BAYESIAN METHODS IN OBSERVING GLOBAL WARMING

Rowan Lavelle, Peter Wu

Indiana University, Statistics Department

ABSTRACT

Global warming, and by extension climate change, are major issues in today's society, however some people still do not believe that global temperatures are rising. This paper is a Bayesian analysis approach looking for indicators to confirm a rising in average global temperatures. We compare and contrast posterior distributions on average temperatures for before and after 1970, which is when global warming began to accelerate rapidly. We compare and contrast posterior distributions over average temperature changes year to year under two different priors to show that the rising temperature signal in the data is so strong that priors have very little affect on the posterior. Finally we use a Bayesian regression approach to compare multiple linear and non linear models to further show indications of rising temperatures throughout the 18th, 19th and 20th century.

Index Terms— Bayesian Analysis, Bayesian Regression, Global Warming, Climate Change

1. INTRODUCTION

Global warming and climate change are considered some of the most important issues of our generation. The implications of rising global temperatures has the potential to affect all facets of life, including agriculture, tourism, manufacturing, and so on and so forth. While some believe that these issues are the biggest threats, others believe that it's a myth.

Within this report, we aim to address a few important questions regarding global warming from a Bayesian perspective, as well as discuss any subject matter implications our findings may have. Our report aims to address 3 main research questions:

1. How drastically has the distribution of temperatures shifted from pre-1970 to post-1970, a year in which global warming was noted to accelerate

2. Observing a prior distribution from the perspective of someone who does not believe in global warming/climate change exists (i.e global temperatures do not significantly change year over year), what can the posterior on the parameters that govern the distribution of temperature change tell us?

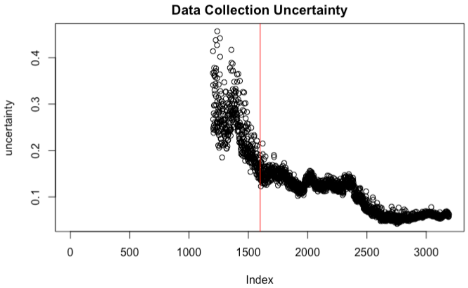


Fig. 1. Uncertainty level of data collection (red line denotes 1880)

3. Can we use a Bayesian regression framework to predict future values of temperature?

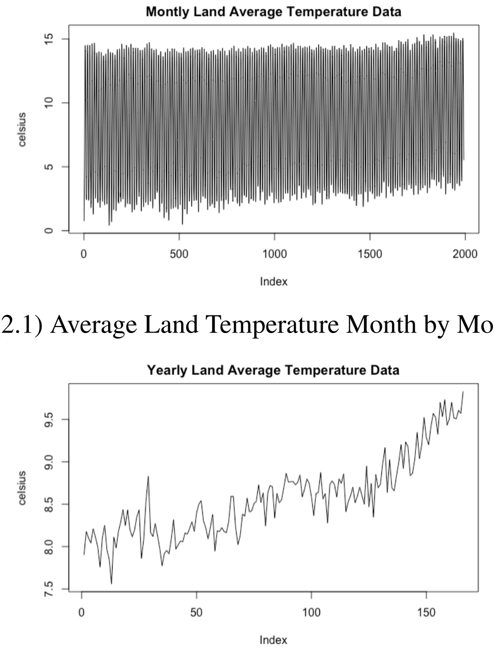
The data utilized during this study was collected from Berkeley Earth [1], which has collected 1.6 billion temperature reports. The dataset itself is collected on the first of every month from 1750 to 2015. The data includes average land temperatures along with their uncertainties as well as land and ocean average temperatures along with their uncertainties. These uncertainties can guide us to only using quality data, as we can see in figure 1 the uncertainties stabilize around 1880 and onward.

For the purpose of our study, we are primarily interested in data from 1850 to 2015, and the only variable that we are interested in exploring is the average global land temperature on a monthly basis. Based on the research question, we also aggregated the data to explore yearly average temperatures, and took the difference in years to explore yearly average temperature changes. Here we work with $n \approx 2,000$ data points month to month, and $n \approx 165$ aggregated year by year data points.

Figure 2.1 shows the raw data as a time series value of the month to month temperature values from 1850 to 2015.

