Data 102 Final Project - Air Quality, Renewable Energy, and Emissions
Group 4: Alex Nii, Ethan Antony Serbanescu, Lucas Czajka, Lydia Jung

1. Data Overview

Table 1: Subregion Resource Mix

This dataset is from the U.S. Environmental Protection Agency (EPA)'s Emissions & Generation Resource Integrated Database (eGRID). Data on generation and capacity by eGRID subregions are part of this dataset, combining a few utility service areas. Every row is a subregion and holds the percentage of electricity generated by different fuel sources. The unit of energy is in terms of megawatts (MW) for nameplate capacity and is in megawatt-hours (MWh) for the net generation column.

This is not a census, but a comprehensive sample of subregions, representative of national generation patterns. The data can be overweighted with larger or more integrated regions. For instance, regions like ERCOT (Texas) manage capacity, possibly skewing results when used for smaller or rural regions.

The granularity is subregion level; lots of facilities are grouped together in each row. This may limit detailed information on heterogeneity within regions. Problems may include measurement error (e.g., mixing-share estimation), and convenience sampling if some of the small grids are excluded.

A significant limitation is the absence of temporal resolution. While simple missing values do not seem to be common in the dataset, there are occurrences of uncertainty indirectly. Columns such as "Other unknown/purchased fuel" are imprecise and could hide interpretations that would be more clear for the reader of the dataset. Overall, data is useful in looking at overall fuel mix patterns as well as between-region energy landscapes.

Table 2: Subregion Output Emissions

This table offers emissions (CO_2, CH_4, N_2O, etc.) which are in pounds per megawatt-hour (lb/MWh) by subregion and is categorized in total and non-baseload. This table is also eGRID-based and complements Table 1 through association of resource mix with environmental impact.

The data is structured as a sample at the level of subregions. The data allows for some strong inferences of comparison but conceivably not facility-level precision. This is because it has high granularity by type of energy. Similar to the first table, larger subregions can drown out trends. This reduces generalizability for area-specific research.

Measurement error could be a problem since the emissions are modeled rather than actually measured. A lack of facility-level emissions or time series limits more nuanced policy or trend analysis. There also exists grid gross loss data, which can prove useful in being aware of what could be system ineffectiveness or inefficiency.

There are no clear missing values, but emissions like CH_4 and N_2O are small and can fall below detection thresholds in some cases, reducing confidence in comparisons. This might be of special concern for models that assign unique weights to variables.

Table 3: State Resource Mix

This table mirrors Table 1 but at the U.S. state level. It includes capacity, generation, and percent contributions from 11 fuel types. Similar to Table 1, it also uses MW and MWh for nameplate capacity and net generation respectively.

This is a sample rather than a census, though it attempts to cover all U.S. states. Comparison with known national averages reveals that some states (e.g., California, Alaska) have unusually high renewable or hydro contributions, which may limit generalizability.

Each row reflects a state's energy mix, so granularity is state-level aggregation. This restricts conclusions about intrastate differences (e.g., northern vs. southern CA). Selection bias is less relevant, but measurement error may exist in how fuel types are categorized. This is especially true for categories like "other fossils" and "unknown". In addition, different states could be doing their reporting differently.

There are no clear missing values, however there are no time stamps and there are some 0.0 values but no NaN, this could cause some problems with our models.

Table 4: State Output Emissions

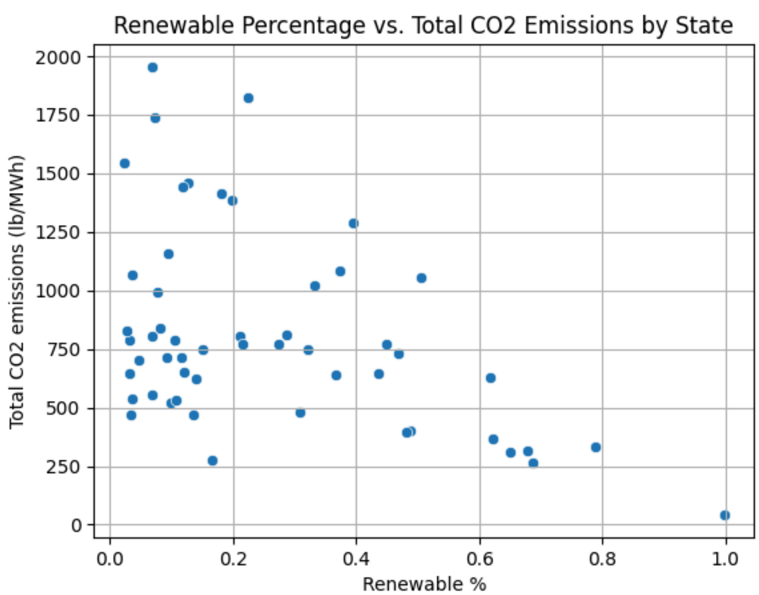
The next table shows average pollutant emissions in lb/MWh on a state level. Metrics are identical to Table 2 but are state-level averages rather than subregions.
This is a sample dataset based on available monitoring or modeled estimates. States with greater coal or gas show greater emissions (e.g., Colorado vs. California). This is a limitation, nevertheless, in that it smooths over intrastate variability and local emissions hotspots.

