Data 102 Final Project

#Group 47

Data Overview

We use publicly available data from the U.S. Energy Information Administration's State Energy Data System (SEDS) (1960–2022) and the EPA's Greenhouse Gas Inventory (1990–2022). These sources provide a comprehensive census of energy use and emissions data for all U.S. states. Each row in our merged dataset represents a state-year pair, capturing energy consumption, population, regional classification, and greenhouse gas emissions.

We used pandas groupby operations to aggregate total greenhouse gas emissions for each state and year. Although both the EPA and SEDS datasets report CO_2 emissions, they rely on different data collection methodologies, resulting in some discrepancies in reported values. However, overall emission trends remain consistent across both sources.

A key limitation of the SEDS data is the lack of transparency in how CO_2 emissions are calculated. Upon closer inspection, we found that fuel-specific emissions variables (e.g., coal, petroleum, natural gas) are likely components of the response variable. Thus, we excluded these variables in research question 2. It is also important to note that the EPA data was used exclusively for question 2.

Since our data is a census, there is no sampling error, but certain systematic exclusions remain, notably, missing or imputed values in early years for some fuel types or GHGs. Additionally, we could not obtain consistent state-level data on policy interventions or political ideology, which limits our ability to control for these important confounders in causal inference.

The dataset underwent extensive cleaning: we reformatted energy units, removed U.S. total summary rows, merged data sources on state and year, and converted population values to numeric. To address the missing data in Research Question 2, we chose to drop observations from 1970 to 1990. The remaining data from 1990 to 2022 provides a sufficiently large time window to train an accurate predictive model. We also calculated the percentage share of energy consumption for each sector, such as residential, industrial, and transportation, to better capture the structural composition of state-level energy use.

Prior Work

Dietz et. al's 2015 paper, "Political influences on greenhouse gas emissions from US states", explores the implications of political ideology and institutional factors on each U.S. state's greenhouse gas emissions using linear regression and multilevel modeling. Similar to our research question, the authors worked with state-level, annual CO_2 emissions from fossil fuel combustion data provided by the Environmental Protection Agency, closely related to the provider of our data, the Energy Information Administration.

Both Dietz and coauthors' work and our method assume a linear relationship between treatment and outcome, and use the state as a unit of analysis, and observe state-level emissions over time. However, the paper cited here estimated hierarchical models to account for potentially unobserved heterogeneity across units. There are two levels: level 1 looks at changes within a state at a base year; level 2 compares states. On the other hand, we use an Ordinary Least Squares regression model assuming unconfoundedness. Even though the authors' method is somewhat within the scope of what we've learned in class with Bayesian Hierarchical Modeling, we opted for outcome regression to better establish causality between baseline emission level and reduction rate.

Costantini et. al's study, Forecasting national CO_2 emissions worldwide, employs both Random Forest Regressor and classical multivariate regression models to forecast CO_2 emissions across 117 countries, utilizing 12 socioeconomic indicators, including energy consumption per capita. The research demonstrates that Random Forest models outperform traditional regression models in capturing complex, nonlinear relationships between predictors and CO_2 emissions. This aligns with our approach in Research Question 2, where we leverage Random Forest to model the nonlinear dependencies between energy usage and CO_2 emissions. While this study operates at a global scale, our analysis focuses on U.S. state-level data, providing more granular insights. Additionally, our feature set includes specific energy sector shares and electricity trade indices, which may offer different perspectives compared to the broader socioeconomic indicators used in the cited study.

