Final Project[1]

Vinícius Miranda

CS146 Fall 2018

[1] Available as a gist here. Thank you for a fantastic course!

Introduction

This project aims to model atmospheric levels of CO_2 from 1958 using data recorded at the Mauna Loa Observatory in Hawaii and forecast measurements between 2019 and 2058. These predictions will also inform an estimate as to when carbon dioxide levels should reach high-risk levels indicating drastic climate change.

Model

The data used is the weekly Mauna Loa data set (Keeling et al., 2001), which contains the date and measurement of CO_2 levels in ppm (parts per million). Let t represent the time in days since measurements started and x_t the levels of CO_2 observed in ppm. Figure 1 presents a subset of this data.

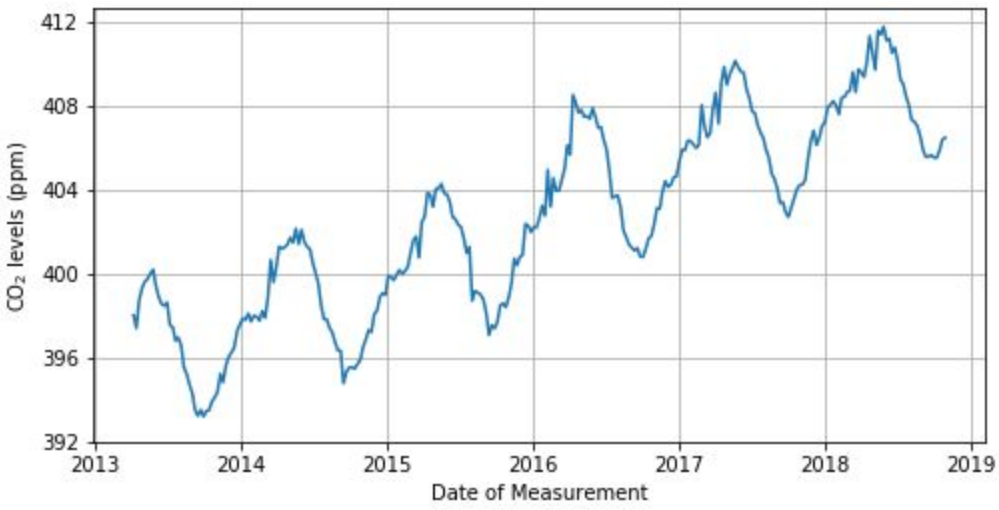

Figure 1. CO_2 levels (ppm) in recent years.

It is visible in Figure 1 that there are three aspects of the data that the model needs to capture. There is a general upward trend, seasonal variation during the year, and noise. One basic model that captures these elements is the following

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

where c_i are unobserved parameters and $p(x_t \mid \theta)$ is the likelihood function.

This model assumes a linear trend, symmetric and cosinusoidal seasonal variation, and normally distributed noise with mean 0. The sections below present an analysis that tests and improves upon this model. However, there are also *ex ante* reasons why skepticism of how good of a fit this model can provide is warranted. First, one might posit that the industrialization has dramatically expanded since the industrial revolution, and hence that atmospheric CO_2 levels probably rose at a quicker pace than what would be shown by a linear trend. Moreover, the periodic pattern in the data (Figure 1) seems to be slightly tilted to the right, an aspect that would not be captured by a symmetric periodic function.

The following sections discuss a more complex approach to modeling the data's trend and seasonality while providing an analysis of some modeling challenges. The blueprint of the model selection approach adopted is:

1. List several candidate models;
2. Generate posterior samples in Stan;
    a. If sampling does not converge, discard the model[2];
3. Plot residuals and compute the RMSE (root-mean-square error);
4. Select the model with the lowest RMSE.

[2] In practice, the candidate models went through a few iterations before being discarded, which would involve adapting priors, changing the structure of parameters such as their data type, and passing the time data in different formats (e.g., days or years since the start of measurements).

Trend

Three candidates models were effectively tested in how much they capture the trend behavior of the data, which can be seen in Figure 2. Two other models were discarded because sampling from them diverged. These models were:

1. **Linear**: $c_0 + c_1 t$,
2. **Quadratic**: $c_0 + c_1 t + c_2 t^2$,
3. **Exponential**: $c_0 e^{c_1 t}$,
4. **Cubic** *(discarded)*: $c_0 + c_1 t + c_2 t^2 + c_3 t^3$,
5. **N-polynomial** *(discarded)*: $c_0 + c_1 t^{c_2}$,

where c_i are unobserved parameters whose posteriors were estimated. Figure 3 plots the three trend curves given by the mean posterior samples of c_i of trend models 1, 2, and 3. Figure 4 shows the residuals under these three models.