Granularity is state-level, represented by the manner in which each row reports aggregate emissions. This may obscure rural versus urban generation sources' differences. Measurement error can be thought of as negligible. For example, GHG figures like CH_4 and N_2O can be hard to accurately estimate. There is also a discrepancy in how states measure non-CO_2 gases. There are no missing values, but data may exclude states with weak reporting infrastructure. A limitation is that non-baseload breakdowns are not provided, which can be found in Table 2

2. Exploratory Data Analysis (EDA)

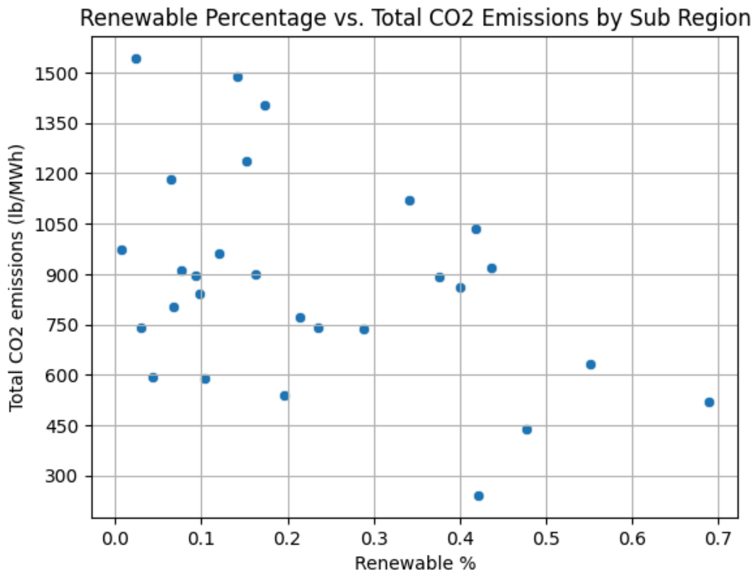
Visualization 1a:

Renewable Percentage vs. Total CO2 Emissions by State



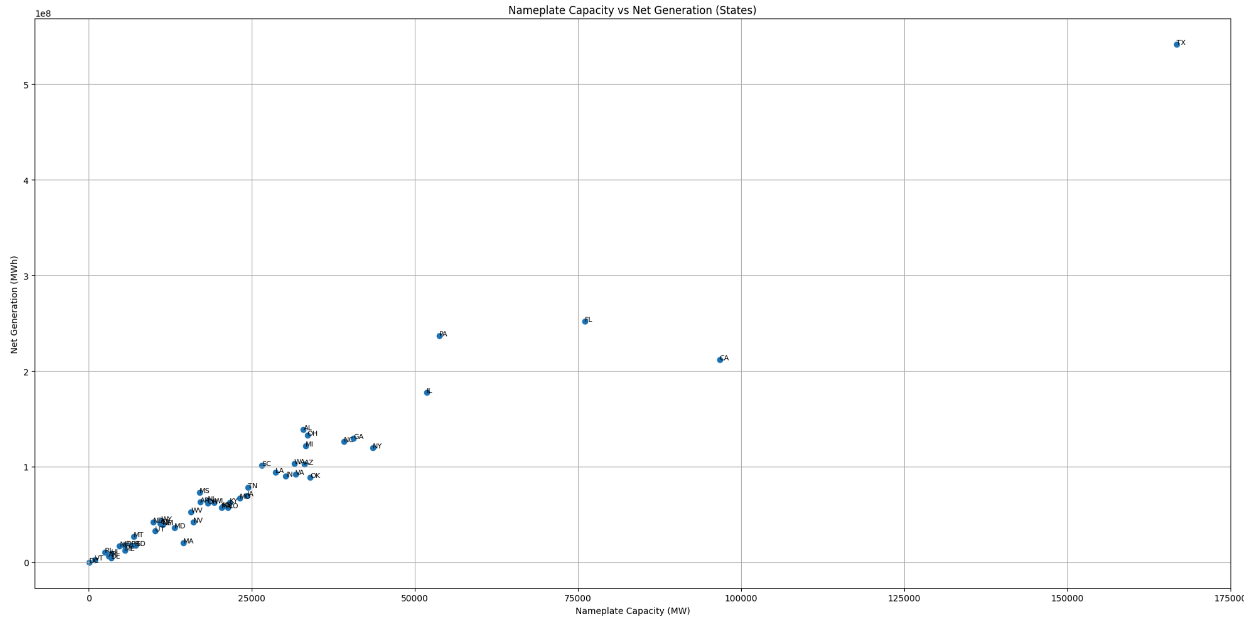
Visualization 1b:

Renewable Percentage vs. Total CO2 Emissions by Sub Region



Visualization 1a and 1b: For both of these scatterplots, we are observing generally a negative correlation between the percent of energy generated from renewable resources and total CO2 emissions (in pounds per megawatt-hour). The relationship is not perfectly linear, but we will begin to account for other factors such as state-size and population in later parts of this project. And regardless, there is a clear downward trend. This helps with our research question because it shows a visible association between renewables and reduced emissions for the causal inference question. This does not prove causality alone but it suggests that we may use renewable percent as a variable and total CO2 as a potential outcome in the causal inference framework.

