Atmospheric Carbon Dioxide Level Modelling

Hackl, Michelle

CS146, Prof. Scheffler

Table of Contents

Abstract

In this assignment I am modelling atmospheric CO_2 levels for the next 40 years, based on past data recorded by the Mauna Loa Observatory in Hawaii, available here. The observatory has been recording atmospheric data in weekly intervals since March 1958. Using past data, we can make predictions about future CO_2 level changes and anticipate when they will reach critical levels to urge policymakers to take action before they are surpassed[1].

Suggested Model

Let's assume that our data is expressed as x_t, where x is the level of CO_2 in the atmosphere measured in pp (parts per million) and t is the number of years since beginning of the recording in decimal. The simplest suggested model was defined by a long-term linear trend and seasonal cosine variation with gaussian noise. Parameterized as such:

$$p(x_t|\Theta) = \mathrm{N}(c_0 + c_1 t + c_2 \cos\left(\frac{2\pi t}{364.25} + c_3\right), c_4)$$

[1] Let's not get into the possibility of runaway greenhouse effects…

Model Extension - Overview

Assuming linearity for the long-term trend is a simplifying assumption that I have found to generate a poor fit to the data. As comparative models I have tested an exponential and polynomial trend, both simple but able to encode the assumption that CO_2 levels are increasing at an increasing rate. For seasonal variation, a cosine model might capture the general seasonal trend, but from closer inspection of the data, there seems to be a skew to the seasonal variation, with peaks being somewhat shifted to the right (Fig. 1). In fact, while the periodic aspect of the data might be modelled well by a cosine model, such a model might be all-together too smooth to account for the high peaks and valleys of the data. While I attempted to work with sawtooth wave models to capture this trend, I found them too difficult to implement in Stan and hence stuck with the simpler models.

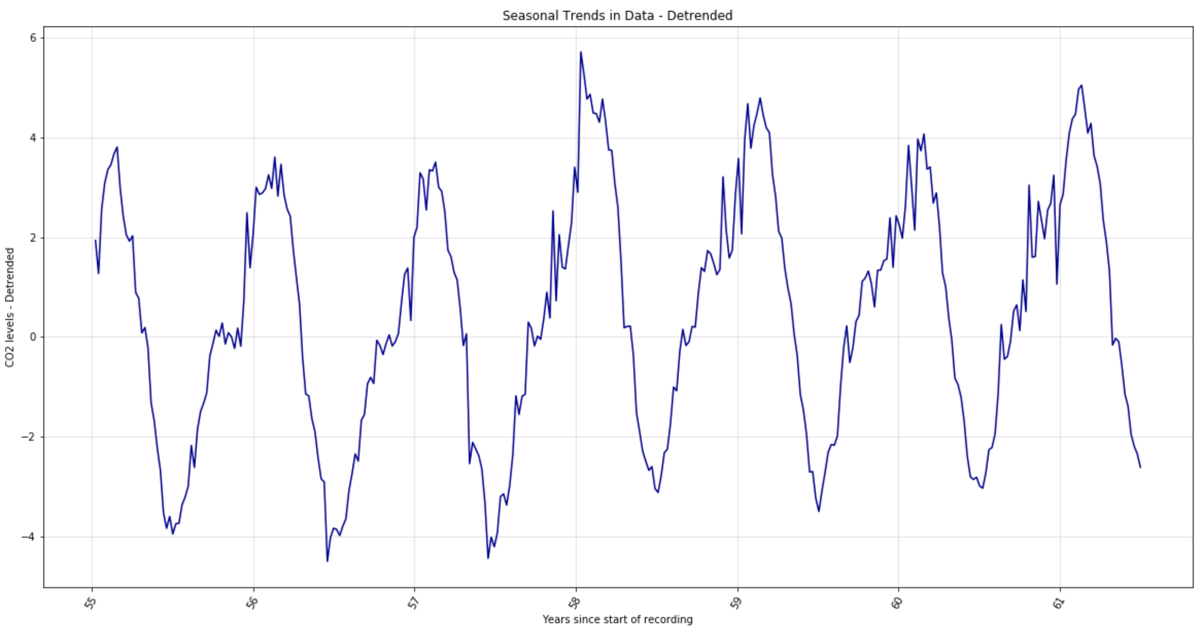


Figure 1. Detrended (polynomial trend model) data showing seasonal variation peaks being right-shifted.

