that the CO2 level is influenced by the parameters such as c_0 , c_1 , c_2 , c_3 , c_4 . However, CO2 level is also influenced by the "days" parameter. However, I don't know how to incorporate it into the graph so I just mention it in the distribution that computes the CO2 level (in red text).

As stated in the prompt, we should try to explore different models other than the linear one. Therefore, I experimented with the quadratic and linear-quadratic models as follows:

Quadratic model:

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

The description of the parameters is exactly the same to that of the linear model, except for c_0 and c_1, which are the unobserved parameters of quadratic long-term trend of the CO2 level.

The priors for the parameters:

By default, unless mentioned otherwise, the prior for parameters is a standard Cauchy distribution with μ = 0 and σ = 1. We are using this distribution because it has thicker tails than the standard normal distribution, while it's also generic enough to capture the model that we don't have a lot of information about.

- c_0 and c_1: unobserved parameters of linear long-term trend of the CO2 level. c_0 is the constant that would be valuable at t = 0, so c_0 should be non-negative to represent the existence of the CO2. c_1 is the linear trend of the CO2, which should be non-negative too given the rise of the CO2 represented in the data exploration step.
- c_2 and c_3: unobserved parameters of the cosine seasonal variation that the CO2 level follows. The CO2 level follows a periodic pattern, with c_2 as the amplitude of the periodic function and c_3 as the phase of the periodic function.
- c_4: the fixed standard deviation of the Gaussian distribution that the noise of the model is drawn from. The Gaussian distribution has μ = 0 and σ = 1, and the

noise is to represent the uncertainty around the mathematical function that we have.

Factor graph for the model:

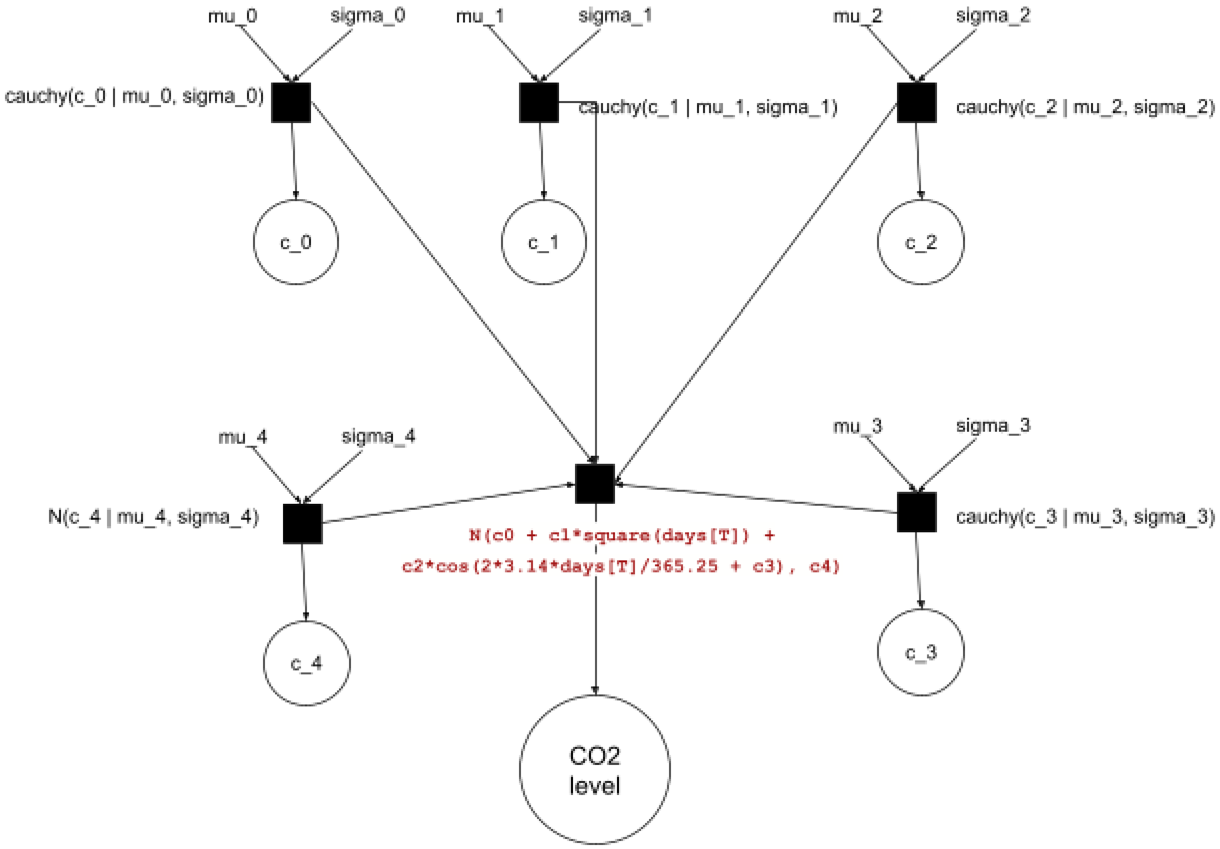

mu_0
sigma_0
mu_1
sigma_1
mu_2
sigma_2
cauchy(c_0 | mu_0, sigma_0)
cauchy(c_1 | mu_1, sigma_1)
cauchy(c_2 | mu_2, sigma_2)
c_0
c_1
c_2
mu_4
sigma_4
mu_3
sigma_3
N(c_4 | mu_4, sigma_4)
cauchy(c_3 | mu_3, sigma_3)
N(c0 + c1*square(days[T]) +
c2*cos(2*3.14*days[T]/365.25 + c3), c4)
c_4
c_3
CO2
level

Figure 5: *Factor graph of the quadratic model*

Essentially, this factor graph is exactly the same as the one for the linear model, except for the distribution that we draw the CO2 level from.

Linear-quadratic model:

$$p(x_t|\,\theta) = N(c_0 + c_1 t + c_{10} t^2 + c_2 cos(2t / 365.25\ + c_3),\ c_4^2)$$

The description of the parameters is exactly the same to that of the quadratic model, except for c_0, c_1, and c_{10} which are the unobserved parameters of linear-quadratic long-term trend of the CO2 level.

The priors for the parameters:

All priors are the same as the quadratic model, except for c_{10} and c_4

- c_{10} has the Cauchy distribution with μ = 0 and σ = 0.2. The σ =2 is lower than the σ = 1 of the distributions of other parameters. The reason for this is that c_{10} is the coefficient of the quadratic term, which has a lot of influence on the final model if it's high. Therefore, I try to keep it low for it to not dominate the whole equation and for the model to converge faster.
- c_4 : the fixed standard deviation of the Gaussian distribution that the noise of the model is drawn from. Other model has it at μ = 0 and σ = 1, but in this model we are having σ = 10 to capture keep up with the large linear-quadratic trend of the model.