We also display the aggregated values to show year to year averages of the raw global temperature values which is seen in figure 2.2. Figures 2.1 and 2.2 show clear rising trend throughout time, both in the seasonal month to month data, and in the aggregated yearly data.

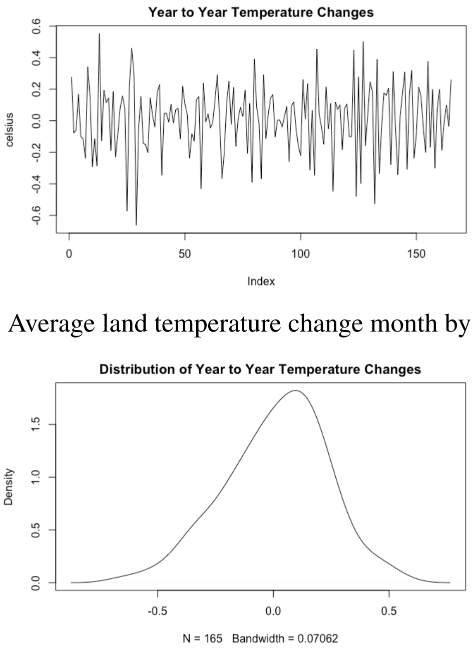
Further, figure 3.1 shows aggregated year to year temperature changes as a time series. We notice that despite the clear rising trend, temperature changes from year to year appear to



(2.1) Average Land Temperature Month by Month

(2.2) Average Land Temperature Aggregated Year over Year

Fig. 2. Temperatures as time series



(3.1) Average land temperature change month by month

(3.2) Average land temperature change aggregated year over year

Fig. 3. Temperature change plots

be fairly stationary, with values centered around roughly zero. This is an interesting phenomenon, as we should be curious about why temperatures are rising if the distribution of temperature shifts are centered around near 0.

We can observe the densities of temperature changes from 1850-2015 in figure 3.2. the temperature change data is clearly from a normal distribution, which justifies our choice in models in the following section.

2. METHODS

2.1. Gibbs Sampler

Across all our proposed research questions, we have decided to utilize a Normal-Normal model to model θ and a Normal-Gamma model to model σ^2. Since a Gibbs sampler method has been extensively utilized throughout the research questions, a primer on the methodology is presented below before addressing the research topics. The assumption here is that our data comes from some distribution $y_1, \dots, y_n \sim Norm(\theta, \sigma^2)$, for the posterior on θ we want to model $p(\theta|\sigma^2, y_1, \dots, y_n)$.

The Normal-Normal model for θ can be denoted as follows. First we see that $p(\theta|\sigma^2, y) \propto p(y|\theta, \sigma^2)p(\theta)$, this then expands into

$$p(\theta|\sigma^2, y) \propto Norm(\theta|\mu_0, \tau_0^2) \times \prod_{i=1}^{n} Norm(y_i|\theta, \sigma^2) \quad (1)$$

Which condenses into $\theta \sim Norm(\mu_n, k_n^2)$ [2] where

$$\mu_n = \frac{\frac{\mu_0}{\tau_0^2} + n\frac{\bar{y}}{\sigma^2}}{\frac{1}{\tau_0} + \frac{n}{\sigma^2}} \qquad k_n^2 = \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right)^{-1} \quad (2)$$

Here θ will be modeling the average temperature change year to year.

We also need to come up with a posterior on σ^2 since the posterior for θ depends on it, but we discard the actual values in the analysis. Here we have $p(\sigma|\theta, y) \propto p(y|\theta, \sigma^2)p(\sigma^2)$, this falls under the Normal-Gamma model which is shown below

$$p(\sigma|\theta, y) \propto Gamma(\sigma|\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}) \times \prod_{i=1}^{n} Norm(y_i|\theta, \sigma^2) \quad (3)$$

Which condenses into $\frac{1}{\sigma^2} \sim Gamma(\frac{\nu_n}{2}, \frac{\nu_n\sigma_n^2}{2})$ [2] where

$$\nu_n = \nu_0 + n \qquad \sigma_n^2 = \frac{1}{v_n}\left[\nu_0\sigma_0^2 + ns^2\right] \quad (4)$$

Where s^2 is the sample variance assuming θ as the mean.