Assumptions. To simplify the modelling approach, I assumed that the parameters of the long-term trend and the seasonal variation can be modelled separately and combined into a single predictive function afterward. This allowed me to run smaller models with less parameters at a time, greatly simplifying the process. To achieve this, I modelled long-term trend first, then used the model residuals as input for the seasonal model.

Priors. Generally speaking, I tried to use simple, broad priors that encode as little bias as possible, but chosen to fit with the constraints placed on the parameters:

- **Beta(1,1)** – This was the chosen uniform prior for parameters that have restrictions between 0 and 1.
- **Gamma(α, β)** – Gamma priors with varying parameters were chosen for parameters that had a constraint to be no less than 0. The α-parameter was chosen to reflect my prior assumptions about what the parameter would be, the scale parameter to match the scale of α.
  i.e. the second parameter of the polynomial model that will be squared can be expected to be rather small, so I made the prior Gamma(0.1, 0.1) while amplitude of the sine model is expected to be around 4, so the chosen prior was Gamma(4, 2).
- **Inverse Gamma(α, β) & N(μ, σ2)** – Those were the priors chosen for the standard deviation of the noise, which was assumed to be positive in all models. I chose an inverse gamma function for situations where the expected standard deviation of the noise was close to 0 and a normal distribution where it was far enough from 0 such that sampling values <0 would not be a concern.

Final Model Visualization. Given the results presented in the following section, I have chosen the final model to be a combination of a second-order polynomial long-term trend and cosine seasonal variation. A factor graph showing choices for priors over all parameters can be seen in Figure 2. Note that h_1 and h_2 are helper-parameters to define c_4.

The final model is parameterized as follows:

$$p(x_t|\Theta) = N\,(\,c_0 \,+\, c_1\,t \,+\, c_2 t^2 \,+\, c_3\,sin(\frac{2\pi t}{365.25} \,+\, c_4),\, c_5)$$

