# Predicting Surface Temperature Change from Agricultural $CO_2$ Emissions

DATASCI207 Fall 2023

Darya Likhareva, Faye Titchenal, Rachel Tripoli , Julia Zhao

Berkeley
UNIVERSITY OF CALIFORNIA

# Motivation

**Primary Motivation:** explore the impact $CO_2$ emissions has on temperature changes, regionally.

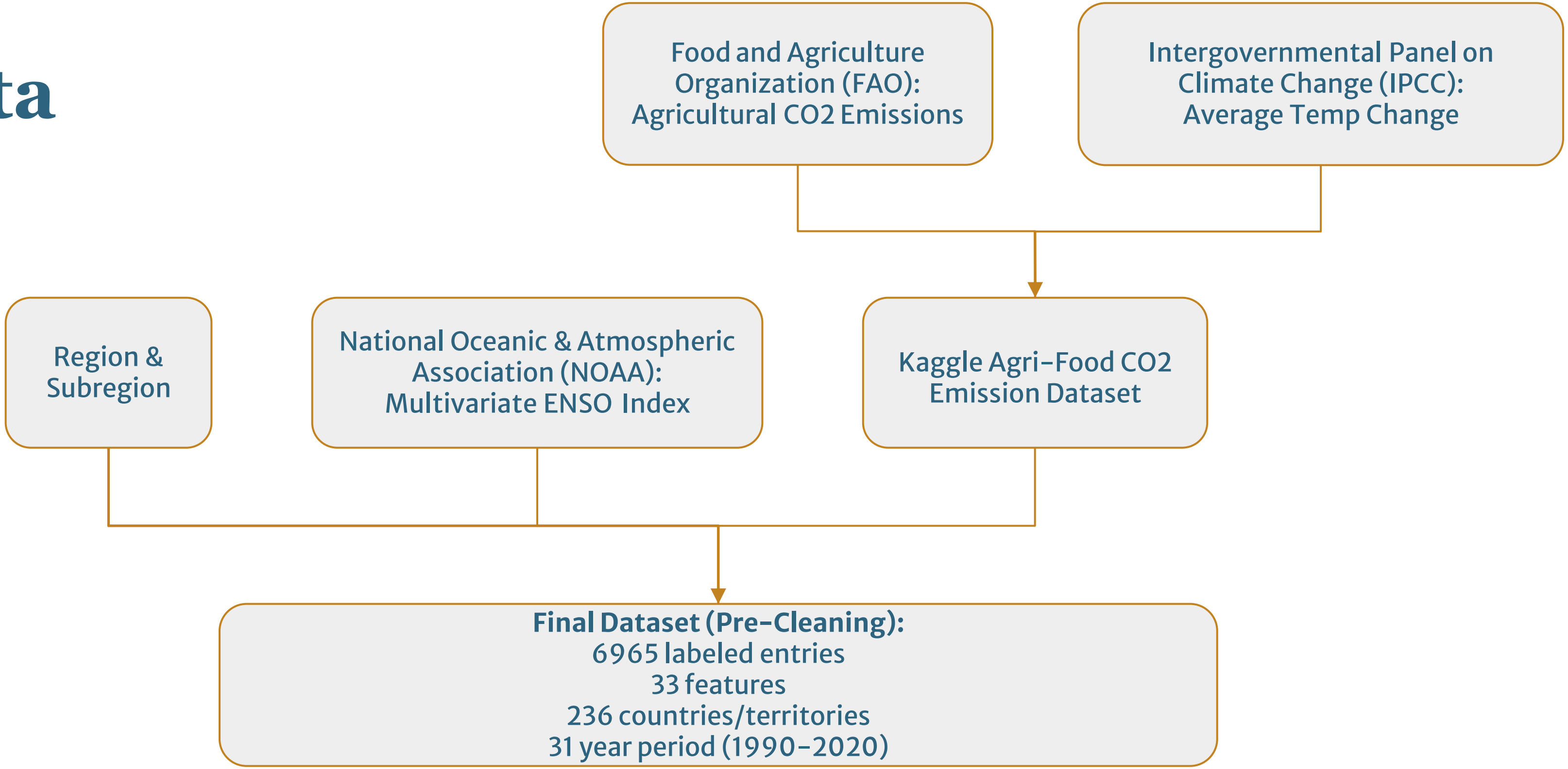
The Intergovernmental Panel on Climate Change (IPCC) and data from Our World in Data collectively show that climate change is a global, rapidly intensifying phenomenon, heavily influenced by human activities, with urgent calls for substantial reductions in greenhouse gas emissions to mitigate its impact

**Past research/models**:
- Short-term: models are based on atmospheric and/or geophysical processes
- Long-term: utilize paleo-climate data to forecast future
- Machine-Learning in climate models is in its infancy
    - Experimental
    - Long term goal in research is to use machine learning to bridge the gap between scales of current models to increase the resolution of accuracy

# Question

Can we predict the average annual land temperature change from annual agricultural $CO_2$ emissions?

