Climate Change: Earth Surface Temperature

Bayesian Machine Learning Project Report

Problem Description

The goal is to derive a robust understanding of long-term climate trends, considering the impact of historical factors on temperature variations. Through Bayesian regression, we aim to provide insights into the global climate dynamics, addressing concerns and contributing to a nuanced perspective on climate change—a topic of significant societal and environmental importance.

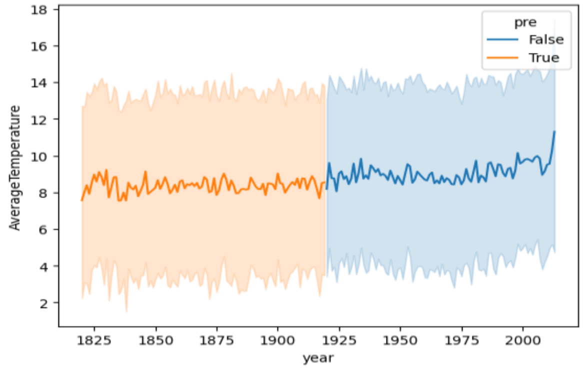
Data Description

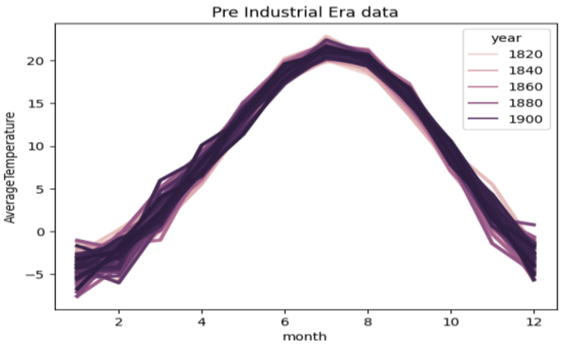
The dataset consists of around 550,000 observations from various countries. To study the effects of environmental factors or industrialization in detail, we focused on the data from the United States. The variables present in the dataset are as follows:

1. dt - starts in 1750 for average land temperature
2. AverageTemperature - average land temperature in celsius
3. AverageTemperatureUncertainty - the 95% confidence interval around the average

Due to unclear reasons about how uncertainty is measured back in 1750, we decided to scale the 'AverageTemepratureUncertainty' column between 0 and 1. Missing values in 'AverageTemperatureUncertainty' are imputed by mean values, and missing values in 'AverageTemperature' column are imputed using linear interpolation method, leveraging the technique to estimate temperature values between known data points.

Further to ease the data analysis, 'dt' column is set to index. We extracted 'Year' and 'Month' from dates and created new columns each. Also, to keep the track of rows we added a 't' column.




Pre Industrial Era data

The right plot above shows yearly maximum average temperature from 1820 to 2000. As there is a slight upward trend in the average temperature, we split the data into pre-1920 (train) and post-1920 (prediction) datasets to examine the change in atmospheric temperature after the start of industrialization.

The left plot above shows the seasonality present in the data. This seasonality is consistent across each twenty year span. Indicating even with the slight increase present across time post 1920 era, the seasonality remains predominantly the same.

Probability Model

The Bayesian linear regression model, as implemented in the provided code, offers a probabilistic framework for analyzing temperature-related data. Despite continuous reparameterization attempts and various adjustments in the sampling size, tuning parameters, and standard deviation, achieving convergence for this model, particularly using NUTS and other traditional sampling methods like Metropolis, has proven to be challenging.

The coefficients in the model, represented by the priors seasonality and trend, aim to capture the intercept and slope of the linear relationship between 'AverageTemperature' and the onset of the Industrial Age. It is assumed that errors follow a normal distribution, where mu denotes the distribution's mean and sigma represents its standard deviation. The likelihood function, denoted as 'likelihood,' expresses observed temperatures in terms of these model parameters. Despite the convergence challenges, efforts have been made to fit the model adequately to the problem by adjusting priors, tuning, and sampling sizes.