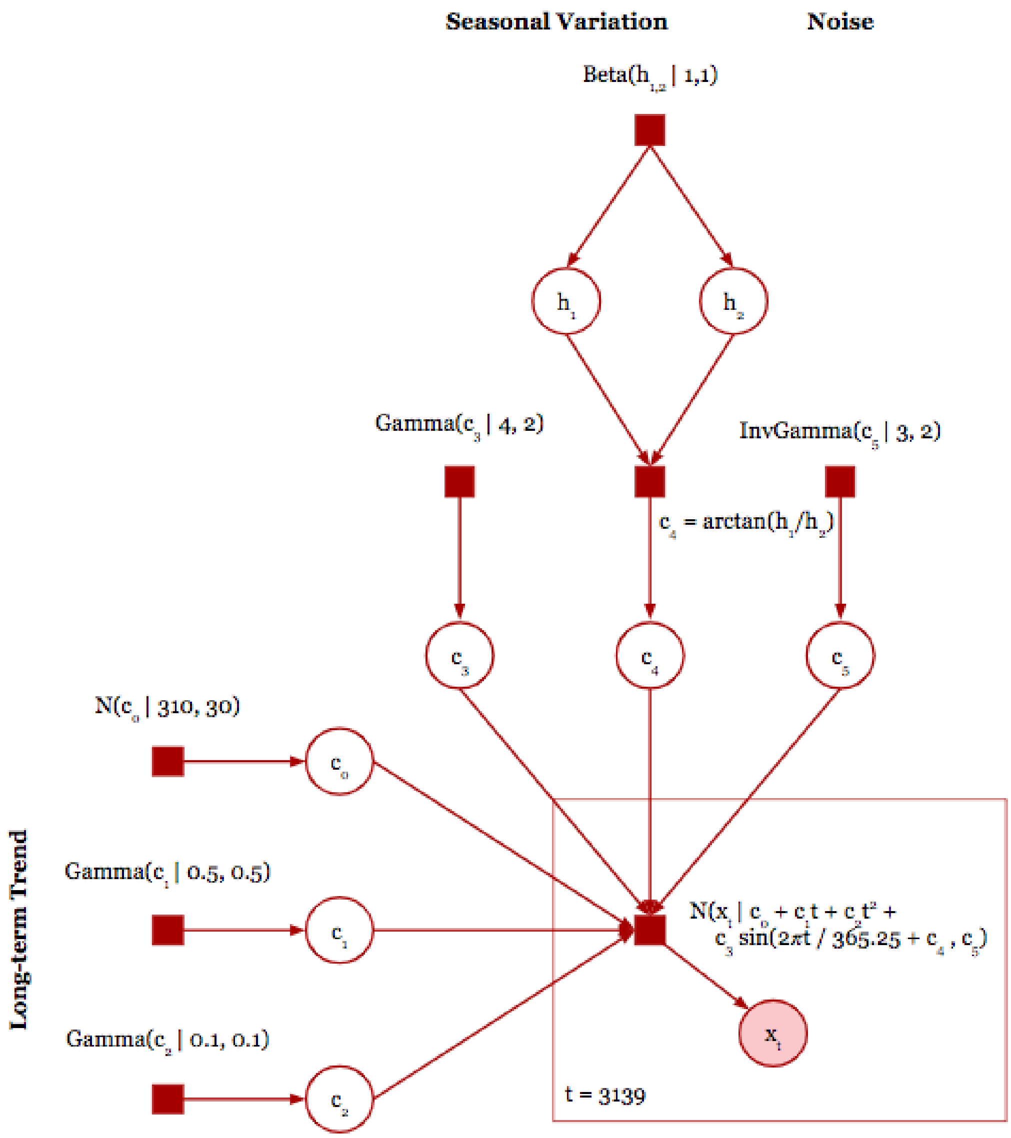


Figure 2. *Factor graph of the final model, showing all choices of priors over the parameters. Observed variables are filled in nodes.*

Model Extension – Inference

The following models were tested for trend and seasonal variation. Chosen models are highlighted.

Trend Models.

- Linear trend: $c_0 + c_1 t$
- Exponential trend: $c_0 + c_1 e^{c_2 t}$
- *Second Order Polynomial trend*: $\mathrm{c}_0 + \mathrm{c}_1 \mathrm{t} + \mathrm{c}_2 \mathrm{t}^2$

Seasonal Variation Models.

- Sine model: $c_0 \, sin(\frac{2\pi t}{365.25} + c_1)$
- Cosine model: $c_0 \, cos(\frac{2\pi t}{365.25} + c_1)$
- Skewed model: $-(1 + e^{t_c}) \arctan\left(\frac{sin(t_c)}{1+e^{-c_1} - cos(t_c)}\right) + c_2$, for $t_c = c_3 \, t + c_4$

  This model was adapted from here. I have added the constant c_3 as a scaling factor for all t-values, since the basic model assumes data between 0-1 on the y-axis for positive values, while c_4 determines the right-left shift of the model.

Model Convergence. For all models, convergence was tested by checking R-hat and N-eff values, ensuring that the sampling converged and that sufficiently independent samples were available for each parameter. At several points those two measures suggested the posterior having converged when pair plots of the parameters showed that the samples consisted of several local clusters (Fig. 3). This indicated that the parameters had multiple equivalent solutions (in particular for the sine and cosine functions) and that my constraints were inadequate to find only a single solution. Changing the constraints correctly would force the model to return only a single solution.

