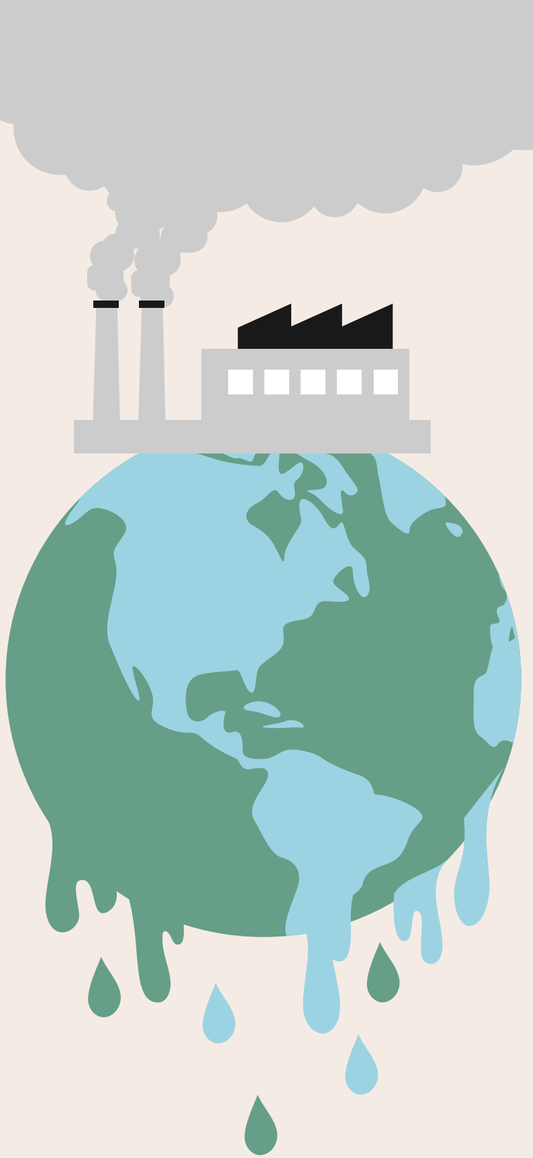
Climate TRACE: Estimating global greenhouse gas emissions from buildings



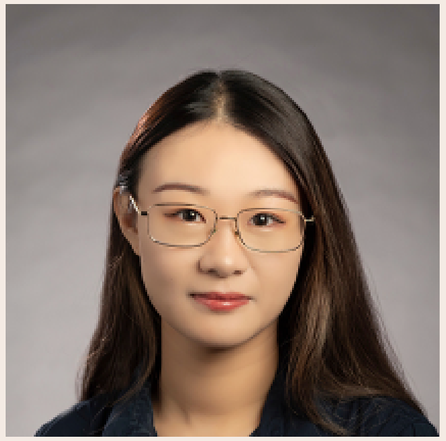
MIDS Capstone 2025
Duke University

Team

Bárbara Flores

Meixiang Du

Jiechen Li



Yulei (Alicia) Xia

Mentor & Client: Kyle Bradbury, Ph.D.

Duke University

Problem Statements

In 2024, global carbon dioxide (CO_2) emissions reached a record 41.6 billion metric tons.[1]
Emissions inventories remain limited—52 countries lack data after 2012, with even larger gaps for subnational and local governments.[2] [3]
Existing data is often outdated, spatially limited, or not free.
Available emission data is at a coarse spatial resolution.

[1]. Source: https://www.livescience.com/planet-earth/climate-change/global-carbon-emissions-reach-new-record-high-in-2024-with-no-end-in-sight-scientists-say
[2]. Source: https://climatetrace.org/news/more-than-70000-of-the-highest-emitting-greenhouse-gas
[3]. Source: Luers, A., Yona, L., Field, C. B., Jackson, R. B., Mach, K. J., Cashore, B. W., Elliott, C., Gifford, L., Honigsberg, C., Klaassen, L., & Matthews, H. D. (2022). Make greenhouse-gas accounting reliable—Build interoperable systems. *Nature*, 607(7920), 653–656.

Problem Statements

In 2024, global carbon dioxide (CO_2) emissions reached a record 41.6 billion metric tons.[1]
Emissions inventories remain limited—52 countries lack data after 2012, with even larger gaps for subnational and local governments.[2] [3]
Existing data is often outdated, spatially limited, or not free.
Available emission data is at a coarse spatial resolution.