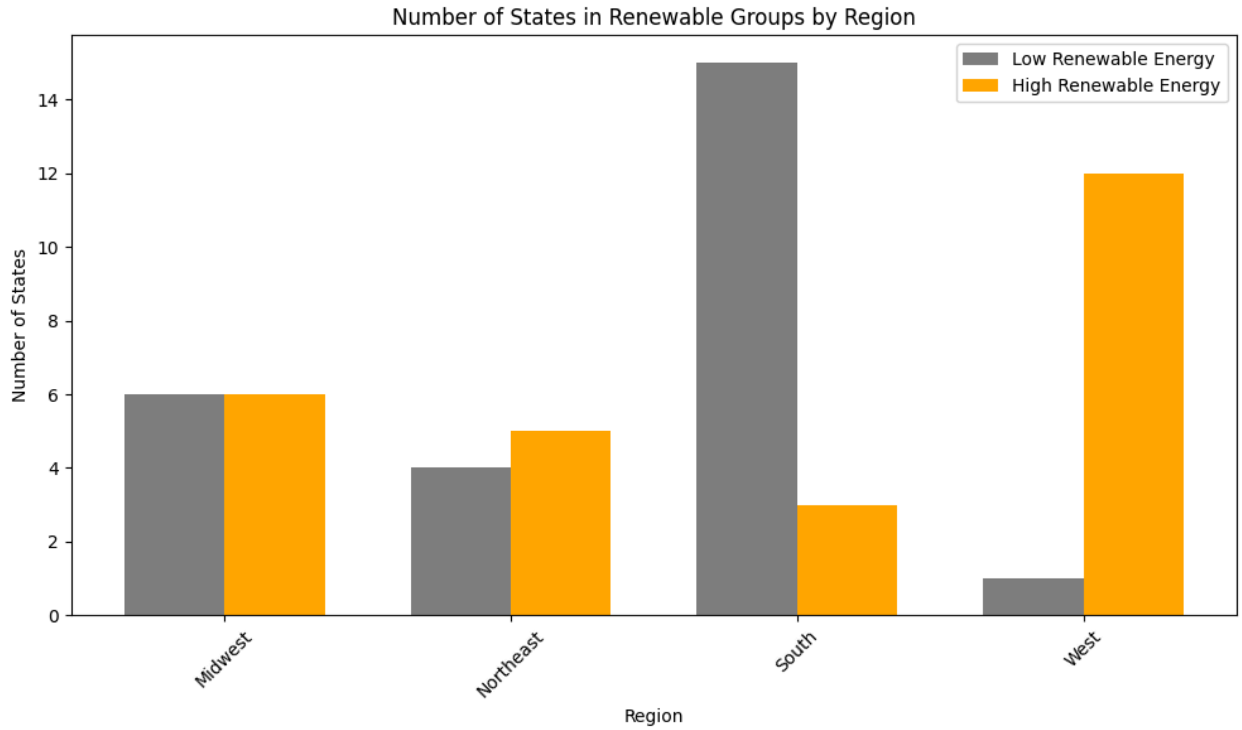
Visualization 2: Nameplate Capacity vs. Net Generation (States)



Visualization 2: This scatterplot shows that on our x axis, we have the maximum capacity of energy a state could produce, while our y axis is actually how much a state produces. The x-axis have the units of Megawatts (MW). While the y-axis of net generation is in megawatt-hours. This plot can show us which states produce more or less energy for the US which can then be correlated with states who also produce more carbon emissions.

Visualization 3:

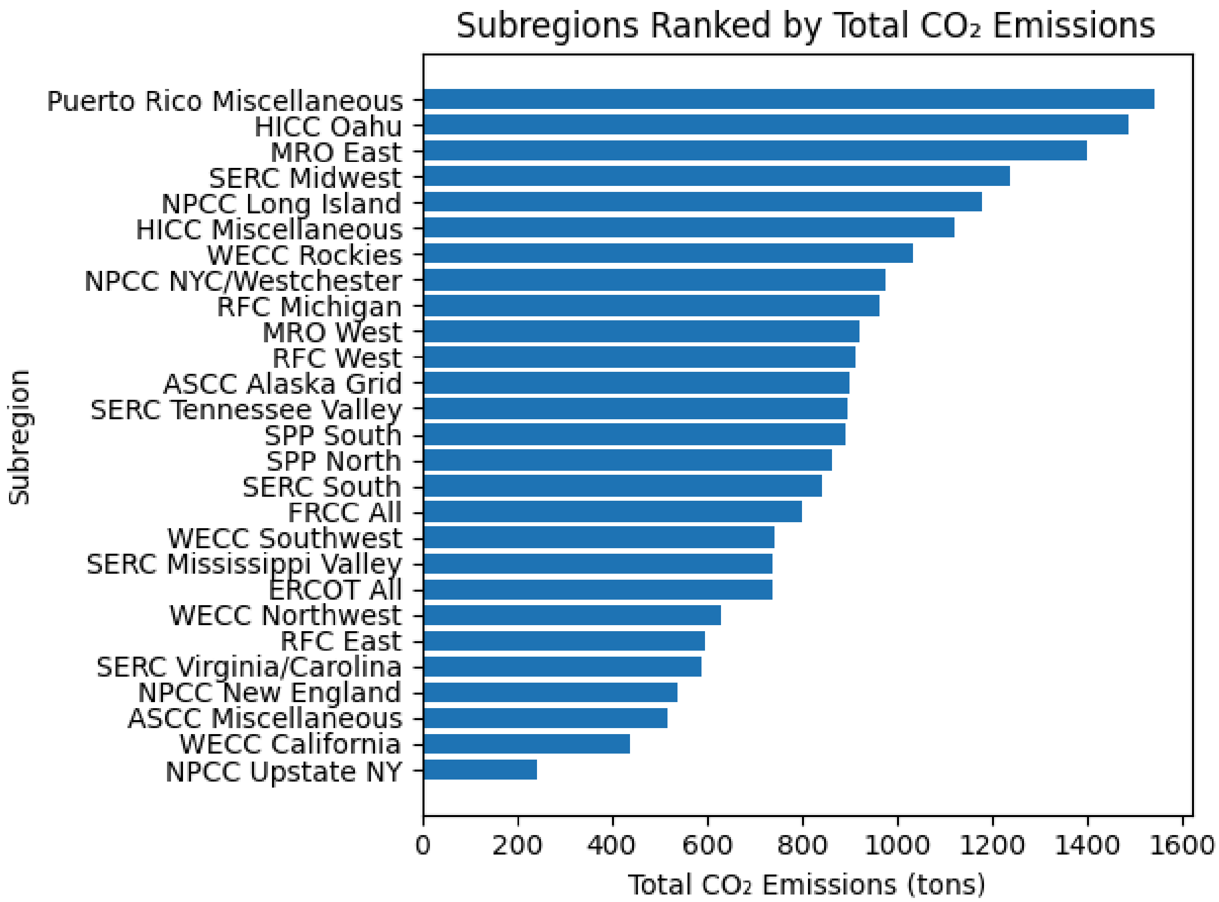
Number of States in Renewable Groups by Region