EDA

Quantitative variables

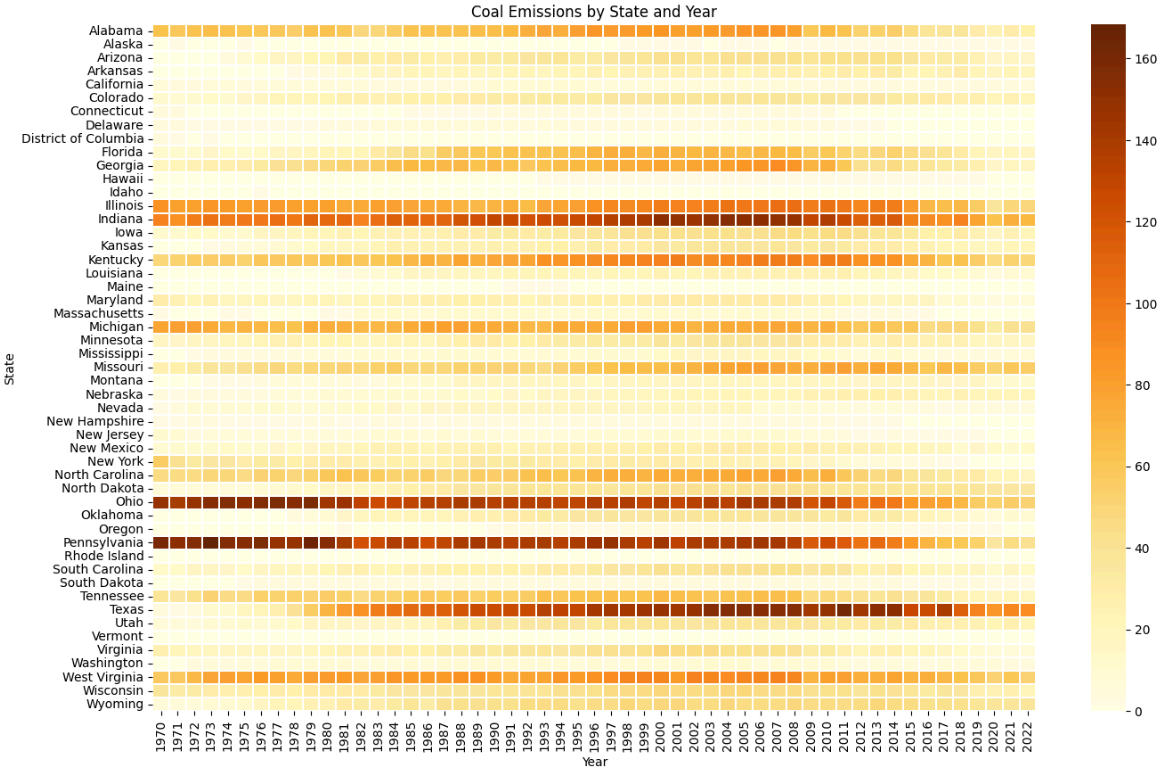


Figure 1: Coal Emissions by State and Year

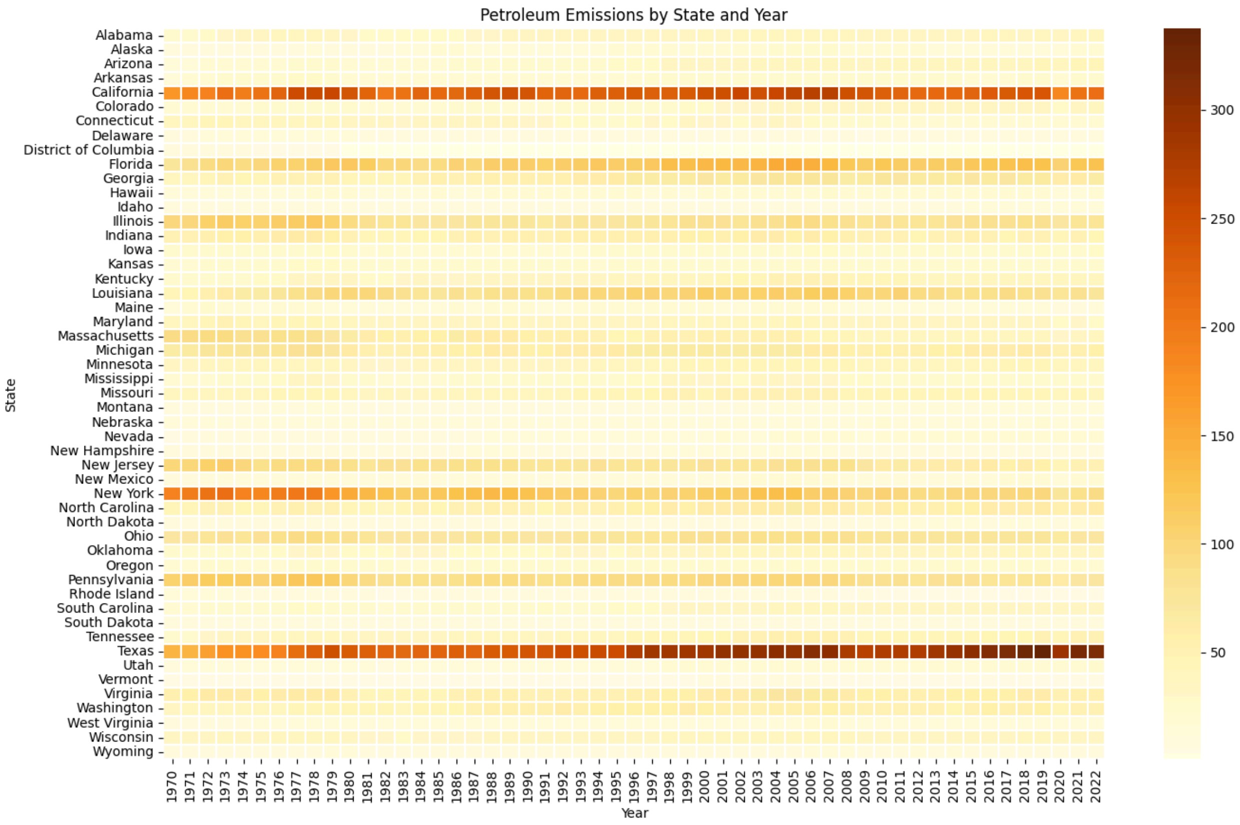


Figure 2: Petroleum Emissions by State and Year

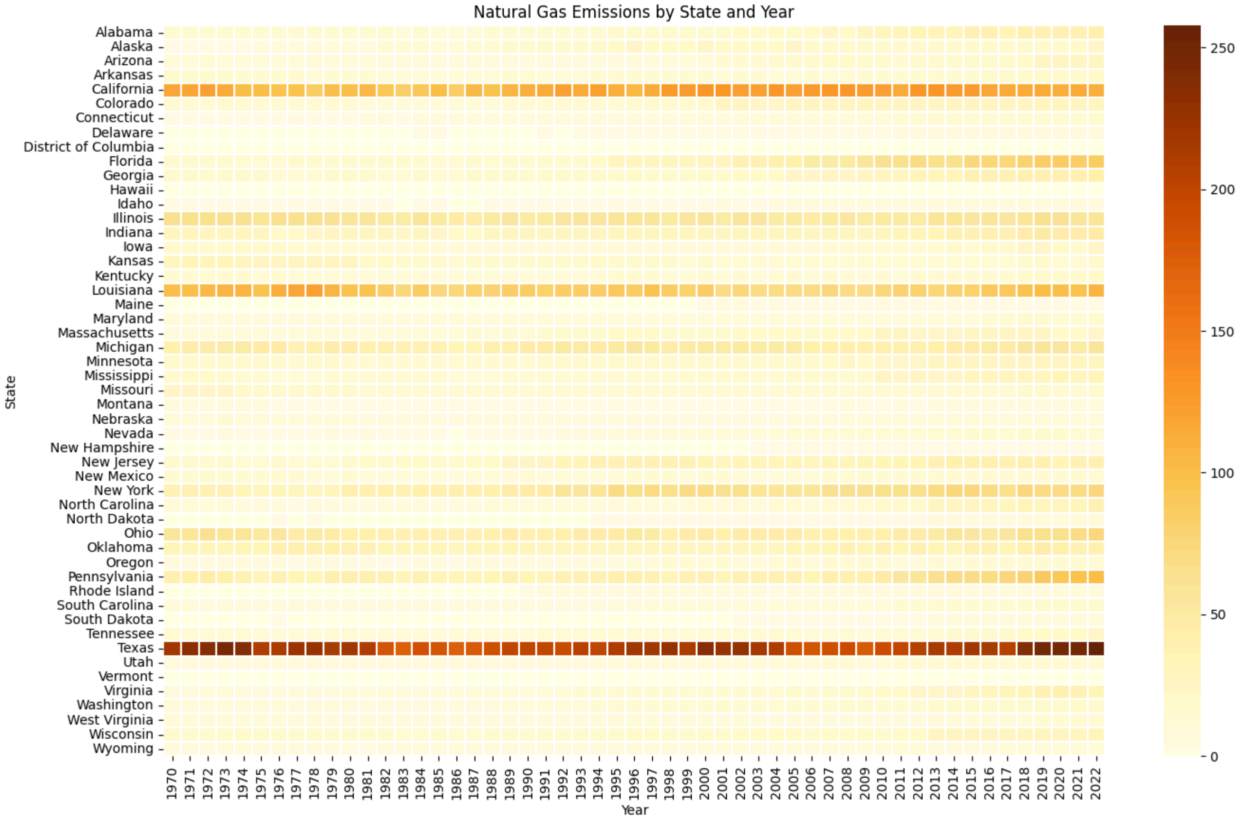


Figure 3: Natural Gas Emissions by State and Year

Figures 1, 2, and 3 depict natural gas emissions by state from 1970 to 2022 on different energy types, which addresses our second research question. We see that Texas has high emissions of coal, petroleum, and natural gas. Coal emissions show a sharp decline in many states starting around the 2000s, suggesting a shift away from coal as a source of energy usage. Petroleum emissions don't show the same trend, with emissions remaining steady and even increasing for some states. Natural gas emission is dominated by Texas, but emission also increases steadily among the top 4 emitting states. The graphs also suggest that when creating a causal model, we should consider emissions by source, not just in total.

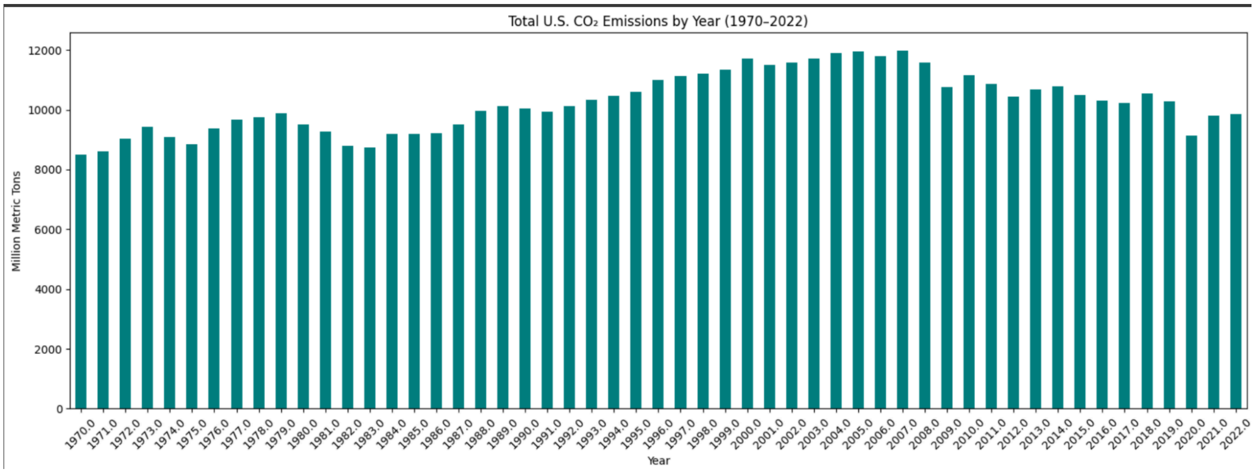


Figure 4: Total U.S. CO2 Emissions by Year (1970-2022)

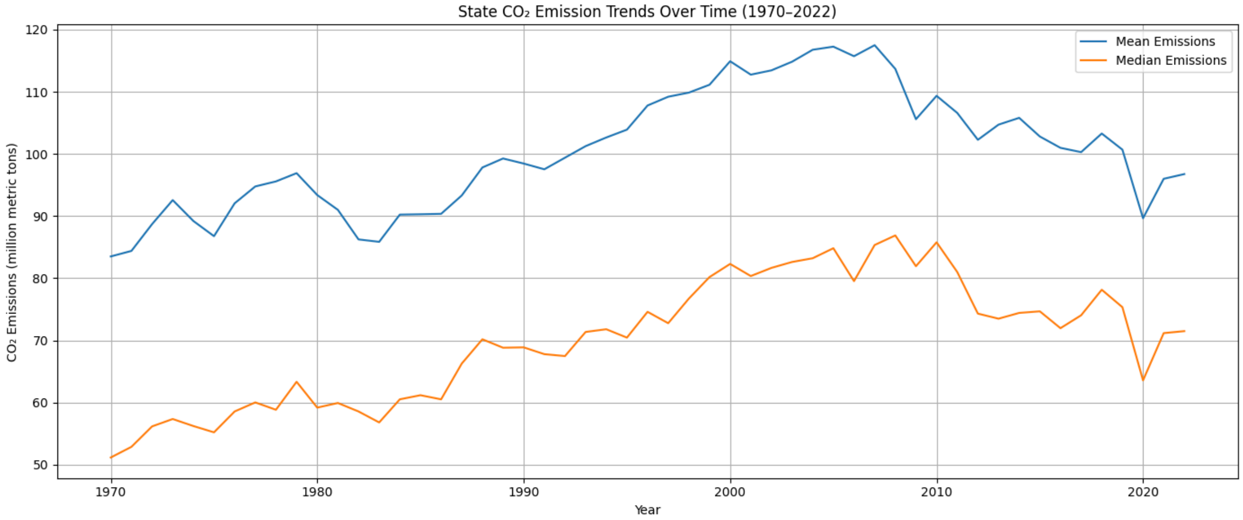


Figure 5: State CO2 Emission Trends Over Time(1970-2022)

Figures 4 and 5 help address research question 2 in the sense that the average CO₂ emissions per state chart identifies which states start with high baseline levels of emissions, including Texas, California, and Pennsylvania. We can also see that the national emissions over time chart shows a non-linear trend where there is a sharp decline after a steady state, which, using a simple linear model may not capture key policy shifts.

Categorical Variables

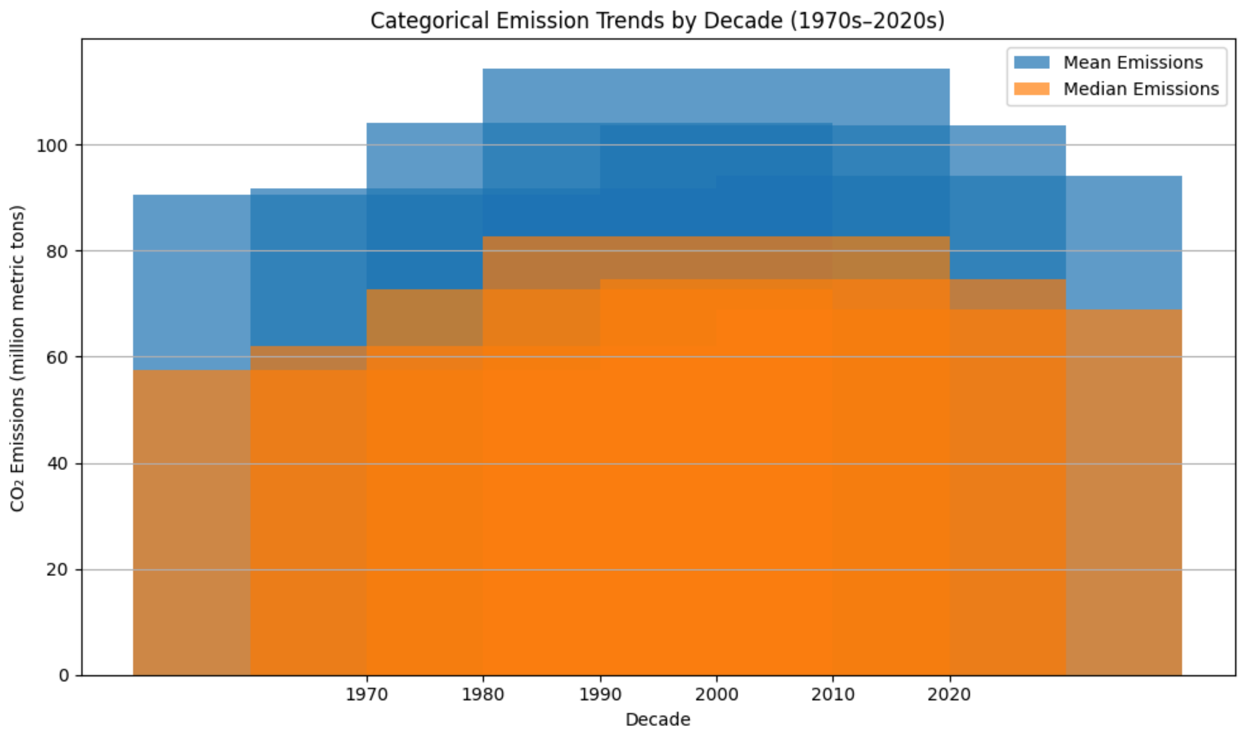


Figure 6: Categorical Emission Trends by Decade (1970-2020)

This grouped bar chart compares the mean and median CO_2 emissions per state by decade from the 1970s to the 2020s. It shows that emissions increased across all states until the 2000s, then declined in the 2010s and 2020s. The gap between mean and median highlights that a few high-emission states disproportionately raise the average. The decadal grouping simplifies the trend and helps us frame "baseline periods" for comparison — e.g., did states with high 1970s–1990s emissions show steeper declines after 2000? Thus, Figure 6 also helps resolve research question 1.

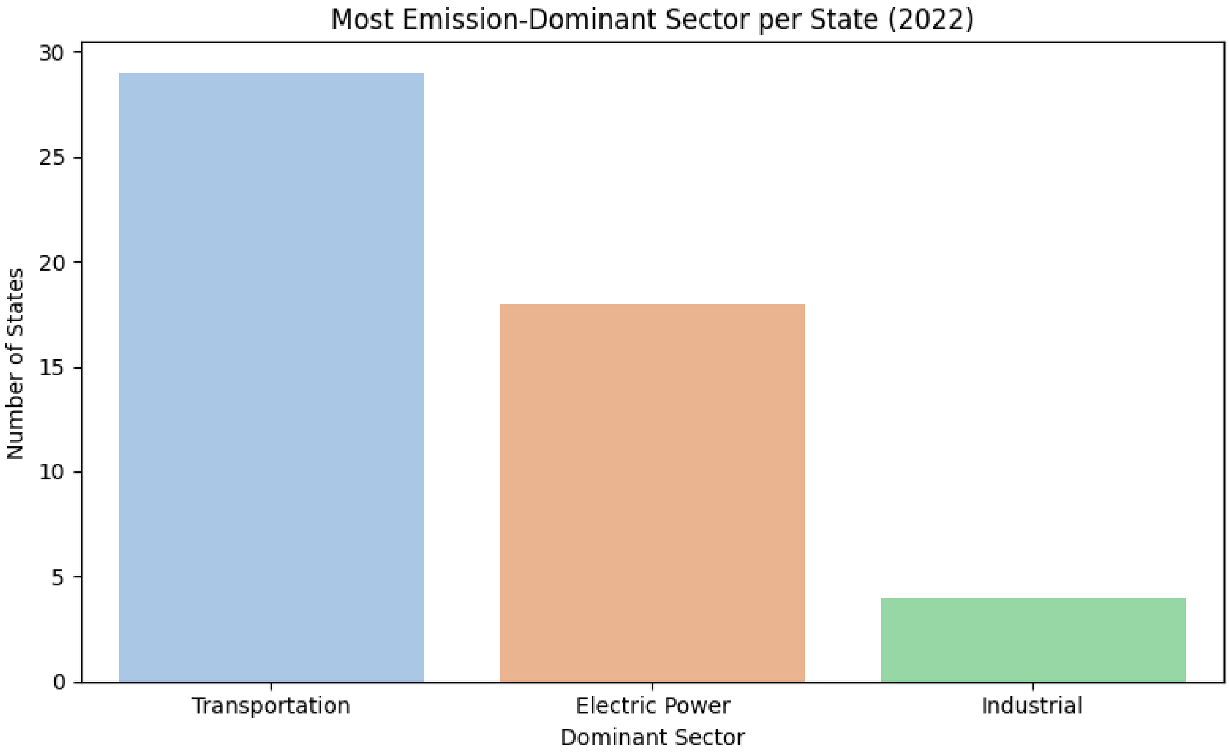


Figure 7: Most Emission-Dominant Sector per State (2022)

Figure 7 helps identify a potential confounder—sector dominance—that should be controlled for in causal analysis of baseline emissions and reduction rates. Most states are dominated by the transportation sector, followed by electric power, with few led by industrial emissions. There is sectoral variation across states, suggesting that emission sources are not uniform. This matters because the type of dominant sector may influence how a state reduces emissions over time (e.g., power-sector reforms vs. transportation policy).