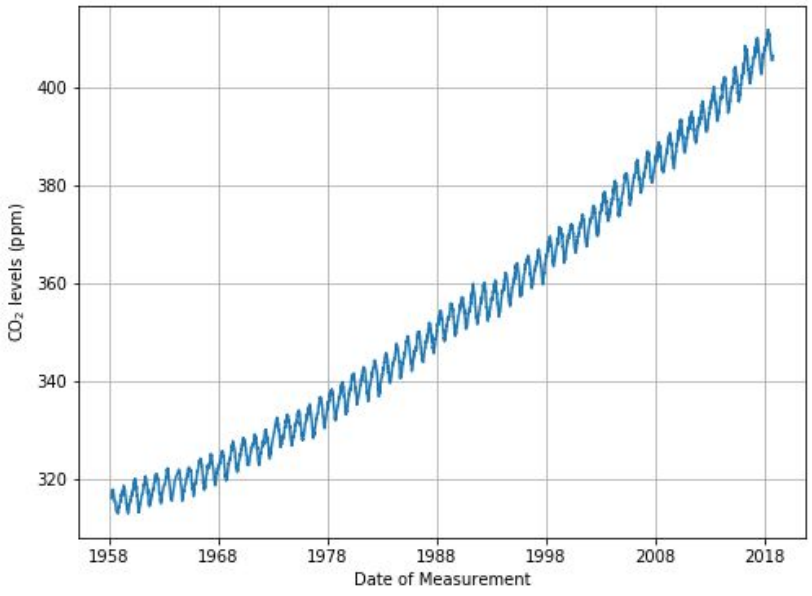

Figure 2. CO_2 levels (ppm) since the start of measurement.

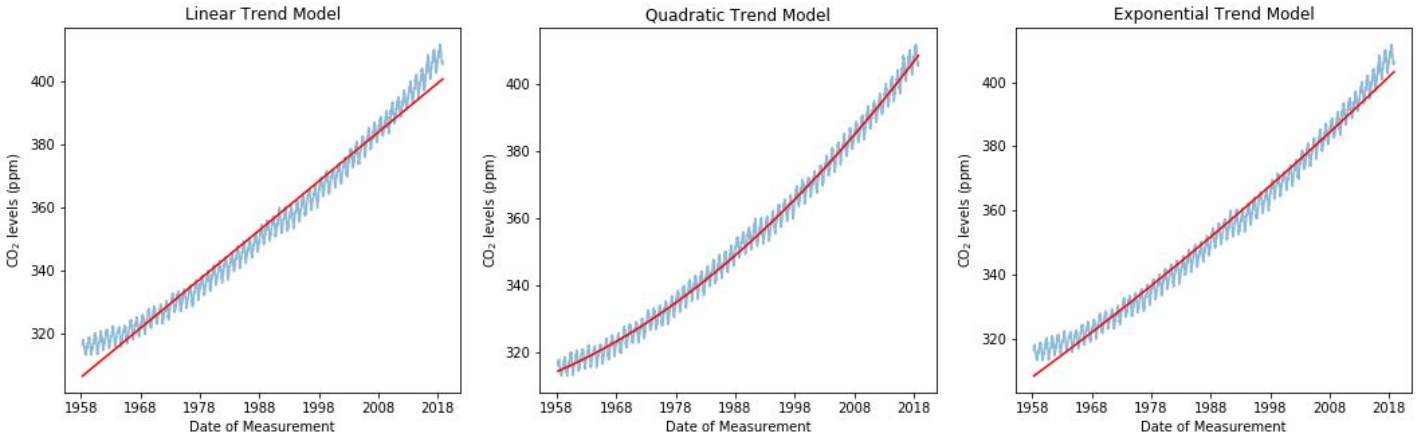

Figure 3. Trend curves for the three models.

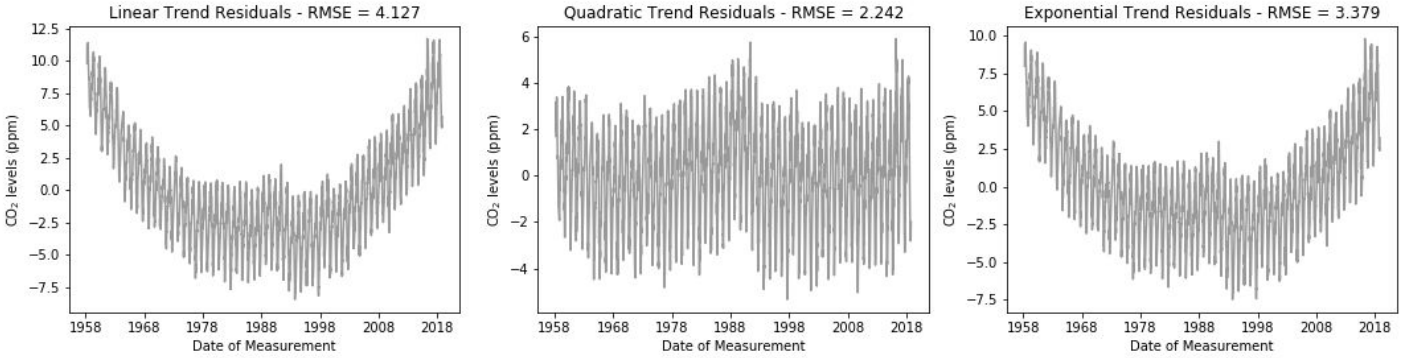

Figure 4. Residuals and RMSEs for the three trend models.