Visualization 3: The double bar graph shows that the West has a high number of states that have a high renewable energy percentage of their resource mix, while regions like the South have more states that have low renewable energy percentage. This regional contrast may reflect policy, geographic, or infrastructure differences in renewable adoption. These patterns are directly relevant to our research question because they suggest that emission rates may also vary by region through their renewable energy usage. If states with more renewable energy have lower emissions, this could point to a significant relationship worth testing statistically. This motivates our hypothesis and can guide which regions may benefit most from renewable energy incentives or emission-reduction efforts.

Visualization 4:

Sub Regions Ranked by Total CO2 Emissions



Visualization 4: As shown in this bar plot, we can see that we are mapping all subregions vs their total CO_2 Emissions per ton. What we can extract from this is which regions are high producers vs which regions are low producers. As we can see NYC and Michigan stand as the two highest producers of total CO_2 emissions while upstate NY and California are the lowest contributors of CO_2 emissions. This helps support our research question as it shows how different subregions even in the same state like NY can have wildly different CO_2 emissions based on just location in the state alone.

3. Research Questions

Research Question 1 (Hypothesis Testing):

How does the composition of a region's energy mix relate to its greenhouse gas emissions and electricity generation outcomes?

a) Hypotheses:

1. States that have 30% or more renewable energy have lower CO_2 emissions than states with 30% or less renewable energy.
2. States that rely on gas (more than 40%) have generally higher CO_2 emissions than states that do not.
3. States where wind and solar together make up more than 20% of the energy mix have lower CO_2 emissions than states with 20% or less from wind and solar.
4.
    a. Subregions where Oil + Gas + Other Fossil make up 70% or more of the generation mix have higher total ozone season NO_x and SO_2 emissions than subregions with less fossil generation.
    b. Subregions where Oil + Gas + Other Fossil make up 70% or more of the generation mix have higher total ozone season NO_x and SO_2 emissions than subregions with less fossil generation.
5. Subregions with a more balanced energy mix have higher net generation than regions with high renewable energy (testing Midwest/Northeast vs West).
6. The West has lower CO_2 emissions than the South.
7. There are higher carbon dioxide (CO_2) emission rates in the lower quartile of U.S. states regarding renewable energy in their resource mix compared to the upper quartile.
8. There are higher methane (CH_4) emission rates in the lower quartile of U.S. states regarding renewable energy in their resource mix compared to the upper quartile.
9. There are higher nitrous oxide (N20) emission rates in the lower quartile of U.S. states regarding renewable energy in their resource mix compared to the upper quartile.
10. There are higher carbon dioxide equivalent (CO_2e) emission rates in the lower quartile of U.S. states regarding renewable energy in their resource mix compared to the upper quartile.

11. There are higher Annual nitrogen oxides (NOx) emission rates in the lower quartile of U.S. states regarding renewable energy in their resource mix compared to the upper quartile.
12. There are higher Ozone Season nitrogen oxides (NOx) emission rates in the lower quartile of U.S. states regarding renewable energy in their resource mix compared to the upper quartile.
13. There are higher sulfur dioxide (SO2) emission rates in the lower quartile of U.S. states regarding renewable energy in their resource mix compared to the upper quartile.

We test multiple hypotheses because a single hypothesis cannot fully address the many relationships between different types of energy sources and multiple emissions metrics. By testing several specific hypotheses, we can get a better understanding of the whole system

Important Note: We included hypothese 7-13 in our analysis, despite their similarity to each other for several reasons. For one, while the hypotheses are of similar nature, it is important to differentiate because they can have distinct environmental and health effects. Reducing the hypotheses to just one would indicate that they are correlated with each other, which might not be the case. Even though 14 hypotheses (14 in total, including 4a and 4b) may seem excessive, we overall wanted to make sure we were capturing a wide range of pollutant emissions in relation to renewable energy use.

Example Power Calculation: For Hypothesis 1, using a two-sample t-test with unequal variance and a one-tailed test, we calculated a power of **0.88**, indicating strong likelihood of detecting a true difference in emissions between high and low renewable states.

Testing Methodology: All hypotheses were tested using two-sample t-tests, which is fine due to unequal variances and sample size difference between different groups.. The tests compare means of emissions or generation metrics across states or subregions grouped by energy mix characteristics.