1. sample $1/\sigma_{(i)}^2 \sim Gamma(\frac{\nu_n}{2}, \frac{\nu_n\sigma_n^2}{2})$
2. sample $\theta_{(i)} \sim Norm(\mu_n, k_n^2)$
3. save $\{\theta, \sigma^2\}_{(i)}$ as a sample from the posterior

Using this Gibbs sampler we can predict

$$\tilde{y}_{(i)} \sim Norm(\theta_{(i)}, \sigma_{(i)}^2)$$

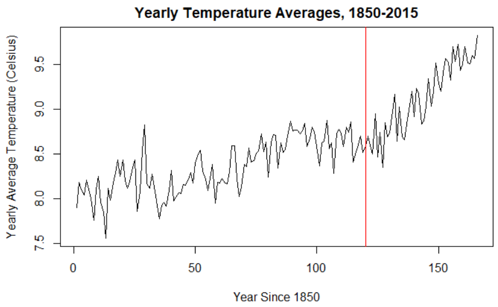


Fig. 4. Yearly averages in Celsius (red line indicates year 1970)

2.2. Comparing Temperatures Pre-1970 vs Post-1970

This research question was proposed partly due to how global warming, and by extension climate change, had come into the general public's view extremely rapidly in the 1970s [3]. Prior to this time period, while some literature was published on global temperature patterns and cycles, it was not nearly as present in the the public conscience as it is today.

Since concerns regarding more drastic shifts (both global warming and global cooling) had begun to become more prevalent in the 1970s in part due to certain more extreme seasons in the decade, we decided to observe and compare how the distribution of temperatures had shifted from before 1970, to 1970 to current day.

While it seems graphically clear (Figure 4) that temperatures have increased throughout the span of the dataset, the goal of this question is to quantify how much the underlying distribution, which we have characterized to be normal, has shifted.

We utilized the Gibbs sampling method denoted by equations (1) and (3) to generate samples from the posterior distribution of θ. Here we use the same prior for both the pre-1970s (θ_{pre}) data as well as the post-1970s (θ_{post}) data in order to remain consistent, allowing focus on how our sampled θ values and posterior distributions will differ from one another. For this prior we set $\mu_0 = 9, \tau_0^2 = 2, \sigma_0^2 = 2$, and $\nu_0 = 150$. This results in a relatively relaxed prior, there is a large range for σ^2 values to take on, and we center the prior for θ around the historical mean of the data. Using this prior will allow the signal from the data to guide the posterior distribution and deliver us the most consistent results.

After generating the prior distribution the Gibbs sampler will generate $10,000$ samples for $(\theta^1_{pre}, \ldots, \theta^n_{pre})$ and $(\theta^1_{post}, \ldots, \theta^n_{post})$. Using these values the goal is to compute

$$p(\theta_{pre} < \theta_{post} | y_{pre}, y_{post}) = \frac{1}{n}\sum_{i=1}^{n} I(\theta^i_{pre} < \theta^i_{post}) \quad (5)$$

This will give us an understanding of how the average temperature change as shifted during these time periods. Furthermore, we compare the posterior expectation and confidence intervals when looking at the distributions. Here we look at

$$E[\theta_{pre}|\sigma^2, y^1_{pre}, \ldots, y^n_{pre}] = \frac{1}{n}\sum_{i=1}^{n} \theta_{pre,i} \quad (6)$$

and

$$E[\theta_{post}|\sigma^2, y^1_{post}, \ldots, y^n_{post}] = \frac{1}{n}\sum_{i=1}^{n} \theta_{post,i} \quad (7)$$

This will also give an indication to how similar these two posterior distributions are, and if one is larger than the other.

2.3. Global Warming Denier Prior

Since there is a lot of controversy surrounding global warming and climate change, it has been associated with an anti-movement that global warming is based on pseudo-science. As such we have decided to see how the posterior distribution on θ, the average temperature change, looks. We do this in two ways, first utilizing a very strong "global warming denier" prior, meaning we assume $\theta = 0$ and is very tightly centered around 0 using a small σ^2 value. We then compare how this posterior looks when we change the prior to be more general, and have a larger range of possible values for θ.