The quadratic trend model fits the data better than both the linear and exponential models, which can be seen in Figure 3 and inferred from the residuals plotted in Figure 4. Its RMSE is hence the lowest, which renders it the preferred trend model.

Seasonality

Figure 5 portrays the seasonality of CO_2 levels in recent years by subtracting the trend from the original data. In comparison with a sine wave, the pattern shows sharp local extrema and is slightly tilted to the right. Three candidate models were successfully tested and two models were discarded due to divergence. They are:

1. **Sine**: $c_0 \sin(\frac{2\pi}{T}t + c_1)$,
2. **Cosine**: $c_0 \cos(\frac{2\pi}{T}t + c_1)$,
3. **Double Sine**: $c_0 \sin(\frac{2\pi}{T}t + c_1) + c_2 \sin(\frac{2\pi}{T}2t + c_1)$,
4. **Triple Sine** *(discarded)*: $c_0 \sin(\frac{2\pi}{T}t + c_1) + c_2 \sin(\frac{2\pi}{T}2t + c_1) + c_3 \sin(\frac{2\pi}{T}3t + c_1)$,
5. **Nested Sine** *(discarded)*: $c_0 \sin(\frac{2\pi}{T}t + c_1 + c_2 \sin(\frac{2\pi}{T}t + c_1))$,

where c_i are the unobserved parameters and T is the period of the data. Both the trend and seasonality models responded better to receiving time data in years since the start of measurement, with the exception of the exponential trend model, which converged when receiving time as a fraction from start to end of measurements.[3] The period T for the data in years is equal to one. Also, the cosine model is expected to have the same explanatory power as the sine model, but it is included for completeness and as a check that the results are reasonable.

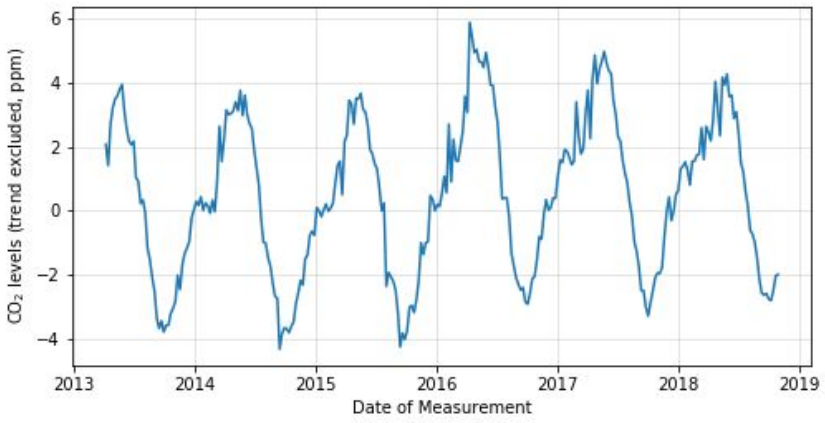

Figure 5. CO_2 levels (ppm) in recent years with the trend subtracted.

[3] As in, the first observation happens at $t = 0$ and the last at $t = 1$, with all others holding fractional time values.

Like before, Figure 6 shows the three seasonality curves given by the mean posterior samples of c_i of the sine, cosine, and double sine models.[4] Figure 7 shows the residuals under these three models.