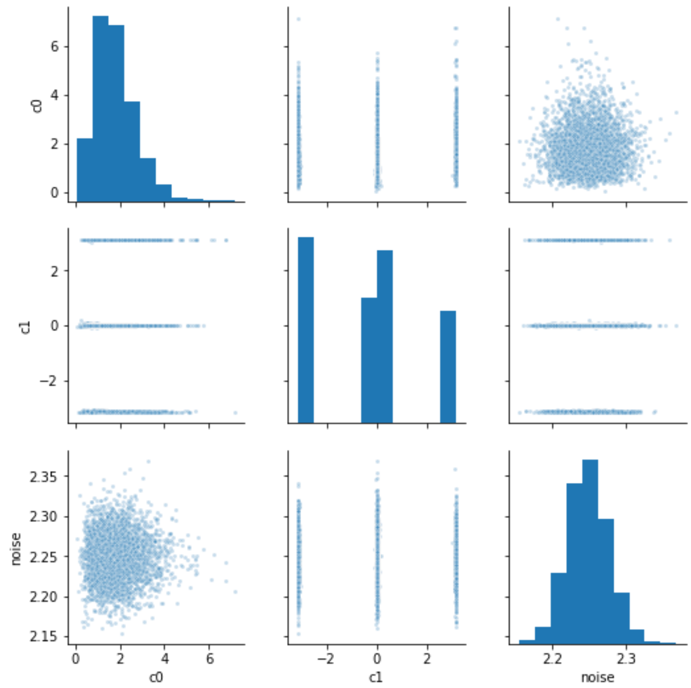
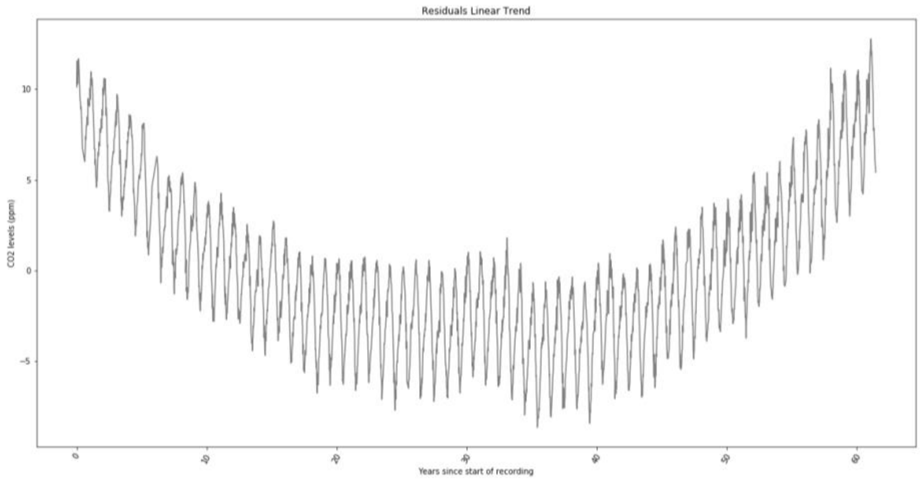


Figure 3. Pairplot of the sine seasonal variation model, showing inadequate parameter constraints. This causes the model to converge, but parameter estimates will be inadequate.

Residuals. In addition to checking for convergence on the parameter posteriors, I used residual plots to estimate how well a model fit the data and if it was misrepresenting the data in systematic ways. Using these plots, I was able to determine that the linear and exponential model were unable to fully capture the long-term trend of the data (Fig. 4).



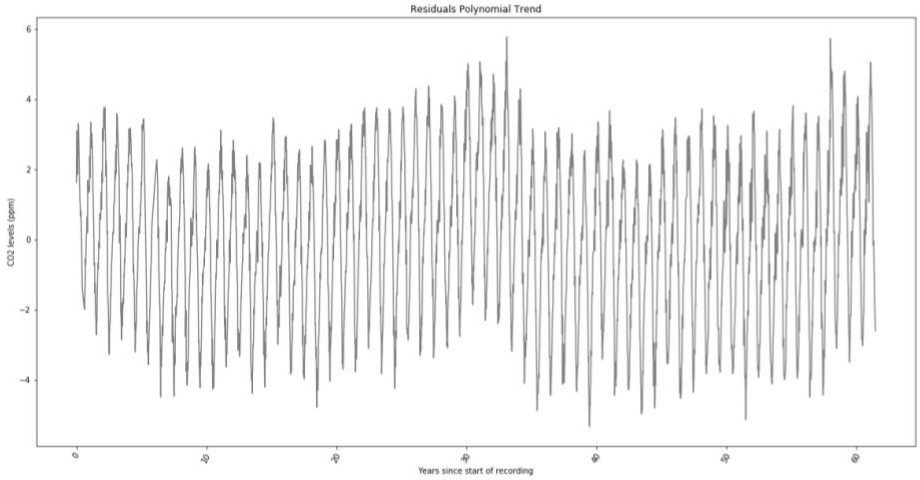


Figure 4. *Residuals of long-term trend models, linear and polynomial model in order. The linear model residuals clearly show that it is unable to capture the full long-term trend of the data, while the polynomial model seems to detrend the data effectively.*

Results

Using the residual plots, the polynomial model was clearly superior at capturing the long-term trend compared to the exponential and linear model (see Fig. 4 above for example). The residual plots of the seasonal models showed that the sine model consistently under-estimated the high peaks of the data, showing large residuals in that area (Fig. 5, first plot). While the skewed model has the overall smallest residuals, it also shows that it periodically falls out of sync with the data (Fig. 5, last plot), making it a poor model choice as well. The cosine model shows the best match to the data. Note, however, that for all models there remains periodic noise, suggesting that they all perform poorly at detrending the data.