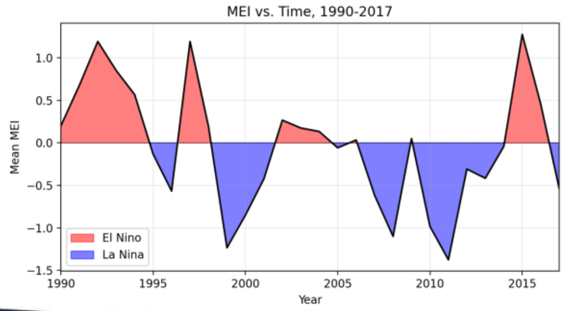
# Data


Food and Agriculture Organization (FAO): Agricultural CO2 Emissions
Intergovernmental Panel on Climate Change (IPCC): Average Temp Change
Region & Subregion
National Oceanic & Atmospheric Association (NOAA): Multivariate ENSO Index
Kaggle Agri-Food CO2 Emission Dataset
Final Dataset (Pre-Cleaning):
6965 labeled entries
33 features
236 countries/territories
31 year period (1990-2020)

# Features

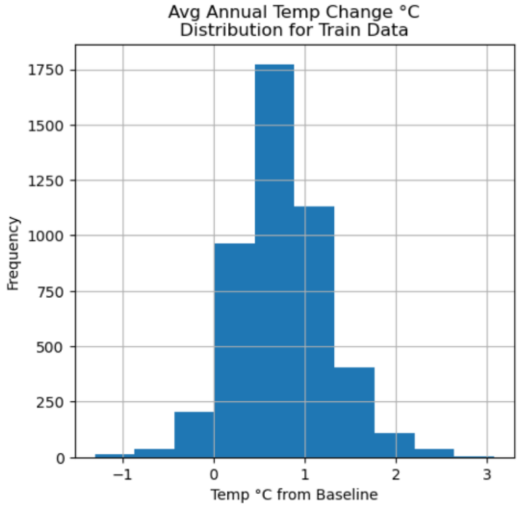
**Inputs:**

1. Country/Territory, Region, Sub-Region
2. Year
3. Agricultural Emissions (in kilotons of $CO_2$):
   - Fires in different ecosystems → Savanas, Forests, etc.
   - Food Systems → Food packaging, transport, & retail
   - Manure Management → Methane production from livestock
   - Industrial Processes and Product Use → Fertilizer manufacturing
   - On-Farm Energy Use → Electricity and fuel for equipment
1. Multivariate El Niño Southern Oscillation (ENSO) Index:
   - Represents fluctuations in sea surface temperature & air pressure
   - Weather patterns directly influence agricultural production




**Output:**

Average Annual Temperature Delta (ºC) → Temperature change from baseline period of 1951-1980



# Pre-processing

1. Combined similar columns
2. Joined with two other datasets to obtain MEI and region/subregion details
3. Evaluated dataset for completeness:
    - Some areas did not have data for the full 31 years
    - Some features contained null values
4. Log transformation of heavily skewed variables
5. Cumulative sum of features
6. Train | Validation | Test Split: ~ 80% | 10% | 10%
    - 1990 - 2014 | 2015 - 2017 | 2018 - 2020
7. Standardize input variables in training dataset