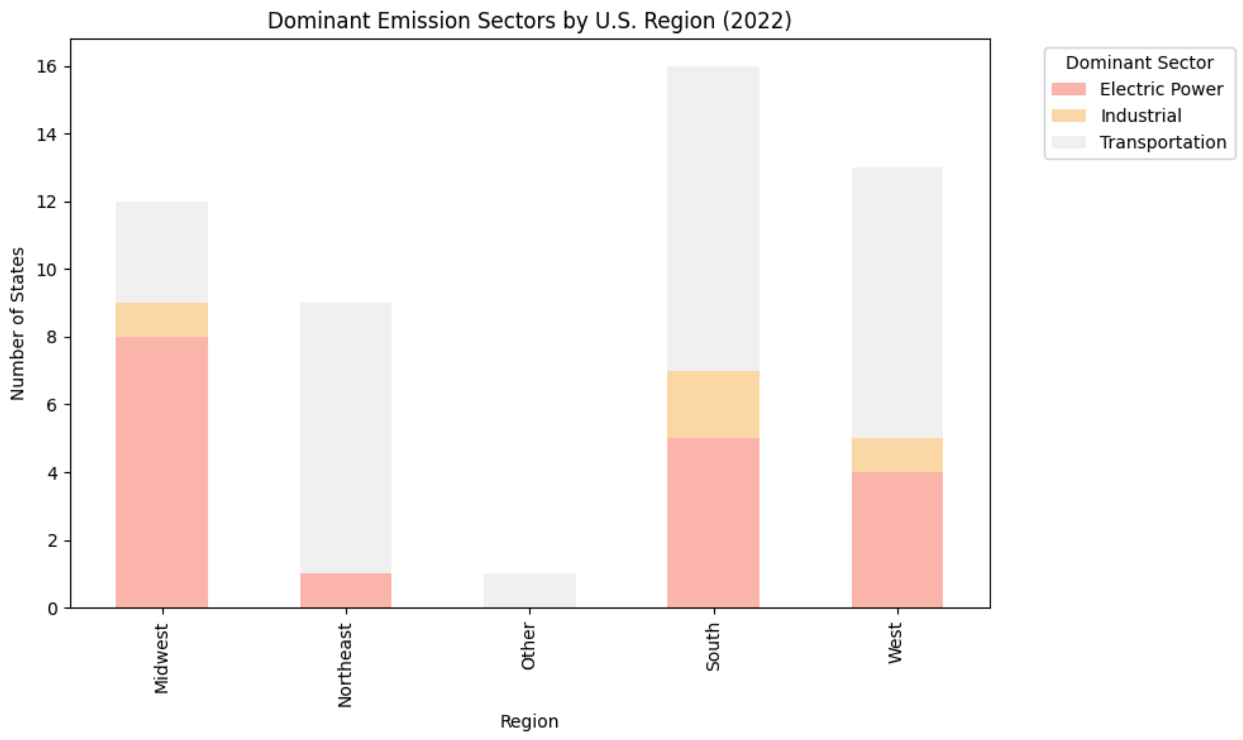


Figure 8: Dominant Emission Sectors by U.S. Region (2022)

This visualization reveals regional patterns in sector dominance, which may introduce a new confounding variable for research question 1. The Midwest is heavily dominated by electric power emissions, while the Northeast and West are led by transportation. The South shows a more even mix, including industrial sources. In regions dominated by electric power, energy usage and emissions might follow a more direct (possibly linear) relationship, driven by coal or gas consumption. In transportation-dominated states, emissions may depend on population density, vehicle type, and urban infrastructure, leading to nonlinear dynamics. This suggests that simple linear models (like GLMs) may oversimplify the underlying structure.

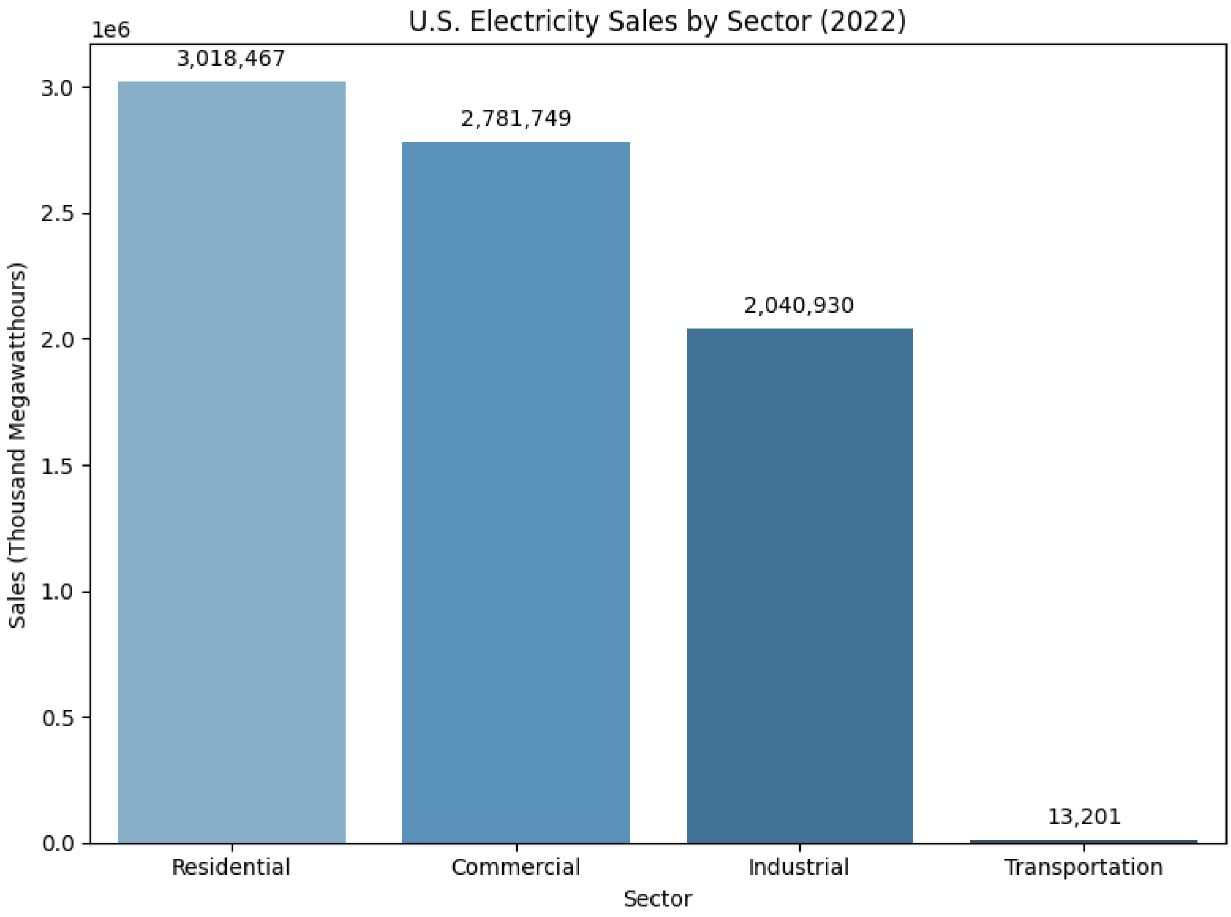


Figure 9: U.S. Electricity Sales by Sector (2022)

From Figure 9, we can see that the residential and commercial sectors account for the highest electricity sales, followed by industrial, with transportation contributing negligibly. This introduces a new confounding variable for research question 1.

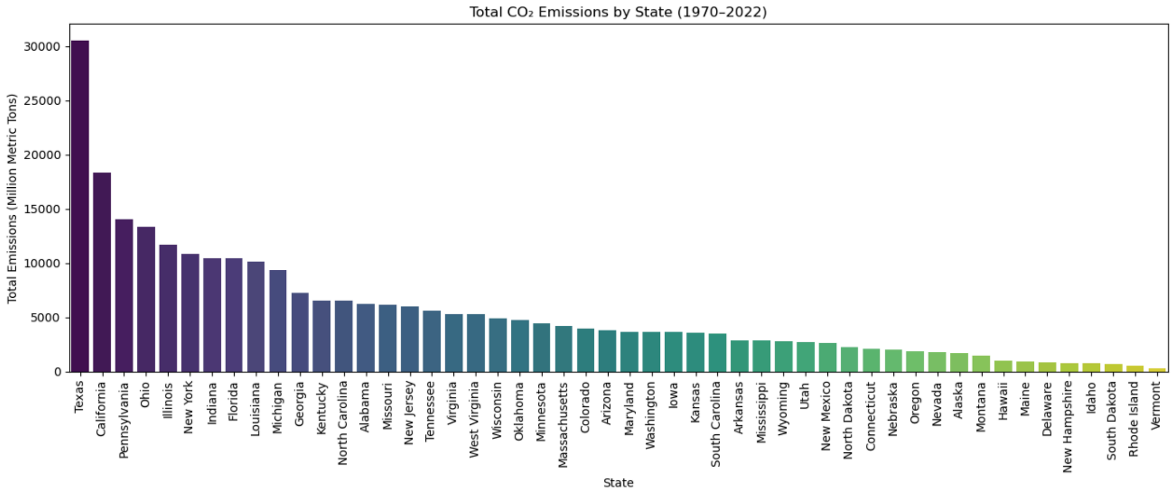


Figure 10: Total CO2 Emissions by State (1970-2022)

Figure 10 shows a clear disparity between states in total CO2 emissions from 1970 to 2022, with Texas, California, and Pennsylvania emitting significantly more than other states. There is a large drop-off in emissions after the top few states, suggesting that emissions are largely concentrated geographically. This will directly help with research question 1 regarding baseline CO2 emissions and their effect on the rate of reduction of emissions over time.

Research Questions

1. Does a higher baseline level of CO_2 emissions causally influence the rate of reduction in emissions over time across U.S. states?
2. Can Bayesian generalized linear models accurately predict carbon emissions from energy usage and related variables, or do nonparametric methods such as random forests uncover nonlinear patterns that improve predictive performance?