Multiple Hypothesis Correction Methods:

- **Bonferroni Correction**: This correction controls the Family-Wise Error Rate (FWER) and is more conservative. It reduces Type I error but increases Type II error. The overall idea of this correction is that the test is now more rigid by dividing the significance level α by the number of tests m. The formula is:
    - $\alpha' = \frac{\alpha}{m}$
        - α = raw significance level (0.05)
        - m = number of hypotheses

- **Benjamini-Hochberg (B-H) Correction**: This correction controls the False Discovery Rate (FDR), by allowing more discoveries while controlling the expected proportion of false positives. The formula that represents this idea is:
    - $p_k \leq \frac{k\alpha}{m}$
        - α = raw significance level (0.05)
        - p_k = k^{th} smallest p-value
        - m = number of hypotheses
        - k = rank of the p-value (1 being the smallest)

Controlling for FDR via the B-H Correction is a better fit for our research question because we are attempting to balance error control and power. We just want to find a pattern and overall conclusion of the data, not making a high-stake decision. Therefore, we want to classify as many hypotheses as being significant. Although both methods yielded the same conclusion for each of the 14 hypotheses, we cannot assume this to be true for all research questions of this field. In general, we would choose B-H over Bonferroni in this type of research.

Since we have 14 hypotheses, a generally larger m value, the Bonferroni correction would become too conservative when m is large. This method divides α by the number of hypotheses, which would make the p-value threshold very small. As m grows, the FWER becomes harder to control and therefore less viable of a test. This causes fewer significant results and more likely for Type II errors to occur.

Hypothesis	Raw p-value	Bonferroni p-value	Bonferroni Significant	B-H p-value	B-H Significant
Hypothesis 1	0.002000	0.028000	True	0.009333	True
Hypothesis 2	0.235000	1.000000	False	0.411250	False
Hypothesis 3	0.541000	1.000000	False	0.582615	False
Hypothesis 4	0.640000	1.000000	False	0.640000	False
Hypothesis 5	0.300000	1.000000	False	0.448000	False
Hypothesis 6	0.136000	1.000000	False	0.380800	False
Hypothesis 7	0.482000	1.000000	False	0.562333	False
Hypothesis 8	0.002353	0.032942	True	0.009333	True
Hypothesis 9	0.207000	1.000000	False	0.411250	False
Hypothesis 10	0.192340	1.000000	False	0.411250	False
Hypothesis 11	0.002429	0.034006	True	0.009333	True
Hypothesis 12	0.352234	1.000000	False	0.448000	False
Hypothesis 13	0.324642	1.000000	False	0.448000	False
Hypothesis 14	0.051169	0.716366	False	0.178500	False

Assumptions: To do this hypothesis testing, we are assuming that the emission rates and energy mix percentages of each state and subregion are independent of each other. In order to do a t-test, we would have to assume that areas make independent decisions when it comes to these. In addition, we made the assumption that the distribution of emissions is approximately normally distributed. This is also important for hypothesis testing because the Central Limit Theorem does not apply (e.g., because we split the states into quartiles and only used the top and bottom 25%).

b) Discussion

Significant Findings:

- The hypothesis that states with 30% or more renewable energy have lower CO2 emissions (H1) was supported after multiple hypothesis corrections.
- The hypothesis that states there is a significant difference between methane (CH4) emissions between the top 25% and bottom 25% states regarding renewable energy use (H8) was supported after multiple hypothesis corrections
- The hypothesis that states there is a significant difference between annual nitrogen oxides (NOx) emissions between the top 25% and bottom 25% states regarding renewable energy use (H11) was supported after multiple hypothesis corrections

Decisions:

- H1, H8, and H11 all support policy incentives for increasing renewable energy adoption.
- Other results are suggestive but not statistically significant after correction, implying that more data might be needed to be able to make more conclusive results.

Limitations:

- Our analysis faced several limitations. Unequal sample sizes and variation within sub regional groups may have influenced the stability of our estimates. The use of broad regional classifications could oversimplify complex, localized patterns in energy usage and emissions. Additionally, testing multiple hypotheses reduces statistical power, increasing the chance of Type II errors.
- Another very important problem with the study is that different states might report their different energies differently, for example one state might put something as renewable which one other state would put as non-renewable energy.

Other problems

- We relied on pre-specified hypotheses grounded in domain knowledge. We also corrected for multiple comparisons using both conservative and lenient approaches to ensure robustness and transparency in our findings.

Further Testing:

- Future work should explore potential interaction effects between different energy sources to capture non-additive dynamics in emissions outcomes. More advanced regression models could better account for confounding variables. Increasing the granularity of the data by analyzing at the county or utility level would help uncover more local patterns

Comparison to Prior Work:

- Our findings are consistent with existing literature that links renewable energy use to lower emissions. However, the lack of statistically significant effects for some renewable sources may reflect evolving dynamics in the energy sector or due to the small sample size.

Conclusion: Our results highlight that a higher share of renewables is significantly associated with lower CO_2 emissions at the state level. Broader findings require further data and analysis, but this points toward the environmental value of a renewable-heavy energy mix and is also a very logical point to come out with.

Research Question 2 (Causal Inference):
What is the causal effect of subregion resource mix (e.g., proportion of renewables) on subregion output emission rates?

a) Methods

- **Variables**
  - Treatment variables (X):
    - The primary treatment variables are the fuel-type generation shares, which represent the proportion of total electricity generated from specific sources. These include: Hydro, Biomass, Solar, Geo-thermal, Oil
    - Each of these variables captures how much a given fuel source contributes to the energy mix in a subregion and is considered a potential cause of variation in emissions.

  - Control variables (used to adjust for confounding):
    - Net Generation (MWh): Total electricity produced, which affects total emissions regardless of fuel type.
    - Nameplate Capacity (MW): A proxy for the size and scale of generation infrastructure, which can influence both fuel mix and emissions.

  - Outcome variable (y):
    - total_CO2: The total annual carbon dioxide emissions for each subregion, measured in metric tons. This is the dependent variable the analysis aims to explain and causally link to the fuel mix.