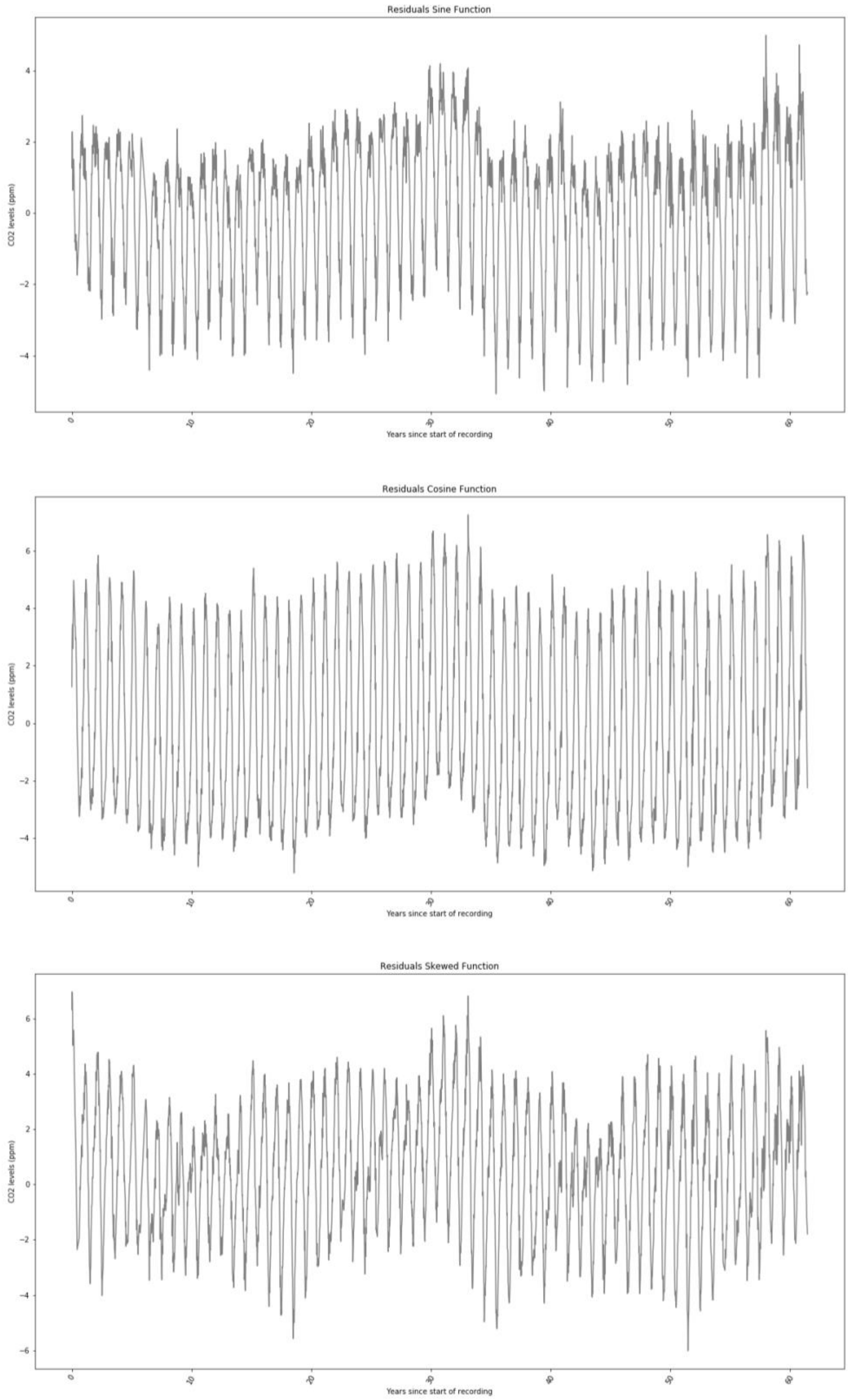


Figure 5. Residuals of seasonal trend models, sine, cosine and skewed model in order. Note the peak residuals of the sine model and the periodic desyncing of the skewed model.

