[CS146] Final Project Report

December 20, 2019

Modeling and forecasting atmospheric CO_2

Report

Taha Bouhoun

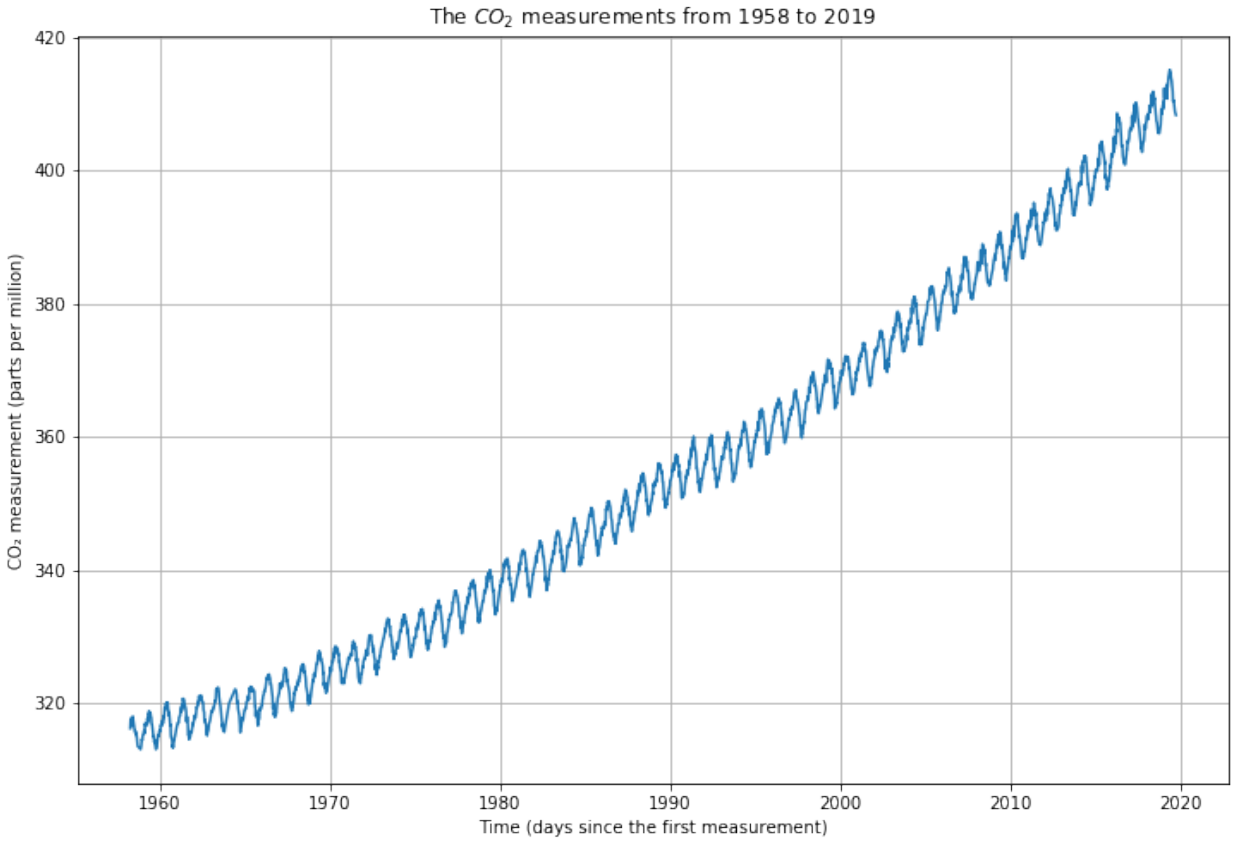
1 Introduction

Since 1958 atmospheric carbon dioxide measurements have been recorded at the Mauna Loa Observatory in Hawaii. CO_2 levels have been increasing steadily since the start of the industrial revolution in the 18th century. Older data are from ice core measurements, not atmospheric measurements. The data from Mauna Loa provide very direct data on atmospheric CO_2, which forms an important part of global climate change modeling[1].

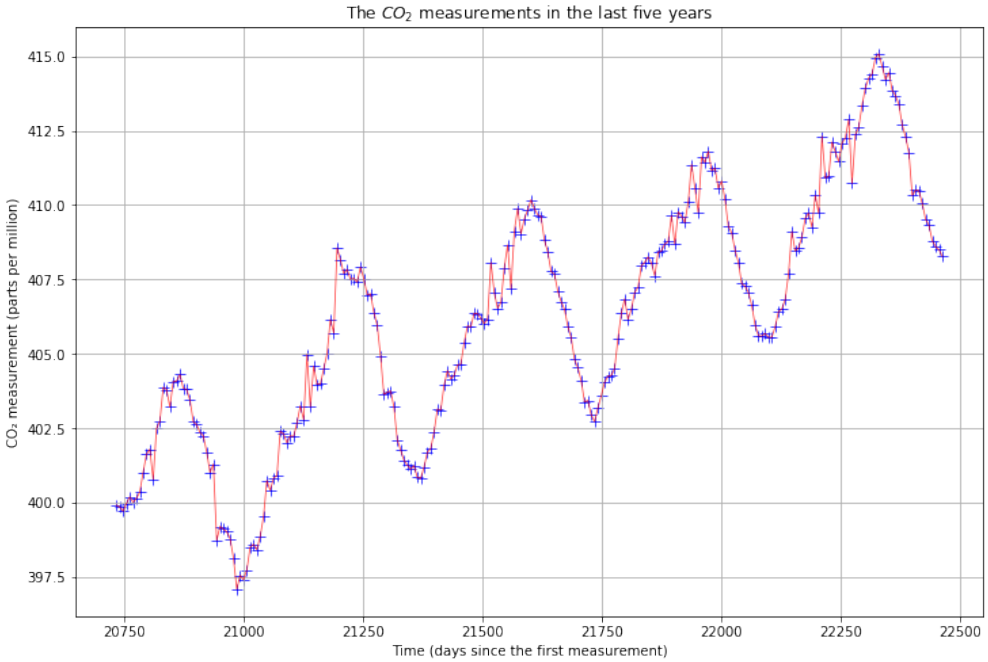
`Snapshot from the Youtube Channel of the American Museum of Natural History [2]`

`Out[32]:`

`Out[6]:`


The CO_2 measurements from 1958 to 2019
CO_2 measurement (parts per million)
Time (days since the first measurement)
420
400
380
360
340
320
1960
1970
1980
1990
2000
2010
2020

Out[7]:


The CO_2 measurements in the last five years
CO_2 measurement (parts per million)
Time (days since the first measurement)
415.0
412.5
410.0
407.5
405.0
402.5
400.0
397.5
20750
21000
21250
21500
21750
22000
22250
22500

2 The assumptions

For this project, the weekly measurements of the Mauna Loa Observatory are used as initial data to train the model based on the Bayesian Framework:

$$Posterior = \frac{likelihood \cdot prior}{data} \rightarrow P(\theta_k \mid X) = \frac{P(X \mid \theta_k) \cdot P(\theta_k)}{P(X)}$$

The data is gathered from the remote islands of Hawaii on top of a volcanic rock and away from large human populations. The aim of this project is to disentangle the human impact on CO_2 levels from the natural cycles of the compound in the atmosphere. As mentioned in the prompt, the differences in seasons between the northern and southern hemisphere don't mitigate the high emissions during the winter since atmospheric mixing is slow.

Our knowledge of the phenomenon depicts our model specification. Indeed, the Bayesian model adjusts our priors as we feed more data into it but specification of the likelihood shapes the parametric model that we believe that better describe the phenomenon.

The vegetation cycle varies in a temporal way. There are the natural germination, growth, and death cycles, whose length varies with the nature of the plant. In addition, we're not accounting for deforestation which could be happening at a higher rate than human efforts to protect the environment.

3 Modeling

3.1 Model One

The example model proposed in the problem statement is constructed as follow:

- Long-term trend: linear $c_0 + c_1 \cdot t$

- Seasonal variation: cosine $c_2 \cdot \cos\left(\frac{2\pi \cdot t}{365.25} + c_3\right)$

- Noise: Gaussian of the form $N \sim (0,\ c_4)$ Combining all the three components gives the following likelihood function:

$$p(x_t \mid \theta) = N\left(c_0 + c_1 \cdot t + c_2 \cdot \cos\left(\frac{2\pi \cdot t}{365.25} + c_3\right),\ c_4^2\right)$$

such as $\theta = \{c_0,\ c_1,\ c_2,\ c_3,\ c_4\}$

Since the prior must be set without consulting the data, the belief about each parameters is:

- c_0**:** intercept has a low bound of 0 since we're confident that the atmosphere has to have a certain amount of CO_2, the prior is set as a Cauchy distribution since it has a heavy tail.

- c_1**:** the coefficient of the linear trend can either be positive, negative, or equals to zero. As a prior, we try not to bias the model towards a positive trend (although true), the data then would depict the coefficient of the slope.

- c_2: the amplitude of the periodic function illustrates the distance between the highest and the lowest measure of CO_2 within one period (i.e. one year). The parameter also follows a Cauchy distribution.