Factor graph for the model:

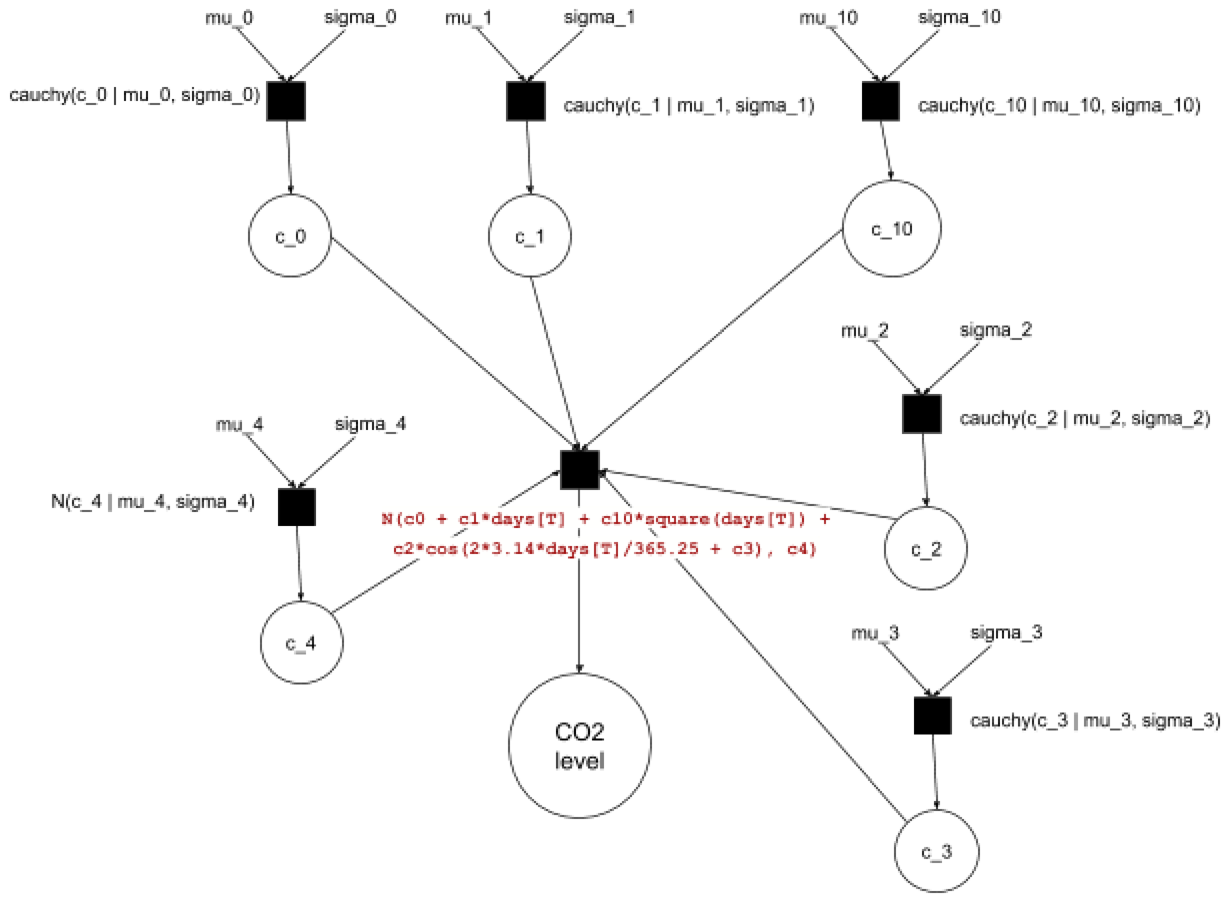

mu_0
sigma_0
mu_1
sigma_1
mu_10
sigma_10
cauchy(c_0 | mu_0, sigma_0)
cauchy(c_1 | mu_1, sigma_1)
cauchy(c_10 | mu_10, sigma_10)
c_0
c_1
c_10
mu_2
sigma_2
cauchy(c_2 | mu_2, sigma_2)
mu_4
sigma_4
N(c_4 | mu_4, sigma_4)
N(c0 + c1*days[T] + c10*square(days[T]) +
c2*cos(2*3.14*days[T]/365.25 + c3), c4)
c_2
c_4
mu_3
sigma_3
cauchy(c_3 | mu_3, sigma_3)
CO2
level
c_3

***Figure 6:** Factor graph of the linear-quadratic model*

Essentially, this factor graph is exactly the same as the one for the linear model, except for the distribution that we draw the CO2 level from and we have extra parameters (c_{10}) to represent the linear-quadratic model.

4. Results

a/ Prediction results according to each model

Figure 7, 8, 9 below are the predictions from the 3 linear, quadratic, linear-quadratic models respectively:

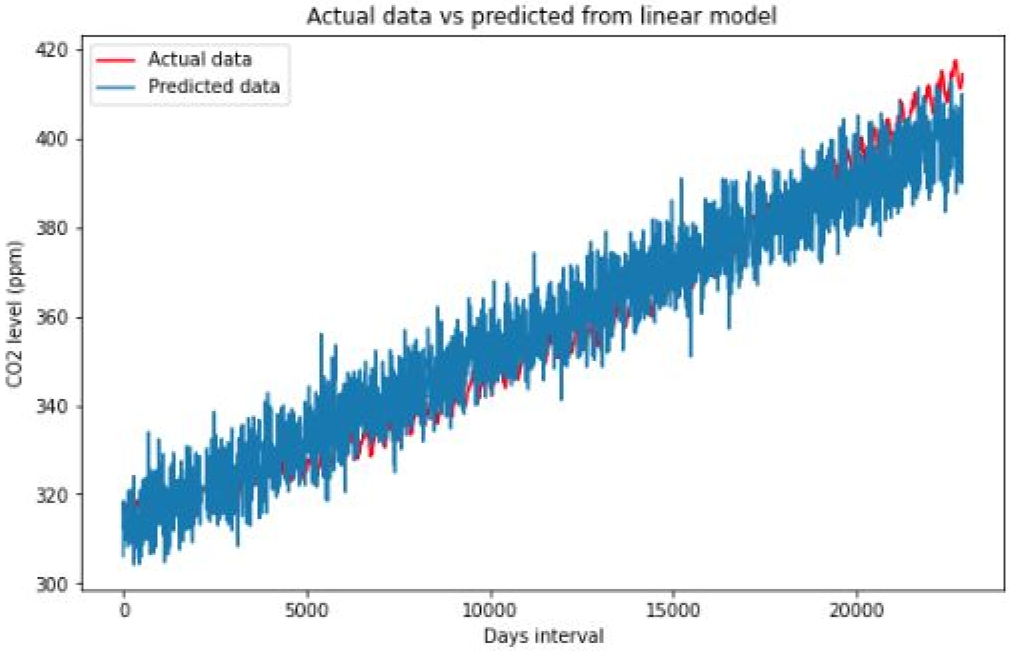


Figure 7: Observed data (in red) and predicted data (in blue) from the linear model

