Modeling and forecasting CO2 from 1958

Introduction

The CO2 level is increasing over time since the industrial revolution in the 18th century. In the 1960s the CO2 growth rate is around 0.6 ppm per year, while 2010s, the growth rate surges to 2.3 ppm per year. This data set is the atmospheric measurement from Mauna Loa Observatory rather than the ice core measurements that are performed traditionally.

In this project, we want to construct a Bayesian model for CO2 level growth and predict CO2 level for the next 40 years till 2060.

Data

The dataset comes from the weekly Mauna Loa data set from the Scripps CO2 program. It contains the date since 1958 and the level of carbon in PPM. The graph below represents the trend of the level of carbon change over time. From the graph, we can see that a periodic trend exists in the model, which represents seasonality. In general, there is more carbon dioxide in the winter and a bit less in the summer, because plants accumulate carbon in summer and release it in winter.

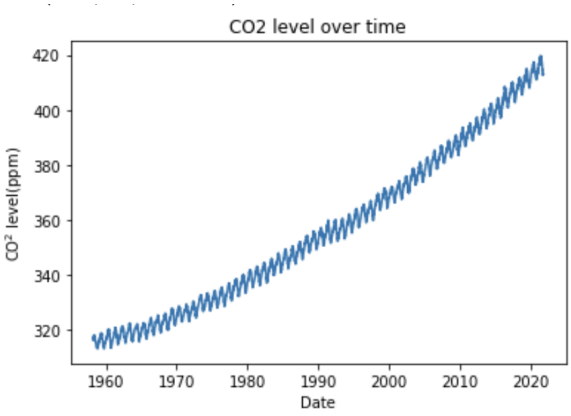

Figure 1. CO2 level growth from 1958 to 2020

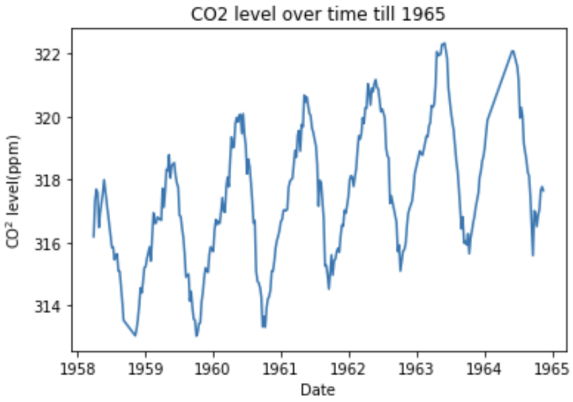

Figure 2. CO2 level growth from 1958 to 1965

Modeling

To build the model, we engineer a new feature, `days_since` that represents the number of days since 1958 March 29th.

date	ppm	days_since
1958-03-29	316.19	0
1958-04-05	317.31	7
1958-04-12	317.69	14
1958-04-19	317.58	21
1958-04-26	316.48	28

Figure 3. CO2 data

We have an initial model

- Long-term trend: linear, $c_0 + c_1 t$
- Seasonal variation (every 365/4 days): cosine, $c_2 cos(\frac{2\pi t}{365.25} + c3)$
- Noise: Gaussian with mean 0 and fixed standard deviation, sigma
- The variables are all unobserved parameters of the model, ci

Combining these three components gives the following likelihood function

$$p(x_t|\theta) = N(c_0 + c_1 t + c_2 cos(\frac{2\pi t}{364.25} + c_3), \sigma^2)$$

where θ represents the set of all unobserved parameters.

However, we assume this model to be flawed because from Figure 1. above, we can see that CO2 seems to be a quadratic growth rather than a linear growth. Therefore, we add the quadratic term c4, over time.

$$p(x_t|\theta) = N(c_0 + c_1 t + c_4 t^2 + c_2 cos(\frac{2\pi t}{364.25} + c_3), \sigma^2)$$

Priors

For the constraints for parameters, we set them based on our observation in Figure 1.