Climate TRACE

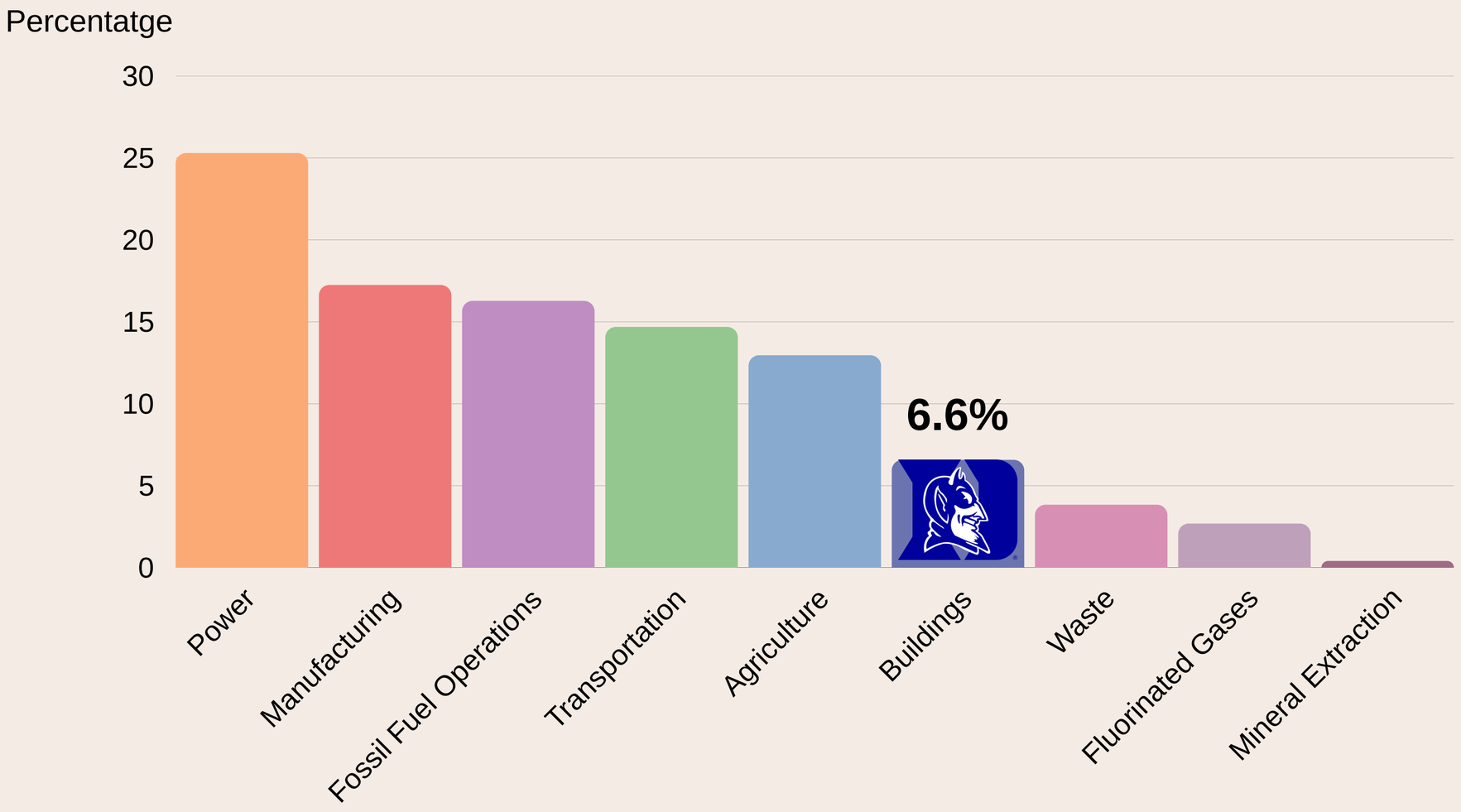
Climate Trace is a global non-profit coalition.
They estimate and track global GHG emissions since 2015.
They provide fast, easy, and free data access to the public.
They provide comprehensive granular data across the sectors.

[1]. Source: https://www.livescience.com/planet-earth/climate-change/global-carbon-emissions-reach-new-record-high-in-2024-with-no-end-in-sight-scientists-say
[2]. Source: https://climatetrace.org/news/more-than-70000-of-the-highest-emitting-greenhouse-gas
[3]. Source: Luers, A., Yona, L., Field, C. B., Jackson, R. B., Mach, K. J., Cashore, B. W., Elliott, C., Gifford, L., Honigsberg, C., Klaassen, L., & Matthews, H. D. (2022). Make greenhouse-gas accounting reliable—Build interoperable systems. *Nature*, 607(7920), 653–656.

Climate TRACE

Global Greenhouse Gas Emissions by Source

Source: https://climatetrace.org/sectors



We focus on the **building sector**, which accounts for **6.6%** of total emissions.

How to Calculate GHG

A x **EUI** x **EF** = **GHG**

A

Estimate from direct observation through satellite imagery

Building Floor Area

EUI

Gathered from statistical assessments and adjusted using machine learning

Energy Use Indensity measures the energy consumption per square meter of building space

Residential

Non-Residential

EF

Gathered from statistical assessments

Equivalent Emissions Factor

Source: Markakis, P. J., Gowdy, T., Malof, J. M., Collins, L., Davitt, A., Volpato, G., & Bradbury, K. (2023). *High-resolution global building emissions estimation using satellite imagery.* Presented at the NeurIPS 2023 Workshop on Tackling Climate Change with Machine Learning.

Dataset

Data Source

- World Bank’s CURB Tool provides comprehensive and reliable data on urban buildings and energy use.

482 Entries

- These **city-level** entries represent 482 distinct locations across the globe, serving as the dependent variable in our analysis.

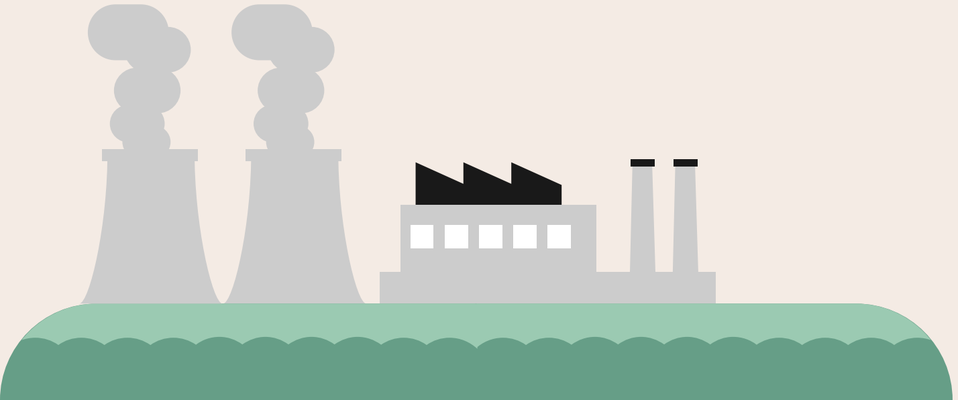
1 Entry = 1 Geographical Location

- Each entry represents a geographical location, defined by a specific latitude and longitude, which is treated as a data point in our calculations.