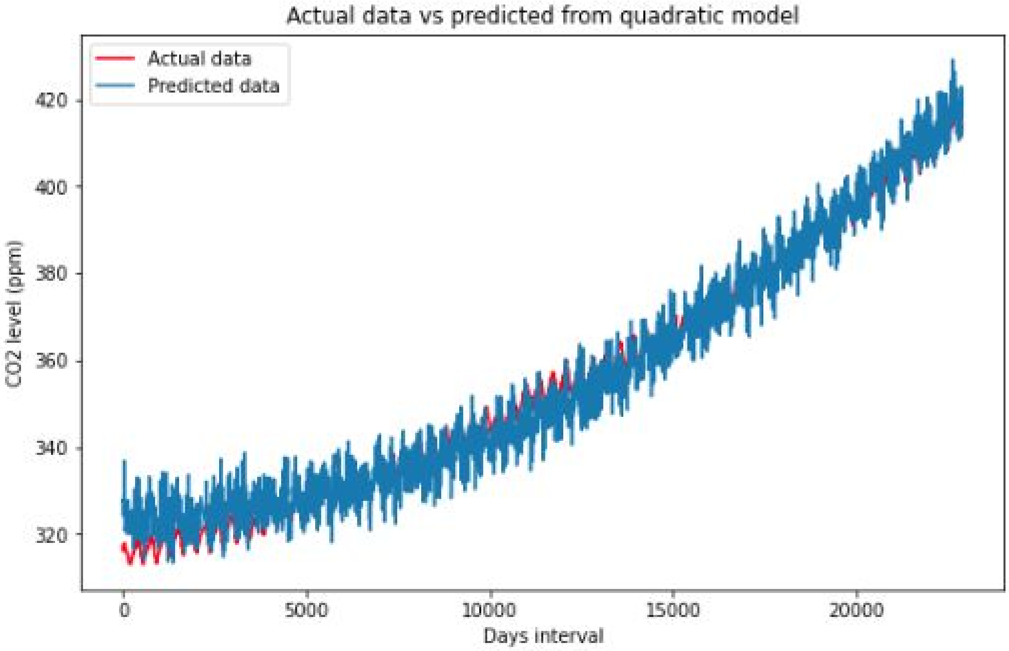


Figure 8: Observed data (in red) and predicted data (in blue) from the quadratic model

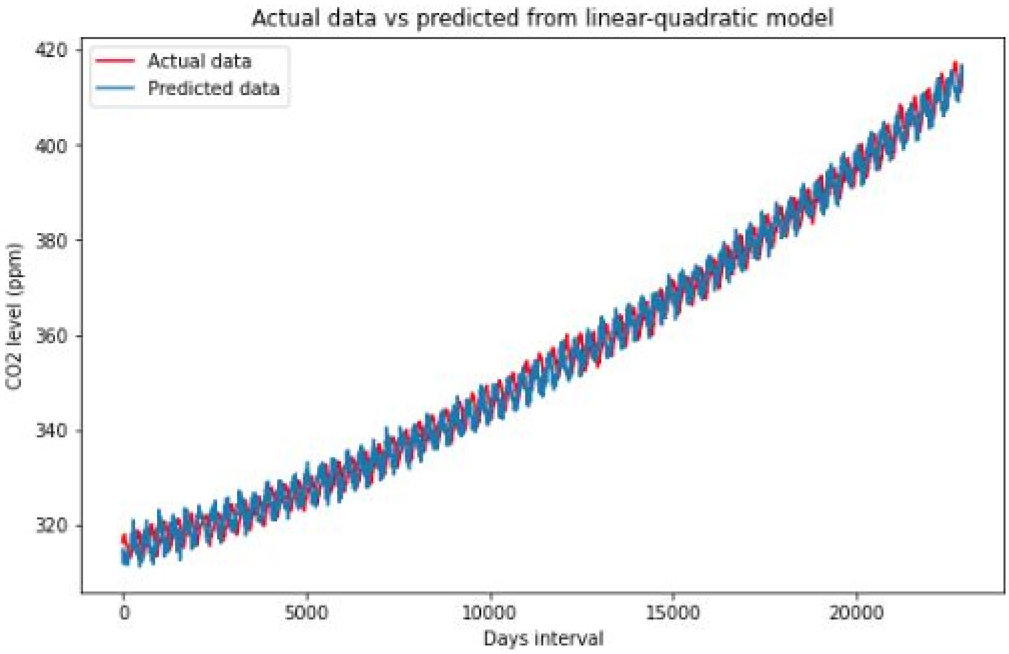


Figure 9: Observed data (in red) and predicted data (in blue) from the linear-quadratic model

At a glance, we can see that the predicted data from linear-quadratic model wrap around the observed data much better than the other 2 models, but let's zoom in the last 500 data points in each model and evaluate again (figure 10, 11, 12):