For the first prior we have $p(\theta) \sim Norm(\mu_0, \tau_0^2)$ with $\mu_0 = 0$ and $\tau_0^2 = 0.001$, this gives a very tight distribution for θ centered at 0. Furthermore we have $p(\sigma^2) \sim Gamma(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2})$ with $\sigma_0^2 = 0.1$ and $\nu_0 = 1000$. This results again in a very tight distribution for σ^2 centered around 0.1. This prior is meant to model how a global warming denier would think the temperature changes year over year.

For the second prior, we take a looser approach and have $\mu_0 = 0$ and $\tau_0^2 = 1$ for $p(\theta)$. This creates a much broader distribution, loosely centered around 0. For $p(\sigma^2)$ we use $\sigma_0^2 = 0.1$ and $\nu_0 = 5$ to get a general prior distribution with a high likely hood for $\sigma^2 < 1.5$. This prior is meant to more generally encapsulate an opinion of someone who would believe that there is changes in temperature, but could not really guess how large.

After the generation of the priors we use the Gibbs sampler which is modeled in equations (1) and (3) for $50,000$ iterations. This results in our values for $(\theta_1, \ldots, \theta_n)$ and $(\tilde{y}_1, \ldots, \tilde{y}_n)$ where $\tilde{y}_i$ is a average temperature change from the predictive distribution. We check the quality of the Gibbs sampler using a auto correlation plot to ensure no dependency.

Using the predictions we look at where $p(\theta|\sigma^2, y_1, \cdots, y_n)$ is centered for both priors. We then find

$$E[\theta|\sigma^2, y_1, \ldots, y_n] = \frac{1}{n}\sum_{i=1}^{n} \theta_i \quad (8)$$

And look at the differences when assuming different priors.

We can also calculate

$$p(\tilde{Y} > 0|\theta, \sigma^2, y_1, \ldots, y_n) = \frac{1}{n}\sum_{i=1}^{n} I(\tilde{y}_i > 0) \qquad (9)$$

and again look at the differences when assuming different priors. We also look at the confidence intervals on the θ samples to see what the upper and lower bounds on the distribution are. Finally using the generated vector $(\tilde{y}_1, \cdots, \tilde{y}_n)$ we can generate what future possible temperature paths look like and see what percentage of them result in a warmer global temperature after $T = 100$ years.

2.4. Bayesian Regression on Average Global Temperatures

To address our last question we build a Bayesian regression model to model the month to month temperature changes, and then use this regression model to predict future temperature values. We compare three different models here. To estimate the posterior distribution on β the model weights we utilize a Monte Carlo sampler, then use $E[\beta]$ as the weight values for predicting and calculating the adjusted R^2 value. The Monte Carlo sampler works as follows.

1. sample $1/\sigma^2 \sim Gamma(\frac{\nu_0+n}{2}, \frac{\nu_0\sigma_0^2+SSR_g}{2})$
2. sample $\beta \sim MultiNorm(\frac{g}{g+1}\hat{\beta}_{ols}, \frac{g}{g+1}\sigma^2[X^TX]^{-1})$

Where $SSR_g = y^T(I - \frac{g}{g+1}X(X^TX)^{-1}X^T)y$ and $\hat{\beta}_{ols} = (X^TX)^{-1}X^Ty$ [2].

This sampler requires a prior with values g, ν_0, σ_0^2. Here we choose a very simple prior setting g to be the number of data points, $\nu_0 = 2$ to loosely center the prior, and $\sigma_0^2 = 1$.

The first model that we propose is a simple linear regression model with

$$\hat{y}_1 = \beta_0 + x\beta_1 \qquad (10)$$

The second model is a harmonic regression model in an attempt to capture the seasonality,

$$\hat{y}_2 = \beta_0 + sin(\frac{2\pi}{12}x)\beta_1 + cos(\frac{2\pi}{12})\beta_2 \qquad (11)$$

Finally the third is a harmonic regression model with an added linear term to gauge if there is a statistically significant linear trend in the seasonality.

$$\hat{y}_3 = \beta_0 + x\beta_1 + sin(\frac{2\pi}{12}x)\beta_2 + cos(\frac{2\pi}{12}x)\beta_3 \qquad (12)$$

We then look at the 95% confidence interval of the posterior distribution for β, this can tell us which parameters in the model are significant. The aim here is to show that confidence interval for the linear term in model 3 does not contain 0, meaning there is a noticeable linear increase in temperature over time. Furthermore we look at the plot of the residuals of models 2 and 3, and compare how they predict, and we take the best model and use it to predict future temperature values.