Goal of the Project

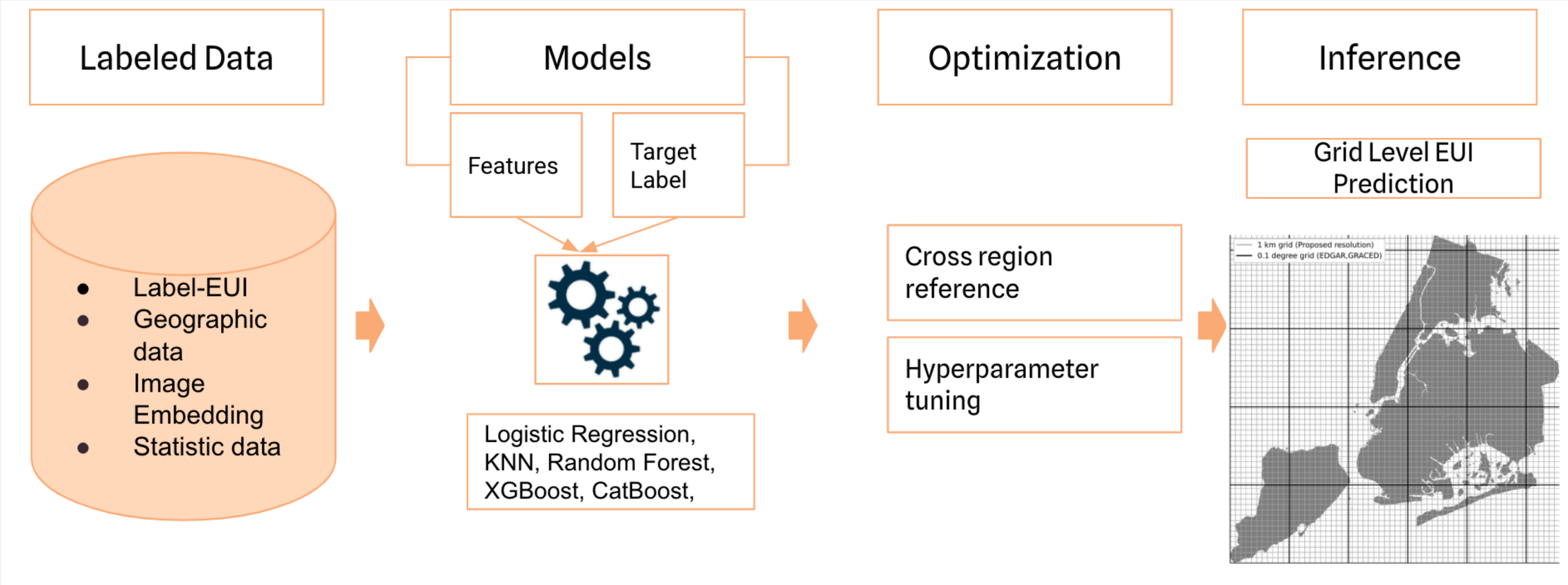
Develop a machine learning model to predict the EUI of buildings using globally available features, enabling the estimation of global greenhouse gas emissions.

Predict EUI



Methodology

- Employing geographic information techniques, image embedding retrieval methods to process features.
- Use supervised learning models for predicting target variables and conduct cross-region evaluation.



**grid image from paper: High-resolution Global Building Emissions Estimation using Satellite Imagery.

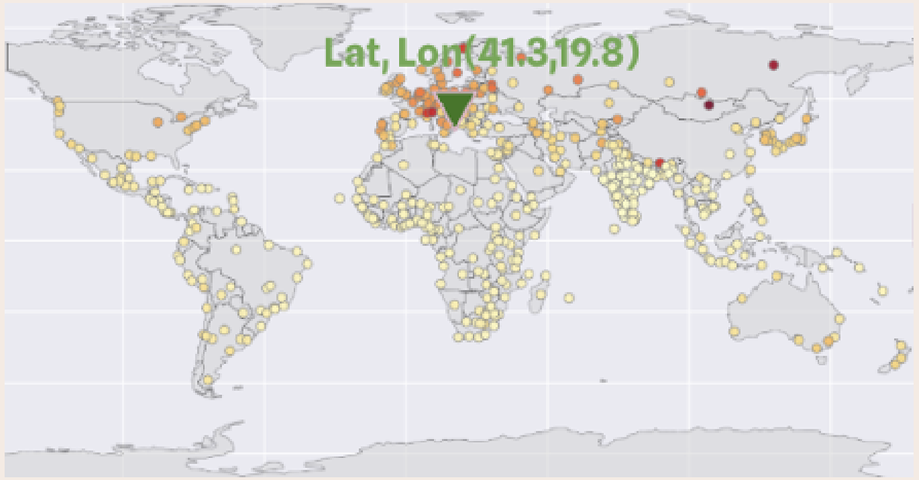
Identify Variables–Feature Map

From previous study, literature review and further interpretation of the project.

Building Geometry Data	**Weather**	**Socioeconomics**	**Policy/Law**
Embeddings from Satellite Image data(Sentinel-2 images)	Temperature, DewPoint Temperature, Heating Degree Days(HDD), Cooling Degree Days(CDD)	Human Development Index(HDI), GDP, Population, Educational Index, Income Index, Urbanization	Paris Agreement