Our first research question aims to establish a causal relationship through outcome regression, assuming unconfoundedness and a linear relationship between treatment and outcome. Causal inference is a good fit for this question, as we want to estimate the causal effect of baseline emission and rate of reduction. Findings from this question will allow for adaptations of environmental policies to take into account how each state's initial level of emissions impacts their ability to efficiently decrease emissions over time, allocating resources and designing strategies to better fit each state. The two biggest limitations of this method, however, are the inability to capture non-linear relationships and a weak unconfoundedness assumption due to the large number of confounders unaccounted for.

Our second research question explores whether Bayesian generalized linear models (GLMs), which assume a linear relationship between predictors and the response, can effectively model carbon emissions, or if nonparametric methods like random forests are better suited to uncover nonlinear interactions in the data. Our intuition for Bayesian GLMs is that they offer interpretability and allow for prior knowledge integration, which can be helpful in policy-relevant domains like environmental forecasting. However, this comes with the risk of introducing prior-driven bias—for example, overly strong or misinformed priors can skew predictions or underrepresent uncertainty. Nonparametric methods like random forests, while often more accurate due to their flexibility, are harder to interpret and may overfit, especially with limited data or when not properly tuned. Ethically, there is a risk that poorly calibrated models may misinform climate policy or disproportionately affect marginalized communities if used for regulatory decisions. Transparency in model assumptions (e.g., linearity vs. nonlinearity, prior strength) and fairness in feature selection (e.g., population vs. energy type) are essential to ensure responsible deployment. Additionally, because carbon emission data often reflect systemic inequalities in energy access and consumption, ethical modeling must consider not just predictive performance, but also the social context behind the data. We expect that while Bayesian GLMs may provide interpretable insights and uncertainty quantification, nonparametric models like random forests will likely outperform them in pure predictive performance by capturing more complex, real-world emission dynamics. The tradeoff will be between interpretability and flexibility.

Research Question 1

Does a higher baseline level of CO_2 emissions causally influence the rate of reduction in emissions over time across U.S. states?

Methods

We use a **causal inference** framework to estimate the effect of a state's baseline CO_2 emissions on its long-term emissions reduction rate. The treatment is defined as a binary indicator: 1 if a state's per-capita CO_2 emissions in the baseline period (1970–1975) are above the national median, and 0 otherwise. The outcome is the annualized reduction rate in CO_2 emissions from 1970 to 2022, capturing the pace of decarbonization over time. To satisfy the unconfoundedness assumption, we condition on several observed confounders that may influence both the treatment and the outcome. These include Energy Use Sector Share (e.g., industrial vs. residential consumption), Geographical Region (U.S. census regions), Energy Source Emissions (e.g., from coal or petroleum), and the Electricity Trade Index. These variables account for structural and geographic differences across states that affect both baseline emissions and reduction trends. We identify potential colliders—such as GDP and average AQI—that are influenced by both treatment and outcome and are therefore excluded from adjustment. Assuming these confounders are sufficient, we treat baseline CO_2 emissions as conditionally independent of potential outcomes, enabling us to estimate causal effects. This is done by computing the conditional effect for each confounder and averaging across the population. Our assumptions and adjustment strategy are represented in the accompanying DAG.

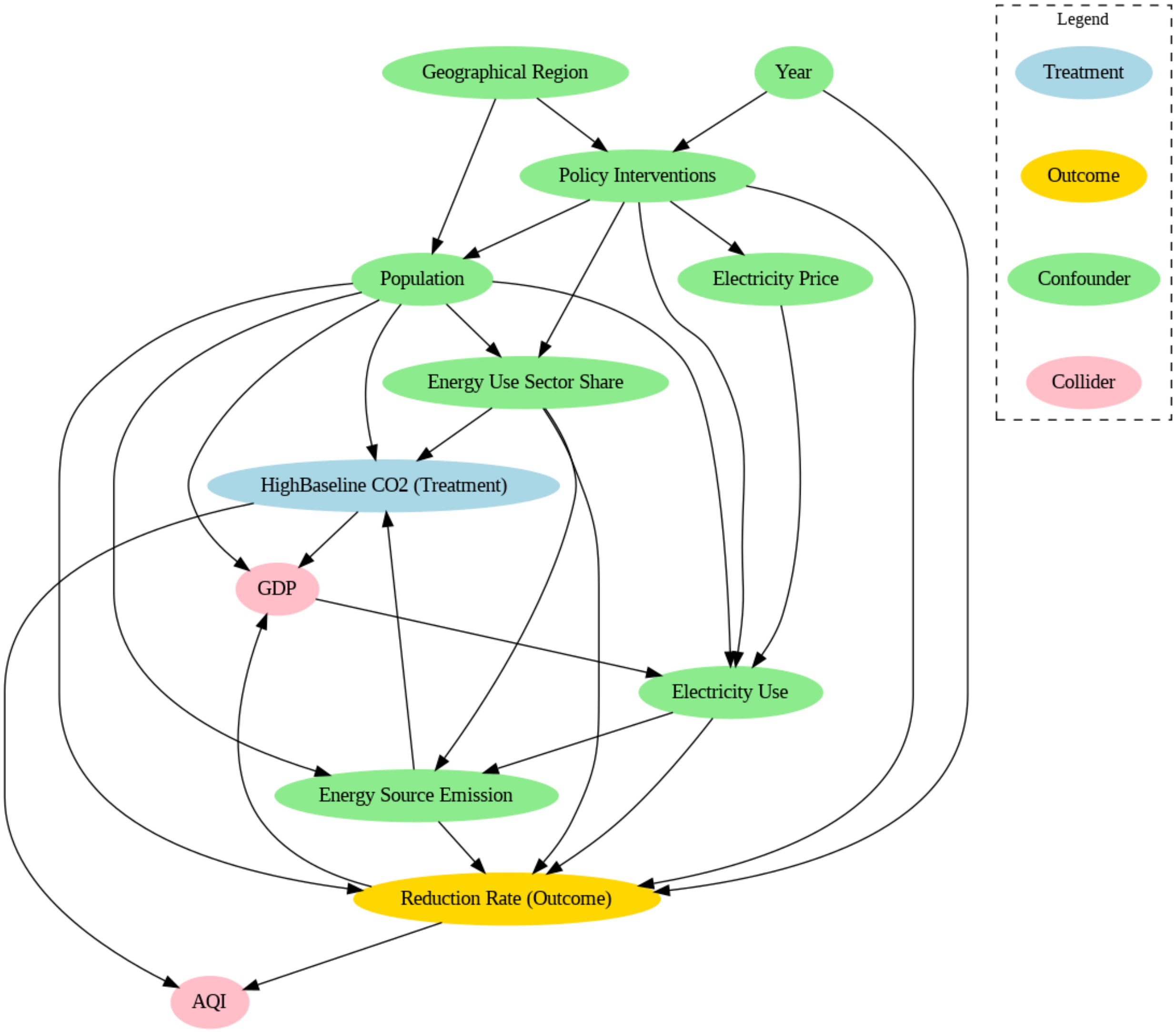


Figure 11: DAG

Results

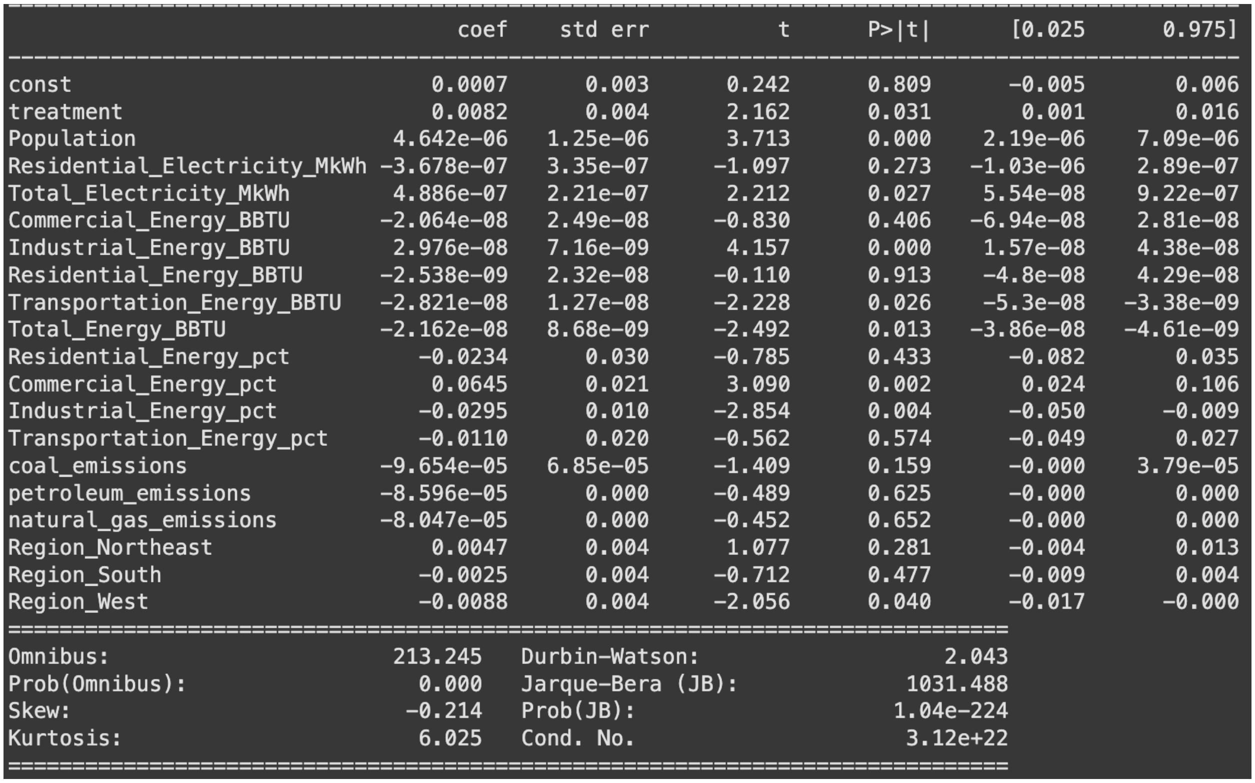
To analyze the association between early baseline CO_2 emissions and subsequent reduction rates, we estimated an Ordinary Least Squares (OLS) regression model conditioned on the confounders mentioned above. The Conditional Average Treatment Effect (CATE) lies in the coefficient of treatment. As seen in figure 13, our CATE is 0.0082. This result is statistically significant at the 5% level as the p-value is less than 0.05 and the 95% confidence interval does not include 0.

We evaluated model performance using Mean Squared Error (MSE) and R^2. Our MSE of 0.039 suggests a moderate prediction error. However, we have very low R^2 and adjusted R^2 values (0.039 and 0.033 respectively). This means that only 3.9% of the variance in our outcome is explained by the model and adding more predictors is only introducing noise to the model. This shows that much of the variation in emission reduction remains unexplained by the included predictors.

We can say that the estimated treatment effect due to the causal relationship between higher baseline CO_2 emissions and future reductions, is conditional on energy usage, emissions, and regional indicators, and the electricity trade index.