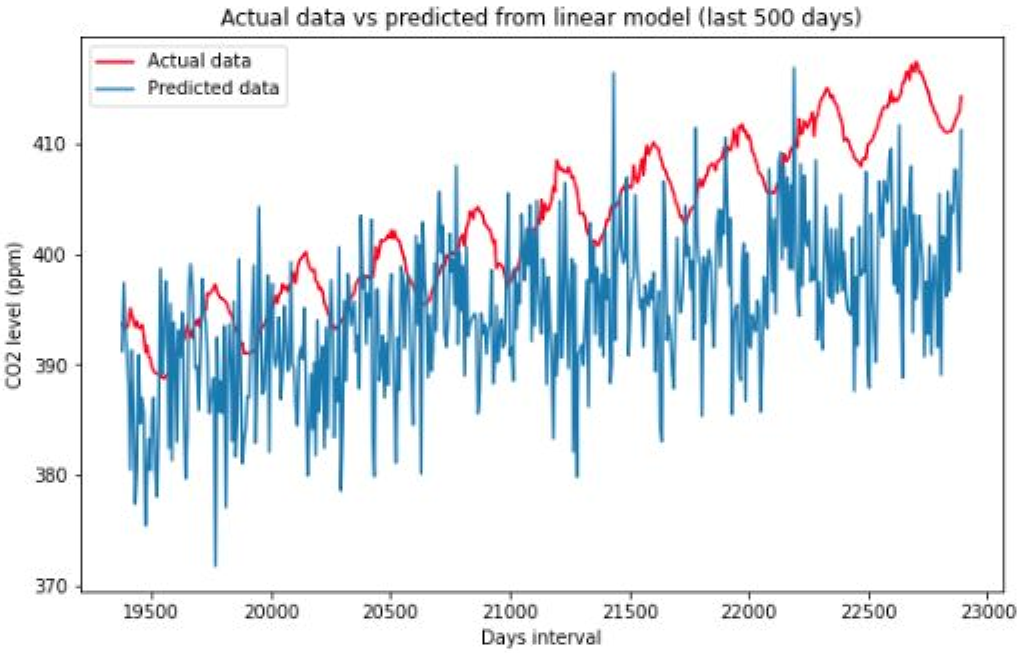


Figure 10: Zoomed in result from the linear model

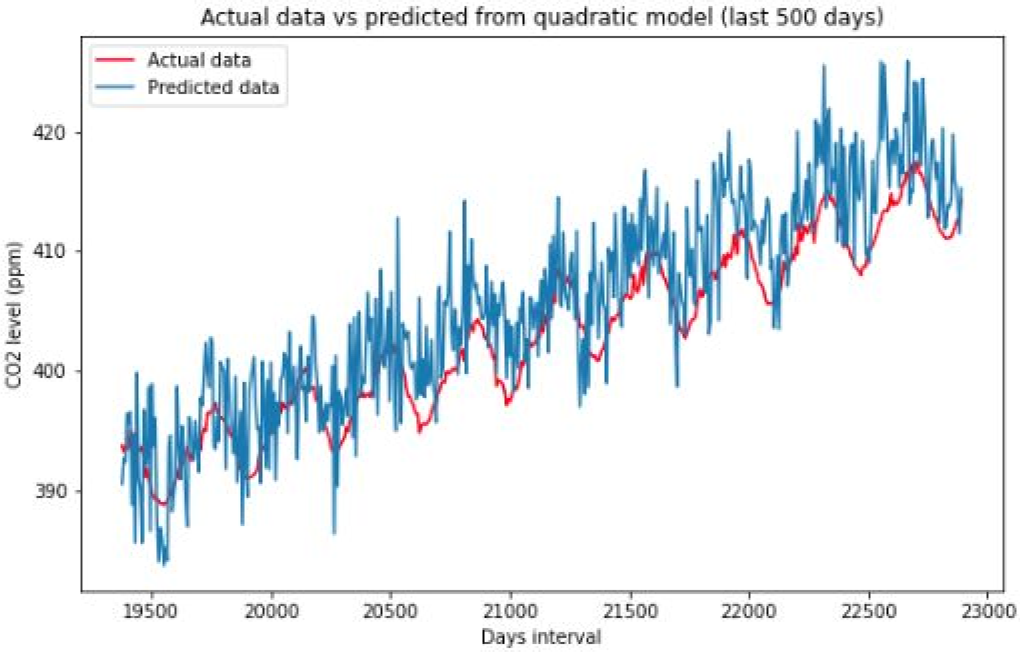


Figure 11: Zoomed in result from the quadratic model

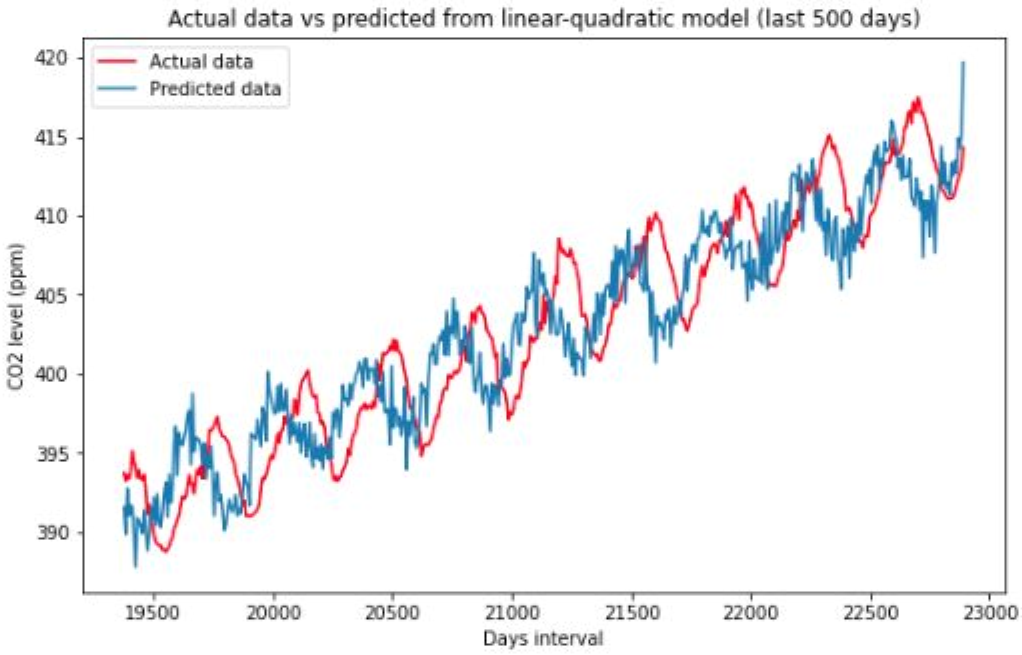


Figure 12: Zoomed in result from the linear-quadratic model

From the 3 figures above, we also see that the linear-quadratic model is most closely correlated with the observed data the most. The shape is strongly similar, but the predicted data is only off by a bit of phase shift, meaning that the trend of the prediction is a bit earlier than the actual observed data.

b/ Choosing the best fitted model (RMSE)

Apart from interpreting the plot by eyes, we can also rely on metrics to evaluate which model is the most accurate. One of the most popular metric to evaluate the performance of regression models is the root mean square error (RMSE):

$\sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$ with $Predicted_i$ as the predicted CO2 value, $Actual_i$ as the observed CO2, and N as the total number of data points we have. As we can see from figure 13, 14, 15 below, the RMSE from the linear-quadratic model is the lowest among the 3 models, hence indicating that the linear-quadratic model is the best fitted model out of all the 3.