Represent Variables–Image Embedding

Locations of Interest



Satellite image

Embeddings for the image

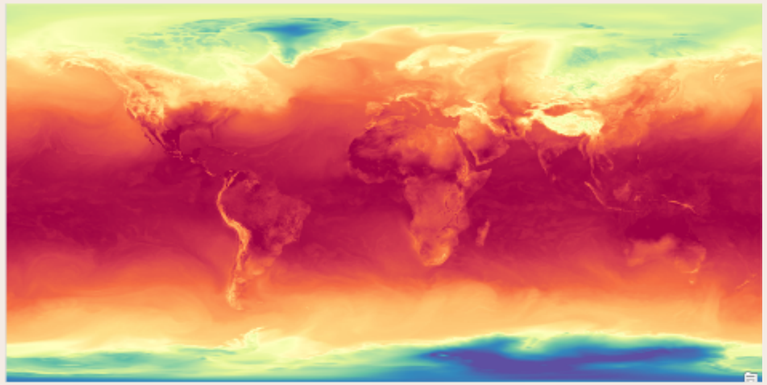
$$A \in \mathbb{R}^{474\times1024}$$

$$A = \begin{bmatrix} 0.1729 & -0.0879 & 0.3115 & \dots & 0.0661 & 0.0046 & 0.1092 \\ 0.2822 & -0.1318 & 0.5555 & \dots & 0.0915 & -0.0186 & 0.0199 \\ 0.2598 & -0.1492 & 0.4791 & \dots & 0.0334 & 0.0208 & 0.0811 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0.2779 & -0.1090 & 0.5454 & \dots & 0.0881 & 0.0216 & -0.0259 \\ 0.2572 & -0.0822 & 0.4835 & \dots & 0.0501 & 0.0377 & -0.0859 \\ 0.3378 & -0.0960 & 0.5385 & \dots & 0.0607 & 0.0020 & -0.0648 \end{bmatrix}$$

1. Retrieve satellite images from Sentinel-2 for the locations of interest.
2. Employ Clay (a pretrained Vision Transformer MAE model specialized in satellite images) to generate embeddings.
3. Apply dimension reduction and/or classification methods to optimize the representational ability of embeddings.

Represent Variables–Geographic information

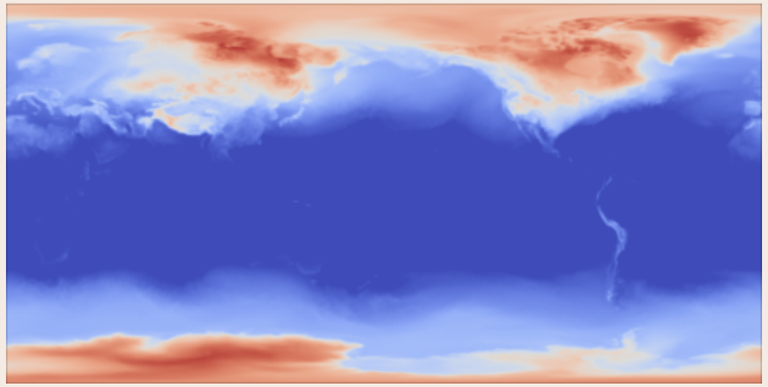
Retrieve Temperature



Low High

Retrieve air temperature at 2m above the surface, data source from European Commission and the Group on Earth Observations.

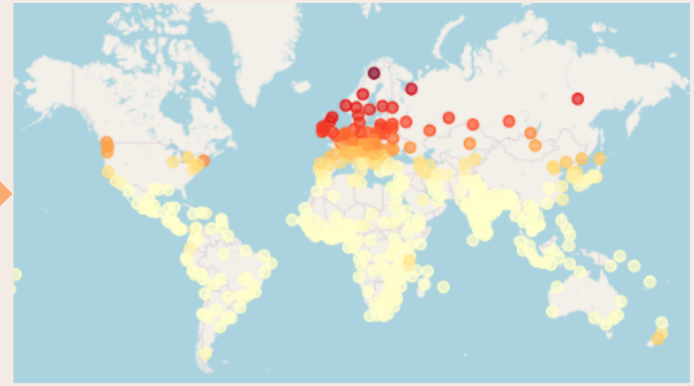
Heating Degree Days



Low High

Calculate HDD to measure the demand for heating energy based on the difference between outdoor temperature and a baseline "comfort" temperature, typically 65°F (18°C).

Heating Degree Days for City



Low High

Assign HDD to the nearest data point(city level) of interest

Experimental Design - Modeling Energy Use Intensity

- **Objective:** Predict Energy Use Intensity

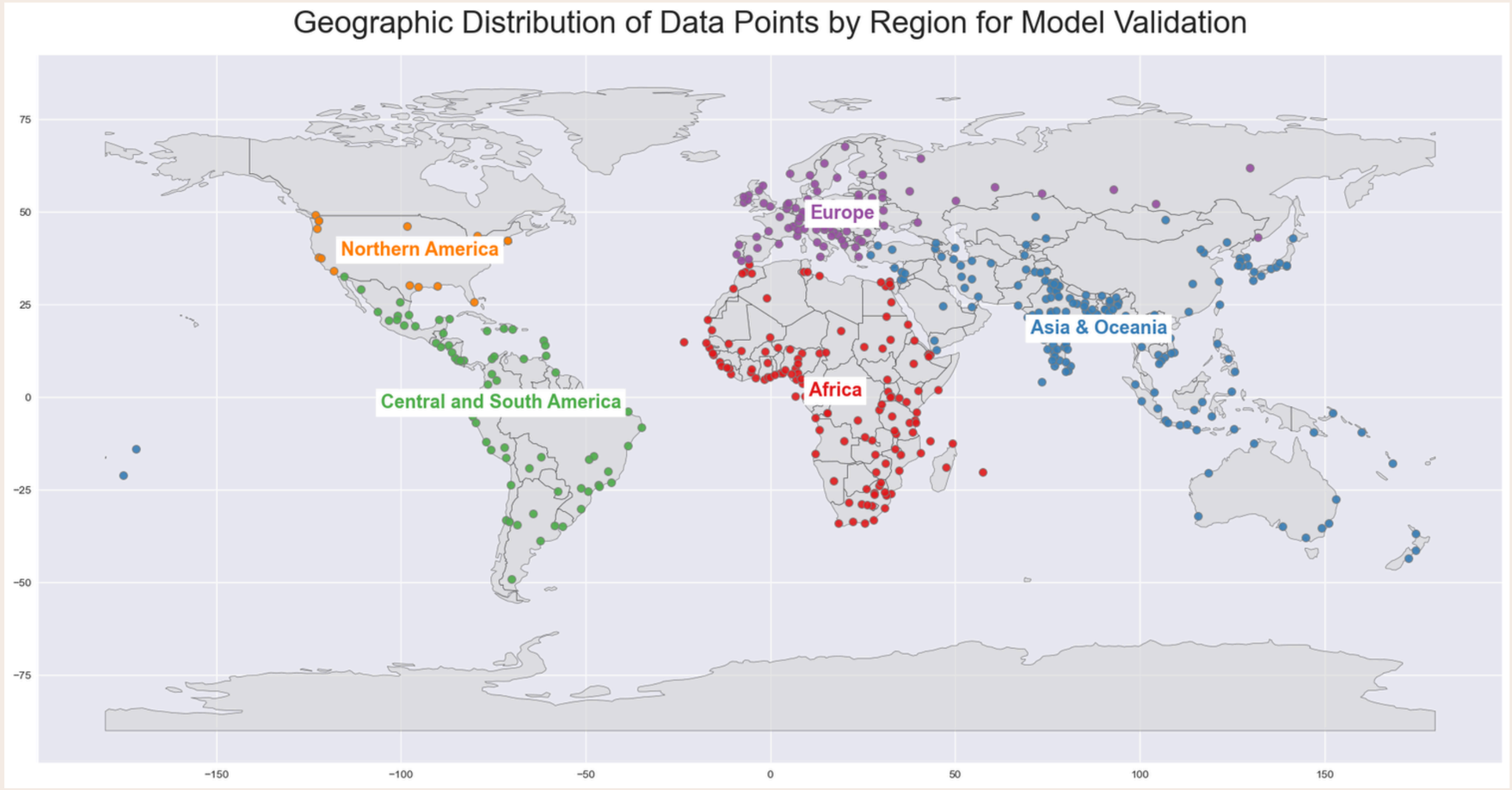
- **Models Used:**
  - K-NN
  - Linear Regression (Lasso and Ridge)
  - Random Forest
  - XGBoost
  - CatBoost