Model	MSE	R^2	Adj. R^2
OLS	0.039	0.039	0.033

Figure 12: Model Performance



```
----------------------------------------------------------------------------------------------
                                   coef    std err          t      P>|t|     [0.025     0.975]
----------------------------------------------------------------------------------------------
const                            0.0007      0.003      0.242      0.809     -0.005      0.006
treatment                        0.0082      0.004      2.162      0.031      0.001      0.016
Population                    4.642e-06   1.25e-06      3.713      0.000   2.19e-06   7.09e-06
Residential_Electricity_MkWh -3.678e-07   3.35e-07     -1.097      0.273  -1.03e-06   2.89e-07
Total_Electricity_MkWh        4.886e-07   2.21e-07      2.212      0.027   5.54e-08   9.22e-07
Commercial_Energy_BBTU       -2.064e-08   2.49e-08     -0.830      0.406  -6.94e-08   2.81e-08
Industrial_Energy_BBTU        2.976e-08   7.16e-09      4.157      0.000   1.57e-08   4.38e-08
Residential_Energy_BBTU      -2.538e-09   2.32e-08     -0.110      0.913    -4.8e-08   4.29e-08
Transportation_Energy_BBTU   -2.821e-08   1.27e-08     -2.228      0.026    -5.3e-08  -3.38e-09
Total_Energy_BBTU            -2.162e-08   8.68e-09     -2.492      0.013  -3.86e-08  -4.61e-09
Residential_Energy_pct          -0.0234      0.030     -0.785      0.433     -0.082      0.035
Commercial_Energy_pct            0.0645      0.021      3.090      0.002      0.024      0.106
Industrial_Energy_pct           -0.0295      0.010     -2.854      0.004     -0.050     -0.009
Transportation_Energy_pct       -0.0110      0.020     -0.562      0.574     -0.049      0.027
coal_emissions               -9.654e-05   6.85e-05     -1.409      0.159     -0.000   3.79e-05
petroleum_emissions          -8.596e-05      0.000     -0.489      0.625     -0.000      0.000
natural_gas_emissions        -8.047e-05      0.000     -0.452      0.652     -0.000      0.000
Region_Northeast                 0.0047      0.004      1.077      0.281     -0.004      0.013
Region_South                    -0.0025      0.004     -0.712      0.477     -0.009      0.004
Region_West                     -0.0088      0.004     -2.056      0.040     -0.017     -0.000
=============================================================================
Omnibus:                     213.245   Durbin-Watson:                   2.043
Prob(Omnibus):                 0.000   Jarque-Bera (JB):             1031.488
Skew:                         -0.214   Prob(JB):                    1.04e-224
Kurtosis:                      6.025   Cond. No.                     3.12e+22
=============================================================================
```

Figure 13: Regression Results

Discussion

Our initial hypothesis was that states with a higher baseline level of CO_2 emissions will see higher reduction in rate of emissions over time in U.S. states. Intuitively, states with higher baseline emissions will likely be the first targets of stricter emission policies, forcing them to improve inefficient systems and shifting away from high-emitting industries (e.g. shutting down old coal plants) that result in higher reduction due to the low-hanging fruit effect. Numerous papers have supported this, including one on political influences on greenhouse gas emissions from US states. In particular, the authors find that “demographic and economic forces can in part be offset by politics supportive of the environment—increases in emissions over time are lower in states that elect legislators with strong environmental records” (Dietz et. al, 2015).

Our regression results show that our hypothesis holds. States with a higher baseline level of CO_2 emissions reduced their emissions at a rate 0.0082 times higher on average per year, conditioned on all potential confounders and exclusion of colliders. Based on our findings, there is a small positive effect of the baseline level of emissions on the rate of reduction.

However, we are not confident that there is a causal relationship between baseline emissions and the rate of reduction due to numerous limitations of our method. Firstly, our unconfoundedness assumption may not hold. For our OLS estimates to yield causal estimates, all potential confounders must be accounted for, which is not the case. Some of the confounding variables not included are states’ responses to emission policies, economic shocks, and states’ political ideologies. We excluded them due to a lack of public data and an inability to control for model complexity. Secondly, binarizing treatment using the median level of CO_2 emission as a threshold results in a significant loss of information, particularly the magnitude of the differences in emissions. There is no distinction between states whose baseline emissions are close to and far from the median if they are all above or all below the threshold. Thirdly, our choice of OLS regression assumes a linear relationship among the variables, which may not fully capture nonlinear dynamics. Lastly, we work with temporal data, which means a time series regression may be a better fit.

Additional data would be immensely helpful to allow us to truly establish or reject a causal relationship between baseline emission level and reduction rate. Data on the potential confounders not included in our model, as previously mentioned, will strengthen our unconfoundedness assumption, further isolating our CATE to treatment alone. Moreover, specific data on each state’s current emission policy, such as emission caps, renewable energy adoption, economic structures, and federal subsidies for emission reduction, will aid us in separating baseline emission levels and ongoing mitigation measures. Data on potential colliders to exclude from our model would also be of great importance in ensuring fit and performance.

Our findings are slightly different from the work we cited, largely because we did not include political ideology as part of our confounders. We assumed that states with higher baseline emission levels would be targeted by progressive (left-leaning) emission policies, but did not explore how states would respond to such constraints. A key limitation of our study is the absence of a measurable variable for state-level policy interventions or political ideology, which likely confounds the relationship between baseline CO_2 emissions and reduction rate. Since we could not directly control for this factor, our causal estimates may be biased. Although we initially considered using year as a proxy, it is part of the outcome definition and thus unsuitable for capturing political or regulatory effects. Moreover, our focus is on causality, whereas the study by Dietz and coauthors seeks correlations between politics and emissions. This means that they may overstate the effect found due to not controlling for confounding variables.

Research Question 2

Can Bayesian generalized linear models accurately predict carbon emissions from energy usage and related variables, or do nonparametric methods such as random forests uncover nonlinear patterns that improve predictive performance?

We aim to predict carbon emissions from energy usage per capita using a range of energy-related and demographic features. Our selected features include population size, energy usage by sector, and electricity usage, all of which are commonly cited determinants of carbon output due to their direct relationship with energy demand and consumption. We incorporate other greenhouse gas emissions data from 1990 to 2022, using publicly available records from the EPA to account for correlated pollutants that may signal broader emissions trends. To control for time-specific effects and technological or policy changes, we also include the year as a feature. Geographical region data can also further refine model accuracy by capturing regional disparities in energy infrastructure, climate policy, and industrial activity. These features were chosen for their relevance, availability in public datasets, and proven significance in environmental modeling literature.

Methods

GLM

We used a Bayesian generalized linear model (GLM) with a Gamma likelihood and log link function to predict carbon emissions based on energy-related features. We used Bayesian GLM because of the non-negative and right-skewed nature of the emissions data (Isidro et al., 2014), which suits the Gamma distribution. We employed normally distributed priors for both the intercept and coefficients (centered at 0 with a moderate variance), allowing for shrinkage without over-constraining the model. The predictors include population size, various types of energy usage, total energy consumption, greenhouse gas emissions, and year, all of which are known drivers of emissions behavior. The data was standardized before inference to ensure coefficients are comparable. Posterior inference was implemented using PyMC with MCMC sampling, and posterior distributions of coefficients were visualized to calculate the influence of each feature and potential model uncertainty. This Bayesian GLM framework allows for probabilistic interpretation of results while still maintaining interpretability through linear coefficients.

We implemented Ridge Regression as our frequentist method due to its effectiveness in handling multicollinearity among predictors, especially given the overlap between different types of energy usage. Ridge adds an L2 penalty to the loss function, shrinking coefficients toward zero and reducing the risk of overfitting. While this approach does not involve probabilistic priors like in Bayesian methods, it acts similarly by discouraging large weights without strong evidence. However, because Ridge uniformly shrinks all coefficients, it may underestimate the effect of truly influential predictors in cases where the data distribution is skewed or imbalanced.

Random Forest

We used Random Forest Regression because it captures nonlinear relationships between variables like energy usage and emissions without requiring strong parametric assumptions, unlike Bayesian GLMs. It is robust to outliers and collinearity, automatically models feature interactions, and provides interpretability through feature importance and partial dependence plots. Compared to other methods like neural networks or SVR, Random Forest is easier to train, more stable, and better suited for medium-sized environmental datasets.

We assume that the relationship between predictors, such as energy use, electricity price, and population, and CO_2 emissions is nonlinear. We also assume that these patterns remain relatively stable across time and between states. Each (state, year) observation is treated as independent, with no significant temporal or spatial autocorrelation (e.g., emissions in one year affecting the next, or neighboring states influencing each other). Finally, we assume that all input variables are accurately measured and reflect the true underlying drivers of emissions.

Model Evaluation

To evaluate the Bayesian GLM's performance in predicting carbon emissions, we used posterior analysis, goodness-of-fit metrics, and cross-validation techniques. We examined the distributions of the intercept, sigma, and beta coefficients to assess the magnitude and uncertainty of each predictor's influence. We calculated the model's R^2 to calculate how much of the variance is explained, and RMSE to measure average prediction error. To identify potential overfitting, we applied cross-validation. These metrics help compare model fit while accounting for model complexity.

We will evaluate the performance of the Random Forest model using Root Mean Squared Error (RMSE) and R^2 (coefficient of determination) on a held-out test set. RMSE measures the average prediction error in the same units as the target (CO_2 emissions), while R^2 indicates how much variance in emissions the model explains. A lower RMSE and higher R^2 suggest better predictive accuracy. We also use cross-validation to assess the model's stability across different data splits and reduce the risk of overfitting.

Results

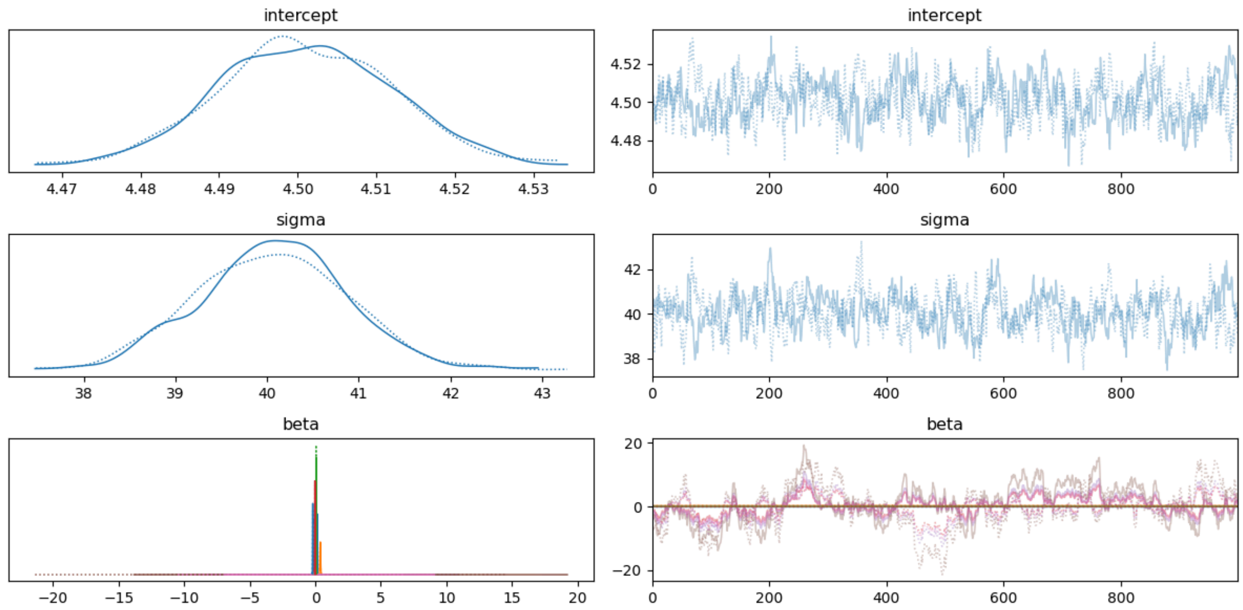


Figure 14: Bayesian GLM Posterior Distributions and Trace Plots for Model Parameters