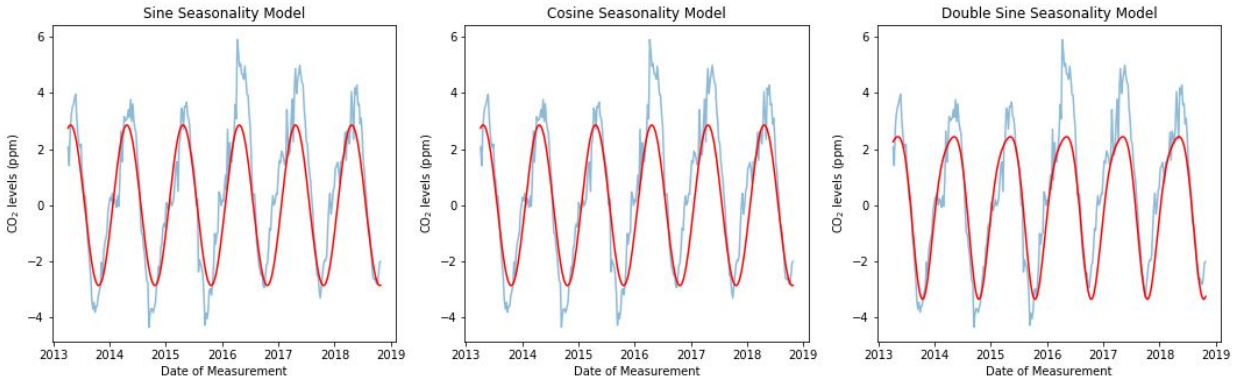

Figure 6. Seasonality curves for the three models.

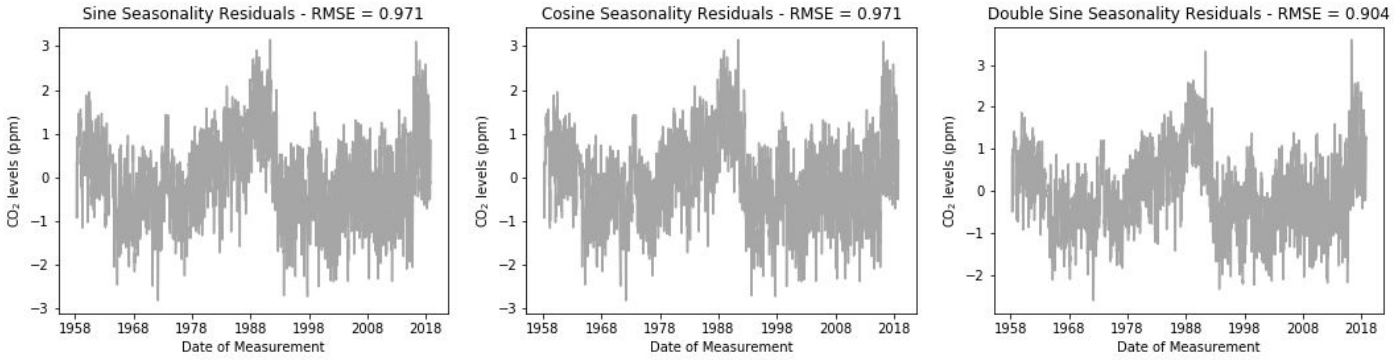

Figure 7. Residuals and RMSEs for the three seasonality models.

The double sine model fits the seasonal aspect of the data slightly better than the simple sine or cosine models. However, its peaks slightly underestimate the peaks of the data. The residuals also have two features worth noticing, their long-term pattern and remaining periodicity. Predictions around 1988 seem to underestimate the data,

[4] Two double sine models were fitted with different parameter constraints. The reasoning for this is provided in the challenges section. I present one the best of these two models.

whereas those around the 2000s slightly overestimate it. Furthermore, the residuals still retain a seasonal pattern, but the period seems to be about half a year (Figure 8).

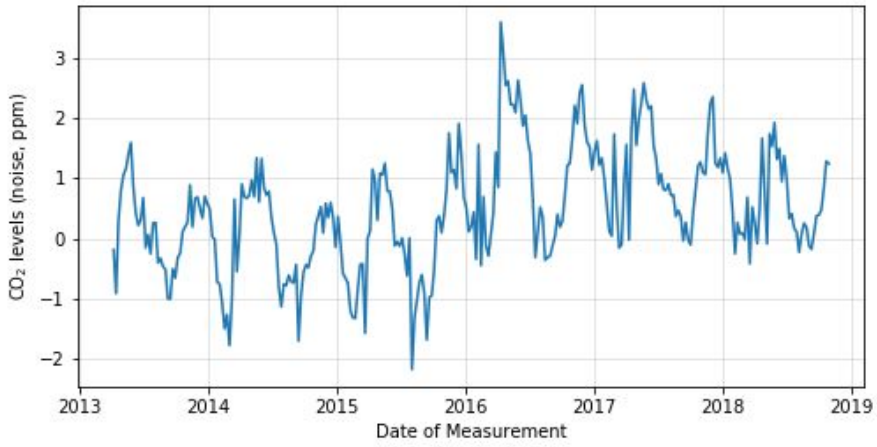

Figure 8. CO_2 levels (ppm) in recent years with the trend and seasonality subtracted.