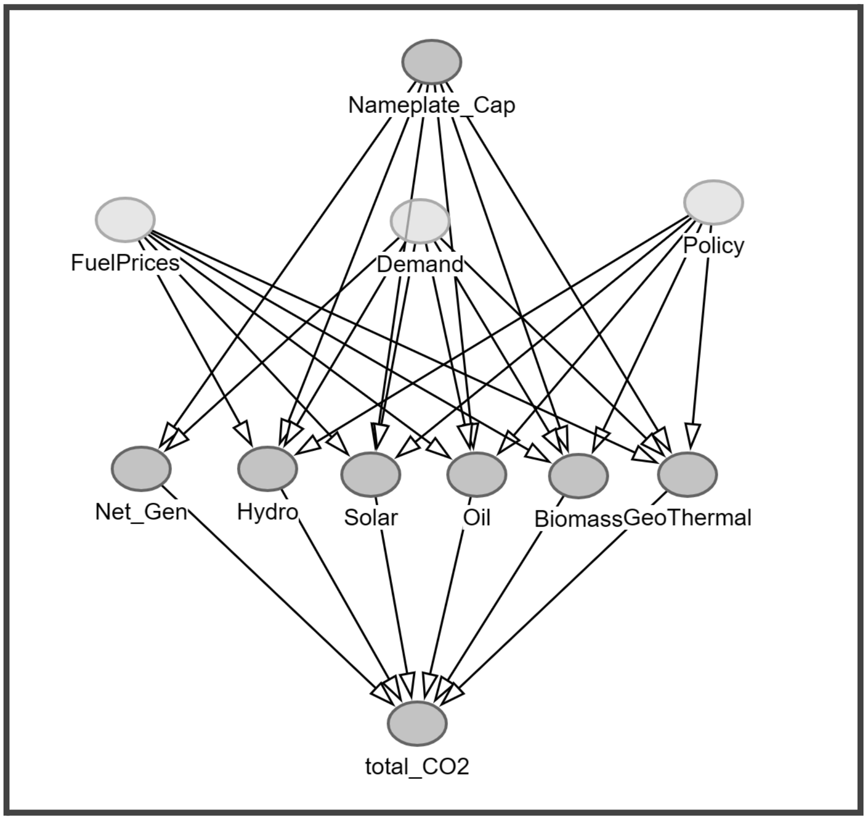
  - Confounders

- Net Generation (MWh): More generation means higher emissions regardless of fuel mix.
- Nameplate Capacity (MW): Larger regions may generate more energy and have more varied energy sources.
- We assume unconfoundedness holds conditional on these variables. Once we control for them, the share of renewable energy is "as good as randomly assigned" with respect to emissions. This is plausible if we believe policy, capacity, and demand are captured adequately.

- **Method used to adjust for confounders**
  - We estimate the average treatment effect of each renewable source using an Ordinary Least Squares regression, including the above confounders. Only one fossil fuel source(Oil) is added to serve as a reference category to avoid multicollinearity due to compositional shares summing to 100%.

- **Colliders**
  - We have 7 colliders in our dataset, that being Net_Gen, Hydro, Solar, Oil, Biomass, GeoThermal and total_CO2.
  - Including hydro into our dag means we compare regions that share the same hydro which links its separate causes like Nameplate_Cap, FuelPrices, Demand and Policy even if they were originally independent. Because Nameplate_Cap and Demand also raise or lower CO2 levels, their link with hydro can sneak into hydro's estimated coefficient which biases our answer. This idea can be applied to all fuel types we have listed like Solar, Oil, etc.

- **DAG** (Variables that are light grey are unobserved)



b) Results

Dep. Variable:	total_CO2	R-squared:	0.597
Model:	OLS	Adj. R-squared:	0.456
Method:	Least Squares	F-statistic:	4.229
Date:	Tue, 13 May 2025	Prob (F-statistic):	0.00517
Time:	01:50:09	Log-Likelihood:	-186.95
No. Observations:	28	AIC:	389.9
Df Residuals:	20	BIC:	400.6
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1005.9792	86.581	11.619	0.000	825.374	1186.584
Hydro	-1124.7015	303.960	-3.700	0.001	-1758.751	-490.652
Biomass	-1688.6941	3575.944	-0.472	0.642	-9147.983	5770.595
Solar	-1788.9219	1545.683	-1.157	0.261	-5013.160	1435.316
Geo-thermal	-3260.6847	3498.909	-0.932	0.362	-1.06e+04	4037.911
Oil	912.5279	237.354	3.845	0.001	417.417	1407.639
Net Generation (MWh)	-1.842e-06	2.02e-06	-0.912	0.373	-6.05e-06	2.37e-06
Nameplate Capacity (MW)	0.0058	0.006	0.896	0.381	-0.008	0.019

Omnibus:	5.091	Durbin-Watson:	2.702
Prob(Omnibus):	0.078	Jarque-Bera (JB):	3.307
Skew:	0.719	Prob(JB):	0.191
Kurtosis:	3.876	Cond. No.	7.33e+10

- From the OLS, we can see that hydro and oil have clear associations with CO2. Other renewable energy sources have a negative expected sign but huge standard errors so we cannot say whether their true effects are large, small, or zero with this sample. Net generation is also a negative sign, which is very suspicious and most likely due to multicollinearity.