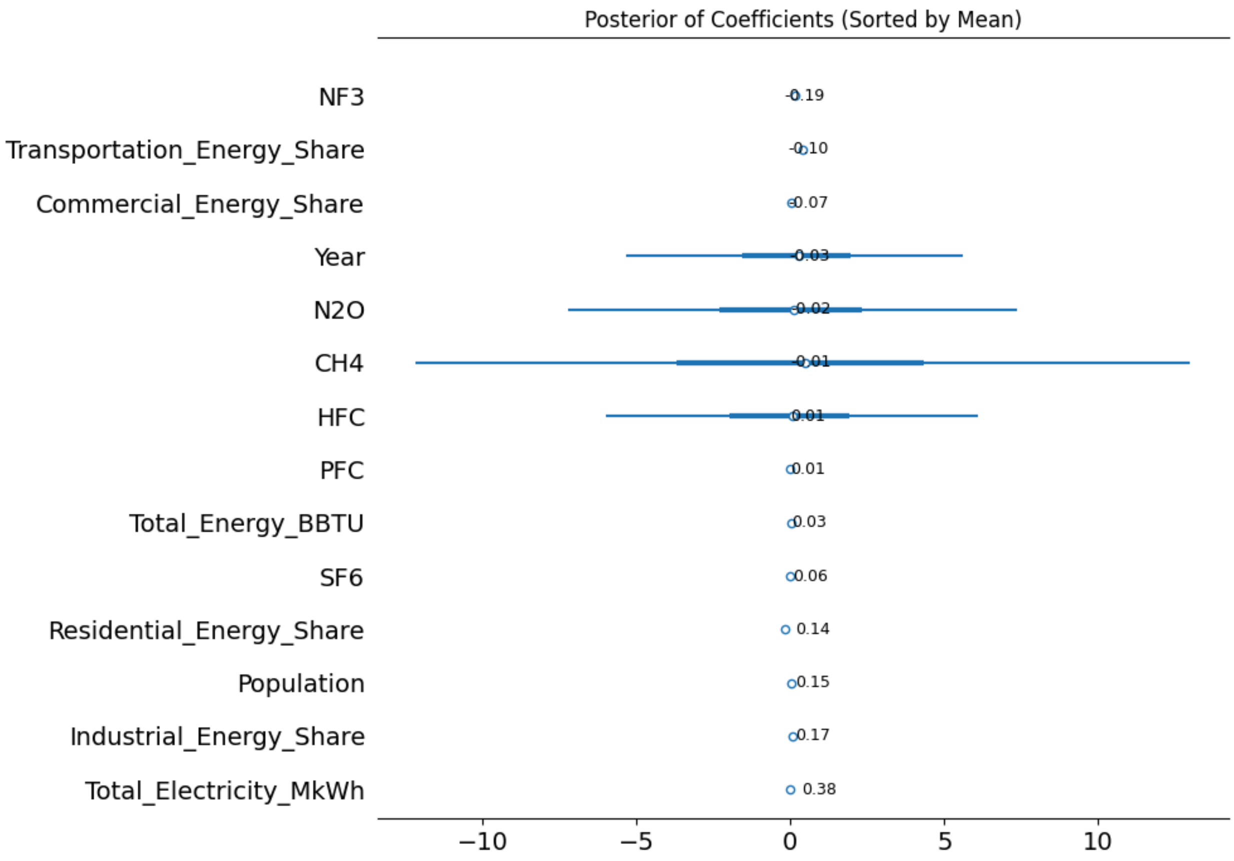


Figure 15: GLM Posterior of Coefficients

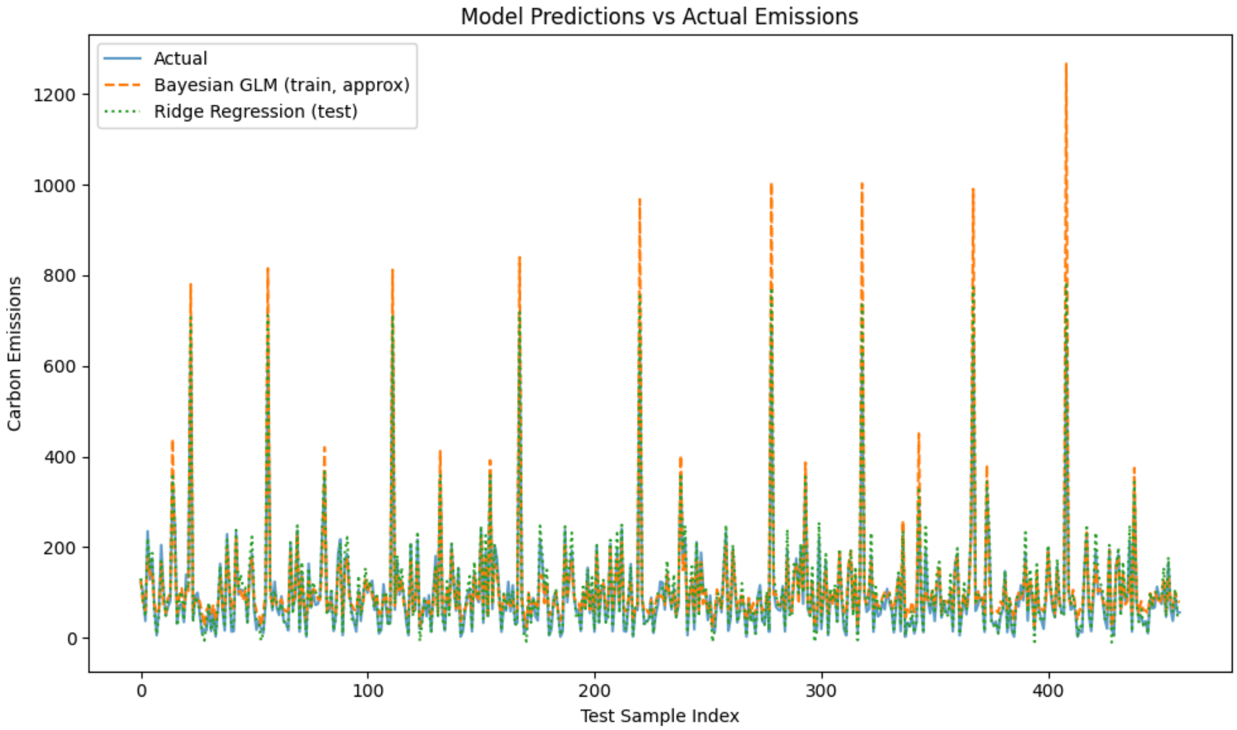


Figure 16: Ridge Regression and Bayesian GLM Predictions vs Actual Emissions

Based on our analysis for Research Question 2, our results suggest that nonparametric methods such as random forests are better suited for capturing the complex, nonlinear relationships inherent in carbon emissions data.

Our Bayesian Generalized Linear Model (GLM) with a Gamma likelihood and log link function was used to model CO_2 emissions using seven standardized predictors: Population, Electricity Usage, Total Energy Usage, Sector-Based Energy Usage, Greenhouse Gas Emissions, and Year. After sampling 2,000 posterior draws with a tuning phase of 1,000 steps and a target acceptance rate of 0.9, we found that the model converged well, with no divergences and smooth trace plots across all parameters.

The posterior distributions for the beta coefficients indicated that Total Electricity Consumption (MkWh) had the strongest estimated positive effect on CO_2 emissions, with a posterior mean of approximately 0.38, followed by Industrial Energy Share (0.17) and Population (0.15). Other variables with positive effects included Residential Energy Share (0.14) and SF6 emissions (0.06). In contrast, features such as NF3 (-0.19), Transportation Energy Share (-0.10), and Commercial Energy Share (-0.07) exhibited negative posterior means, suggesting an inverse association with emissions. Although the posterior means reflect estimated directional effects, most 95% HDIs overlapped with zero, indicating uncertainty about the precise magnitude and sign of several predictors.

On the other hand, our implementation of Ridge regression achieved near-perfect R² scores (0.96), but a high RMSE score of 422.24. Both models—Bayesian GLM (orange dashed line) and Ridge Regression (green dotted line)—generally follow the same pattern as the actual values (solid blue line), indicating that both models are capturing the overall trend. Bayesian GLM tends to exaggerate peak emissions, frequently overshooting actual values at sharp emission spikes. This may be due to overfitting on high-emission outliers or poor generalization to extreme values. Ridge Regression predictions are closer to the actual values for extreme cases. It seems to better manage outliers, likely due to its regularization, but may slightly underpredict peaks. This underscores the need for using visual checks along with summary statistics like R². The random forest model (shown below) outperformed in predictive flexibility, revealing nonlinear interactions that neither linear model could visualize.

While Bayesian GLMs offer transparency and probabilistic interpretation, their linear assumptions limit performance on variable environmental data. Nonparametric models like random forests, although less interpretable, seem to be more capable of uncovering complex relationships that influence emissions, providing better use for prediction-focused environmental modeling.

Results are attached below for our Random Forest model.

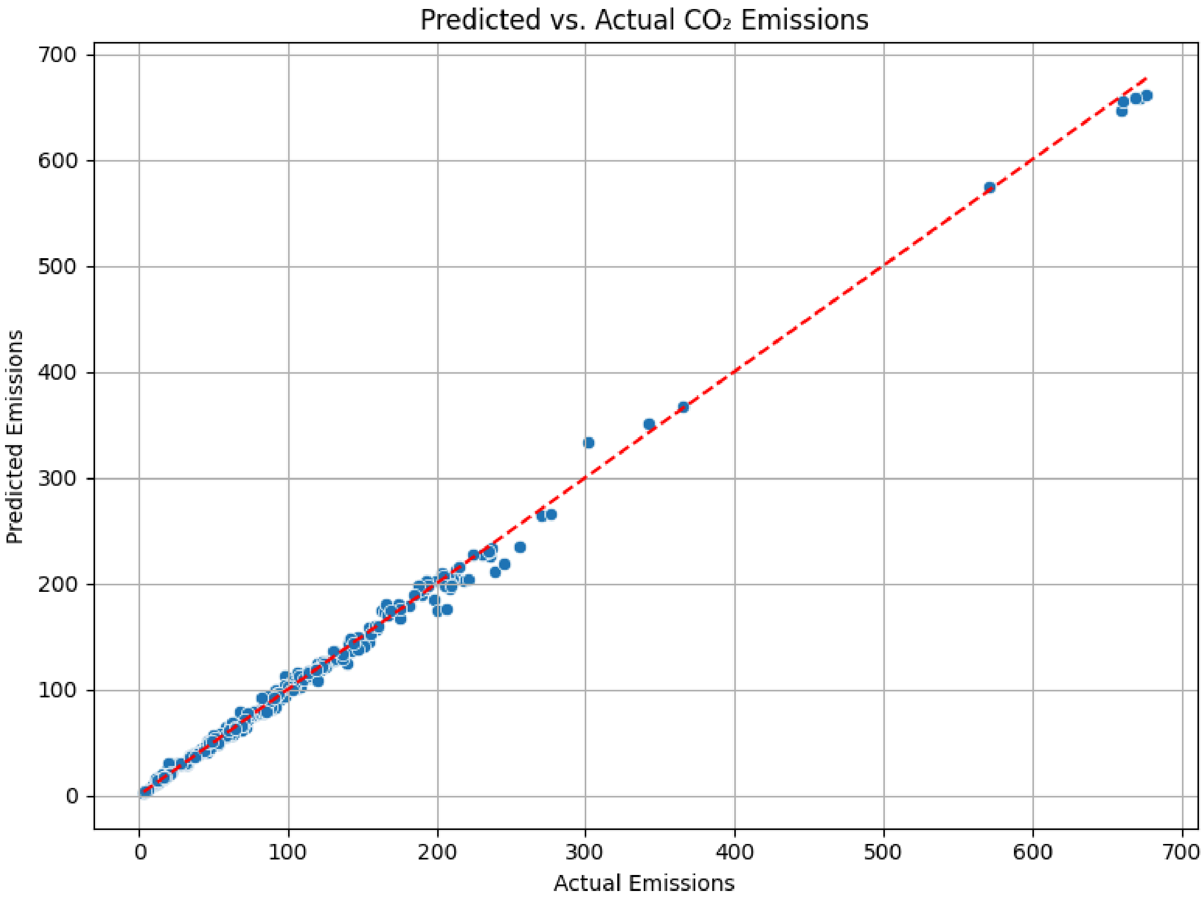


Figure 17: Random Forest Prediction Result vs. Actual Values

Figure 17 shows that the Random Forest model performs very well, as most points lie close to the red 45-degree line, indicating high prediction accuracy. This strong alignment supports the model's high R^2 value. Figure 18 below reveals that most residuals are centered around zero, showing low bias overall. However, we observe some dispersion and mild heteroscedasticity for higher emission values, suggesting that prediction errors slightly increase with emission magnitude.