To address the lack of convergence, various strategies have been employed, such as adjusting priors and reparameterizing the model. These modifications aimed to enhance the convergence of the posterior distributions of the model parameters, including 'alpha,' 'beta,' and 'sigma.' However, despite these attempts, achieving satisfactory convergence and stability in the model's posterior inference has remained elusive.

In summary, while continuous reparameterization and adjustments in sampling size, tuning parameters, and standard deviation have been made to improve model convergence, the Bayesian linear regression model's full convergence and stability have not been achieved yet. This ongoing challenge persists in accurately characterizing the uncertainty and estimating the relationship between 'AverageTemperature' and the Industrial Age through this model.

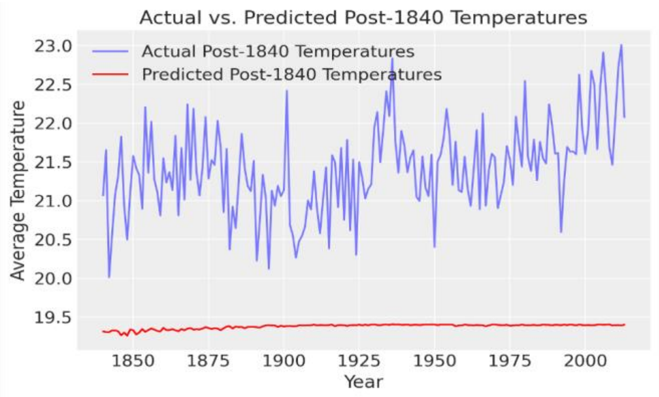
Approach

In tackling the problem of Bayesian linear regression to analyze temperature-related data, several approaches were employed to model the relationship between 'AverageTemperature' and the onset of the Industrial Age. Initially, efforts were directed towards formulating a comprehensive Bayesian linear regression model. This involved defining prior distributions for the model parameters, including intercepts, slopes, and error terms, aiming to capture uncertainties and quantify the linear association between the variables. The utilization of different sampling methods, such as NUTS (No-U-Turn Sampler), Metropolis, and was pivotal in exploring the posterior distributions of the model parameters and handling complex probabilistic calculations. These methods, if converged, would allow for a probabilistic characterization of the linear relationship and enable the assessment of uncertainties associated with the model predictions.

Though continuous reparameterization attempts were made, manipulating priors, tuning parameters, and adjusting sampling sizes to enhance convergence and stability within the model. These adjustments aimed to improve the efficiency of sampling algorithms and achieve convergence in the posterior distributions of the model parameters. Trying to even take in on the maximum temperatures of each year, and the maximum and minimum, in order to reduce the sample variation. Despite these efforts, challenges in achieving convergence persisted, necessitating an iterative process of refinement and exploration of different approaches. Each approach was aimed at refining the model, addressing convergence issues, and gaining deeper insights into the uncertainty and relationship estimation between 'AverageTemperature' and the Industrial Age.

Results

Through the observed data points we were able to note that there was an increase in post 1920. With the prior predictive model it was seen that the observed data in comparison to the prior predictive was greater. When sampling, the model was unable to converge with a r-hat that is around 1.2 and effective sample size that was less than 400 for multiple coefficients.



We had implemented another method in an ensemble model that included ridge regression and bagging. This was the only model that presented us with the observed and predicted comparisons for the post era.

We implemented this so that it could deal with uncertainty robustly, but this may have not been the best model to implement, but was the only one that delivered results.

Conclusions

In order to identify underlying patterns and an increasing trend in the global average temperature, we fitted the model using time series from Bayesian regression and included the components "seasonality" and "linear trend."

Due to divergence and mismatch in data frame dimensions we were unable to plot the posterior predictive observations which would help in showing uncertainty in the model.