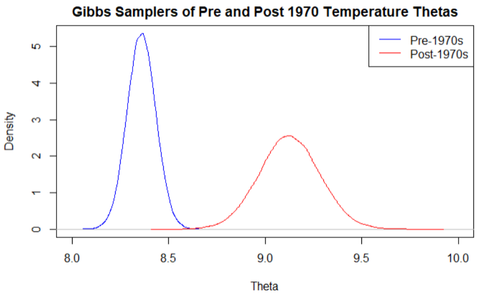


Fig. 5. θ Values of Pre-1970 vs Post-1970 Temperatures

3. RESULTS

3.1. Results for Comparing Temperatures Pre and Post 1970

After running the Gibbs sampler on our 2 datasets (pre-1970 and post-1970), using the same prior applied to both samplers we generate our 10,000 values of θ. We can compare the posterior distributions between the sampled θs.

Figure 5 shows that using the same prior set for both samplers results in two completely different posterior distributions. While the pre-1970 θ distribution demonstrates a tight, peaked distribution with thin tails, centered around values 8.35, with a 95% confidence interval of 8.21 to 8.50. We compare this to the broader post-1970 θ distribution which contains almost no overlap with the pre-1970 θ values, centered around value 9.13, with a 95% confidence interval of 8.82 to 9.43. When computing (5), overlap between the θ values of the 2 distributions was found to be essentially 0. Given this almost nonexistent overlap between these two distributions, we can conclude with a very strong degree of certainty a significant rise in temperature values from pre-1970 to post-1970, as well as higher variance in year to year temperature values post-1970, as demonstrated by the broader distribution, corroborating prior information regarding global warming contributing to more extreme climate shifts year to year.

3.2. Results for Global Warming Denier Prior

Using the global warming denier prior the Gibbs sampler runs successfully which is corroborated by the auto correlation plot showing no dependency's in the Markov chain. When using the more relaxed prior we also see a successful run of the sampler with no correlation in the chain. Table 1 shows the posterior expectation (8) and 95% confidence interval as well as the probability of the predictive distribution being above zero (9) under each prior. Figure 6 shows the predictive distributions overlaid on the actual, and Figure 7 shows what 100 possible future paths look like when using the relaxed prior.

As we can see both priors lead to distributions where zero is contained within the confidence interval. The distributions appear very similar however under the relaxed prior the cen-

	Confidence Interval on θ			$p(\tilde{Y} > 0 \mid \theta, \sigma^2, y_1, \cdots, y_n)$
	2.5%	50%	97.5%	
Strict Prior	-0.02977820	0.007556693	0.04494966	0.50724
Relaxed Prior	-0.02307734	0.01174582	0.04630086	0.51672

Table 1. Temperature change posterior distribution comparison under two priors

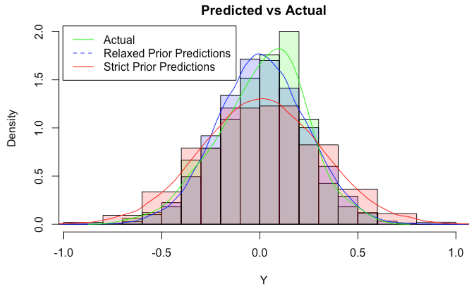


Fig. 6. Histogram of predictive temperature change distributions

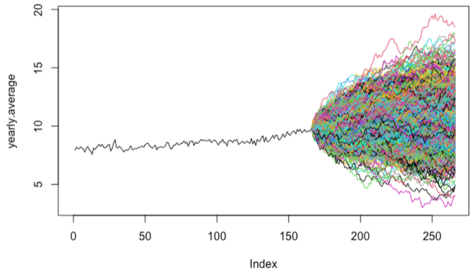


Fig. 7. Monte Carlo predictions of future temperature paths

Model 1	Model 2	Model 3
0.004	0.9862	**0.9933**

Table 2. Regression adjusted R^2 values

ter if slightly shifted away from zero. Both priors also lead $p(\tilde{Y} > 0|\theta, \sigma^2, y_1, \cdots, y_n)$ to be about 50%, where the true given data has $p(y > 0) = 0.55$. These results are a good indication that the prior does not have much affect on the posterior, meaning our data has a very strong signal.