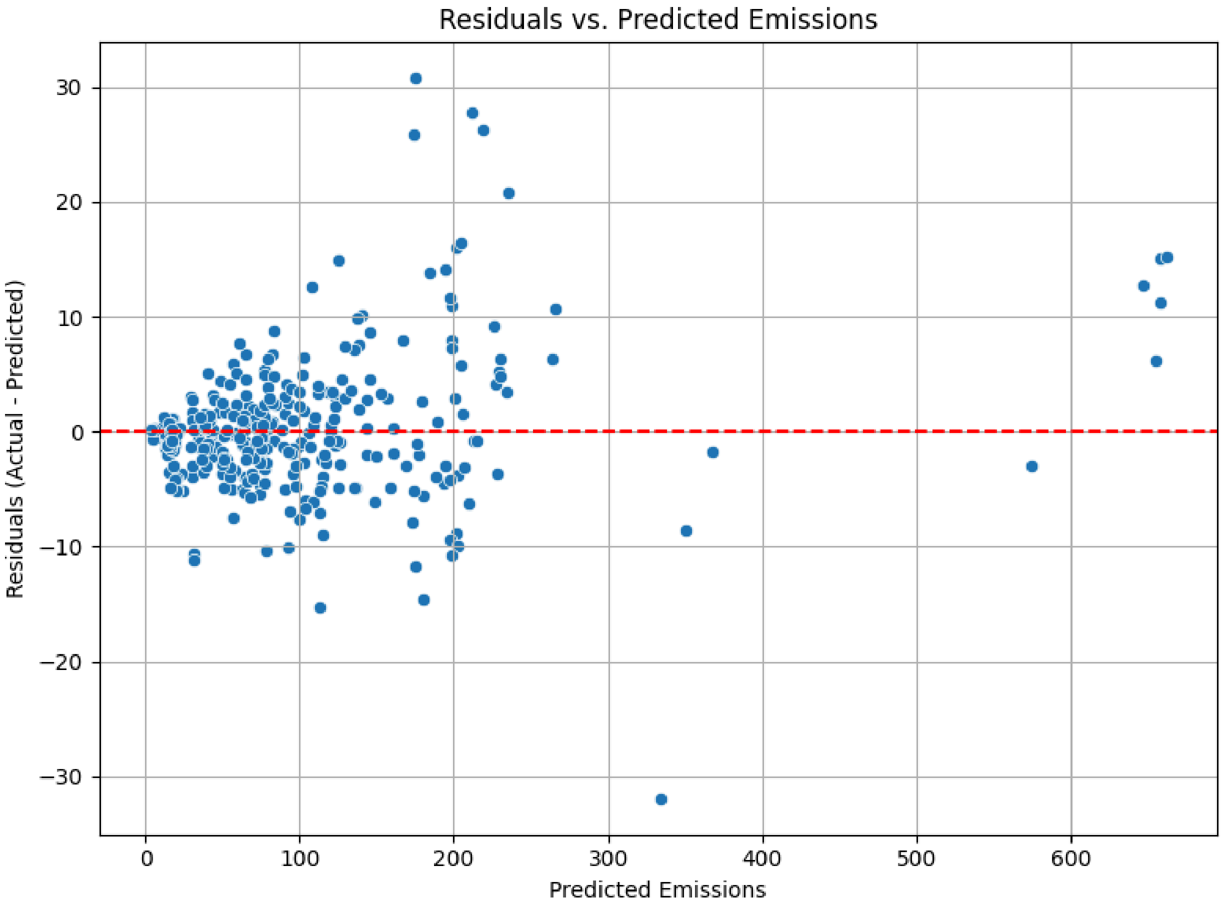


Figure 18: Random Forest Residual Plot

Discussion

Bayesian GLM Interpretation:

The findings from our Bayesian GLM highlight that Total Electricity Consumption and Industrial Energy Share are the most influential predictors of CO_2 emissions, with posterior means of approximately 0.38 and 0.17, respectively. These results are intuitive, as higher electricity consumption and industrial energy use typically reflect greater overall energy demand. Population and Residential Energy Share also showed notable positive associations (0.15 and 0.14), suggesting that both demographic and household-level energy usage play significant roles in driving emissions. In contrast, variables such as NF3 (-0.19) and Transportation Energy Share (-0.10) had negative posterior means, which may reflect differences in emission intensity across sectors or fuel types. The small negative coefficient for Year (-0.03) could indicate slight improvements in emission efficiency over time.

Random Forest Interpretation:

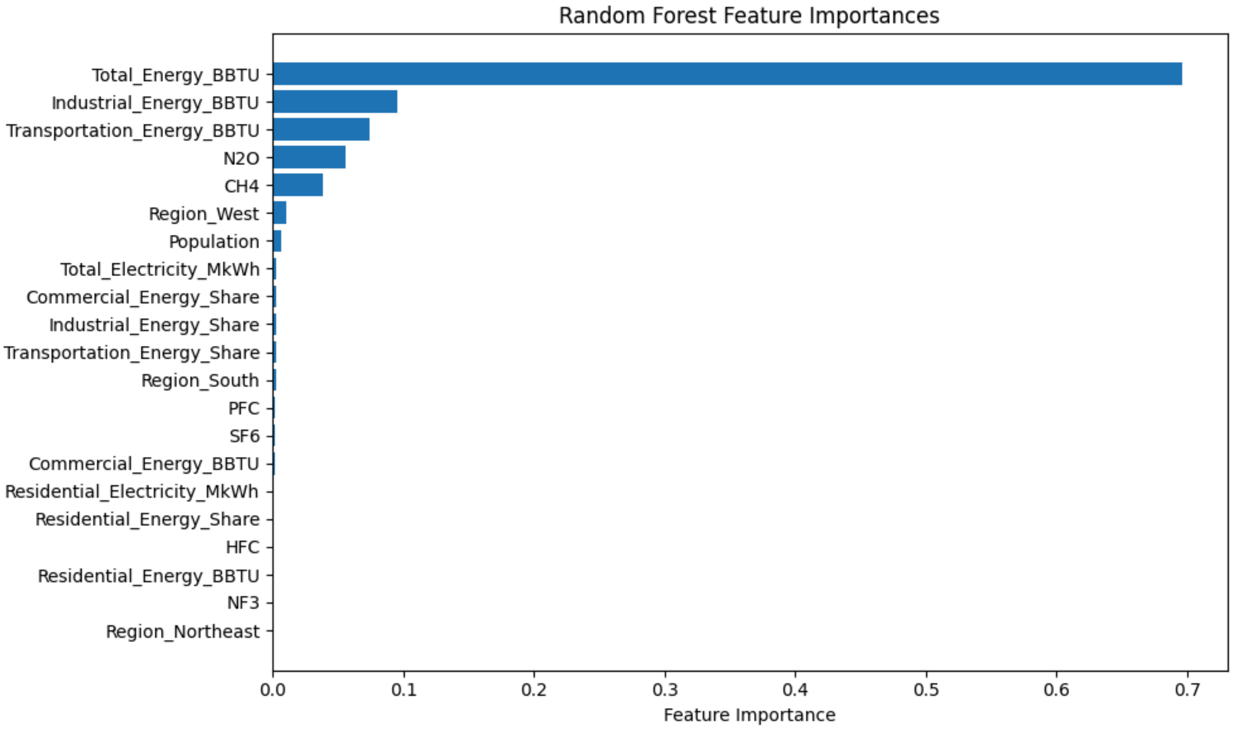


Figure 19: Random Forest Feature Importances

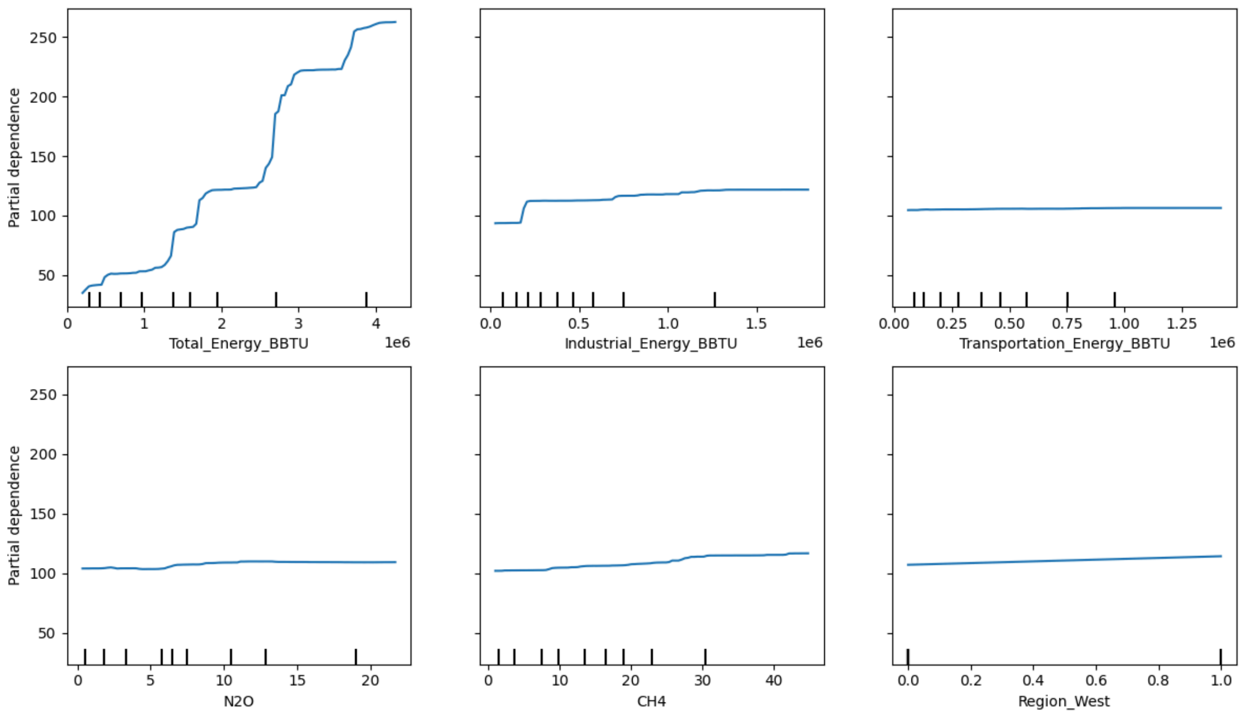


Figure 20: Partial Dependence Plot (Top 6 Features)

Parameter	Value	Interpretation
max_depth	20	Limits how deep each tree can grow. A depth of 20 allows complex splits, capturing detailed patterns, but not too deep to overfit badly.
max_features	sqrt	At each split, the model considers only the square root of the total number of features. This introduces randomness, helping reduce correlation between trees and improving generalization.
min_samples_leaf	1	Each leaf (end node) must have at least 1 sample. This allows trees to fully split, possibly capturing fine-grained details, but may lead to overfitting.
min_samples_split	2	A node must have at least 2 samples to be split. This is the lowest allowed value, allowing maximum branching.

n_estimators	200	The forest has 200 trees, improving prediction stability through averaging, though with a greater computational cost.