The original dataset lacks additional possibly related variables such as CO2 emissions, industry-based influencing factors. The accuracy of the posterior prediction may have also been impacted by the absence of these predictors and their interaction effect. In addition, we did not have computers that were able to handle the complexity of sampling. We continuously ran out of RAM and had random crashes occur.

In conclusion, if we had chosen a model with more predictor variables and software that could handle the computational expensive nature of sample steps like No U turn Hamiltonian Monte Carlo, we may have produced meaningful results that showed the predictive power and uncertainty of our selected Bayesian models.

References

Berkeley Earth (2023). "Climate Change: Earth Surface Temperature Data". Kaggle. [Online]. Available: https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data/data. Accessed: December 2023.

Dennis Kimutai Koech. 2022. "Title of the Article or Webpage." Section.io. [Online]. Available: https://www.section.io/engineering-education/missing-values-in-time-series/. Accessed: December 2023.

Vincent, Benjamin T. (2023). "Counterfactual inference: calculating excess deaths due to COVID-19," PtMC Team (Eds.). DOI: 10.5281/zenodo.5654871.

Github to access code (ReadMe explains pertinent files purpose): https://github.com/psa7rm/Bayesian-6040-Project/tree/main

Appendices

Chart A: Graphic Diagram of bayesian model used on monthly data. Student T being implemented in an attempt to improve robustness and convergence.

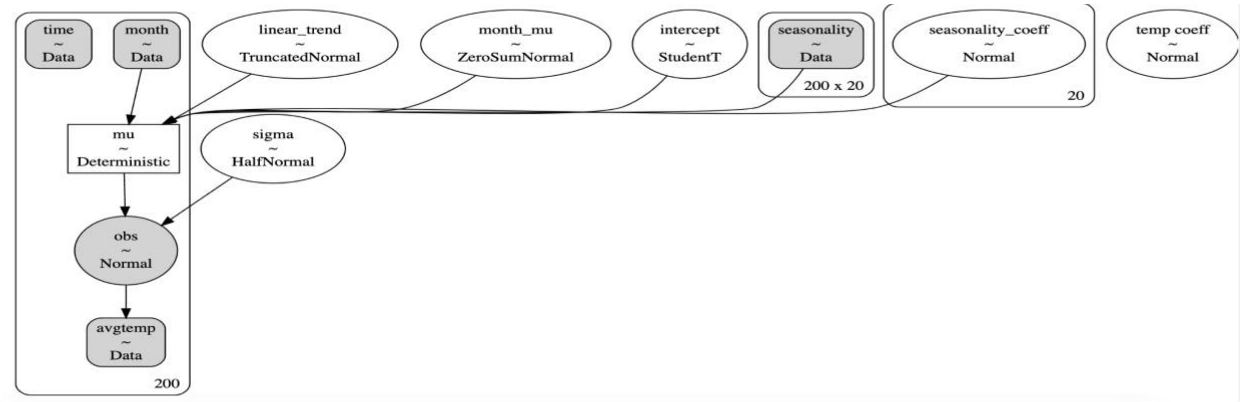
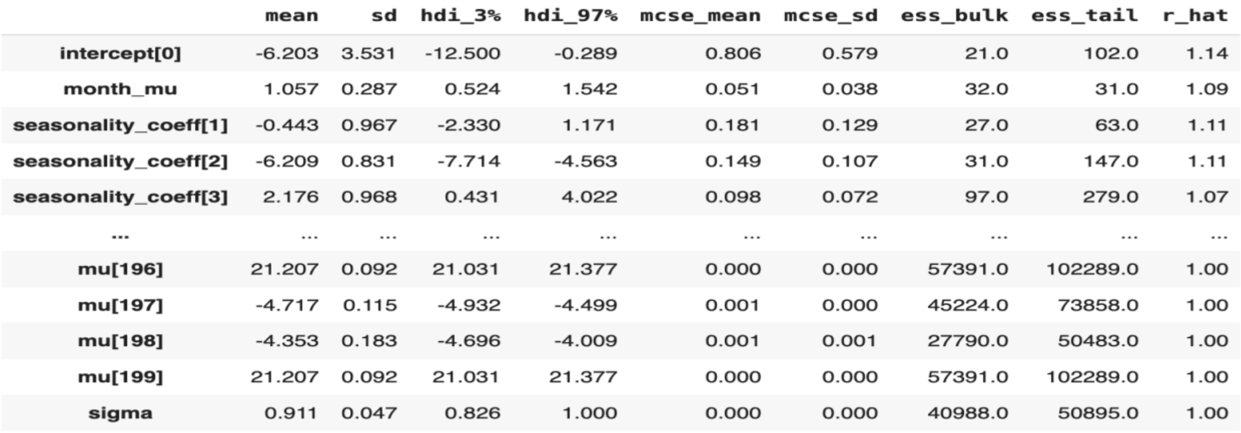
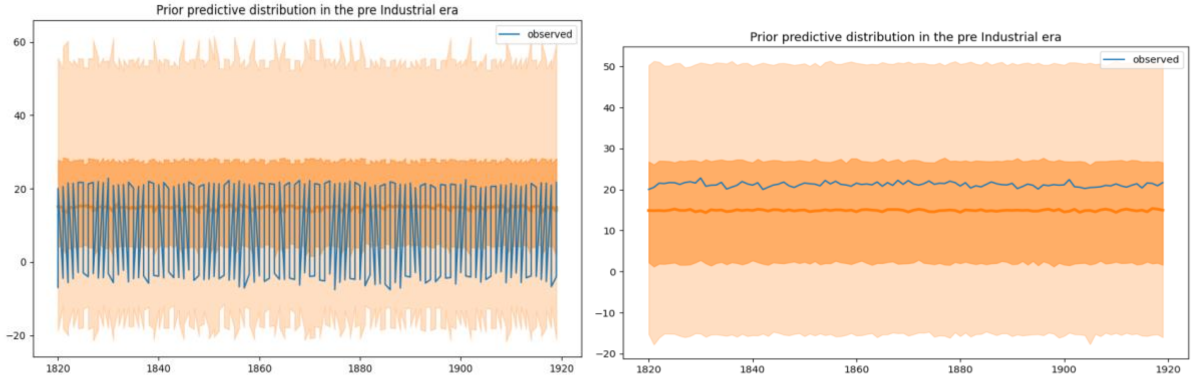