- the constant term, c_0: It is positive because we can see an intercept above 300, so we set the Cauchy distribution that has a mean of 300.
- the linear term, c_1: It is positive because we observe a positive trend.
- the quadratic term, c_4: Since the trend is roughly concave up, c_4 should be positive.
- amplitude, c_2: Since seasonality exists, the amplitude is positive. We find that if the amplitude is smaller than 1, the seasonality effect will be ignored. Hence, we set the minimum value as 1.
- sigma, σ: Sigma is the standard deviation that is positive by default.

```
parameters {
    real<lower=0> c0; //intercept
    real<lower=0> c1; //coef for linear term
    real<lower=0> c2; //amplitude
    real<lower=0, upper=1> c3; //constant within the periodic term
    real<lower=0> c4; //coef for quadratic term
    real<lower=0> sigma;
```

Figure 4. parameter settings for the model

In addition, we notice that taking out limits will decrease the run time but drastically deteriorate the model quality.

For example, if we only add `<lower=0>` to intercept (c0), amplitude (c2), sigma, and c3, the result is wacky and we have a low effect size(Model version 8).

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff
c0	2.64	1.73	2.44	0.28	0.29	2.18	5.09	5.93	2
c1	0.15	0.71	1.01	-1.42	-0.78	0.41	1.03	1.21	2
c2	1.36	0.67	0.95	0.28	0.48	1.23	2.28	2.72	2
c3	0.52	0.17	0.24	0.14	0.29	0.6	0.73	0.73	2
c4	-7.0e-6	3.8e-5	5.4e-5	-6.4e-5	-5.4e-5	-2.1e-5	4.3e-5	7.8e-5	2
sigma	2.7	1.91	2.7	0.35	0.87	1.58	4.75	7.29	2

Figure 5. Parameter output model version 8

Backcasting & Forecasting

We use the model to perform backcasting and forecasting.

- Backcasting: We use the model to backcast the data we already observe and check if the model fits well.
- Forecasting: 40 more years till 2060.

Results

Our result shows that the intercept, c_0 , and amplitude, c_2 is the dominant term for prediction.

- c_0 is 314.75. This indicates on 1958 March 29th, we have a 314.75 CO2 level in ppm.
- c_1 implies that every day CO2 level increase by 0.002 ppm.
- c_2, the maximum variation among different seasons, is 2.63 ppm.
- c_3, the adjustment of the seasonality period is 0.0034.
- c_4 represents that the quadratic growth overtime is 0.0000001.
- Sigma: The standard deviation for noise, sigma, is 1.28 pm.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
c0	314.75	1.4e-3	0.07	314.62	314.71	314.75	314.8	314.89	2353	1.0
c1	2.0e-3	3.1e-7	1.3e-5	2.0e-3	2.0e-3	2.0e-3	2.1e-3	2.1e-3	1878	1.0
c2	2.63	2.5e-3	0.03	2.57	2.61	2.63	2.65	2.69	168	1.04
c3	3.3e-4	1.3e-5	3.3e-4	1.4e-5	9.6e-5	2.3e-4	4.7e-4	1.2e-3	636	1.0
c4	1.0e-7	1.3e-11	5.6e-10	9.9e-8	10.0e-8	1.0e-7	1.0e-7	1.0e-7	1817	1.0
sigma	1.28	3.3e-4	0.02	1.25	1.27	1.28	1.29	1.31	2484	1.0

Figure 6. Parameter output for the final model

The graph below showcases our prediction. We can see that 95% interval captures the fluctuation and growth of CO2 level. Since we know that CO2 levels of 450 ppm are considered high risk for dangerous climate change, we can see this will happen by 2040, if actions aren't taken to address CO2 emission.