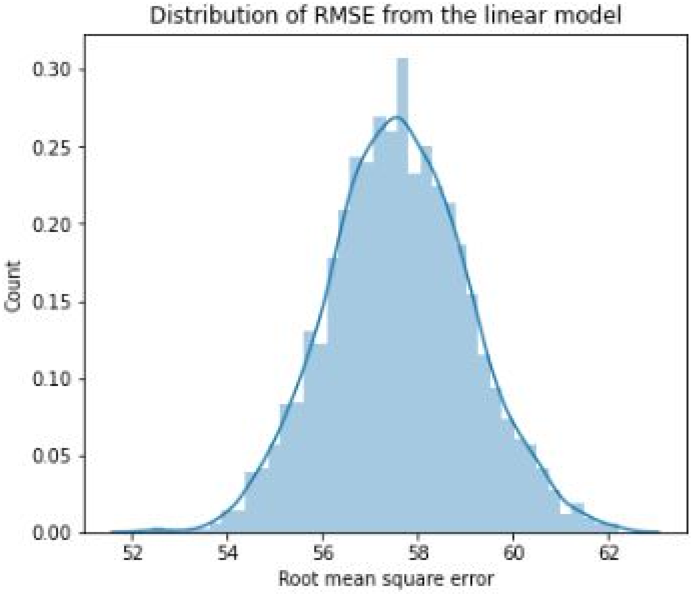


Figure 13: RMSE from the linear model

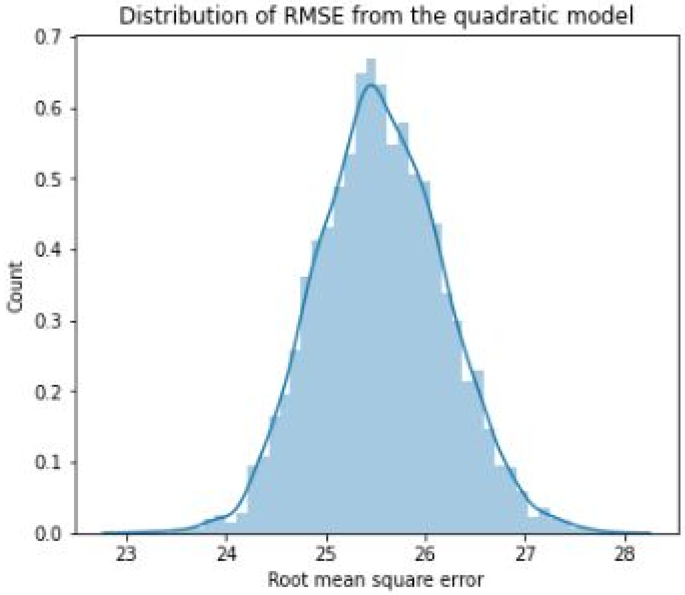


Figure 14: RMSE from the quadratic model

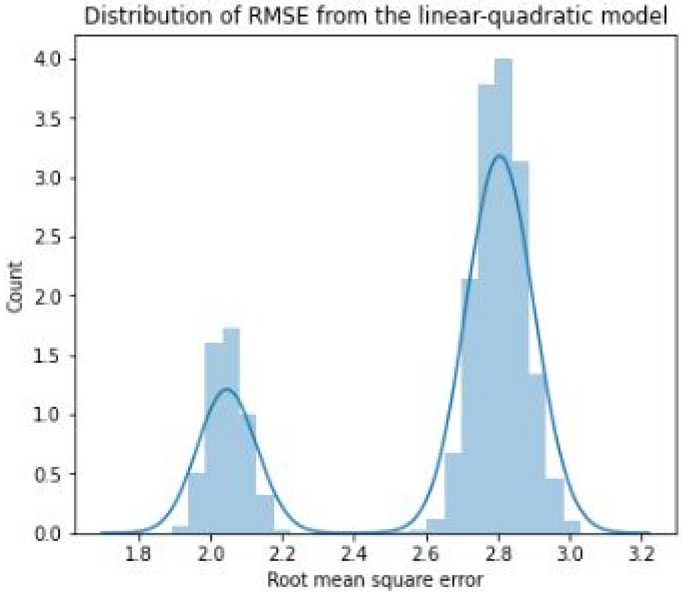
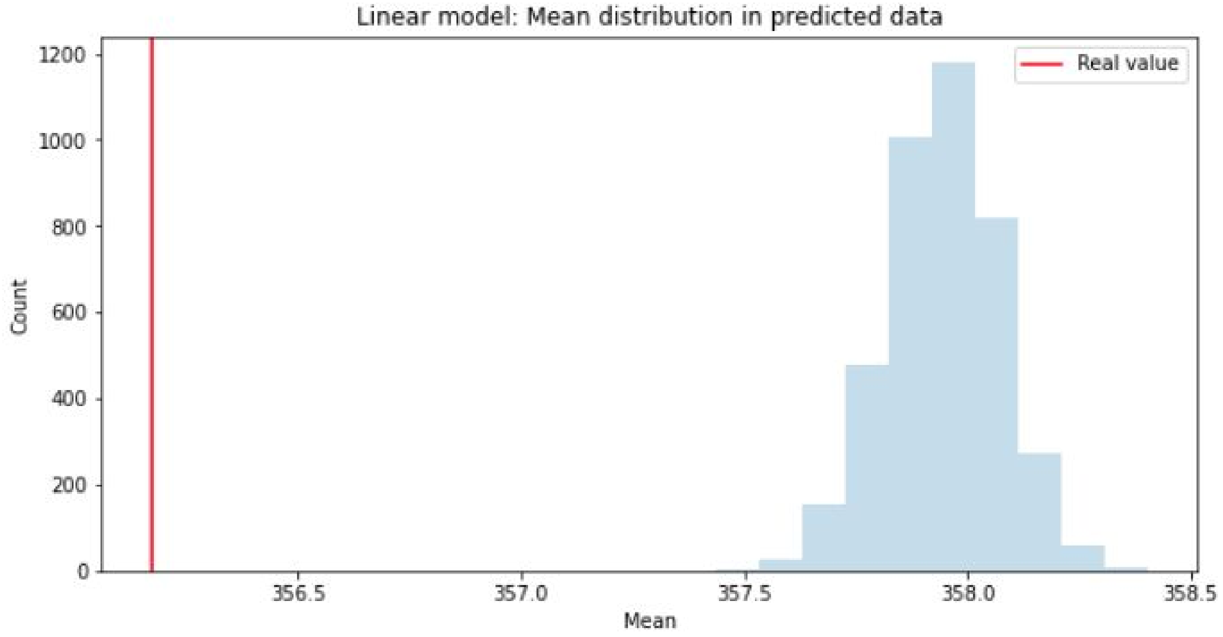


Figure 15: RMSE from the linear-quadratic model

c/ Predicted distribution evaluation

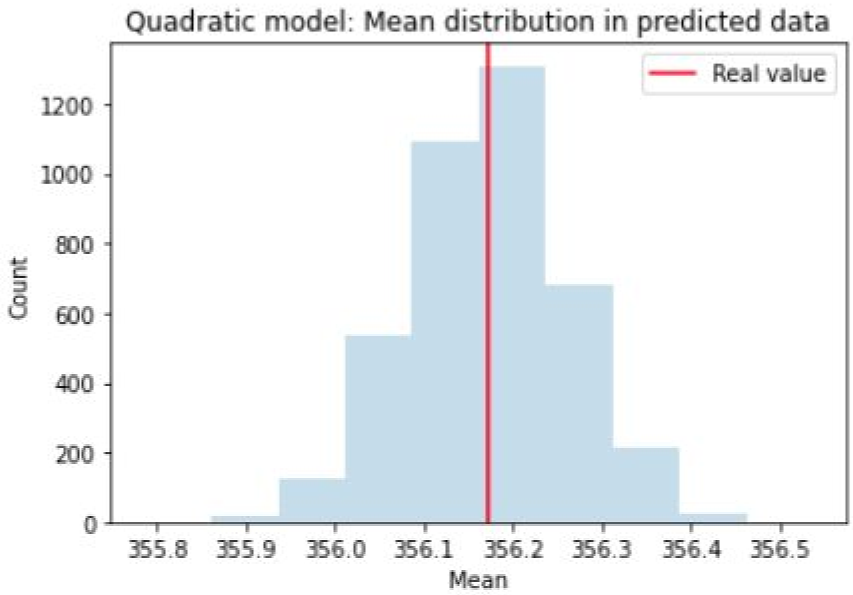
To evaluate the distribution generated from each model, we can compare the test statistics (mean and standard deviation) of the predicted data to that of the observed data. Figure 16, 17, 18 represents the mean of the predicted values from the 3 models respectively, and figure 19, 20, 21 represents the standard deviation.