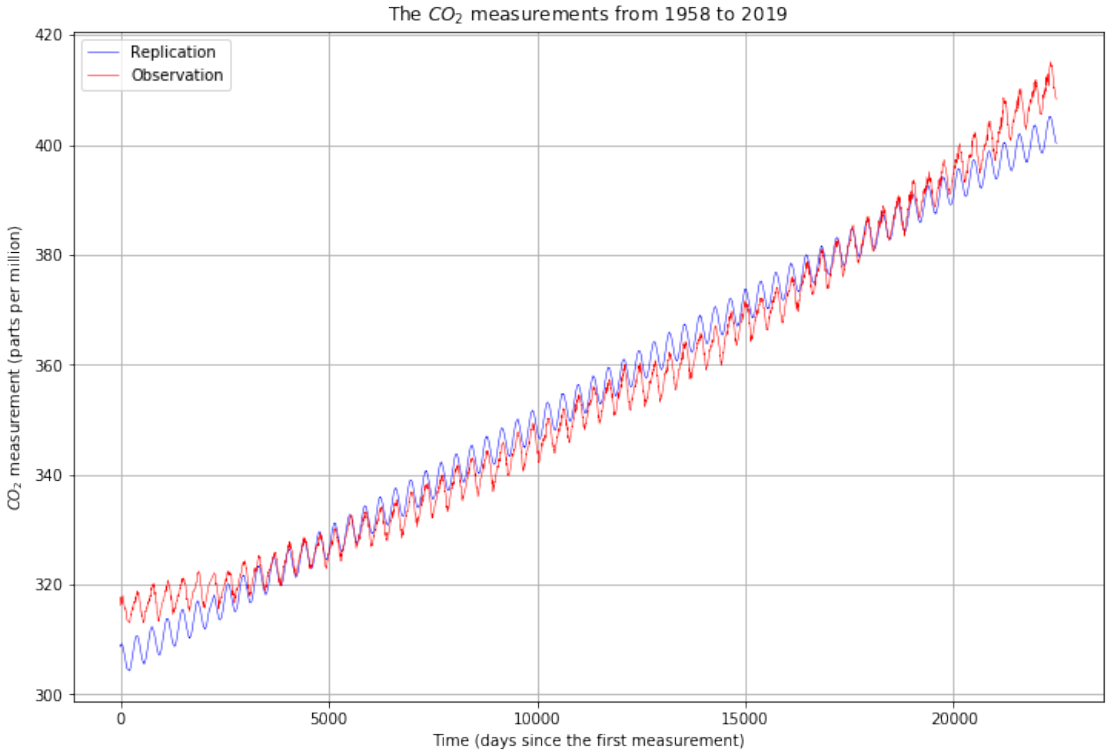
- c_3: the phase of the period function represent the shift. The prior belief is that it has to be bounded at a certain range but we can identify them on the second iteration of the improved model.

- c_4: the variance of the noise is constrained to be positive but we have little information about its magnitude, thus, we set a Cauchy prior for this parameter.

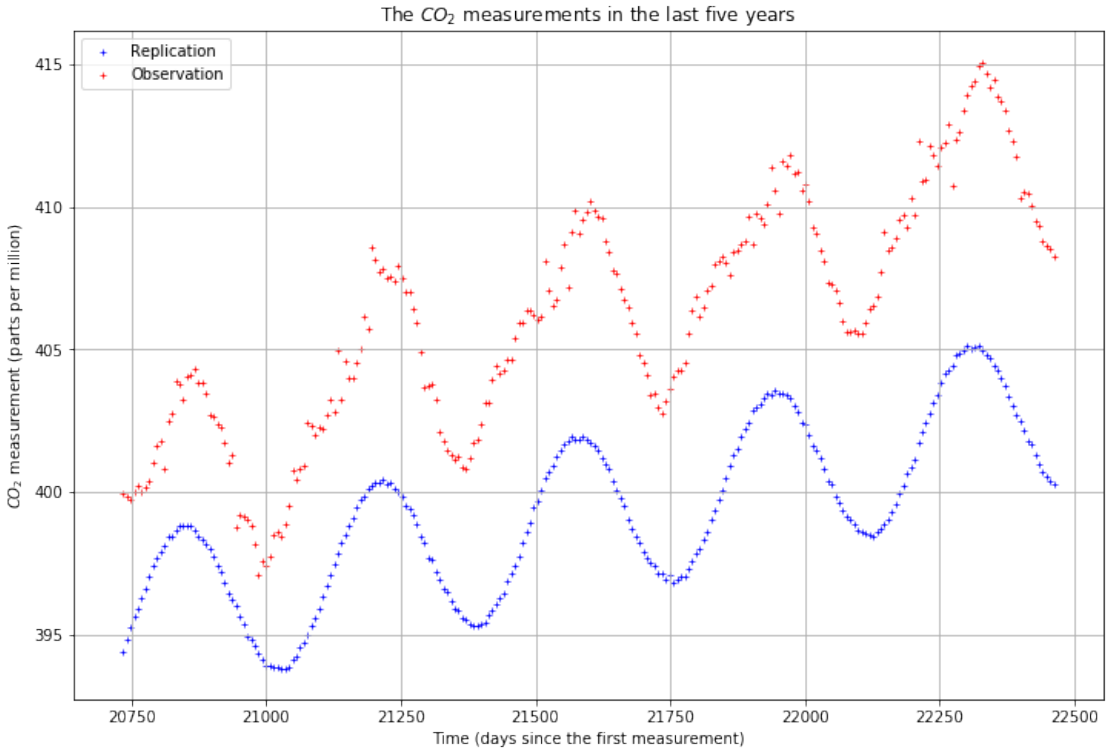
3.1.1 Results of Model One

- The long trend model doesn't align with the observations' trend which suggest that the trend is not linear. We can notice the shape of the residual graph which shows how the error is not centered at zero throughout the time period of the measurements.
- The Stan output shows that the Rhat for both the amplitude and the phase are greater than 1 which suggests that the Markov Chains didn't mix well and the samples are highly correlated leading the number of effective sample size to be equal to 2. As a result, we can conclude that the proportion of variance within each Markov Chain over the overall variance of the four chains didn't converge. (Appx. Model One)

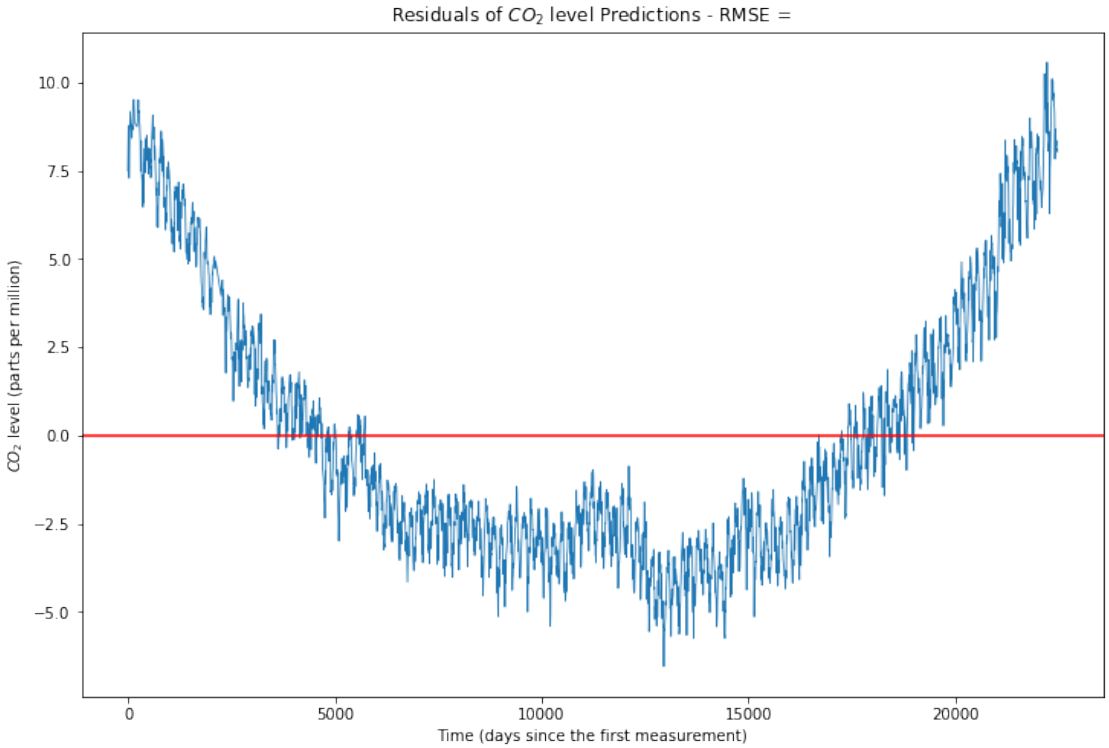
Out[8]:



Out[10]:


The CO_2 measurements in the last five years
Replication
Observation
CO_2 measurement (parts per million)
Time (days since the first measurement)
415
410
405
400
395
20750
21000
21250
21500
21750
22000
22250
22500

Out[11]:


Residuals of CO_2 level Predictions - RMSE =
CO_2 level (parts per million)
Time (days since the first measurement)
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
0
5000
10000
15000
20000

3.2 Model Two

The example model proposed in the problem statement is constructed as follow:

- Long-term trend: quadratic $c_0 + c_1 \cdot t + c_2 \cdot t^2$

- Seasonal variation: cosine $c_3 \cdot \cos\left(\frac{2\pi \cdot t}{365.25} + c_4\right)$

- Noise: Gaussian of the form $N \sim (0,\ c_5)$

Combining all the three components gives the following likelihood function:

$$p\left(x_t \mid \theta\right) = N\left(c_0 + c_1 \cdot t + c_2 \cdot t^2 + c_3 \cdot \cos\left(\frac{2\pi \cdot t}{365.25} + c_4\right),\ c_5^2\right)$$

such as $\theta = \{c_0,\ c_1,\ c_2,\ c_3,\ c_4,\ c_5\}$

We can update our priors based on the results of the first model:

- c_0 : the intercept can be the first measurement in 1958) as a Normal with μ = 315 and σ = 5.