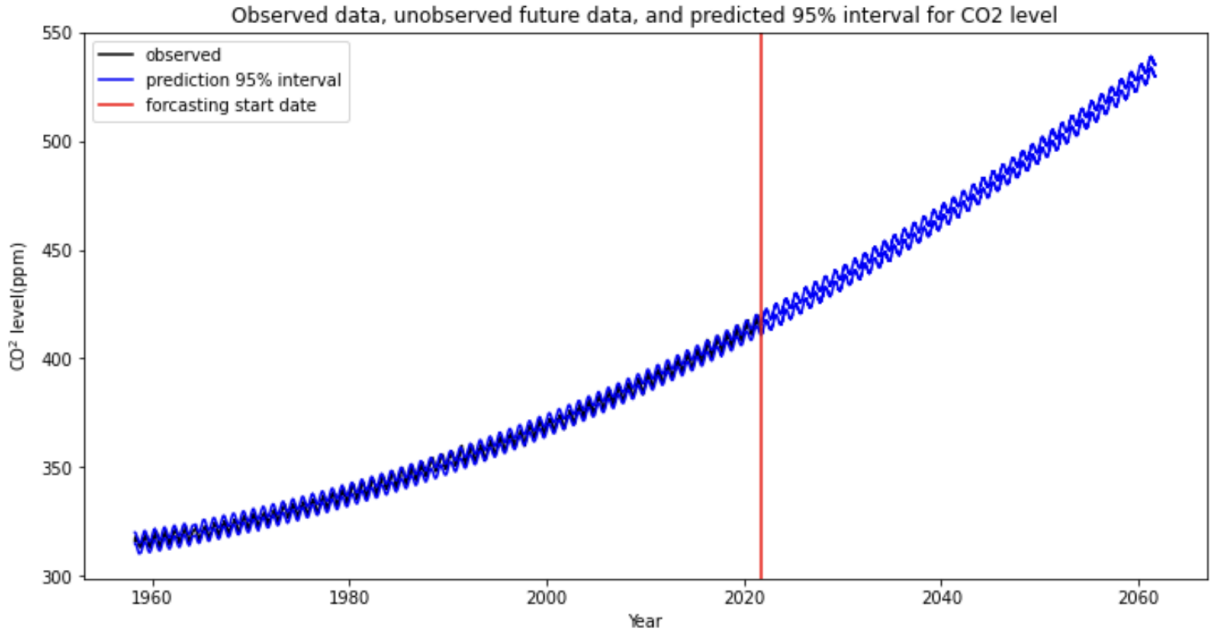

Figure 7. CO2 level prediction till 2060

Assessing the quality of the model

In Figure 6., we use `n_eff` and `Rhat` as the main indicator to evaluate our model. For `n_eff`, we need hundreds of samples to present the numbers of effective samples is enough. For `Rhat`, we prefer it to be close to 1 which means that different Markov Chains merge well and have a similar result that's why the ratio is close to 1.

We can see that we have satisfying results for `n_eff` for all parameters except c_2, and it's `Rhat` isn't the best too.

Autocorrelation

We observe the autocorrelation for parameters. From the graphs, we can see the autocorrelation is close to 0, meaning that our samples are independent of each other.