With that said we can see in figure 6 that the strict prior leads to a much worse modeling of the data, the predictive distribution has much heavier tails and not as high of a peak. When looking at figure 7 we can see that the temperature paths form a nice normal distribution out from the end of our data. When comparing the generated temperature paths under each prior, it was found that under the relaxed prior about 70% of these result in an increased temperature, and under the strict prior about 60% of them result in an increased temperature from the start point.

When looking at all of these values, the posterior distributions are very similar, and only lightly contain zero on the left rails, and taking into account the randomly generated paths we can see that our predictive distributions are leaning in favor of increasing temperatures.

This is an interesting experiment because as shown in the preliminary data analysis the year over year temperature changes are relatively static around zero. The data seems to bounce around noisily between 0.6 and -0.6, however from the randomness we do see a slight tilt towards a positive non zero average. Both of our posterior distributions, and posterior expectations seem to corroborate these observations.

3.3. Bayesian Regression on Average Global Temperatures

The Monte Carlo sampler was used to draw 1,000 values for β for each of the three described models. Table 3 shows the posterior expectation and 95% confidence intervals for each of the models, table 2 shows the adjusted R^2 values that each model achieved. Figure 8 shows the regressions of models 1 and 2 overlaid on the data and Figure 9 shows the residuals for models 2 and 3 plotted against each other.

Some very interesting results came out of this experiment, we can see that model 1 picks up a increasing linear trend in the data, this can be seen in the confidence interval for β_1 which does not include zero, making the linear term useful. With that said, the R^2 value for model 1 is very poor, since the model is basically just predicting the mean of the data.

More interesting results come out of comparing model 2 and model 3. We see that both of them achieve very good adjusted R^2 values. None of the confidence intervals for the β values in either model contain zero, making each term a useful predictor. This is a good result because we can see a slight increase in the adjusted R^2 value for model 3 when we include the linear term, and the linear terms distribution does not contain zero. This is evidence that global average temperatures are increasing during the time period that we are analyzing. The harmonic terms $sin(\frac{2\pi}{12}x), cos(\frac{2\pi}{12}x)$ do a very good job of picking up the yearly seasonal trends within the data. In figure 8 we can see the slight visual difference between model 2 and model 3.

To even further reinforce the importance of the linear term, looking at figure 9 we can see that model 2 is not picking up any kind of linear increase in temperature. This is evident from the residuals, we see that model 2 under predicts the data in the first half, and over predicts the data in the second half of the set. Model 3 does not struggle with this which can be seen from the straight line and normally

	Confidence Interval on β											
	β_0			$x \cdot \beta_1$			$sin(\frac{2\pi}{12}x) \cdot \beta_2$			$cos(\frac{2\pi}{12}x) \cdot \beta_3$		
	2.5%	50%	97.5%	2.5%	50%	97.5%	2.5%	50%	97.5%	2.5%	50%	97.5%
Model 1	7.75	8.44	9.18	0.00019	0.0023	0.0045	-	-	-	-	-	-
Model 2	9.06	9.11	9.16	-	-	-	-3.32	-3.24	-3.16	-4.89	-4.82	-4.74
Model 3	8.41	8.51	8.60	0.0018	0.0021	0.0024	-3.30	-3.24	-3.17	-4.89	-4.82	-4.76

Table 3. Confidence intervals for regression weights

distributed residuals around zero.

The one thing that the model 3 does not pick up on well is the noise at the peaks of the seasonality, this can be seen in figure 8 by the points that fall above and below the models predictions.