Full Model

The random fluctuation of the data is modeled by a Gaussian with mean 0 and unknown standard deviation. The trend, seasonality, and noise models are combined to form the likelihood function

$$p(x \mid \theta) = N(c_0 + c_1 t + c_2 t^2 + c_3 \sin(\tfrac{2\pi}{T} t + c_4) \; + \; c_5 \sin(\tfrac{2\pi}{T} 2t + c_4), c_6^2)$$

where θ represents all unobserved parameters. This model received the data in years since the start of measurement, and hence the period is equal to 1. The following priors were used:

I. $c_0 \sim N(300, 10)$. Historic levels of atmospheric CO_2 were close to 300 ppm before and during the 1900s. The normal distribution is used because of its thin tails,

which imply that there is high confidence that the concentration at the start of measurements should not be far from 300 ppm.

II. $c_1 \sim N(0,\ 10)$ and $c_2 \sim N(0,\ 2)$. There is no *a priori* knowledge of the coefficients besides that they should be small, and hence a normal prior centered around 0 is used. The coefficient of the quadratic term should be smaller than the coefficient of the linear term due to larger t^2 values, so the standard deviation of its prior is smaller.

III. $c_3, c_5 \sim N(0, 4)$. The absolute value of the trend residuals commonly peaks at 4. The amplitudes of the periodic functions should thus be around, but not much larger, than this value. Furthermore, because $c_1 \sin(x) = -c_1 \sin(x + \pi)$, the prior of c_5 is restricted to either only the positive or negative real numbers so that its posterior has only one mode, generating two spin-off models (this point is elaborated below in the Challenges section).

IV. $c_4 = \arctan(\frac{d_0}{d_1})$, where $d_{0,}\ d_1 \sim N(0,\ 1)$. d_0 and d_1 are implemented as elements of a unit vector is Stan, i.e., $\sqrt{{d_0}^2 + {d_1}^2} = 1$. Their priors are determined implicitly.

V. $c_6 \sim InvGamma(3, 2)$. The inverse gamma has a positive support, which directly respects the requirements that c_6 be positive given that it is the standard deviation of a normal distribution. The hyperparameters 3 and 2 distribute most of the probability mass of the inverse gamma between 0 and 1, corresponding to the intuition that the residual error be relatively small.

The corresponding factor graph is shown in Figure 9.

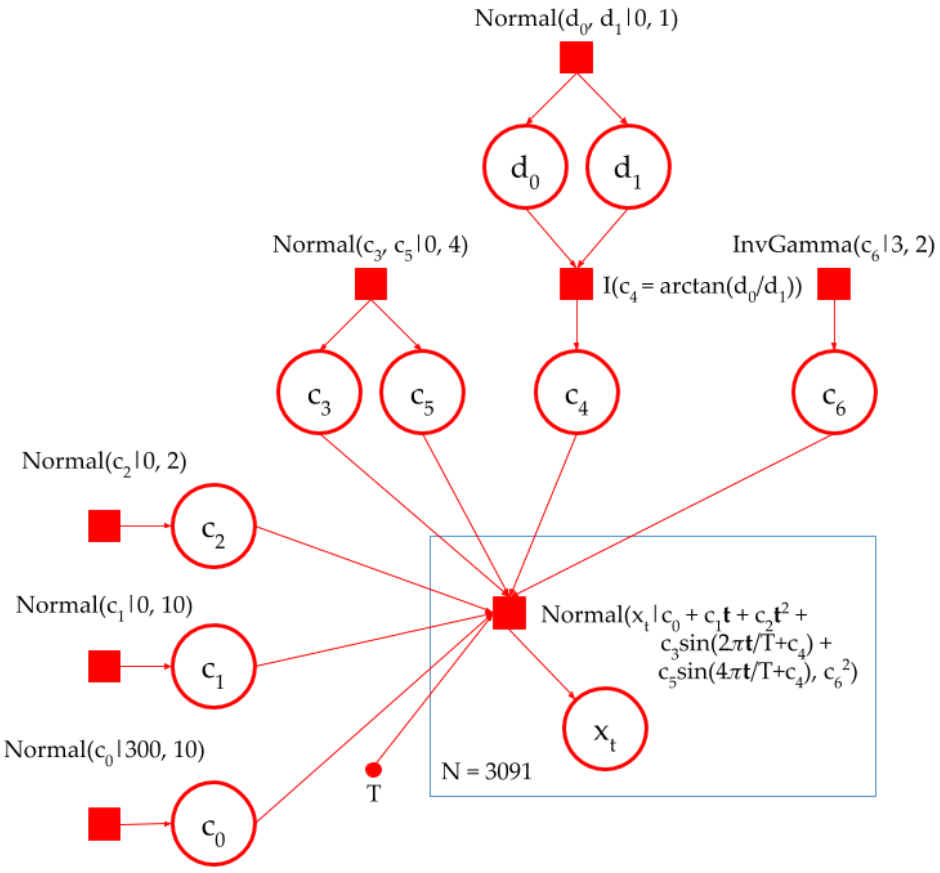

Figure 9. Factor Graph of the full model.