- **Feature Selection:** Removed less important features based on feature importance analysis.

- **Hyperparameter Tuning:** Grid Search for MAPE optimization.

- **Validation Strategy:** Region-based split for validation

Experimental Design - Regional Validation Approach



- One challenge with global data is the difficulty of extrapolating results due to regional variations.
- To address this, we will **validate** our predictions at the regional level and estimate Energy Use Intensity while considering these effects. This approach helps us identify biases and enhances the model's robustness for better **extrapolation**.
- We de are considering 5 regions

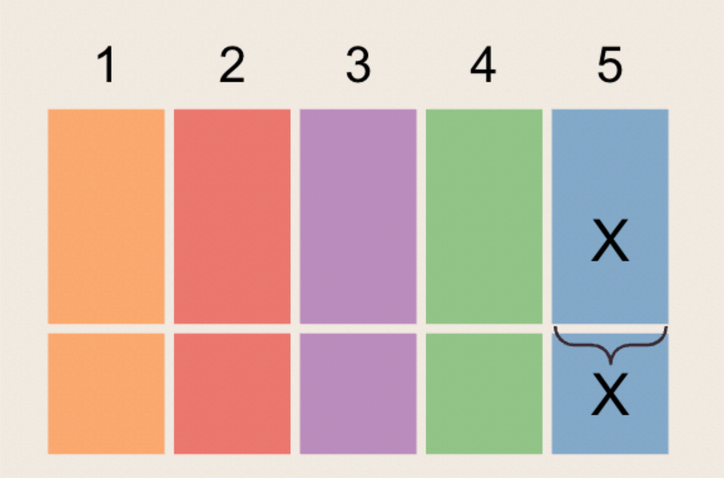
Experimental Design - Regional Validation Approach

Region

Train 80%

Test 20%

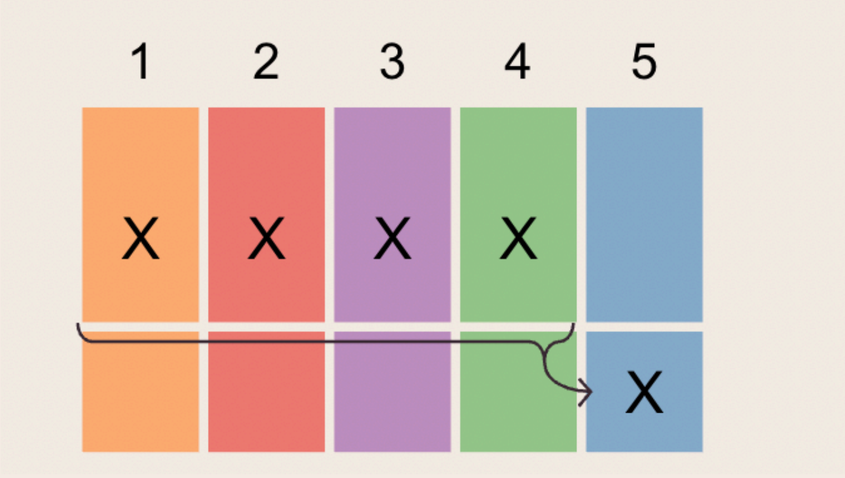
Within-Domain



We train our model on 80% of the data from each region and test it on the remaining 20% of the same region.

To obtain a global result, we calculate the average of our evaluation metrics across the 5 regions.

Cross-Domain



We train our model using 80% of the data from 4 regions and test it on the 20% of the 5th region. We repeat this to evaluate the model's extrapolation.

The global result is the average of all outcomes.