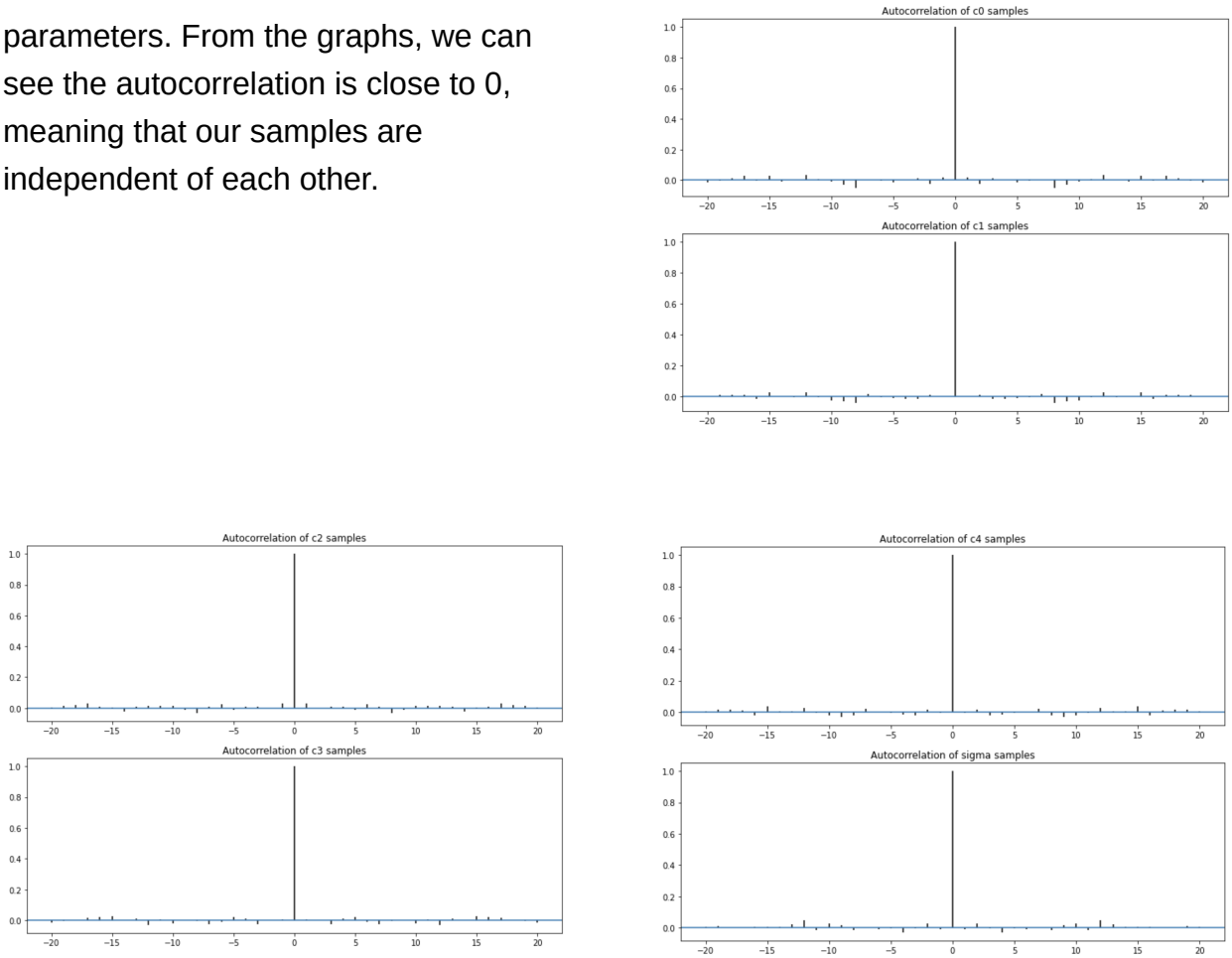

Figure 8. Autocorrelation plot for parameters (c0, c1, c2, c3, c4, sigma)

Pairplot

We include a pair plot for all the parameters. If our sampling is non-biased, we will ideally get random scatters between different pairs of parameters.

Here, we can see that $c_3$is right-skewed toward 0, and its mean is almost 0 too. The effect of c_3 might be insignificant that most of the value-centered towards 0.

In addition, we see negative correlation between $c_1$and c_0, c_0 and c_4, and $c_1$and c_4 that worth future investigation.