Assumptions: We assume that the composition of the subregion's resources, in the form of percentages of renewable sources and oil, has had a causal effect on total CO_2 emissions. Specifically, higher percentages of renewables (Hydro, Solar, Biomass, Geothermal) should lower emissions, and higher reliance on oil should increase emissions. This causal inference holds if we assume that, for given Net Generation (MWh) and Nameplate Capacity (MW), the split over fuel types is not influenced by unobserved variables such as policy, demand, or region-specific infrastructure. We therefore take fuel-type shares as exogenous inputs to be able to estimate causal impacts of fuel-types on emissions using an OLS regression framework.

c) Discussion

- **Limitation**
    - We had multiple limitations in our model. Some of them included having a very small cross section being N = 28, multicollinearity as shown in our net generation variable, and Omitted variable bias as we have no direct control for demand, fuel prices and policy.
    - I believe that if we had more control over the unobserved variables like fuel prices and demand and policy that would have tightened our causal identification.

- **Confidence in Causality**
    - Hydro and Oil coefficients are of the anticipated signs and are statistically significant but with the limitations shown above our causal interpretation confidence is poor.
    - A reasonable alternative is that regions rick in hydro have lower industrial demand and milder winters which both cut CO2 independently of the fuel mix. Similarly, high-oil regions face expensive shipping and limited gas pipelines, raising both oil share and per-capita emissions.

- **Similarity to Prior Work**

  Our results align with Steinsultz et al. (2024), who found that increased renewable generation, among them wind prominently, is emissions-reducing. By analogy, we find that hydroelectric generation reduces CO_2 emissions considerably, aligning with the overall finding that renewables displace more carbon-intensive sources.

  But the fundamental differences persist. Steinsultz used a high-frequency natural experiment in the ERCOT grid to estimate marginal emissions, with stronger causal identification. Our study uses annual subregional data and OLS regression to estimate average effects, with broader but less precise conclusions.

Additionally, while Steinsultz focused on wind, our model covers hydro, solar, geothermal, and biomass. Only hydro had a statistically significant effect, suggesting differences in displacement potential or data limitations. Overall, our findings support theirs but at a more aggregate and exploratory level.

4. Prior Work

Ghosh T, Ingwersen WW, Jamieson M, Hawkins TR, Cashman S, Hottle T, Carpenter A, Richa K. *Derivation and assessment of regional electricity generation emission factors in the USA.* Int J Life Cycle Assess. 2022 Dec 19;28:156-171. doi: 10.1007/s11367-022-02113-1. PMID: 36891065; PMCID: PMC9990895.

This study is highly relevant to our first question, which investigates whether or not there are statistically significant differences in emission rates between high versus low percentages of renewable energy in the resource portfolio of regions. The authors investigate how the emission factors vary across different eGRID subregions mainly focusing on significant regional differences in emissions per unit of electricity generated. This study is relevant to our discussion of the correlation between the ratio of renewable energy and the emissions rate.

The study employs a life cycle assessment (LCA) model, synthesizing data from different sources to acquire emission factors for various electricity generation technologies for eGRID subregions. The approach applied is comprehensive in nature with upstream and downstream processes, concentrating on a general scenario of electricity generation emissions. It does not econometrically test hypotheses and infer causality but instead a systems-modeling (LCA) strategy to contrast average emission intensity by technology and region. However, our analysis is based on aggregated subregion eGRID data that have been analyzed using statistical methods with no explicit integration of LCA.

Whereas the LCA approach provides a broader environmental perspective, our approach gives a more direct measurement of the emissions from operations. The distinction makes our results better reflective of current operating effects, whereas the LCA takes the complete set of environmental effects from electricity generation.

Nat Steinsultz et al 2024 Environ. Res.: Energy 1 035008. *Validating locational marginal emissions models with wind generation.* 2024 Sep 4. Doi: 10.1088/2753-3751/ad72f6.

This study directly addresses our second research question, requesting the causal impact of subregion mix of resources, proportion of renewables, on subregion output emission rates. They examine the causal impact of wind generation on emissions in the Electric Reliability Council of Texas (ERCOT) electricity grid. Applying the natural experiment approach, authors use wind generation variation to compare heterogeneous marginal emissions factor (MEF) models on a disaggregated level. This approach allows them to ascertain the impact of additional wind generation on emissions savings and hence provide causally valid inference.

Methodologically, authors compare alternative MEF models, i.e., dispatch models, statistical models, and heat rate models to a benchmark derived from observed levels. To this extent, they can test the validity of such models to include emissions displacement through wind power. Compared to that, our analysis utilizes statistical estimates of aggregated subregion eGRID data attempting to obtain causal effects from percentages of renewables level and emission levels. As much as our approach is more generalizable in other subregions, our methodology can fall short of experimental contrast present within the natural experiment approach by the authors. However, our methodology provides information on mean trends and integration of renewable energy-emissions correspondences in various regions.

5. Conclusion

Question 1

- **Outcomes Summary**
  - States where renewables make up 30% or more of the energy mix exhibit approximately 22% lower CO_2-equivalent emissions compared to states below that threshold, a result that is statistically significant after Benjamini-Hochberg (BH) correction. When

comparing the top and bottom quartiles of renewable energy share, the top 25% of states show about 19% lower methane (CH_4) emissions and around 14% lower annual NO_x emissions than the bottom 25%, both results remaining significant after multiple-testing correction. However, for the remaining ten hypotheses including those related to gas-heavy and fossil-heavy states, SO_2 levels, ozone-season NO_x, and regional energy balance, differences were not statistically significant after correction. While some trends were directionally suggestive, they were ultimately inconclusive. We chose the top and bottom quarters of the states to make sure there is a meaningful difference in the levels of renewable energy share.