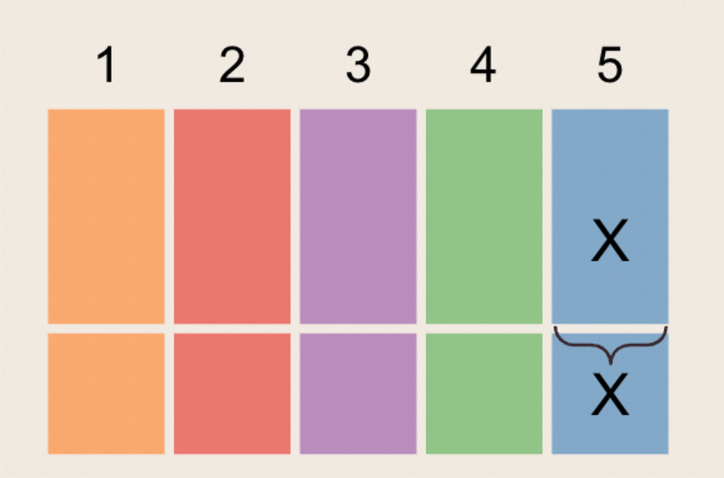
Experimental Design - Regional Validation Approach

Region

Train 80%

Test 20%

Within-Domain

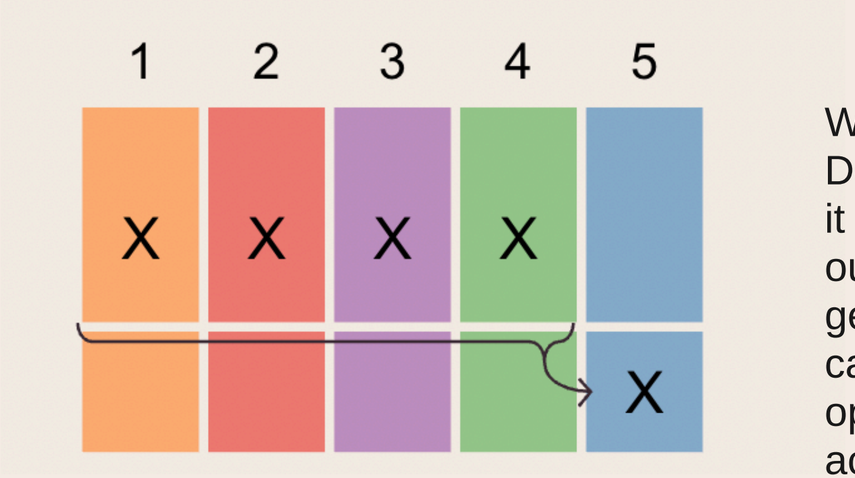


We train our model on 80% of the data from each region and test it on the remaining 20% of the same region.

To obtain a global result, we calculate the average of our evaluation metrics across the 5 regions.

Optimistic Validation

Cross-Domain



We train our model using 80% of the data from 4 regions and test it on the 20% of the 5th region. We repeat this to evaluate the model's extrapolation.

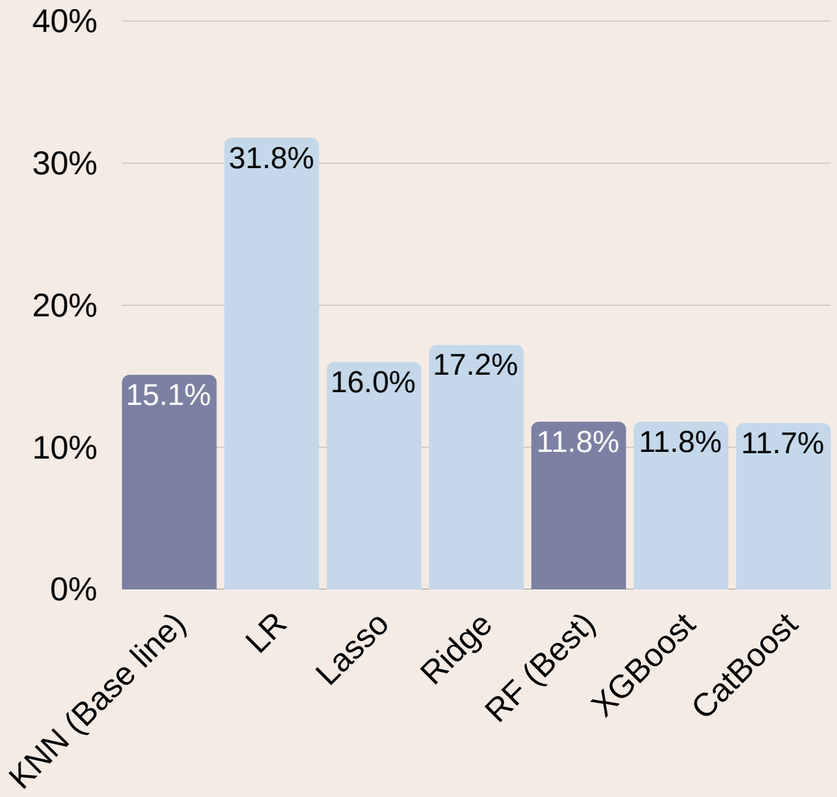
The global result is the average of all outcomes.

Conservative Validation

We prioritize Cross-Domain validation as it better represents our model's real-world generalization capabilities and optimize our models accordingly