We can see that the mean of the predicted values is also approximately the same as the mean of the real value in the quadratic and linear-quadratic models (figure 17 and figure 18) with the red line around the mean of the distribution and p-value as around 0.5 (0.4915 and 0.4985). Meanwhile, the mean of the predicted values from the linear model (figure 16) is way too far away from the real mean, with the p-value as 0.



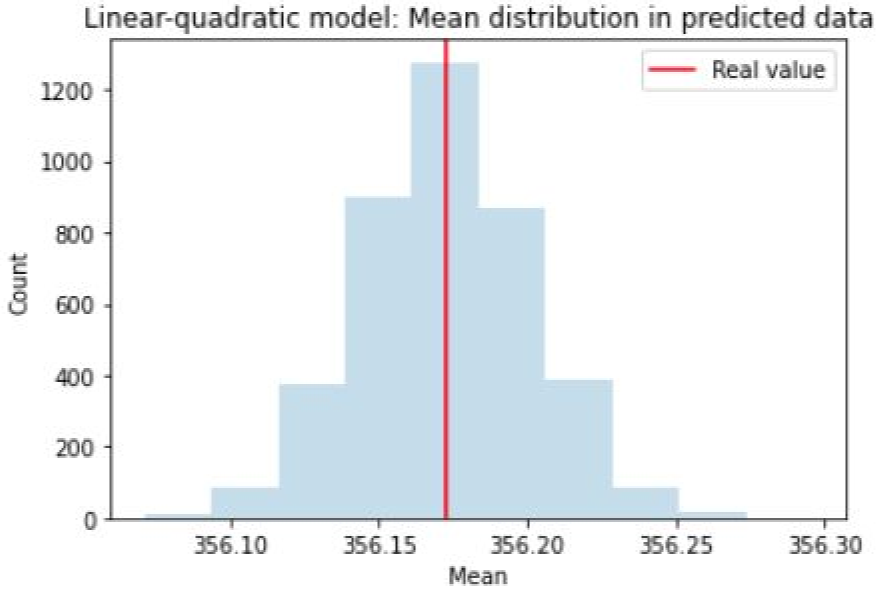
```
p-value is: 0.0
```

Figure 16: Distribution of the mean in the predicted data of the linear model



```
p-value is: 0.4915
```

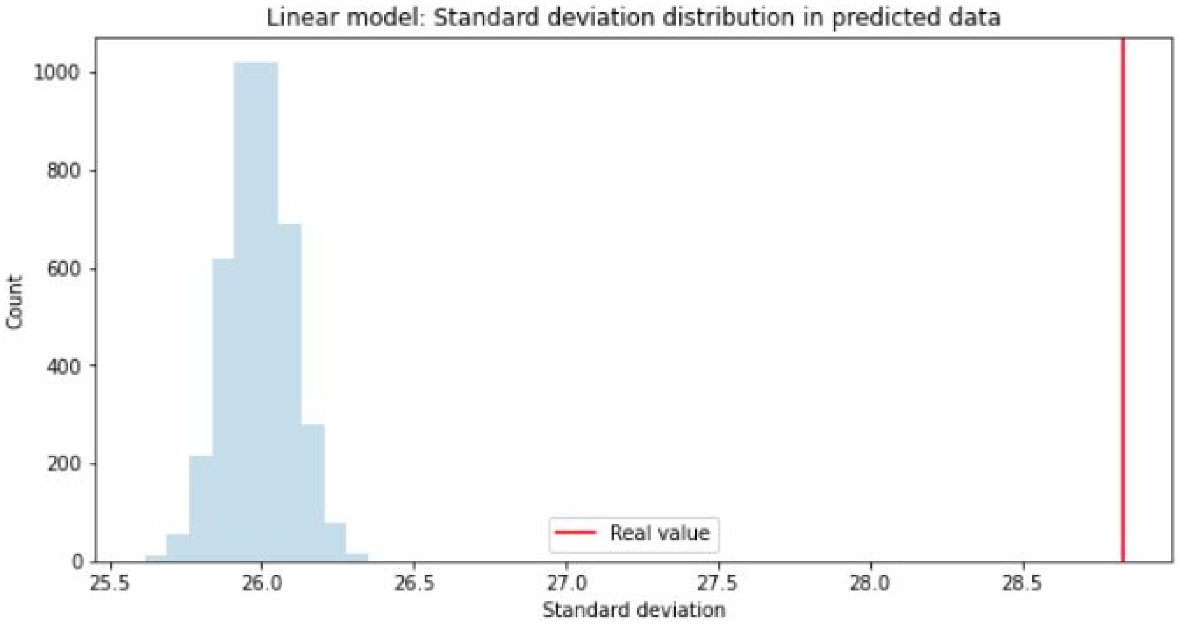
Figure 17: Distribution of the mean in the predicted data of the quadratic model



```
p-value is: 0.4985
```

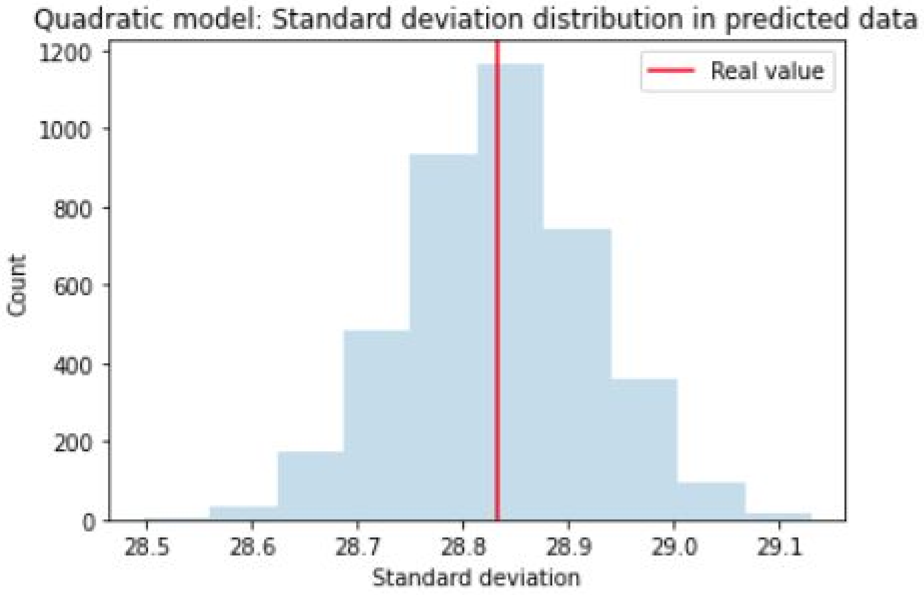
Figure 18: Distribution of the mean in the predicted data of the linear-quadratic model

With the similar analysis as above, we can also see that the standard deviation of the predicted values in the quadratic and linear-quadratic models (figure 20 and figure 21) is closer to the real standard deviation with the red line around the mean of the distribution and p-value as around 0.5 (0.4955 and 0.49175). Meanwhile, the standard deviation of the predicted values from the linear model (figure 19) is way too far away from the real mean, with the p-value as 1.