Figure 21: Random Forest Parameter Interpretation

Figure 21 summarizes each parameter’s function and trade-offs. However, Random Forest parameters don’t directly map to interpretable real-world relationships the way coefficients in linear models do. We use Figures 19 and 20 to interpret results indirectly. From Figure 19, the model's top features—Total_Energy_BBTU, Industrial_Energy_BBTU, and Transportation_Energy_BBTU—have the highest importance scores, indicating that total and sector-specific energy consumption are strong drivers of CO_2 emissions. This aligns with domain knowledge that fossil fuel energy use contributes heavily to carbon output. Partial dependence plots (Figure 20) further confirm a positive nonlinear relationship, especially for Total_Energy_BBTU, showing that as energy consumption increases, CO_2 emissions rise rapidly. Conversely, variables like residential energy, sector shares, and regional dummies had low importance.

Bayesian GLM Evaluation and Limitation:

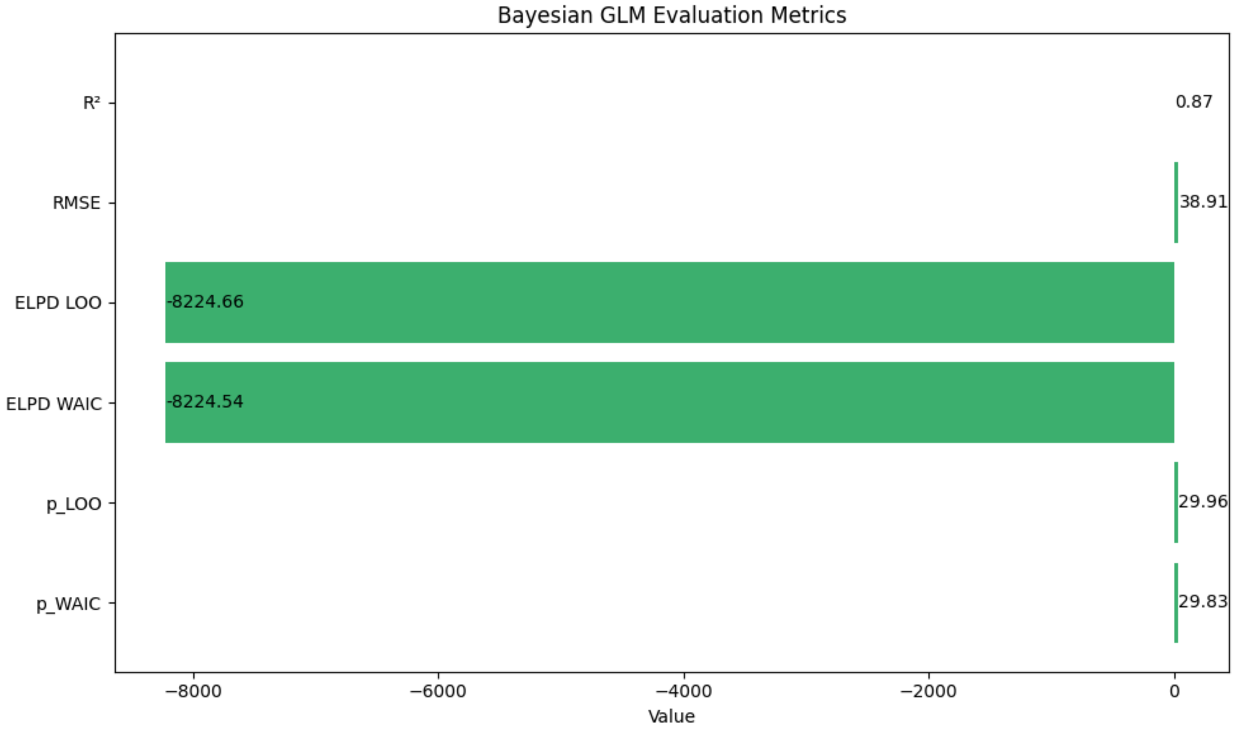


Figure 22: Bayesian GLM Model Evaluation

Since our Bayesian GLM was designed for prediction, we evaluated its performance using several appropriate Bayesian model diagnostics and performance metrics. Because traditional metrics like AIC and BIC are not typically applied in Bayesian settings, we used WAIC (Widely Applicable Information Criterion) and LOO (Leave-One-Out cross-validation), which are more appropriate for probabilistic models. Both LOO (-8224.66) and WAIC (-8224.54) values suggest strong model fit (meaning less expected predictive error), with all Pareto k diagnostic values falling in the "good" range. This indicates the model generalizes well to new data and is not overly sensitive to individual observations.

Additionally, the R^2 value of 0.87 demonstrates that the model explains approximately 87% of the variance in CO_2 emissions—a very strong predictive result. The RMSE of 38.91 also suggests relatively low average error in predictions across the dataset. These metrics confirm that the Bayesian GLM is highly effective at capturing and predicting trends in emissions from the included features.

That said, predictive success in this context comes with caveats. For example, the year variable was included as a continuous numeric feature rather than one-hot encoded, which may attribute importance to "being later in time" rather than to actual structural changes in emissions behavior. Some features (e.g., sector energy shares and total energy usage) may be highly collinear. This can make it hard for the model to assign distinct contributions to each, even if the overall prediction is accurate. This can flatten or broaden posteriors and reduce coefficient interpretability. Besides, although WAIC and LOO are powerful tools, they are internal cross-validation-like methods. Without a true holdout test set, we can't directly evaluate generalization on unseen data.

In future iterations, incorporating a test set or conducting more rigorous cross-validation would provide additional assurance about out-of-sample performance. Moreover, adding features beyond energy usage, such as industrial output, renewable energy penetration, or emissions policies, could make the model more robust and interpretable.

In summary, the Bayesian GLM successfully achieves its intended goal of accurate prediction, and the use of appropriate evaluation tools supports its strong performance. However, its structure and the nature of the data limit the reliability of coefficient interpretation.

Random Forest Evaluation and Limitation:

We chose R^2 and RMSE as goodness-of-fit measures. R^2 quantifies the proportion of variance in the dependent variable explained by the model. To look into the overfitting issue, we also introduced a 5-fold cross-validation method. RMSE measures the average magnitude of prediction errors, in the same units as the response variable. Lower RMSE means better fit. RMSE is useful for absolute evaluation and can reflect overfitting when there's a large gap between training and cross-validation RMSE.

Method	Average R^2	Average RMSE
Random Forest	0.995	6.412

Random Forest (Cross Validation)	0.79	43.33
Random Forest after Grid Search	0.996	5.98
Random Forest after Grid Search (Cross Validation)	0.86	38.31

Figure 23: Random Forest Model Performance

The Random Forest model performed well on training data ($R^2 = 0.995$, RMSE = 6.41), but cross-validation revealed moderate overfitting (CV $R^2 = 0.79$, RMSE = 43.33). After tuning with Grid Search, model accuracy slightly improved ($R^2 = 0.996$, RMSE = 5.98) with better generalization (CV $R^2 = 0.86$, RMSE = 38.31). Despite the high R^2 level, the RMSE is still large and nontrivial in our research case. The RMSE indicates the average prediction error of the model in the same unit as the emissions (e.g., million metric tons of CO_2). Cross-validation revealed a much higher RMSE, suggesting overfitting and reduced generalization. This means that in real-world applications, the model's predictions could deviate by up to ±40 $MtCO_2$, which is significant.

As mentioned earlier, while feature importance gives a ranking of variable relevance, it doesn't explain the direction or size of influence. This makes it difficult to draw clear causal or policy-relevant conclusions. Random Forests are excellent at interpolation—predicting within the range of observed data—but struggle with extrapolation. For example, if a new state-year pair has an energy usage higher than any seen during training, the model's predictions may be unreliable. From feature importance analysis, we notice that Random Forests tend to favor continuous variables and features with many unique values. For example, Total_Energy_BBTU dominated the importance chart, partly because of its scale and range, potentially overshadowing meaningful categorical features like region or sector shares. Lastly, despite grid search improving performance, with 200 trees and deep branching, the model becomes computationally intensive, and this limits scalability to larger or real-time applications.

Comparison with Prior Work:

The study by Costantini et. al (2024) is methodologically similar to ours, as both use Random Forests to predict CO_2 emissions from energy-related variables. Their study had a higher R^2 of around 0.95. Both studies identify energy consumption as the most predictive factor, but our partial dependence plots reveal stronger nonlinear effects. Unlike their model, some expected features in our study, like electricity price, were less important. This is likely because Costantini et. al's study operates on national-level global data and our different choices of features. Another key takeaway is that they took income level into account and detected a non-constant dynamic of CO_2 emissions worldwide.

Conclusion

To conclude our analysis, we addressed two research questions: (1) whether a higher baseline level of CO_2 emissions causally influences the rate of reduction in emissions over time across U.S. states, and (2) whether Bayesian Generalized Linear Models (GLMs) or nonparametric methods like Random Forests more effectively predict carbon emissions from energy usage and related variables. For RQ1, we found a statistically significant positive relationship between high baseline CO_2 emissions and greater reduction rates, but acknowledge that limitations such as unobserved confounders (e.g., policy interventions, political ideology) and low model R^2 (0.039) caution against strong causal claims. For RQ2, our Bayesian GLM had a high R^2 (0.912) with a low RMSE (31.64), but its linear assumptions limited the flexibility. On the other hand, Random Forests highlighted nonlinear patterns and had higher predictive capabilities ($R^2 = 0.996$, RMSE = 5.98 on training), though cross-validation generalization concerns ($R^2 = 0.86$, RMSE = 38.31). Our Ridge regression achieved near-perfect R^2 scores (0.96), but a high RMSE score of 422.24, indicating poor performance during high-emission spikes despite accurately capturing overall trends. This indicates that while Ridge Regression effectively models baseline carbon emission patterns, it struggles with outlier values, making it less reliable for forecasting sharp emission peaks in the context of predictive climate policy.

One key limitation of our analysis is the lack of consistent state-level data on policy interventions, industrial output, and political ideology, which may have acted as confounders but were not accounted for. Domain knowledge from energy policy experts may have guided the inclusion of such variables and guided our understanding of emission drivers. Our conclusions are sensitive to modeling choices—while the Bayesian GLM provided interpretable coefficient estimates, the random forest model displayed nonlinear patterns the GLM missed, highlighting the impact of model assumptions on the results. Although we used a comprehensive U.S. state-year census from 1990–2022, our findings may not generalize beyond this scope into international contexts without additional data. Overall, the analysis is robust within its scope but is restricted to capturing complex, policy-driven dynamics of emissions.

Building on our findings, future studies can incorporate additional variables such as policy intervention timelines, state-level regulatory strength, and renewable energy adoption rates to bolster our understanding of emission drivers. Based on our results, our call to action is for policymakers to prioritize reductions in coal consumption and target energy-intensive sectors with customized interventions, especially in highly populated states. This recommendation is feasible through existing environmental regulatory agencies like the EPA and state energy commissions, although pushback from stakeholders in fossil fuel industries might cause delays. The policy impacts would benefit communities facing the brunt of climate-related harm, while energy providers or regions economically dependent on coal might face economic and social challenges.

References

Dietz, T., Kalof, L., & Stern, P. C. (2015). Political influences on greenhouse gas emissions from US states. *Proceedings of the National Academy of Sciences,* 112(27), 8254–8259. https://doi.org/10.1073/pnas.1417806112

Costantini, L., Laio, F., Mariani, M.S. et al. Forecasting national CO2 emissions worldwide. Sci Rep 14, 22438 (2024). https://doi.org/10.1038/s41598-024-73060-0

Isidro A. Pérez, M. Luisa Sánchez, M. Ángeles García, Marta Ozores, Nuria Pardo, Analysis of carbon dioxide concentration skewness at a rural site, Science of The Total Environment, Volumes 476–477, 2014, Pages 158-164, ISSN 0048-9697, https://doi.org/10.1016/j.scitotenv.2014.01.019.