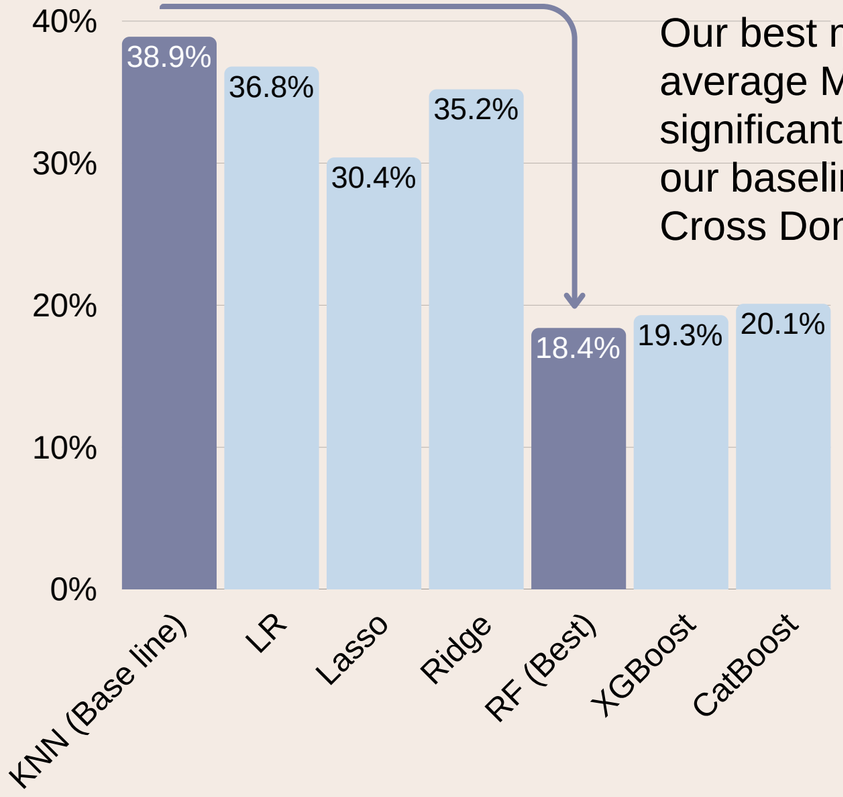
Results - Average MAPE across regions and building types

Within-Domain



Optimistic Validation

Cross-Domain

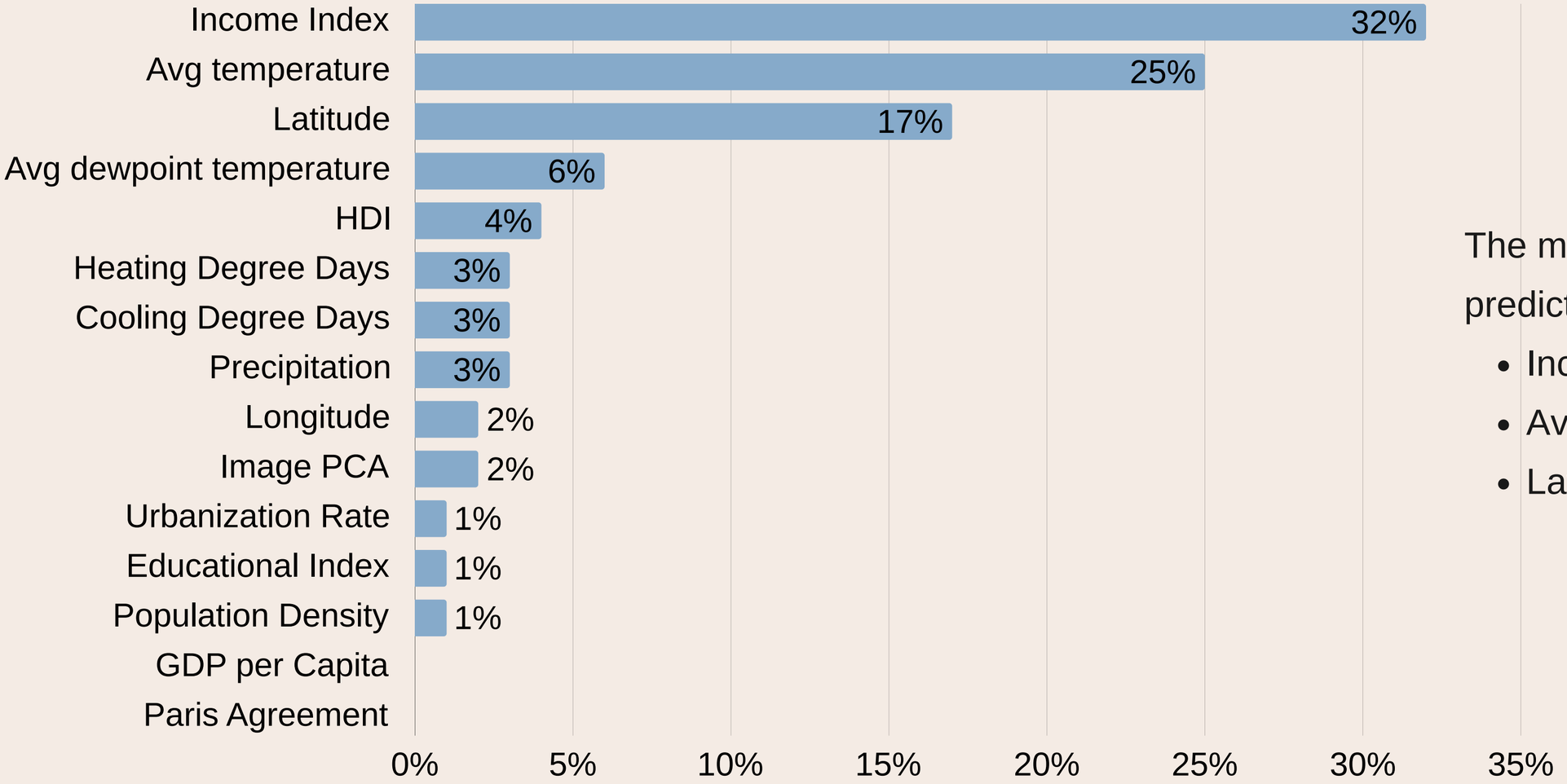


Conservative Validation

Our best model achieves an average MAPE of 18.4%, significantly improving upon our baseline of 38.9%. in Cross Domain validation

MAPE: Mean Absolute Percentage Error. Is the average of the absolute percentage errors between predicted and actual values. A **lower MAPE indicates better** model performance.

Results - Feature Importance Using Random Forest



The most important variables in predicting EUI in buildings are:

- Income Index
- Average Temperature
- Latitude

Heating Degree Days: quantify heating demand based on temperatures below a set threshold

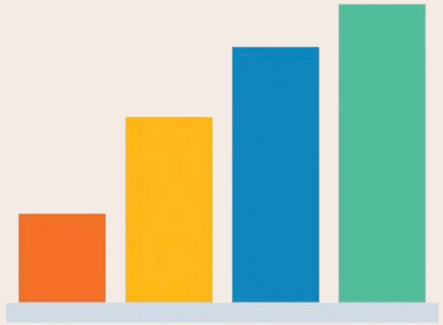
Cooling Degree Days: quantify cooling demand based on temperatures exceeding a set threshold

HDI Human Development Index: measures a country's development based on life expectancy, education, and income per capita.