Challenges

Modeling a Tilted Periodic Function

The main shortcoming of the simple sinusoidal seasonality model is that it did not account for the tilt in the data. There are several functions which are both periodic and tilted, but some are not easily modeled in Stan. For example,

$$y = \ -\sin[x-(\pi/2\ -\alpha)y]$$

is a skewed trigonometric function whose local extrema are displaced by α (hypergeometric, 2014). However, y cannot be easily isolated in this equation, which

renders its representation in Stan difficult. Instead, a combination of sine functions was employed to represent the seasonality. The underlying intuition of this approach is that the infinite series

$$\sum_{k=1}^{\infty} \frac{\sin(kx)}{k}$$

is a sawtooth wave (robjohn, 2017). Finite approximations of the sawtooth function would, therefore, be tilted periodic functions.

Modeling Periodic Variables

The phase parameter of a periodic function is a periodic variable by definition. It is important to model this feature explicitly in Stan because otherwise the posterior has multiple modes and the chains do not mix. This problem was tackled by declaring an intermediary variable, a unit vector of size 2, and modeling the phase as the arctangent of the quotient of the unit vector elements. This strategy restricts the value of the phase to the interval $[-\pi, \pi]$.

Sampling from a Complex Model

By far, the greatest challenge was divergence in sampling due to multimodality in the posterior. The reason was this is a hard challenge is because if sampling diverges, there is no clear debugging strategy. For example, when the phase is not modeled as a periodic variable, sampling diverges and Stan is unable to produce good results (i.e., Rhat is large and the effective sample size is close to zero). In other cases, sampling will take an inordinate amount of time and the process will have to be stopped before results are available.

This problem was encountered not only when modeling the phase, but also in the double sine model. As mentioned above, the identity $c_1 \sin(x + c_2) = -c_1 \sin(x + c_2 + \pi)$ entails that the posteriors for c_1 and c_2 will not be unique if c_1 can take negative values.

However, this generates a problem in the double sine model because once the phase of the second sine function is restricted to be equal to the first one (as is necessary for the sawtooth approximation), the correct solution for the posterior may require that the amplitude of the second sine be negative. However, because the model is also doing inference on the phase, allowing for negative values of the amplitude renders the solution indeterminate. This happens to be the case for modeling CO_2 levels, where the parametrization of the full model above that best fits the data corresponds to a positive amplitude for the first sine and a negative amplitude for the second sine.

This problem was solved by splitting the model into two submodels, one which restricts the second amplitude to be positive and one which requires it to be negative. This approach eliminates the bimodality of the posterior. This approach would not generalize well for better approximations of the sawtooth wave, given that 2^{n-1} submodels will be required for approximations with n sine functions. An alternative solution might be to model a Bernoulli variable k, which would function as a multiplier for the sine phase shift, i.e., $c_1 \sin(x + c_2 + k\pi)$. The model would then restrict all amplitudes to be positive, but do inference on k to check whether the solution requires the phase to be shifted or not. This approach was not attempted.

Inference and Analysis

Weekly predictions are generated for the time period ranging from January 4th, 2019 to March 29th, 2058, one hundred years after measurements started in Mauna Loa. At the latter date, the model predicts the atmospheric CO_2 concentration of 522 ppm. The 95% confidence interval is $[520.1,\ 524.6]$. The level of 450 ppm is usually considered a mark of high risk for drastic climate change. The model predicts this threshold to be surpassed with high confidence (97.5%) in March 11th, 2035. Figure 10 presents these predictions.