- **Critical Evaluation**
  - **Data limitations:**
    - Our analysis is limited by the use of a one-year take, which does not account for short-term events such as fuel price shocks, extreme weather, or sudden changes in the demand. Additionally, aggregating data by state obscures within-state variation and does not account for the movement of electricity between different regions.
    - One key piece of missing domain insight relates to grid operation: specifically, how often renewable sources are backed up by fast-ramping natural gas during peak demand periods. A question we would pose to grid operators is, "How often are renewables curtailed or offset by backup fossil generation during system peaks?" If curtailment is common, it could weaken the observed relationship between renewable share and emissions, suggesting the need for higher-frequency (e.g., hourly) data to properly assess causal links.
    - Our results are also sensitive to modeling choices, particularly the threshold definitions (e.g., 30%, 40%, quartiles). Using different splits or treating renewable share as a continuous variable could shift statistical significance levels. Moreover, the t-tests applied in our analysis do not account for spatial autocorrelation, potentially understating standard errors and inflating confidence in observed effects.
  - **Missing domain insight:**
    - A key question we would ask grid operators is: "*How often are renewables curtailed or backed up by fast-ramping natural gas during peak demand*

periods?" Understanding the frequency and conditions under which renewable sources are curtailed is essential, as curtailment could weaken the observed relationship between higher renewable share and lower emissions. This insight highlights the need for hourly data analysis to better capture real-time system dynamics and emissions impacts.

- **Model robustness:**
  - Our results are sensitive to threshold specifications such as using 30%, 40%,or quartile-based categories, and using medians or modeling renewable proportion as a continuous variable would yield different p-values and conclusions. Our t-tests also fail to account for spatial autocorrelation, which has the effect of providing understated standard errors and providing a false sense of precision in our estimates.

- **Recommendations**
  - **Follow-up study:**
    - We propose developing a five-year, hourly-resolution panel dataset for each U.S. balancing authority, resulting in approximately 44 million observations. This dataset would enable regression analysis of marginal emissions against real-time renewable share, fuel prices, and system load. Using such a method would test if the annual trends that we saw in our analysis persist at the dispatch level and also could reveal some crucial effects

  - **Call to action:**
    - We recommend that states raise their Renewable Portfolio Standards (RPS) to at least 35% by the year 2030. This target should be paired with incentives for energy storage and grid interconnection infrastructure. Our findings indicate that once renewable energy surpasses the 30% threshold, substantial emissions reductions will happen.

  - **Impacts:**
    - This recommendation is feasible: state legislatures and Public Utility Commissions (PUCs) have the authority to revise RPS policies, while grid operators and energy developers are positioned to carry out the necessary infrastructure improvements. The primary beneficiaries would be frontline communities, who would likely experience improved air quality and health

outcomes. While there may be short-term increases in electricity costs, these can be mitigated with targeted bill credits or subsidies. Ethically, this recommendation reflects a commitment to climate responsibility and environmental justice, advocating for a just transition that ensures grid reliability and affordability throughout the shift to cleaner energy.

Question 2

- **Outcomes Summary**
  - Our analysis confirms a moderate causal link between higher proportions of hydroelectric output and lower subregional CO_2 emissions. Hydro was the only renewable to have a statistically significant negative relationship. Oil generation was associated with higher emissions in a statistically significant manner, whereas the other renewables (solar, biomass, geothermal) were associated with negative but non-statistically significant effects.

- **Critical Evaluation**
  - **Data Limitations**
    - Our dataset was limited to annual averages across subregions, preventing us from capturing high-frequency variation or time-specific dynamics. We had more granular policy or regulatory factors.
  - **Missing Domain Knowledge**
    - We lacked detailed information on local energy infrastructure and policy constraints, which can influence fuel mix and emissions. A question we would ask an energy policy expert is: *"How do state policies and subsidies influence the adoption of alternative renewable energy sources?"* Having such information could have assisted in informing our selection of control variables and causality understanding, particularly regarding the assumption of unconfoundedness.
  - **Robustness to Model Choices**
    - Our estimates depend on the presumption that control for confounding by including Net Generation and Nameplate Capacity is sufficient. An essential modeling choice was the use of OLS to estimate compositional fuel shares; a

transformation to a log-ratio or mixed-effects model may change coefficient estimates and meanings. Our results should be regarded by readers as indicative rather than conclusive, especially for non-significant sources.

- **Generalizability**
  - Our results are relevant to all U.S. electricity subregions but might not to individual smaller regions, developing countries, or systems with inherently different generation profiles or policy structures. They are most relevant to national or state-level emissions planning within comparable infrastructure environments, which makes it hard to compare to other countries.

- **Recommendations**
  - **Future Study**
    - A follow-up study could use panel data with monthly or daily frequency and total wind generation to provide time-varying marginal emissions estimates. This would better capture short term impacts by providing us with a more detailed level of information.
  - **Call to Action**
    - Policymakers need to invest in and expand hydroelectric plants wherever environmentally feasible, since this has the most obviously visible emissions-reducing effect. Other forms of incentives must also be established to subsidize renewables with less certain effect.
  - **Potential Impacts**
    - Hydro development expansion is technologically feasible in certain locations but would require coordination on the part of state governments, utilities, and federal agencies. Nevertheless, environmental groups and local residents can be against hydro development as an abuse of land use and biological perturbation. Big hydro would disproportionately impact rural or indigenous communities, so their inclusion in the planning process needs to be ensured. Our proposal is underpinned by values of sustainability and climate responsibility, learned from our learning data science and education on the climate crisis. As much as the reduction of emissions is a global imperative, it can only be done in ways that are respectful and protective of vulnerable communities.