GDP Gross Domestic Product: is the total value of goods and services produced in a country.

Results

- Ensemble models outperformed individual algorithms, with **Random Forest** achieving the **best results** overall
- Our best model achieved an **average MAPE of 18.4%**, representing a **53% improvement from the baseline** (38.9%) in cross-domain validation
- Socioeconomic factors proved crucial - **Income Index** is the most influential predictor **(32% importance),** followed by **Temperature (25%)**



Implications

- **Filling the Knowledge Gap:** This project fills the knowledge gap of local and municipal emissions inventories by having accurate building emissions data, crucial for estimating and reducing global GHG emissions.

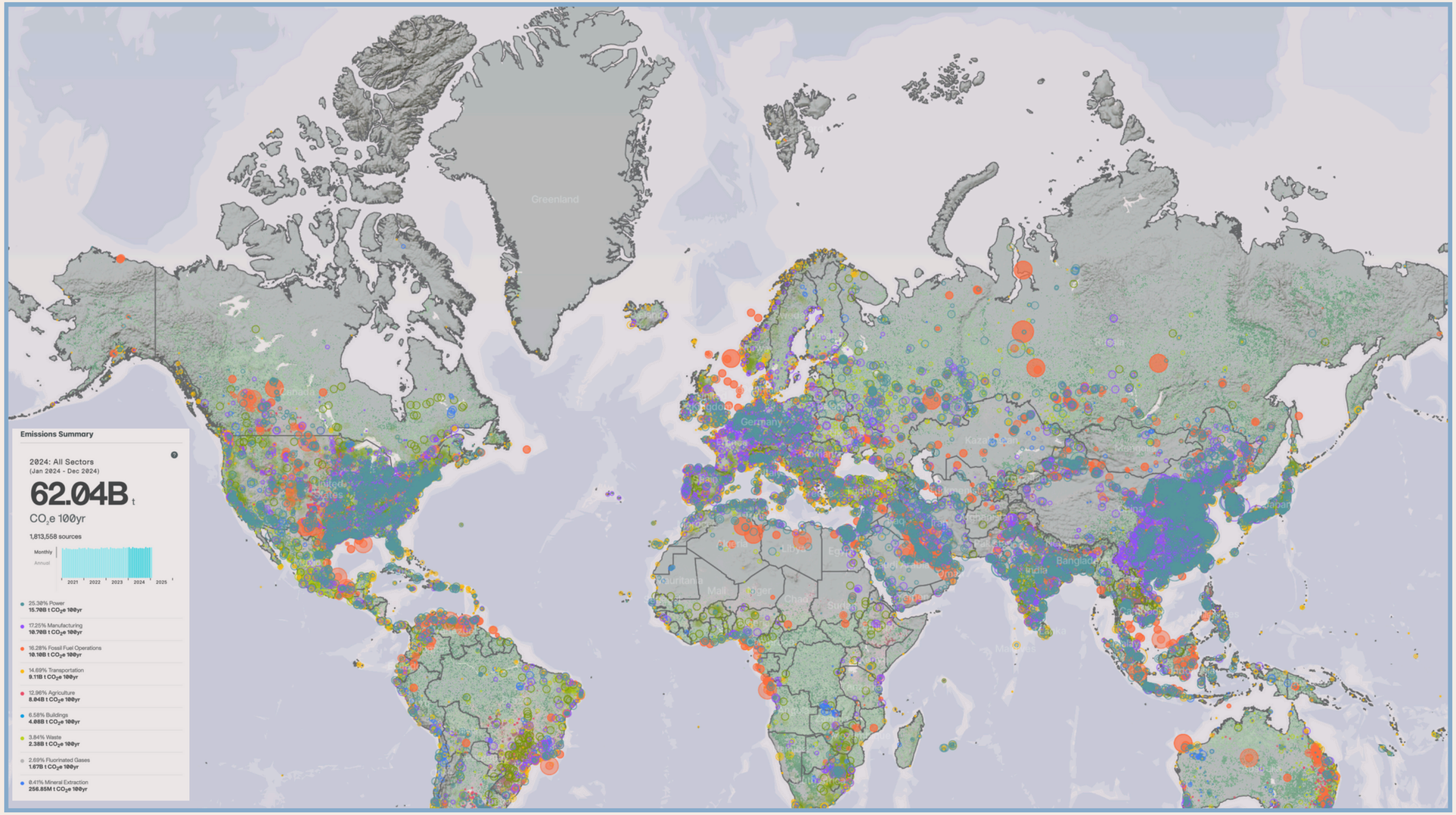
- **Informing Climate Policy:** Access to accurate emissions data enables policymakers to develop targeted, data-driven strategies for reducing GHG emissions.

- **Expanding Data Sources:** Given data limitations, we explored additional public EUI datasets. However, U.S. city-level data offered no new insights beyond our original dataset, underscoring the challenge of obtaining comprehensive global building energy data.

Limitations

- **Data Sparsity:** With only 482 data entries, the dataset is insufficient to accurately represent global trends and predict the EUI for the entire world.
- **Lack of Building-Specific Features:** Key factors such as age, materials, and energy codes are missing due to the limited granularity of the dataset.
- **EUI Data Constraints:** Derived from a single, possibly imprecise time point and centered on major cities (overlooking suburban/rural areas). Aggragation data that applying US-based building-use assumptions globally and converting units adds further uncertainty, limiting overall accuracy and usefulness.

Future Work



Global Scale Extrapolation

Estimating GHG Emissions

Exploring New Features and Model Refinements

Thanks!

We are ready for your questions!