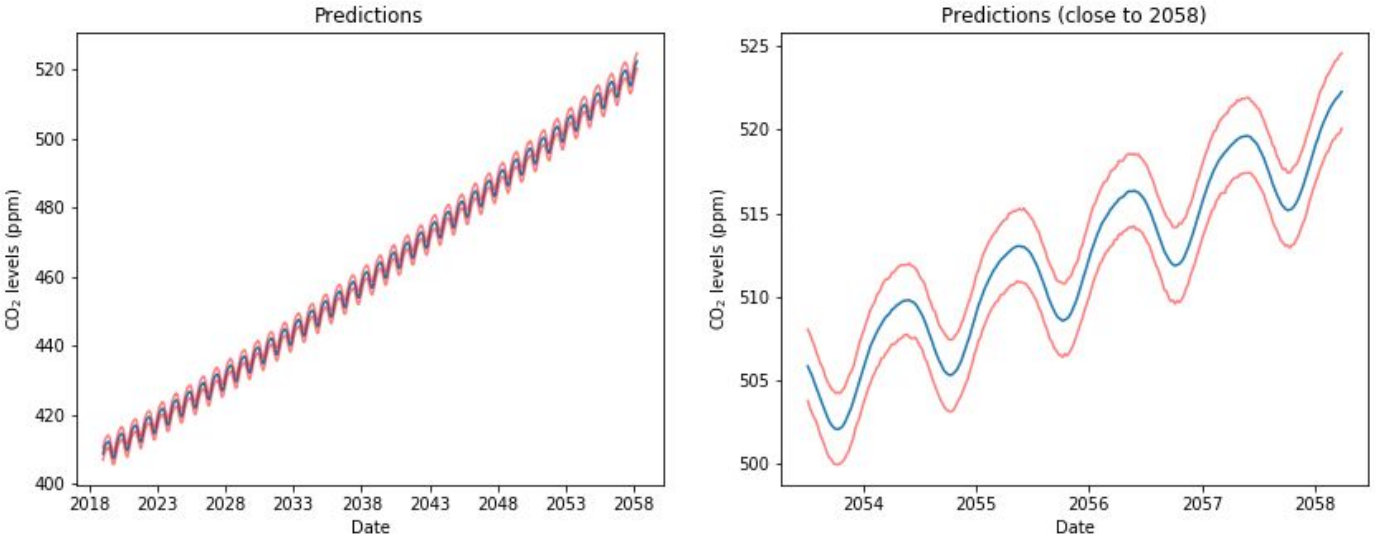

Figure 10. Predictions until 2058 and zoomed view on latter predictions. The blue line shows the mean predictions and the red lines the 95% confidence interval.

The shortcomings of this model can be analyzed by how well it fits recorded data and how reasonable are its predictions. Figure 11 in Appendix A shows posterior predictive checks for several test statistics to investigate whether data sets generated by the model have characteristics similar to those of the actual data. One can see that the model does a generally good job, especially with the global features of the data.

The mean, standard deviation, skewness, and maximum value of the data are all well replicated by generated data sets. The model fails in accounting for the minimum value, range, and kurtosis and of the data. Although this is certainly a problem, these three statistics are interconnected. The range is given by the maximum value minus the minimum value, so the model is poor under the range statistic for the same reason it is poor under the minimum value statistic. Furthermore, the kurtosis is a measure of the "tailedness" of a distribution. Because the model generates data sets with lower minimum values, it is also natural that the absolute value of their kurtosis is higher than the actual data's.

These shortcomings might be remediated by improving the trend fit so that the first values of the generated datasets are more representative of the data. This goal could potentially be accomplished by having a fractional degree in the polynomial trend function. Finally, it should also be noted that the model successfully passes the posterior predictive check that checks whether the errors are normally distributed around zero, as was assumed. Hence, the model seems to account well for at least some of the most common statistics of the data.

There is more reason to be concerned about the predictions. Intuitively, one should expect the certainty of the predictions to decrease, and hence that the confidence intervals become wider, as the model produces forecasts further into the future. As seen in Figure 10, that is not true. The model produces forecasts for 2058 with similar certainty than for 2019. The main reason for this phenomenon is that the prediction x_t is independent of the prediction at x_{t-1}, an extension of the model's underlying assumption that data points are independent and uncorrelated. This is clearly not the case, given that if x_t is larger than x_{t-1}, there is a higher probability that x_{t+1} will be larger than x_t than not.

Another way of expressing the reasoning above is to say that the data is autocorrelated, i.e., that the value of a data point is correlated with the value of the preceding data points. One solution to this problem is to employ autoregressive models, which would account for this feature of the data. One prominent example is Holt-Winters Exponential Smoothing, which is a time series forecasting method which is able to account for both the trend and seasonality of the data (Kalekar, 2004). This model has been applied to the Mauna Loa dataset, but not under a Bayesian statistical framework (Krzaczek & Yates, 2017). Implementing this model is Stan would be a daunting task, but simpler exponential models have been successfully executed before (Smyl & Zhang, 2015). The main advantage of autoregressive models is that predictions further into the future "inherit" the uncertainty of earlier predictions, so confidence intervals get wider as time passes.

Despite the unwarranted certainty the model conveys in its predictions, its output clearly gives reasons for worry. The 450 ppm threshold is predicted to be surpassed within two decades, and CO_2 levels in 2058 could be well above 500 ppm. The implications in terms of climate change could be catastrophic. Investments in the prevention of climate change need to occur in much greater scale and with much greater urgency, lest there be too little time to avoid it.

References

hypergeometric. (August 25, 2014). Skewed Trigonometric Function. Retrieved from https://math.stackexchange.com/q/908438

Kalekar, P. S. (2004). Time series forecasting using holt-winters exponential smoothing. *Kanwal Rekhi School of Information Technology, 4329008*, 1-13.