- c_1**:** the linear coefficient

- c_2**:** the quadratic coefficient

- c_3**:** the amplitude

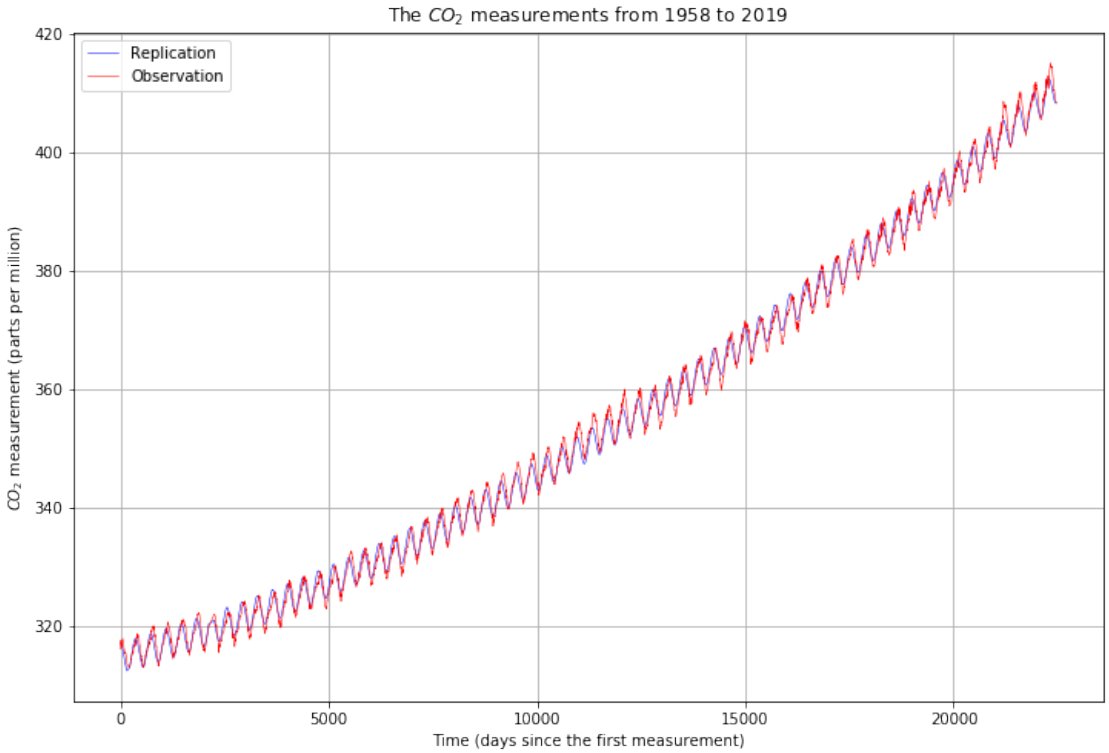
- c_4**:** the phase of the periodic function

- c_5**:** the variance of the noise is a Cauchy distribution

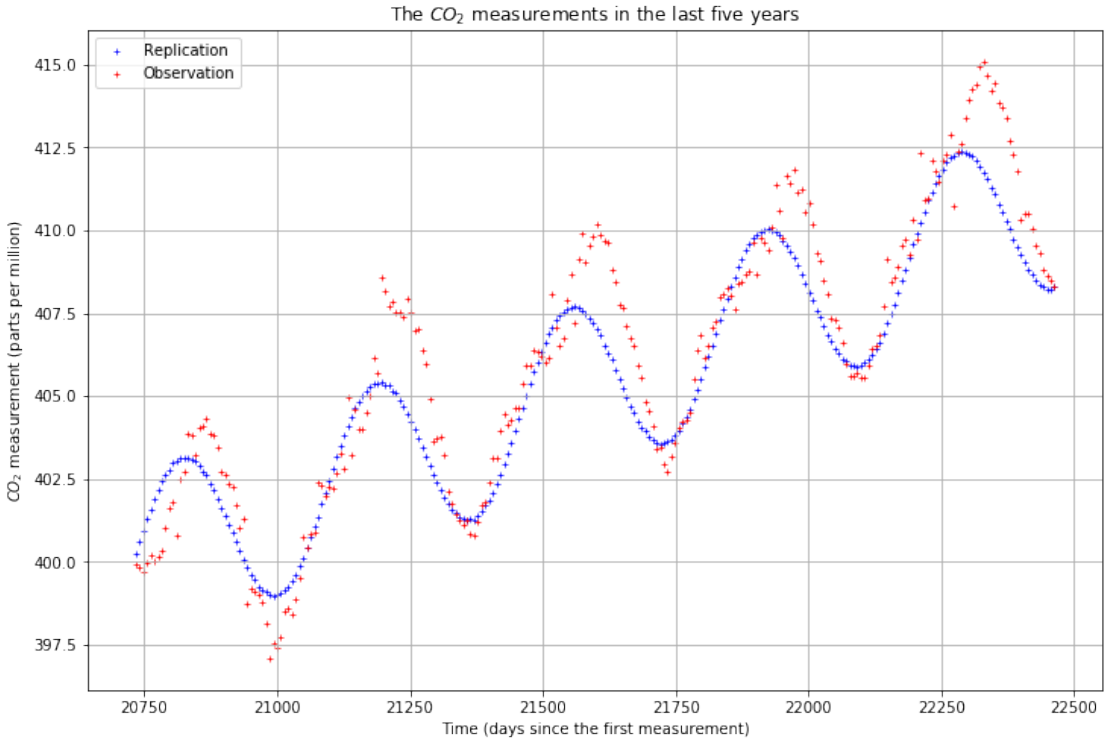
3.2.1 Results of Model Two

- The long term trend aligns well with the observations trend. The residual plot shows that the error is centered around zero throughout the measurements period with minor anomalies towards the center. Initially, the idea of trying an exponential model is suitable for this context since we're modeling the natural growth of CO_2 but the result of the quadratic model were good enough to be suitable for the modeling process.
- The Stan output shows that all parameters have an Rhat that is exactly equal to 1 as well as larger enough independent samples (in the order to thousands for most parameters). **However**, upon a close look at the five-years interval graph, we can notice that the model's values doesn't stretch enough to hit the tip of the periodic trend. In other words, the seasonality is not symmetrical and the rise of CO_2 is higher than the decrease in a one-year period. The gap is concerning because it's repetitive, thus, we can't argue that it is just noise and we have to investigate a better parametric model that would capture the gap.

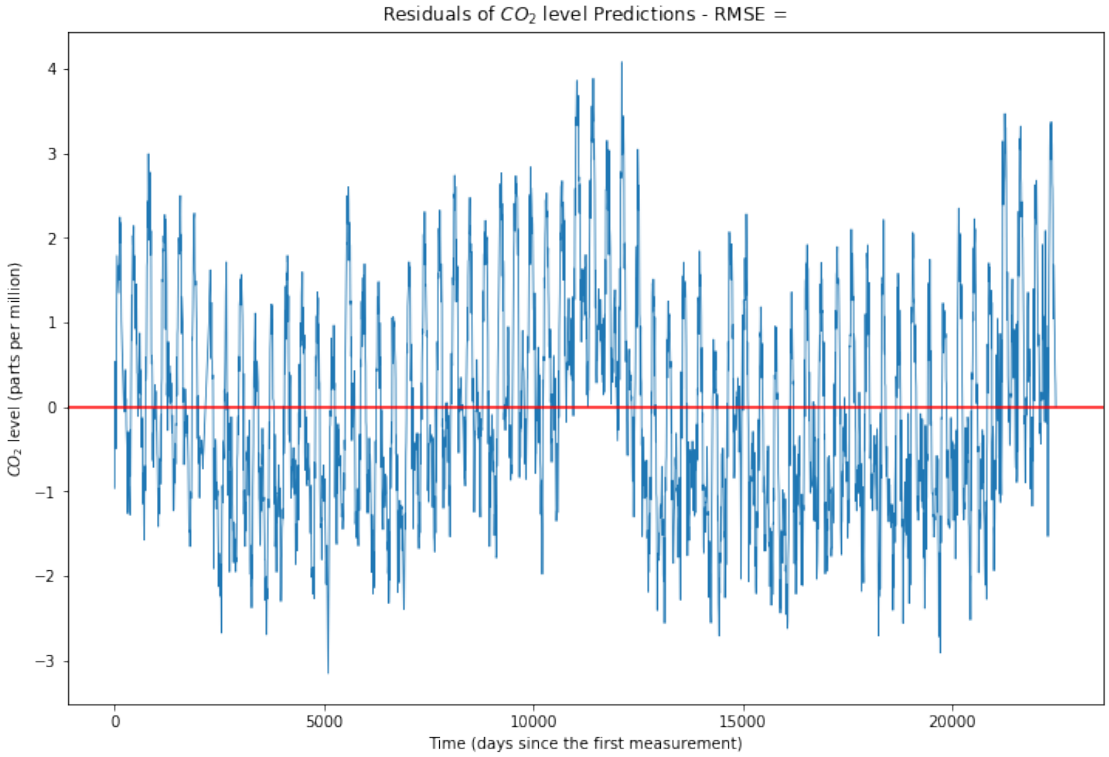
`Out[12]:`


The CO_2 measurements from 1958 to 2019
Replication
Observation
CO_2 measurement (parts per million)
Time (days since the first measurement)
420
400
380
360
340
320
0
5000
10000
15000
20000