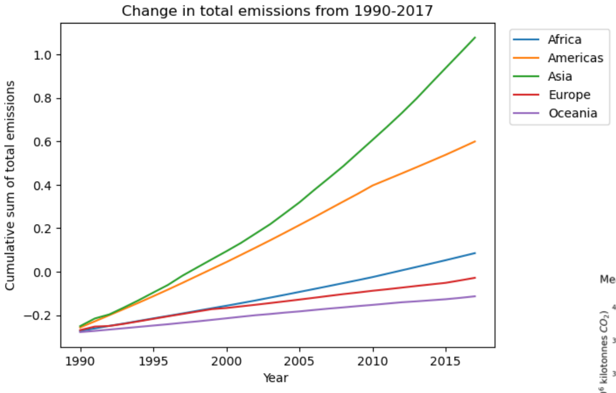
Start:

```
(6965, 31)
```

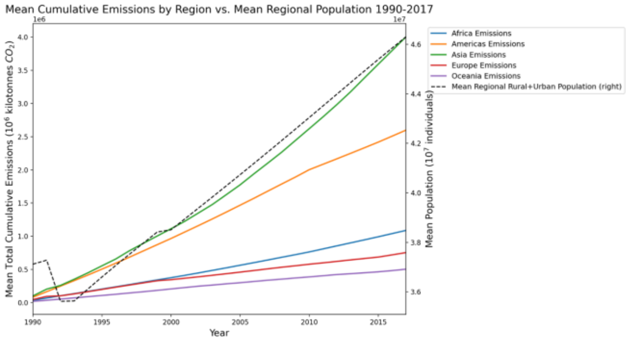
End:

```
(4681, 48)
(570, 48)
(570, 48)
```

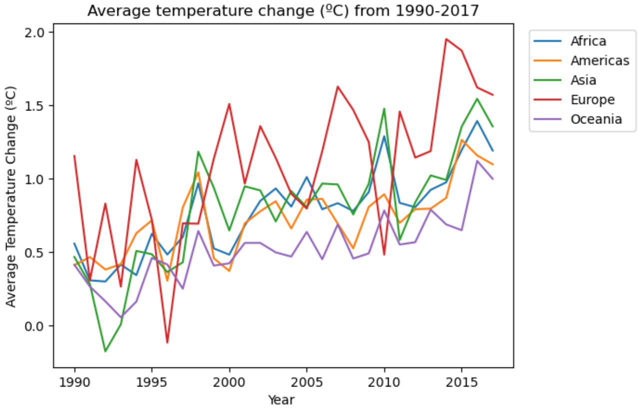
# EDA




Asia shows a steep upward trend, indicating a significant increase in total emissions over time




Here as well, Asia shows increased emissions by population level

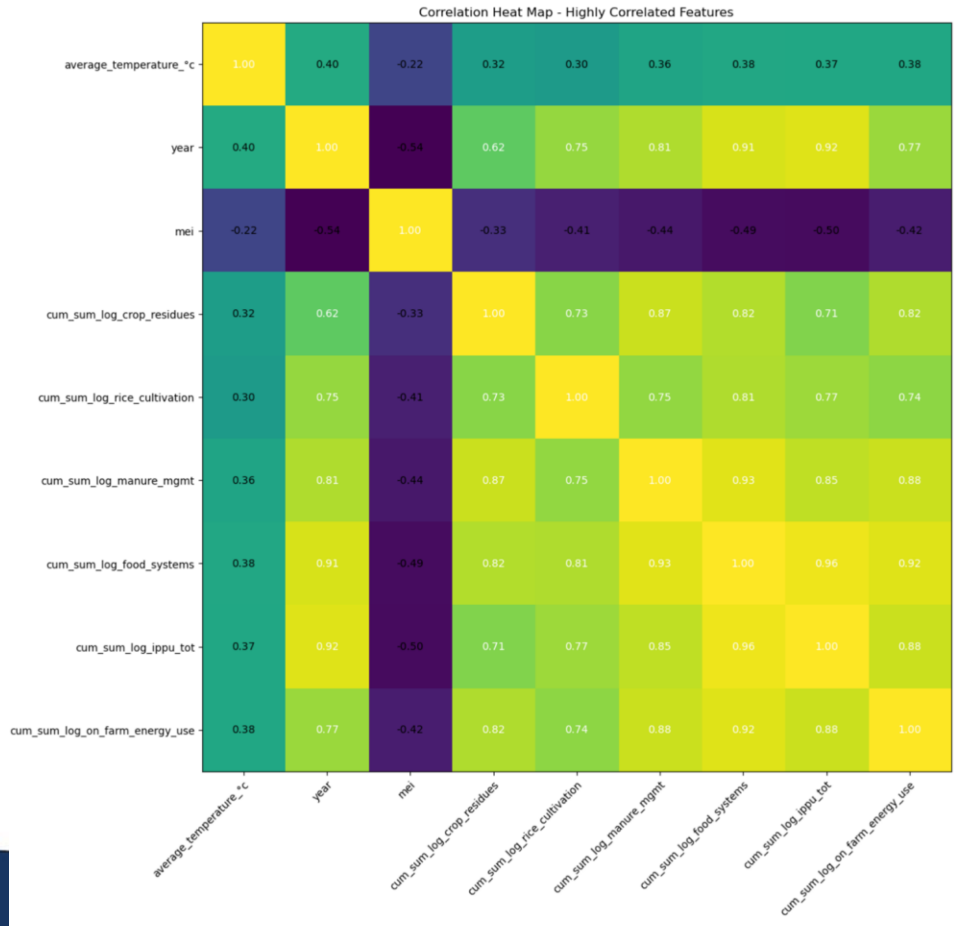



Despite fluctuations, there is a general upward trend in average temperature change for all regions