Future Predictions. Having chosen a final model, we can make predictions for the upcoming decades (Fig. 6). According to this model, we predict 524ppm of CO_2 in the atmosphere 40 years from the last measurement, with 95% confidence intervals being [517, 530]. On conservative estimates, using the lower bounds of the confidence intervals, we are expected to surpass critical levels of CO_2 in the atmosphere mid 2036, less than 2 decades from now, but our model predicts this could happen as early as the beginning of 2033.

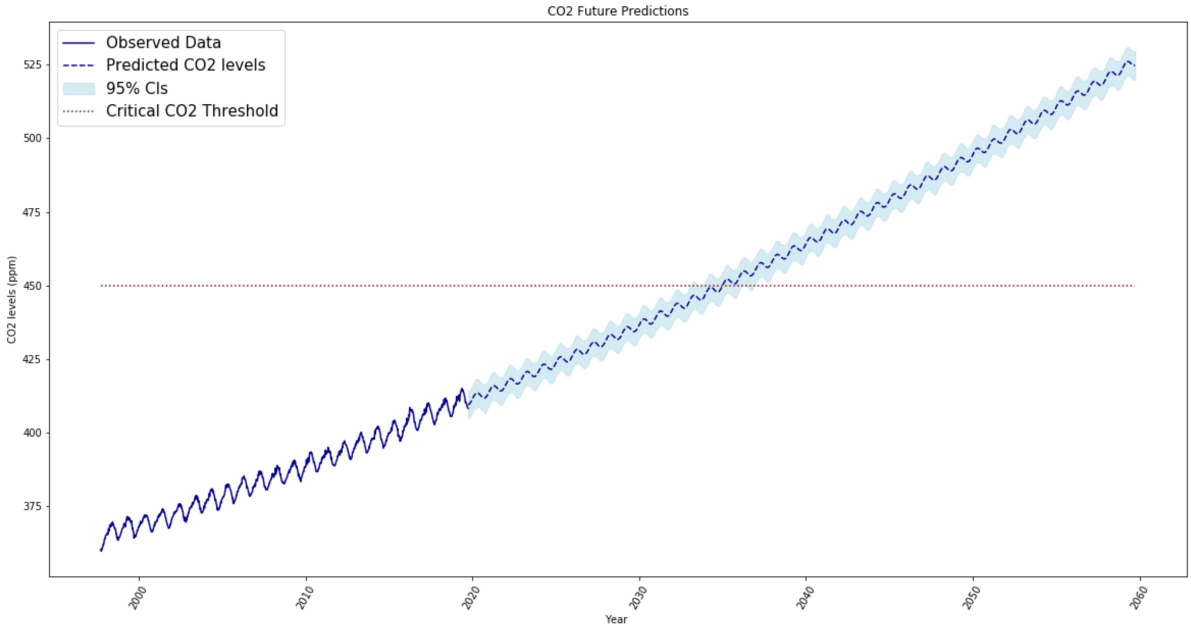


Figure 6. Future predictions of CO_2 levels in the atmosphere for the next 40 years. The red line indicates 450ppm, what is considered a critical threshold.

All-together, this model suggests that we are rapidly running out of time to implement greenhouse gas interventions. Rising CO_2 levels are closely associated with the rising temperature levels, which is in line with the scientific consensus on global warming .

Model Critique

Lastly, I would like to evaluate how well the model fits the observed data, the trends in residuals that I have observed and how accurate my forecast model is, illuminating potential modelling flaws.

Posterior Predictive Checks (PPCs). Several PPCs were implemented to analyze how well the model we have created is able to replicate the data we have already observed. If it does so well, it gives us additional confidence in its predictive abilities. I performed six PPCs, namely *mean*, *standard deviation*, *25th and 75th percentile*, *minimum* and *maximum* (Fig. 7). Of those, the model only fails to represent the data well with regards to the minimum values, suggesting that our model does not match the initial data values of the time-series well. If this is due to the polynomial model being, for example, bent upward too far, this might mean our model will overestimate trends in the long-run.