Out[14]:


The CO_2 measurements in the last five years
Replication
Observation
CO_2 measurement (parts per million)
Time (days since the first measurement)
415.0
412.5
410.0
407.5
405.0
402.5
400.0
397.5
20750
21000
21250
21500
21750
22000
22250
22500

Out[15]:


Residuals of CO_2 level Predictions - RMSE =
CO_2 level (parts per million)
Time (days since the first measurement)
4
3
2
1
0
−1
−2
−3
0
5000
10000
15000
20000

3.3 Model Three

The example model proposed in the problem statement is constructed as follow:

- Long-term trend: quadratic $c_0 + c_1 \cdot t + c_2 \cdot t^2$

- Seasonal variation: sine $c_3 \cdot \sin\left(\frac{2 \cdot \pi \cdot t}{365.25} + \phi\right) + c_4 \cdot \sin\left(\frac{4 \cdot \pi \cdot t}{365.25} + \phi\right)$

- Noise: Gaussian of the form $N \sim (0,\ c_5)$

Combining all the three components gives the following likelihood function:

$$P\left(x_t \mid \theta\right) = \left(c_0 + c_1 \cdot t + c_2 \cdot t^2 + c_3 \cdot \sin\left(\frac{2 \cdot \pi \cdot t}{365.25} + \phi\right) + c_4 \cdot \sin\left(\frac{4 \cdot \pi \cdot t}{365.25} + \phi\right),\ c_5^2\right)$$

such as $\theta = \{c_0,\ c_1,\ c_2,\ c_3,\ c_4,\ c_5, \phi\}$

We can update our priors based on the results of the first model:

- c_0**:** the intercept can be the first measurement in 1958) as a Normal with μ = 315 and σ = 5.

- c_1**:** the linear coefficient as a Normal distribution centered at 0.

- c_2**:** the quadratic coefficient as a Normal distribution centered at 0.

- c_3**:** the amplitude of the first sine function as a Cauchy distribution.

- c_4**:** the amplitude of the second sine function as a Cauchy distribution.

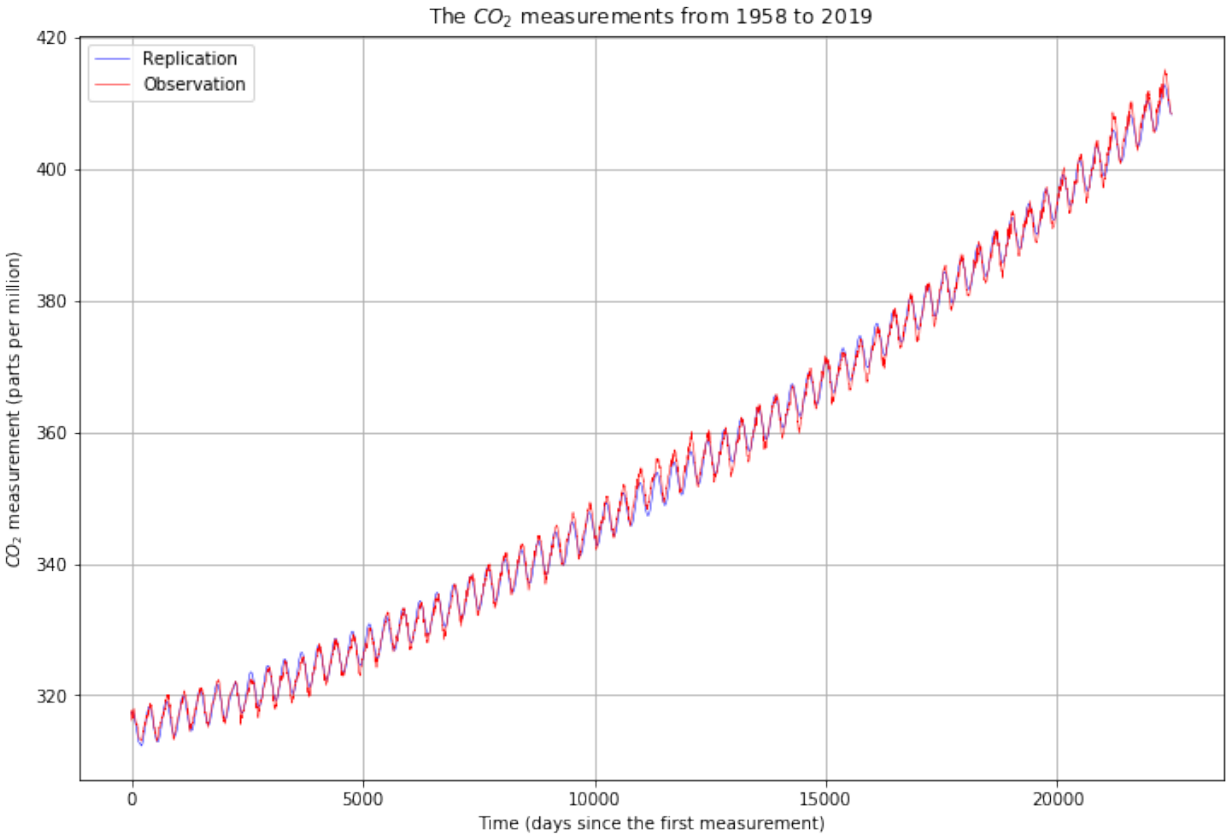
- c_5**:** the variance of the noise is a Truncated-Normal centered at 0.

- ϕ**:** the phase of both periodic functions as a Cauchy distribution.

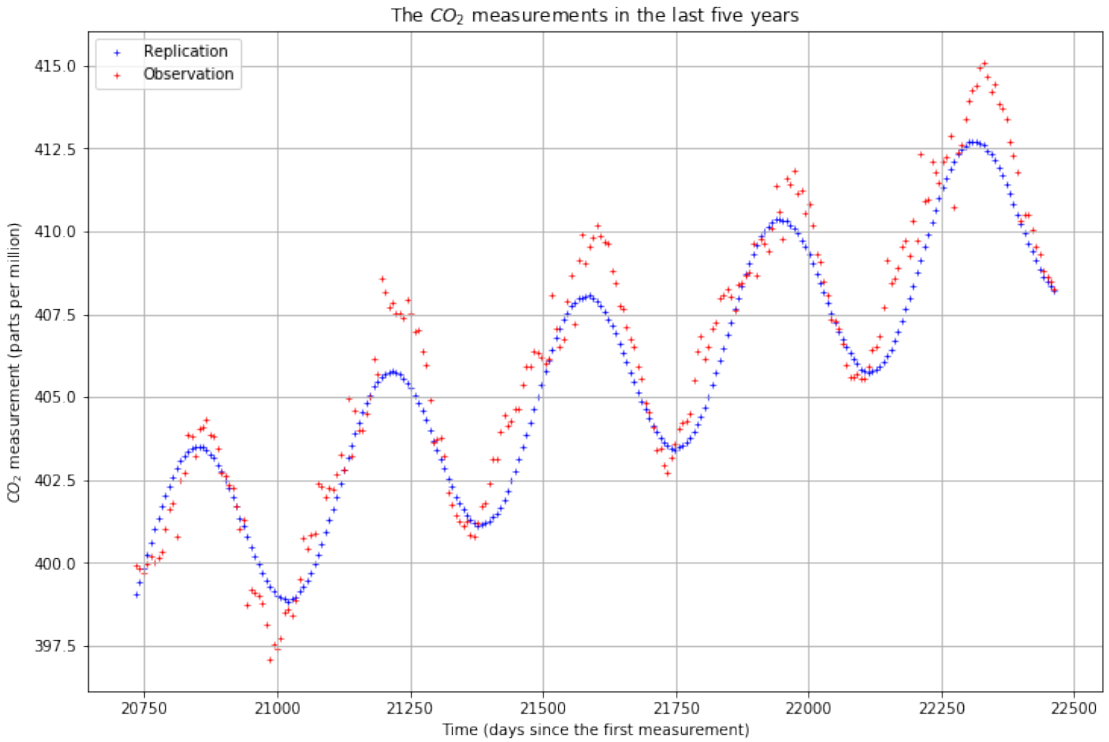
3.3.1 Results of Model Three

- The adjusted periodic function pictures two seasonality:
  1. The yearly seasonality: which represents the natural rise and fall of CO_2 concentration in the atmosphere. These changes are due to the vegetative cycle of trees.
  2. Semi-yearly: leaving the period of the second sinusoidal as a free parameter would distort the output of the Stan simulation. As a first attempt, the second period is set to represent semi-yearly patterns based on the belief that energy use is high during winter and autumn.
- The Stan model illustrated that all parameters had an Rhat that is exactly 1, furthermore, the independent samples (n_eff) were of the order of thousands for all parameters. As a result, the auto-correlation graphs (Appx. Model Three) shows no correlation between the samples.
- For this iteration, the model covers better the shape of the seasonal trend of the observations. The five-years plot shows that -within a period- the rise of the CO_2 is less steep but longer than the fall of CO_2 which is steeper but to shorter.