Keeling, C. D., Piper, S. C., Bacastow, R. B., Wahlen, M., Whorf, T. P., Heimann, M., & Meijer, H. A. (2001). Exchanges of atmospheric CO2 and $^{13}CO_2$ with the terrestrial biosphere and oceans from 1978 to 2000. Global aspects, SIO Reference Series, No. 01-06, Scripps Institution of Oceanography, San Diego.]

Krzaczek, L. A.,, & Yates, P. A. (2017). A Statistical Analysis of Atmospheric CO2 Levels at Mauna Loa. *Ball State Undergraduate Mathematics Exchange, 11*(1).

robjohn. (September 20, 2017). Equation of a "tilted" sine. Retrieved from https://math.stackexchange.com/q/2431811

Smyl, S., & Zhang, Q. (2015). Fitting and extending exponential smoothing models with stan. In *International Symposium on Forecasting*.

Appendix A — Posterior Predictive Checks

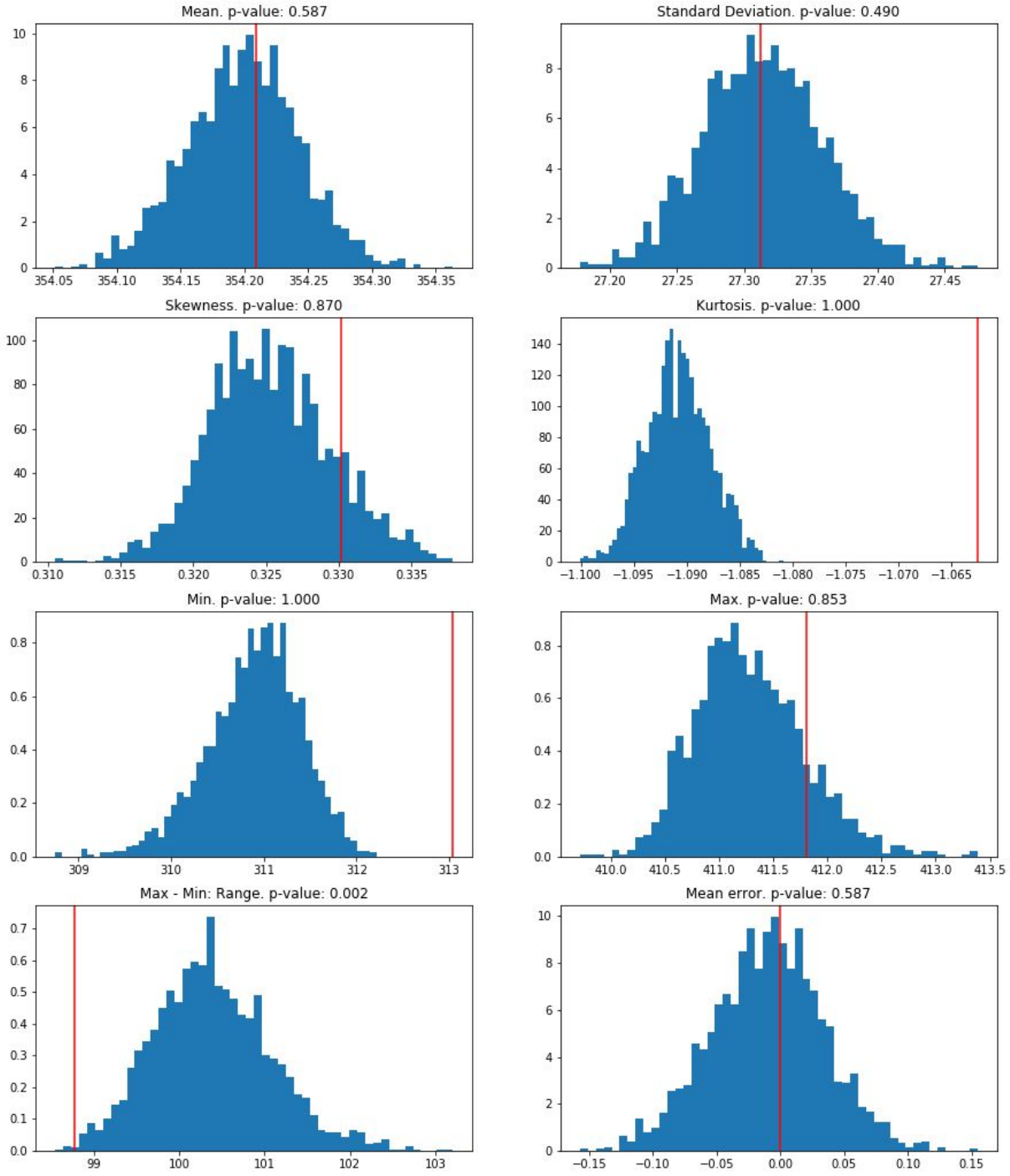

Figure 11. Posterior predictive checks for seven different test statistics.