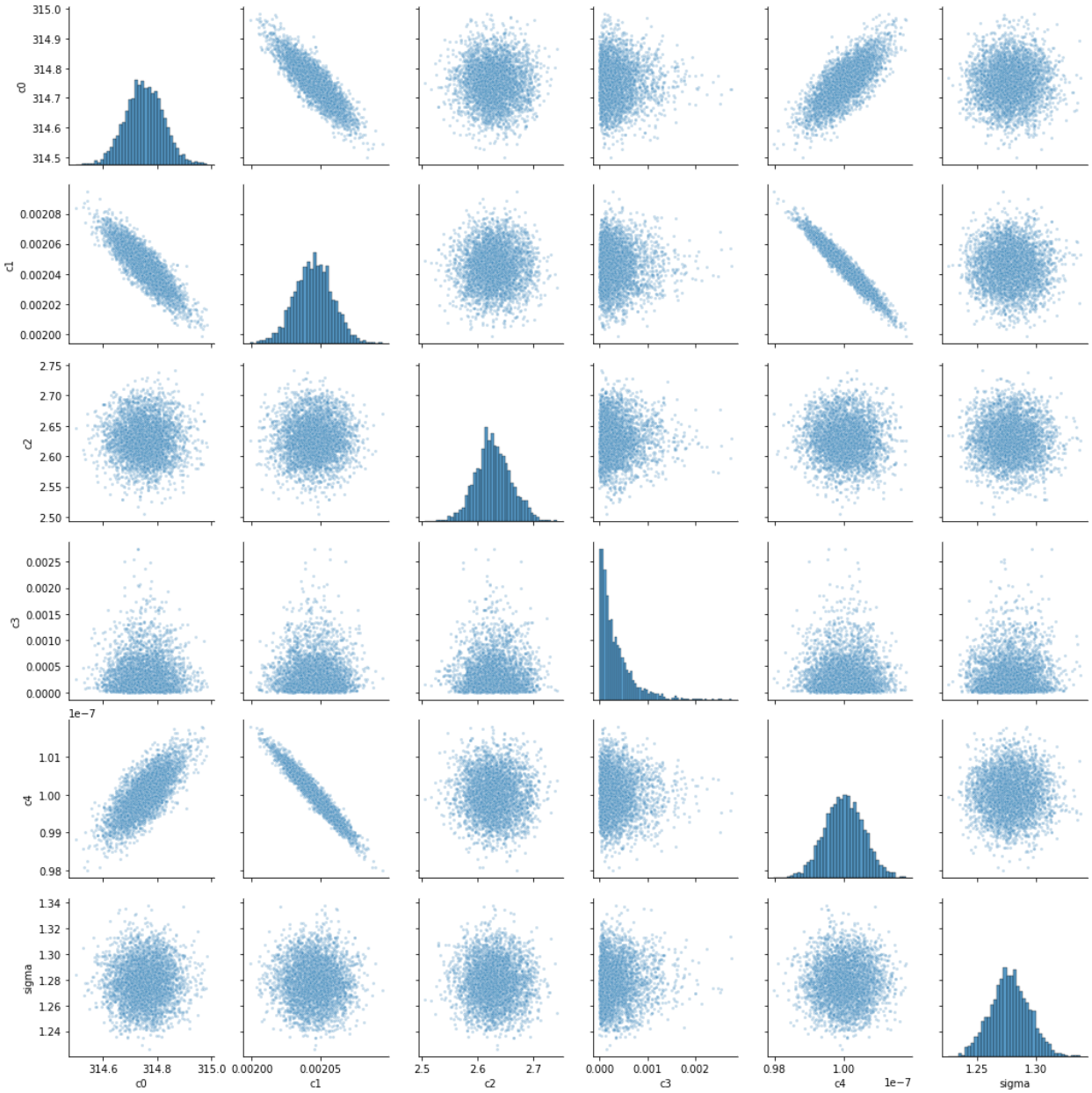

Figure 9. Pairplot for parameters (c0, c1, c2, c3, c4, sigma)

Limitation

Though the model successfully captures the CO2 level growth before 2020, it might not be able to predict 2060 accurately due to several limitations. For instance, due to COVID, global CO2 emissions decreased by -17%, which might affect the CO2 level. In addition, with the raising awareness of global warming, CO2 emission policies might be changed that slow down the growth rate.

Appendix

All the 8 testing models can be found: here

Python notebook for the final model: here

References

- When seasonality meets Bayesian: Decomposing seasonalities in Stan: here
- Climate Change: Atmospheric Carbon Dioxide: here
- Temporary reduction in daily global CO2 emissions during the COVID-19 forced confinement: here
- Civil disobedience movements such as School Strike for the Climate are raising public awareness of the climate change emergency: here
- CO2 – Why 450 ppm is Dangerous and 350 ppm is Safe: here