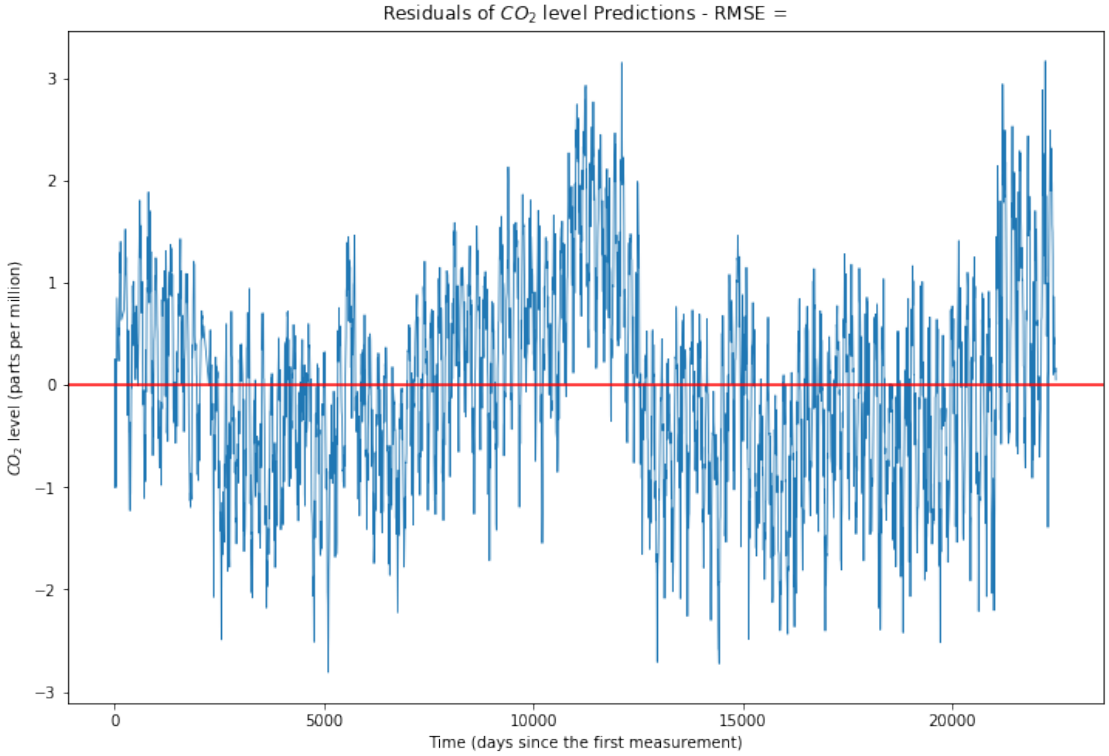
Out[16]:


The CO_2 measurements from 1958 to 2019
Replication
Observation
CO_2 measurement (parts per million)
Time (days since the first measurement)
420
400
380
360
340
320
0
5000
10000
15000
20000

Out[17]:


The CO_2 measurements in the last five years
Replication
Observation
CO_2 measurement (parts per million)
Time (days since the first measurement)

Out[18]:


Residuals of CO_2 level Predictions - RMSE =
CO_2 level (parts per million)
Time (days since the first measurement)

3.4 Stan Output of Model Three

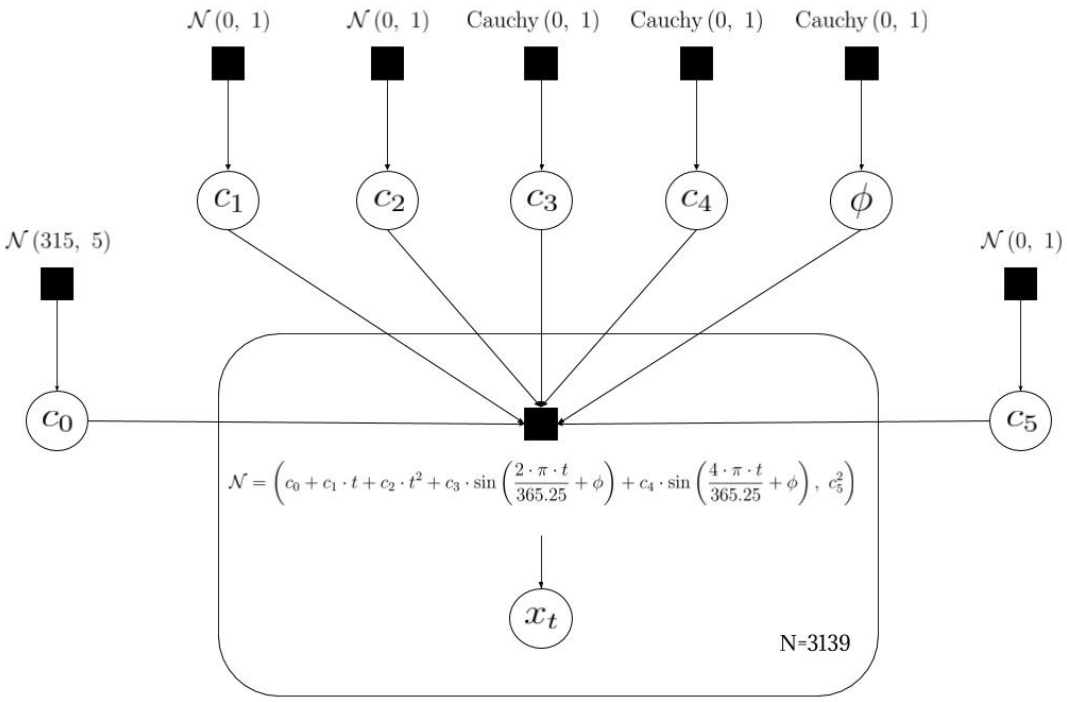
Inference for Stan model: anon_model_42637ee5484d5134fe0f739f9b3d4c36.
4 chains, each with iter=2000; warmup=1000; thin=1;
post-warmup draws per chain=1000, total post-warmup draws=4000.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
c_0	314.57	1.1e-3	0.05	314.46	314.53	314.57	314.61	314.67	2667	1.0
c_1	2.1e-3	2.3e-7	1.1e-5	2.1e-3	2.1e-3	2.1e-3	2.1e-3	2.1e-3	2294	1.0
c_2	9.7e-8	9.5e-12	4.7e-10	9.6e-8	9.7e-8	9.7e-8	9.7e-8	9.8e-8	2427	1.0
c_3	2.86	4.4e-4	0.03	2.81	2.84	2.86	2.88	2.91	3327	1.0
c_4	1.2e-3	1.9e-5	1.2e-3	3.2e-5	3.7e-4	8.6e-4	1.7e-3	4.5e-3	4128	1.0
c_5	0.99	2.0e-4	0.01	0.96	0.98	0.99	0.99	1.01	3774	1.0
ϕ	1.15	1.5e-4	8.8e-3	1.14	1.15	1.15	1.16	1.17	3305	1.0

Samples were drawn using NUTS at Thu Dec 19 19:58:29 2019.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains
(at convergence, Rhat=1).

3.5 Factor graph of Model Three

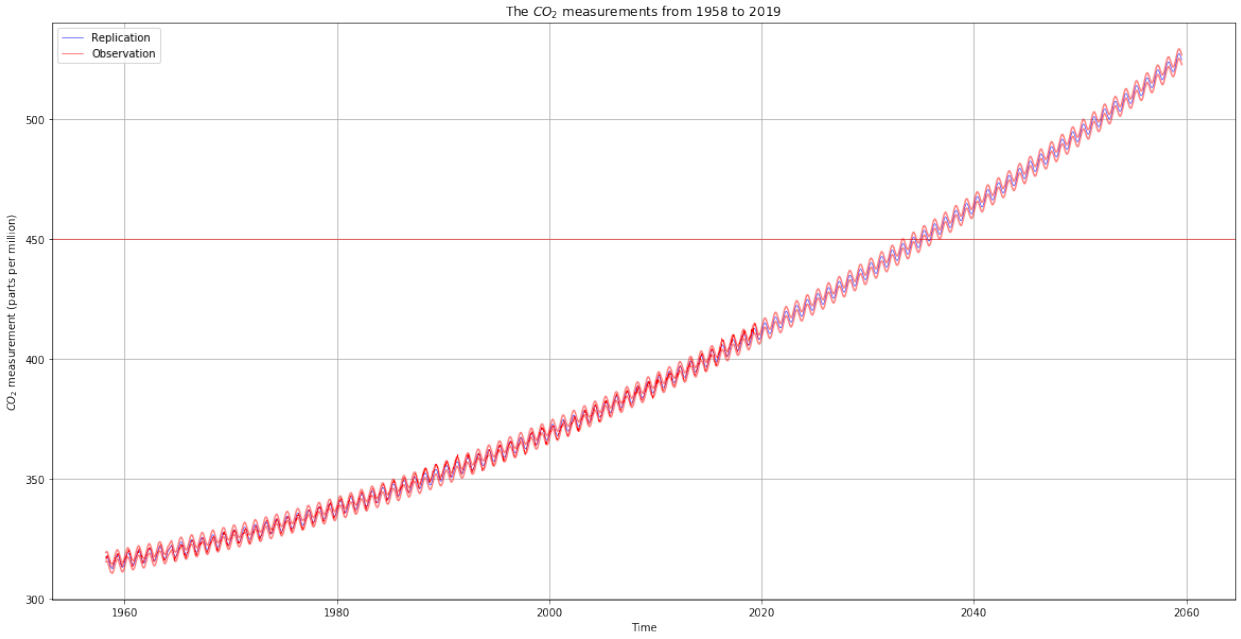
Out[24]:



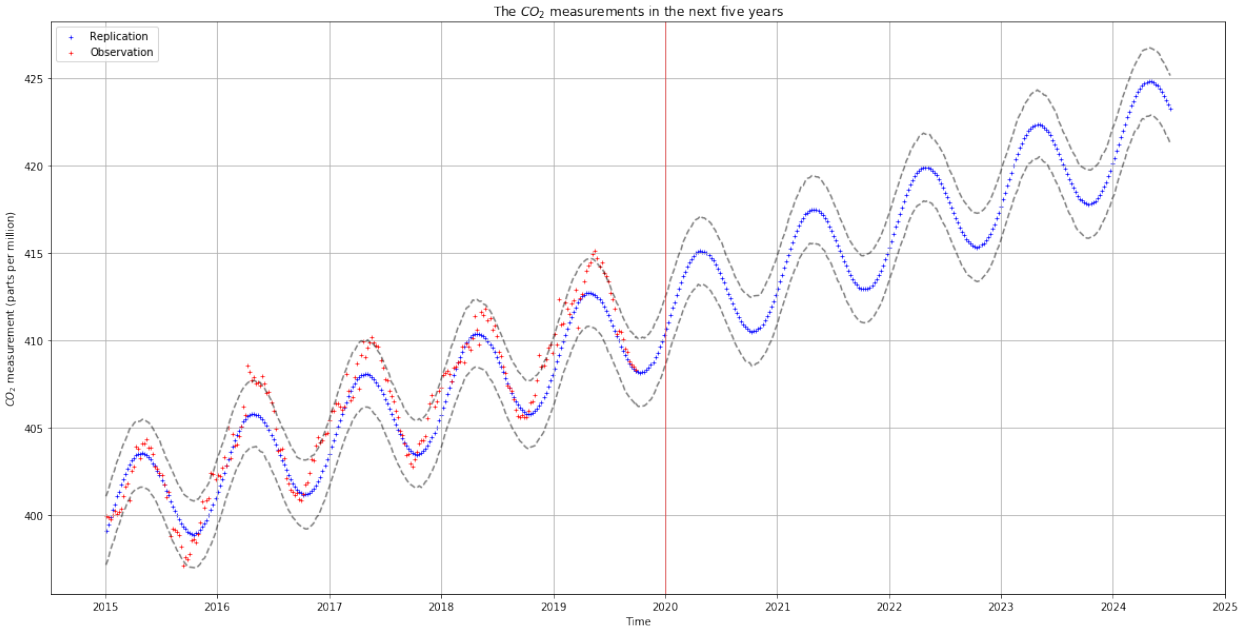
4 Future predictions and confidence intervals

The model constructed for the prediction is a result of the mean of the generated samples of each of the parameters. This methodology would mean that the confidence interval is not dependent on time but on the parameters chosen for the model. Therefore, we notice that the confidence interval doesn't stretch over time, despite being the plausible scenario in this case. Nonetheless, we can trace the value at which the level of CO_2 in the atmosphere is considered to be dangerous. According to the prediction graph, the 450 ppm level hits the upper bound of the projected level by March 2033. Then after one year, the predictive model hits 450 ppm as a mean in the third week of March 2034.

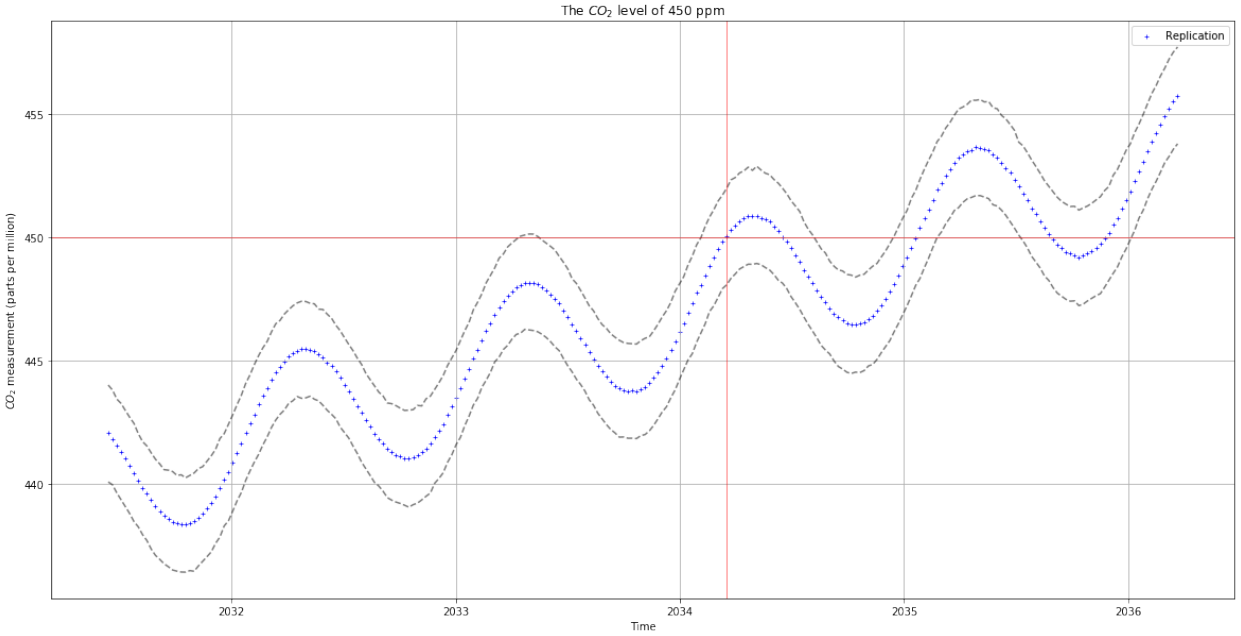
Out[26]:



Out[27]:


The CO_2 measurements in the next five years
Replication
Observation
CO_2 measurement (parts per million)
Time
425
420
415
410
405
400
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025

Out[28]:


The CO_2 level of 450 ppm
Replication
CO_2 measurement (parts per million)
Time
455
450
445
440
2032
2033
2034
2035
2036

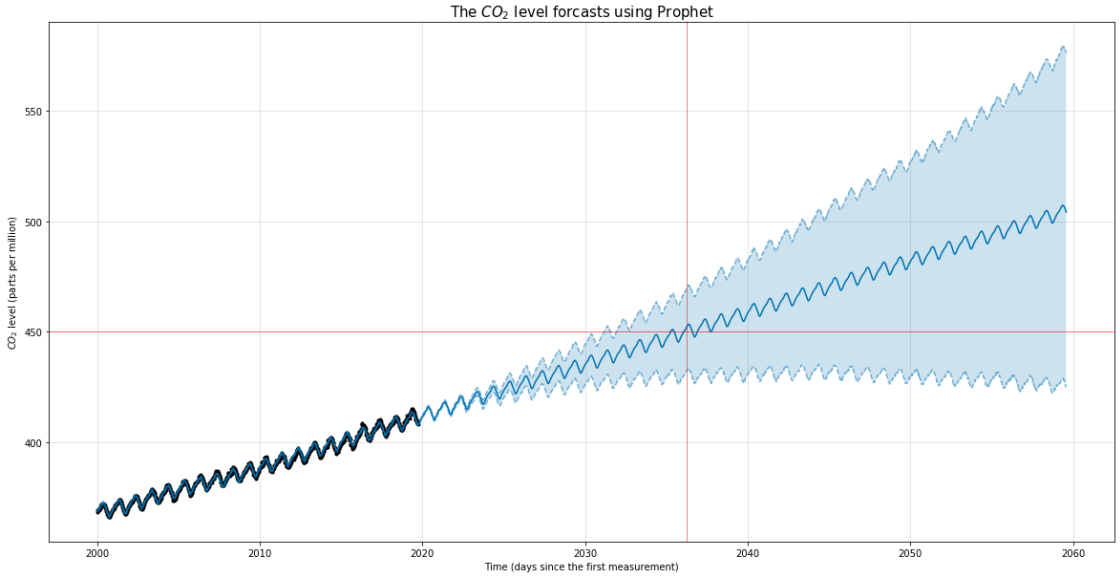
5 Prophet by Facebook

Prophet is a time series forecasting package and its back-end is implemented in Stan [6]. As an alternative to the Stan model constructed above, we can exploit the finding of this simulation to strengthen/question our previous results. Prophet uses the following methodologies to model long-term trend and seasonality:

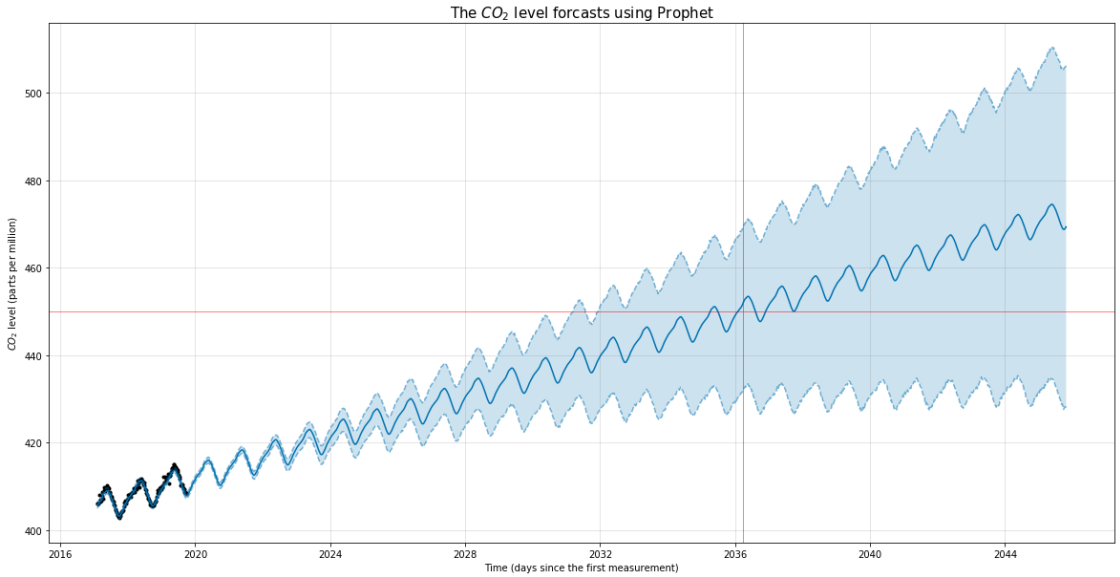
- **Non-linear trend:** Prophet uses a logistic growth model. At each specific moment in time, the change in growth rate is modeled using a vector of rate adjustments.

- **Seasonality:** Prophet models seasonality using a standard Fourier series.

Out[29]:



Out[31]:



5.0.1 Results of Prophet Model

Based on the prediction by Prophet, the anticipated date for CO_2 levels to hit dangerous level is by the end of March 2036 (two year later than the previous model). However, the upper bound of the Prophet model is expected to hit the 450 ppm mark as early as June 2031. In this model, the confidence intervals stretch over time, giving more window for the worst and best case scenario.

6 Inference and further improvements

Keeling's measurements are one of the strongest evidence of the impact of human activity on CO_2 levels. The historical data pictures the rising trend over more than 60 years, but people seem to misinterpret seasonality to think that Climate Change is not a reality. Having cold winters is not an argument that the planet is fine, but we need to examine the trend relative to previous winters (same applies for summers). The urgency of this problem doesn't only lay on its complexity but also on its irreversible nature. The analysis of the predictive models shows that we're on the verge of hitting critical levels of CO_2 by the beginning of the third decade of the 21st century which is in 10 years. The level of confidence grows significantly to almost 100% by 2035.

Further improvement on the model can include the time parameter in the confidence interval formula to picture the rise on uncertainty over time. Furthermore, we can check the validity of the measurements by fitting the model through another data measurements from a different observatory. The expectation is that the consistent/global trends are going to appear in all CO_2 observations regardless of the location of the measurements.

Finally, as worrying this situation might be, I'm confident that in light of the recent climate strikes that took place in over 150 countries, the phenomenal widespread of the #trashtag challenge, and naming Greta Thunberg as TIME person of the year 2019, the rise of environmental consciousness will empower us, humans, to find a way out (not from this planet hopefully).

7 Appendix

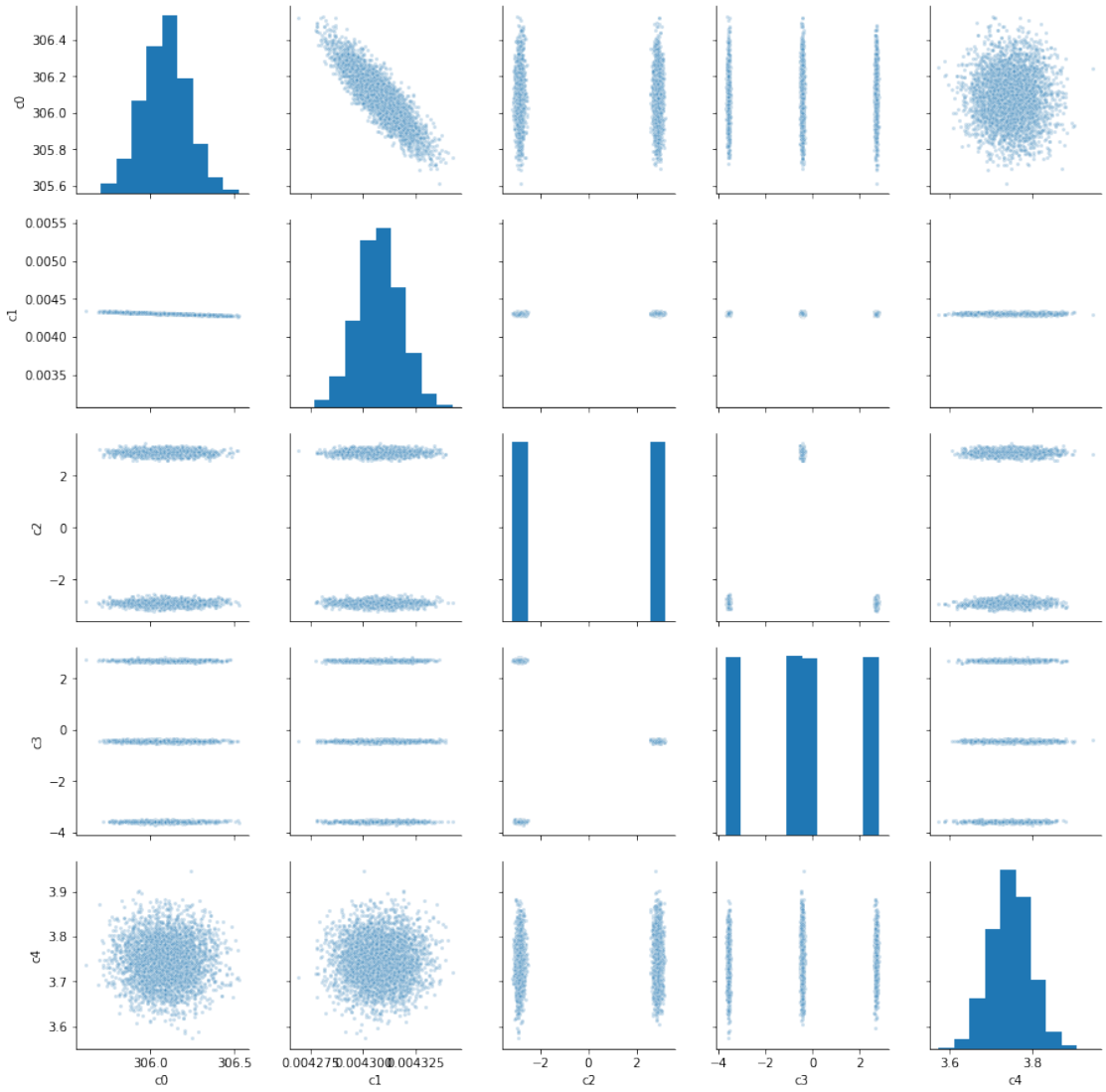
The Jupyter notebook contains the outputs of the Stan simulation as well as the auto-correlation plots for each of the three models suggested above. Link: https://github.com/Tahahaha7/Modern_Computational_Statistics/blob/master/Forecasting%20Carbon%20Dioxide/%5BCS146%5D%20Final%20Project%20Code.ipynb