Chart B: Summary diagram of bayesian model showing low effective sample size and high r-hat.



	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
intercept[0]	-6.203	3.531	-12.500	-0.289	0.806	0.579	21.0	102.0	1.14
month_mu	1.057	0.287	0.524	1.542	0.051	0.038	32.0	31.0	1.09
seasonality_coeff[1]	-0.443	0.967	-2.330	1.171	0.181	0.129	27.0	63.0	1.11
seasonality_coeff[2]	-6.209	0.831	-7.714	-4.563	0.149	0.107	31.0	147.0	1.11
seasonality_coeff[3]	2.176	0.968	0.431	4.022	0.098	0.072	97.0	279.0	1.07
...	...	...	...	...	...	...	...	...	...
mu[196]	21.207	0.092	21.031	21.377	0.000	0.000	57391.0	102289.0	1.00
mu[197]	-4.717	0.115	-4.932	-4.499	0.001	0.000	45224.0	73858.0	1.00
mu[198]	-4.353	0.183	-4.696	-4.009	0.001	0.001	27790.0	50483.0	1.00
mu[199]	21.207	0.092	21.031	21.377	0.000	0.000	57391.0	102289.0	1.00
sigma	0.911	0.047	0.826	1.000	0.000	0.000	40988.0	50895.0	1.00

Chart C: Plot of prior predictive model of minimum and maximum ‘AverageTemperature’ [left] and only maximum ‘AverageTemperature’ in right.


Prior predictive distribution in the pre Industrial era
observed
Prior predictive distribution in the pre Industrial era
observed