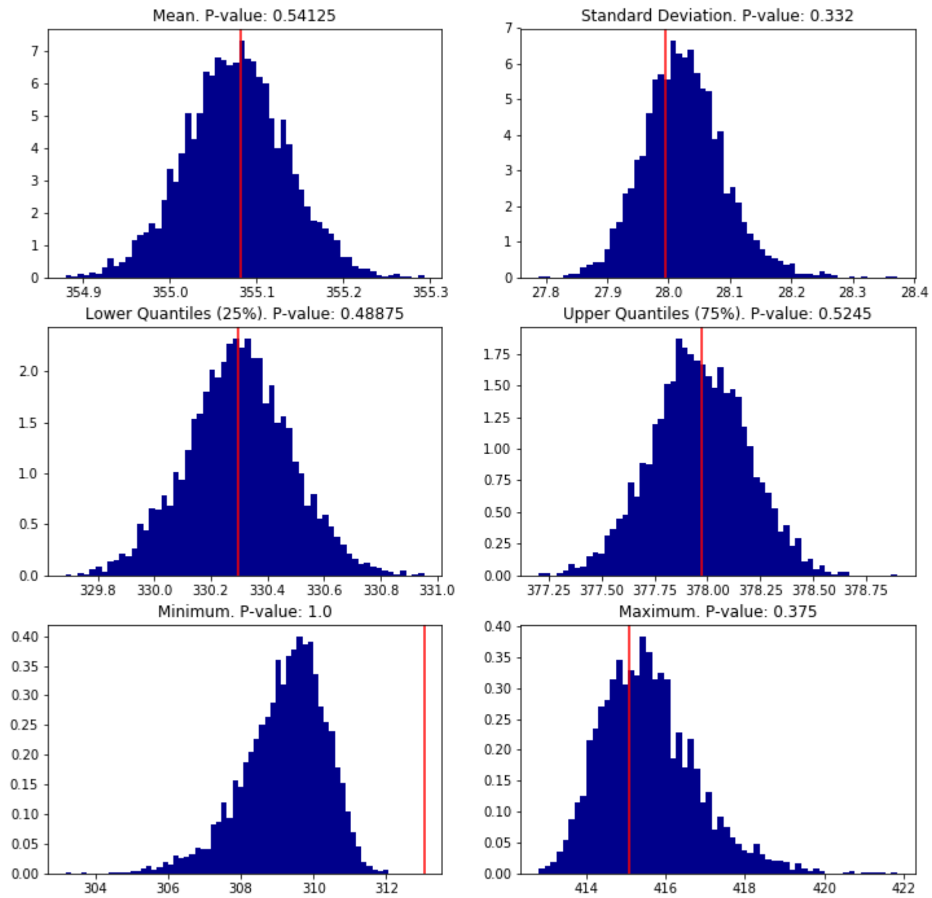


Figure 7. Six different Posterior Predictive Checks performed on the final model.

Residuals - Seasonal Trend. From the residual plots above (Fig. 5), we can see that for all seasonal models residual plot shows clear periodic trends left over. In an ideal case scenario, we would like for the seasonal model to be able to pick up on these patterns and reflect them, suggesting that neither model does a particularly good job at representing the full seasonal trends.

Uncertainty. Lastly, this model does not have widening confidence intervals as we get further from the present moment and hence does not fully represent the uncertainty of predictions correctly. In an ideal case scenario, the model we devised would be autoregressive in its error term, such that the standard deviation of the noise is dependent on x rather than being a static value. In such a model we could observe widening confidence intervals as time increases. I found this too difficult to implement in Stan to include it in this assignment.

References

Goldblatt, C., & Watson, A. J. (2012). The runaway greenhouse: implications for future climate change, geoengineering and planetary atmospheres. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *370*(1974), 4197-4216.

Israel, R. (2016, December 21). Robert Israel's answer to "asymmetrical sine wave". *Mathematical Stack Exchange.* Retrieved from https://math.stackexchange.com/questions/1779870/asymmetrical-sine-wave

Keeling, C. D., Piper, S. C., Bacastow, R. B., Wahlen, M., Whorf, T. P., Heimann, M. and Meijer, H. A. (2019). Weekly in-situ CO_2 data set. *SIO Reference Series, No. 01-06, Scripps Institution of Oceanography, San Diego, 88 pages, 2001*. Retrieved from https://scrippsCO2.ucsd.edu/assets/data/atmospheric/stations/in_situ_CO2/weekly/weekly_in_situ_CO2_mlo.csv

NASA. (2019). Scientific consensus: earth's climate is warming. *Global Climate Change.* Retrieved from https://climate.nasa.gov/scientific-consensus/