7.1 Pair plots

7.1.1 Model One

```
In [20]: Image.open('pairp_model1.png')
```

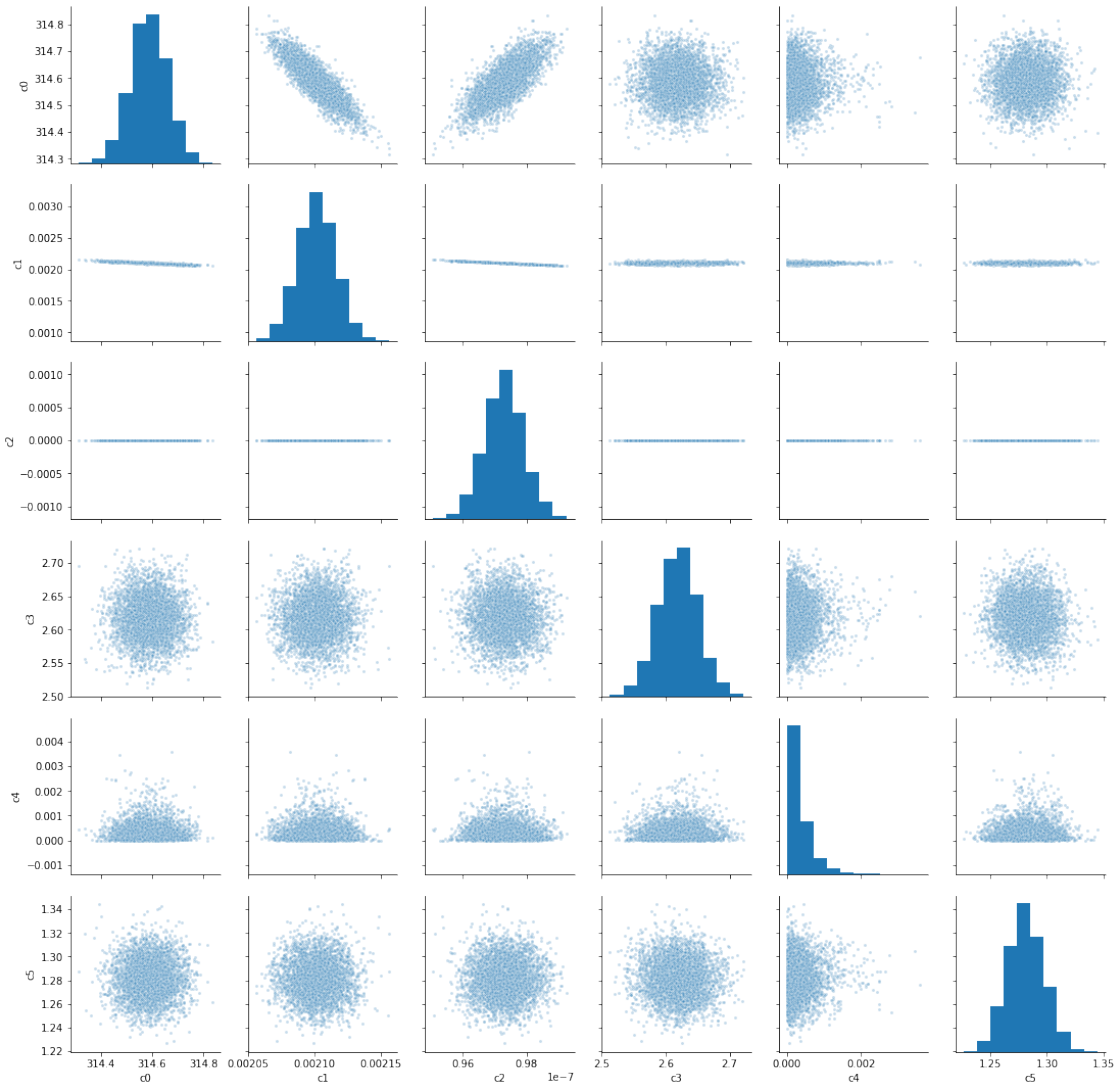
Out[20]:



7.1.2 Model Two

```
In [21]: Image.open('pairp_model2.png')
```

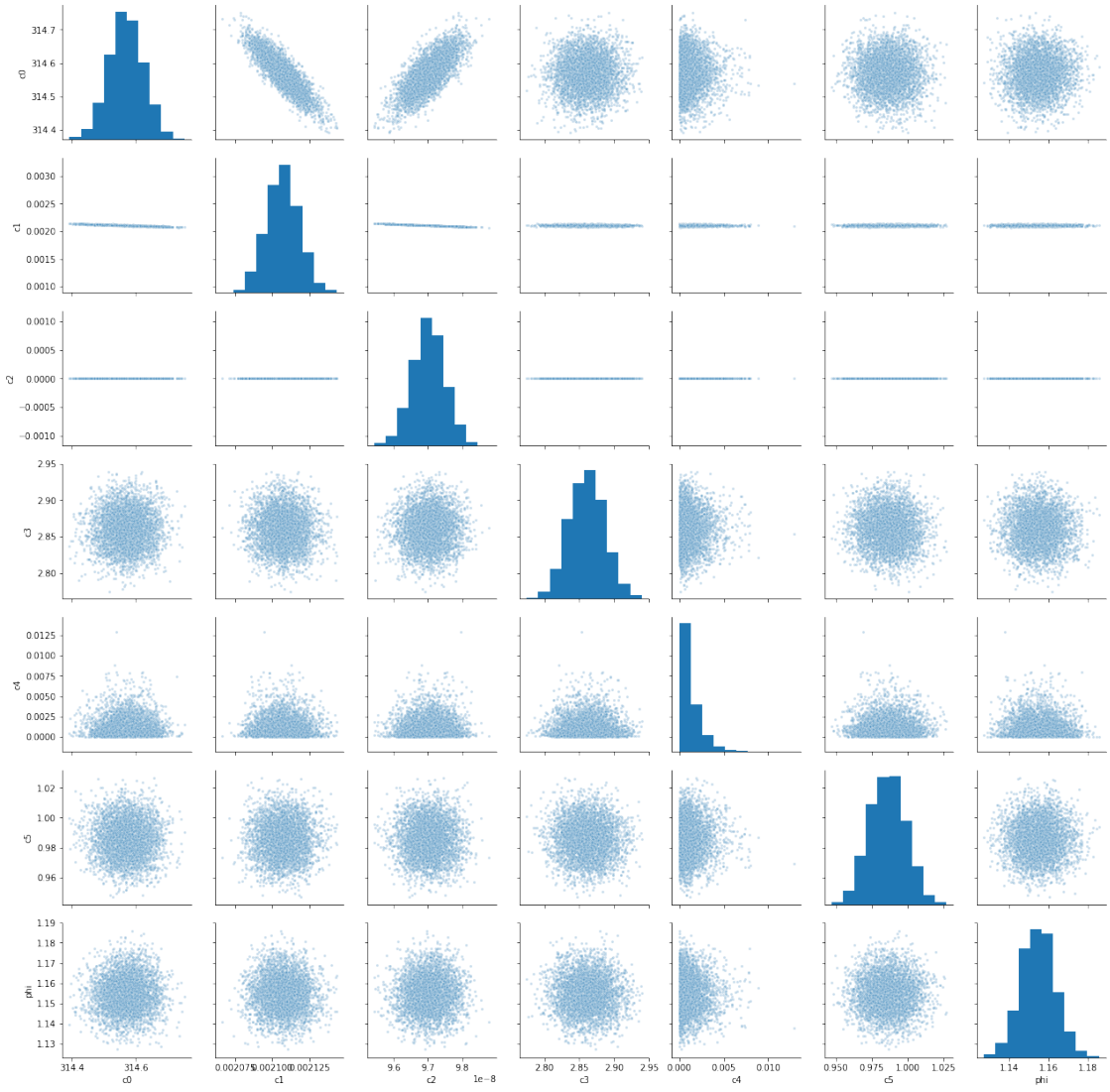
Out[21]:



7.1.3 Model Three

```
In [22]: Image.open('pairp_model3.png')
```

Out[22]:



8 HC Applications

1. **sampling**: Using the Stan package to sample from the posterior distribution and generate the parameters that best fit the observations. The process also includes the interpretation of the Stan output in terms of the independence of samples which is the goal of using a Hamiltonian Monte Carlo

2. **confidenceintervals**: Constructing the confidence intervals of the predictive model based on the CI of the parameters generated by Stan. The process however shows that the confidence intervals are independent of time, whereas in reality, we expect it to widen over time as illustrated by the model from Prophet.

3. **descriptivestats**: Interpreting the results of the Stan model and relating the findings to the auto-correlation and pair plots as well as evaluating the fitness of the model and independence of the samples.

4. **gapanalysis**: In the process of approximating the observations' graph, gap analysis was used to imitate the seasonal and long-term trend. The initial stage was the model proposed in the prompt, the final stage was Model Three which is a result of tweaking and adjustments.

9 References

1. Scheffler C., (2019). CS146 Final Project: Modeling and forecasting atmospheric CO_2 from 1958 until 2058. Retrieved from: https://course-resources.minerva.kgi.edu/uploaded_files/mke/00148349-5633/cs146-final-project.pdf
2. American Museum of Natural History (2014). Science Bulletins: Keeling's Curve – The Story of CO_2. YouTube video retrieved from: https://www.youtube.com/watch?v=0Z8g-smE2sk
3. SCRIPPS Institution of Oceanography (2019). The Keeling curve. Retrieved from: https://scripps.ucsd.edu/programs/keelingcurve/
4. Stan Development Team (2019). Stan Functions Reference: Version 2.21. Retrieved from: https://mc-stan.org/docs/2_21/functions-reference/index.html
5. Prophet (2019). Forcasting at scale. Facebook Inc. Retrieved form: https://facebook.github.io/prophet/
6. Shioulin (2017). Taking Prophet for a spin. Cloudera Fast Forward Labs is a machine intelligence research company. Retrieved from: https://blog.fastforwardlabs.com/2017/03/22/prophet.html
7. The PyMC Development Team (2018). PYMC3 documentation example: CO2 at Mauna Loa. Retrieved from: https://docs.pymc.io/notebooks/GP-MaunaLoa.html