A fourth model was considered for this problem

$$\hat{y}_4 = \beta_0 + x\beta_1 + x^2\beta_2 + sin(\frac{2\pi}{12}x)\beta_3 + cos(\frac{2\pi}{12}x)\beta_4 \quad (13)$$

however it would seem after the analysis that the exponential variable did not add any significant value to the model, as zero was comfortably contained within the confidence interval for this variable. The idea behind this model was to see if there was any sort of exponential increase in temperatures through the seasonality, but model 3 shows evidence for it just being a linear increase.

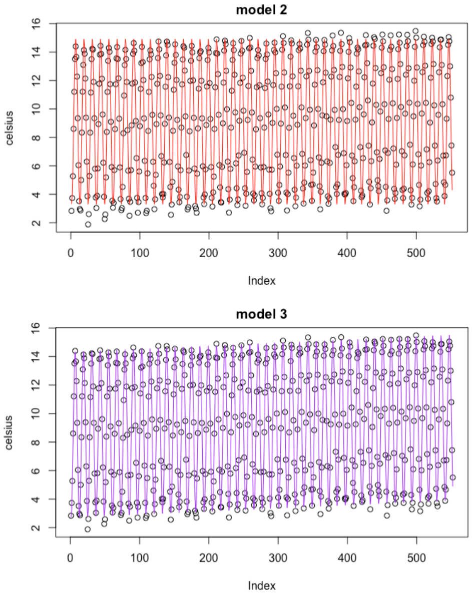


Fig. 8. Regression model predictions

4. CONCLUSION

Overall, across all research questions, we have essentially been attempting to determine the extent to which global warming has progressed throughout the years. We have successfully established not only that global warming is a real trend given the data even with a very strong prior assumption that it did not exists, but also has been occurring for quite some time. We have further established the degree to which global warming has affected average global yearly temperatures, and have also been able to predict a continuous rising trend in yearly average temperatures.

While this research was fairly robust in answering the proposed research questions, the natural extension in addressing global warming is addressing the role that humans and industrialization and other human activities have had in global warming. We can try to further address questions such as, ”Has human activity been the main driver of global warming for the past century?”, as well as ”Can we predict how future temperature values will respond if levels in human activity related to global warming and climate change stay constant?”. It would also be interesting to pair this dataset with a dataset that contains human factors as described. We could then preform a Bayesian analysis to see if we can find correlation between the two data sets, and see where things interact.

During this analysis we only used one of the descriptor variables available to us. The rest of the dataset has the same variables, but is collected country by country. It could be interesting to analyse the differences between countries, if countries have different starting times for when temperatures began increasing, or if they have increased at different rates.

When comparing the posterior distribution on θ under different priors we came across the results that although it seems that the change in temperature is relatively static year to year, over the long run the distribution leans just right enough to have the temperature increase. Something that could be done to continue analyzing this problem in the future is to do a Bayesian analysis on the Markov chain that would denote the rising and falling of temperatures year to year. This method would most likely yield strong results to further reinforce what we had found in that section. We would be looking for some sort of distribution that shows that increases in temperature occur more often in the chain than decreases, and the stationary matrix would have the increase state at a higher probability than the decrease state.

Another form of future work could be done in the regression analysis. The data seems to have noisy peaks, and the best preforming regression model did not learn anything about them. There may be some other way to formulate the regression model that would allow for variability in amplitudes of the sine and cosine waves, which would then further

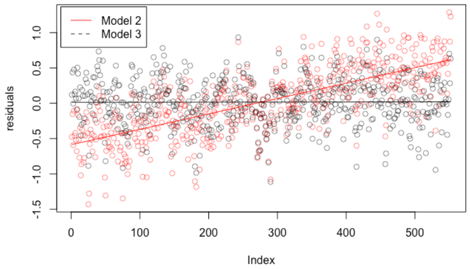


Fig. 9. Regression model residuals

increase the adjusted R^2 value. With that said, complicating the model more to account for the variability in peaks could also lead to overfitting, since the data may just be noisy, and there is no underlying trend governing the size of the peaks.

5. REFERENCES

[1] E. Berkeley, "Climate change: Earth surface temperature data." Kaggle.

[2] P. Hoff, "A first course in bayesian statistical methods." Springer.

[3] "Climate change indicators: U.s. and global temperature." United States Environmental Protection Agency.