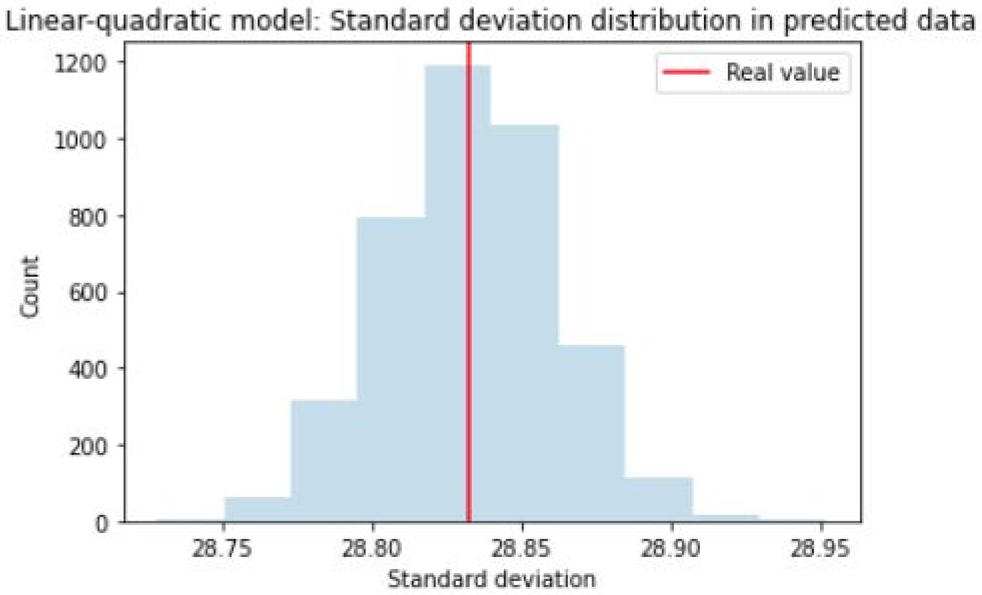
```
p-value is: 1.0
```

Figure 19: Distribution of the standard deviation in the predicted data of the linear model



```
p-value is: 0.4955
```

Figure 20: Distribution of the standard deviation in the predicted data of the quadratic model



```
p-value is: 0.49175
```

Figure 21: Distribution of the standard deviation in the predicted data of the linear-quadratic model

d/ Prediction till 2060 with 95% confidence interval

The prediction in figure 22 shows that the CO2 level will continue rising and surpass 500 ppm atmospheric CO2 level by 2060. This is an alarming rate because as we talked in the introduction, a CO2 level at 450 ppm is already dangerous for climate change.

The zoomed in prediction in figure 23 denotes the uncertainty around the prediction. However, both the lower and upper bound of the 95% confidence interval still indicate the rise of the CO2 level.

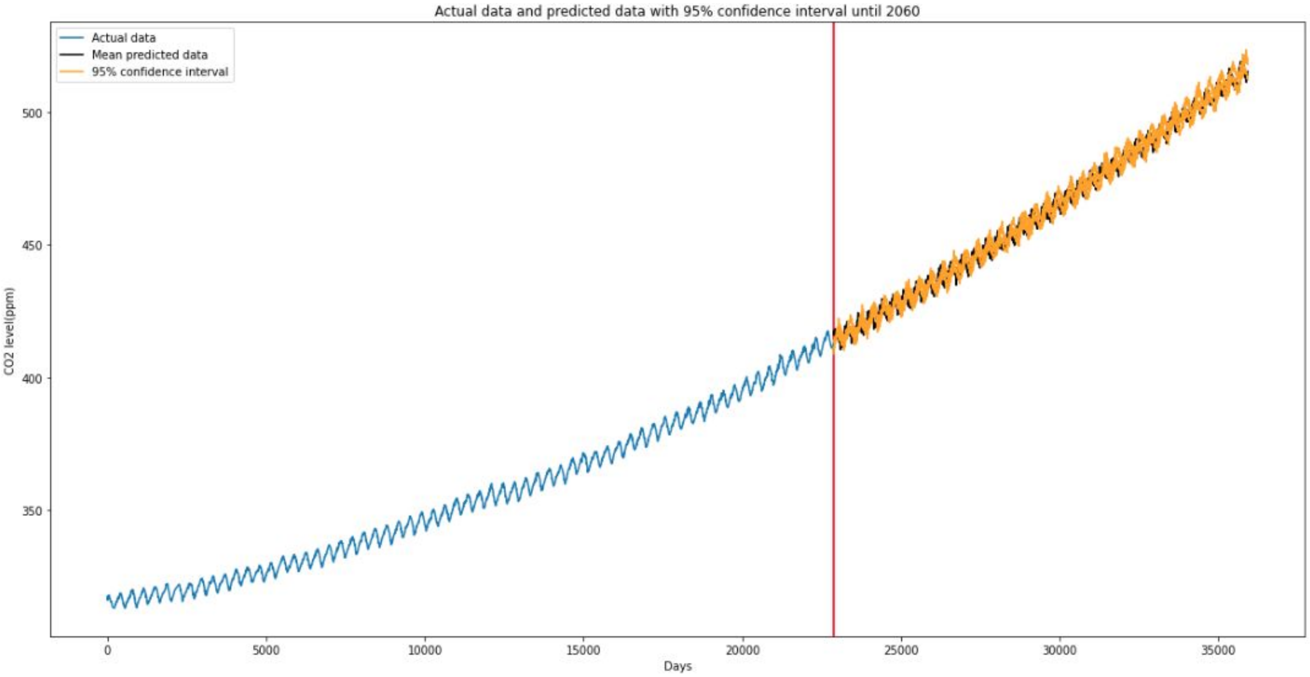


Figure 22: Prediction of CO2 level trend until 2060 with 95% confidence interval, with the vertical line marking the time as of now

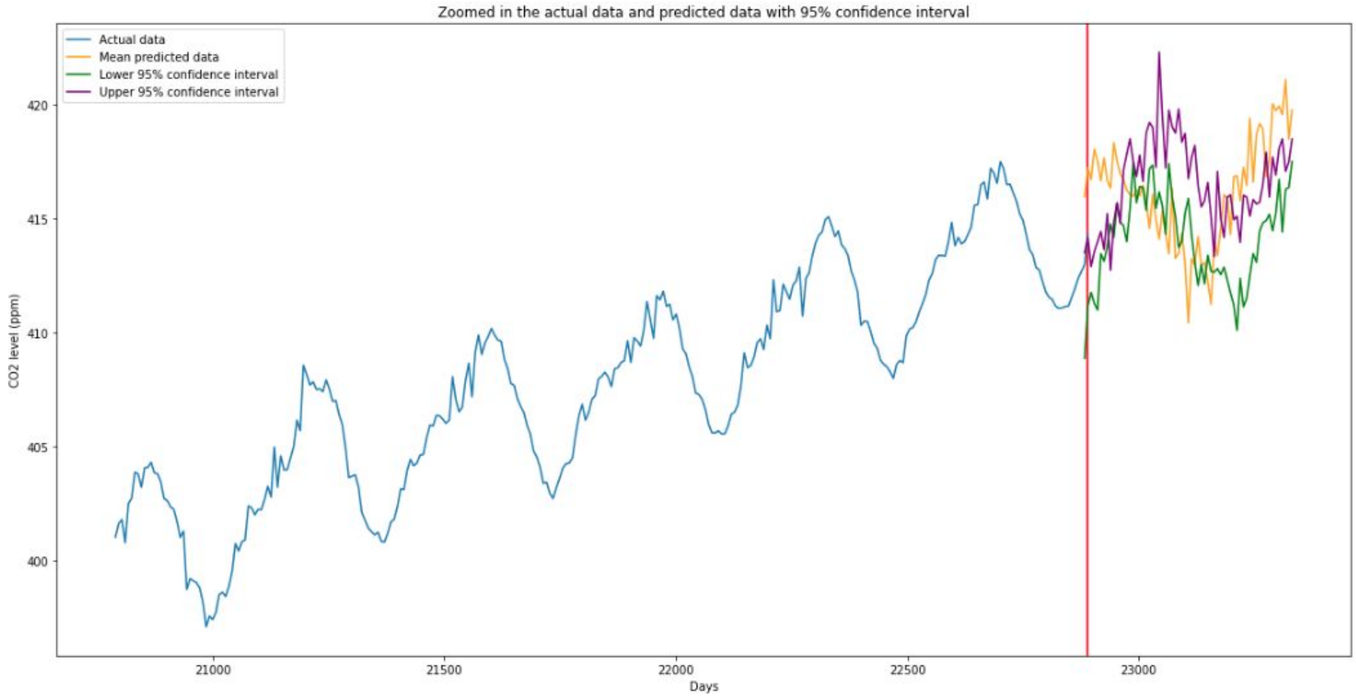


Figure 23: Zoomed in prediction of CO2 level trend until 2060 with 95% confidence interval, with the vertical line marking the time as of now

e/ Prediction with 450ppm threshold with 95% confidence interval

Our goal is to predict when we would reach the 450 ppm threshold, and figure 24 below shows that we will reach that threshold:

- At the mean day of January 7, 2033
- At the lower 95% confidence interval day of March 4, 2034
- At the upper 95% confidence interval day of March 26, 2033

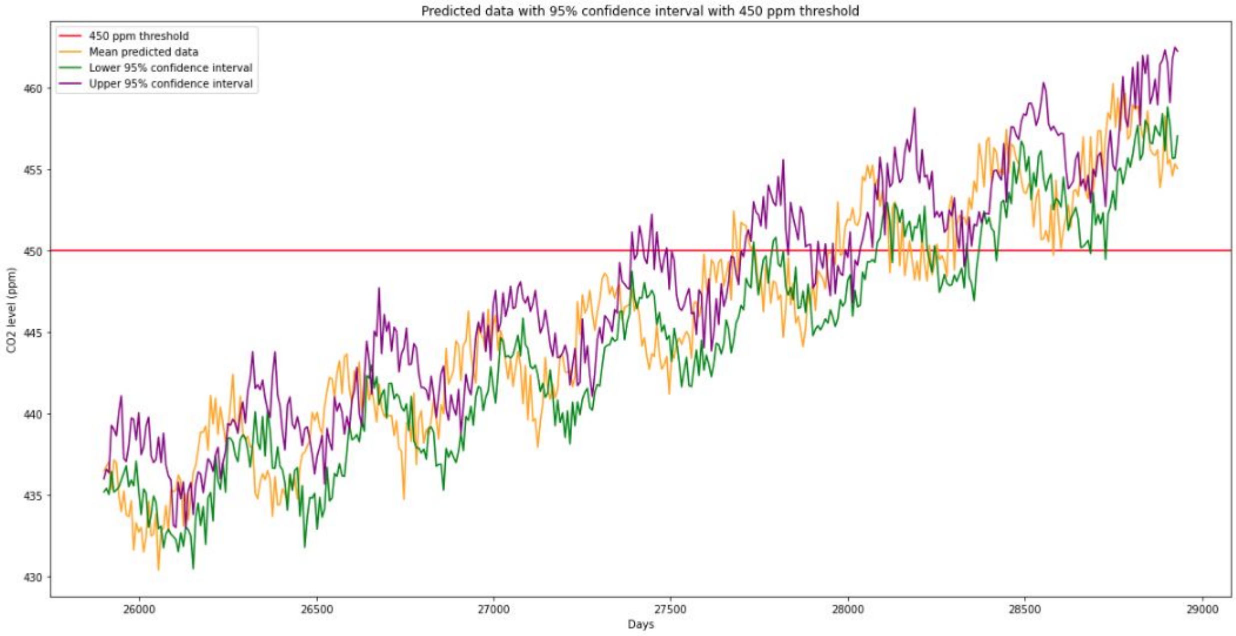


Figure 24: Predicted data with 95% confidence interval with 450 ppm threshold

However, as pointed in the discussion above, this is just a rough estimate of the CO2 level trend and not an accurate prediction. In order to have better results, we need to try more models and compare the results (here we only do 3 models). Also we need to check for biases that can happen along the way, such as overfitting or underfitting the model, so that we can choose the most appropriate model

5. Conclusion[2]

In this report, we have explored the CO2 dataset from the Mauna Loa Observatory in Hawaii. We try to explain the relationship between the data points by statistical model, and use these models to predict future trends of the CO2 level in the

[2] #organization: This report has a table of contents at the beginning and is divided into 5 parts (introduction, about the dataset, modeling, results, conclusion) with clear title for each one. Thanks to this structure, a busy reader can examine

atmosphere. It is helpful to know how we should expect the CO2 level to rise, and know when the CO2 passes the threshold of 450 ppm for us to prepare to protect ourselves from climate change. Though much is still needed to have a model with better accuracy, with this preliminary result we can see that we are only roughly 13 years away from the threshold, and each of us and policy makers need to join forces to slow down the CO2 increase as much as possible.

Appendix

Appendix A: Dataset from Mauna Loa Observatory

The dataset can be found here:

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

Appendix B: Python code

The code for this report can be found in this Google Colab notebook:
https://colab.research.google.com/drive/1-gKL_hP5rt0zpzCjSjZCUdKtY-f7q8_1?usp=sharing