# Feature Selection

Highest correlations between input variables and average temp:

- Cum_sum_log_on_farm_energy_use
- Cum_sum_log_food_systems




Berkeley UNIVERSITY OF CALIFORNIA

# Modeling Approach

Baseline Model: Linear Regression with one input feature

Model 1: Regression Tree

Model 2: Random Forests

Model 3: XGBoost Tree

Model 4: Multiple linear regression

Model 5: Feed forward neural network

# Metrics

Mean Absolute Error (MAE):

- Robust to outliers & interpretable

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Root Mean Squared Error (RMSE):

- Penalizes large errors & interpretable

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}|y_i - \hat{y}_i|^2}{N}}$$

2 Sample T-test:

- Compare errors between models to determine statistical significance

Berkeley
UNIVERSITY OF CALIFORNIA

# Final Results

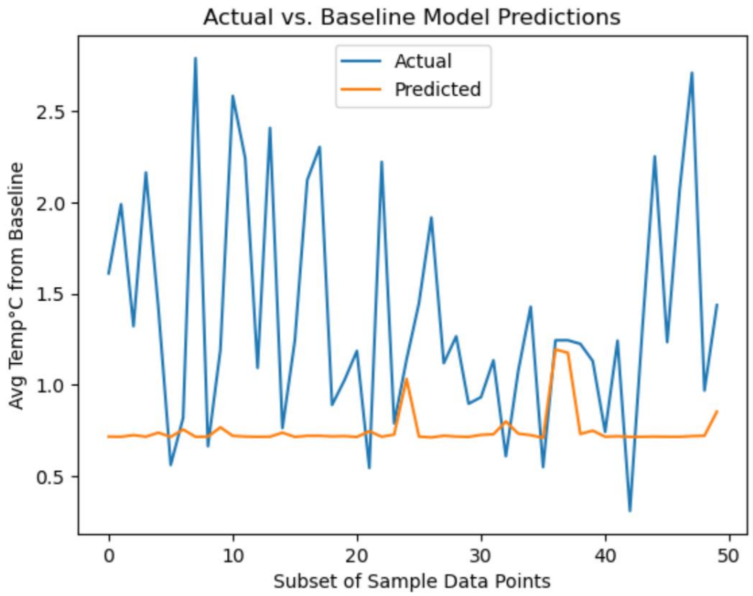
- Linear Regression model demonstrated best performance
- 2 Sample T-test indicates statistically significant difference between Linear Regression and FFNN errors
- On average prediction is 0.38ºC from actual temperature

| | MAE | RMSE |
|---|---|---|
| Linear Regression | 0.382 | 0.502 |
| FFNN | 0.411 | 0.522 |
| XGBoost | 0.483 | 0.645 |
| Regression Tree | 0.519 | 0.697 |
| Random Forest | 0.521 | 0.690 |
| Baseline | 0.743 | 0.916 |

# Experiments

# Baseline Model

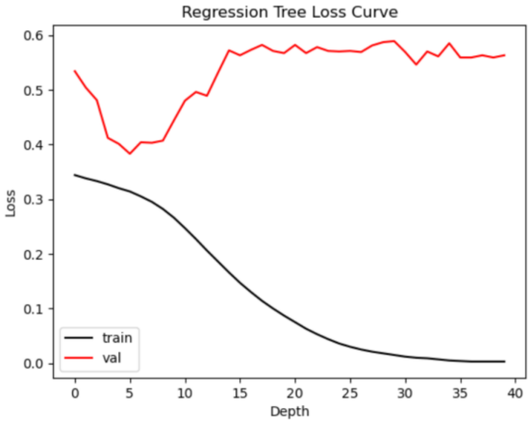
- Linear regression with single input feature (Cumulative Sum Total $CO_2$ Emissions)
- Established baseline loss for comparison on future models
- MAE = 0.743ºC

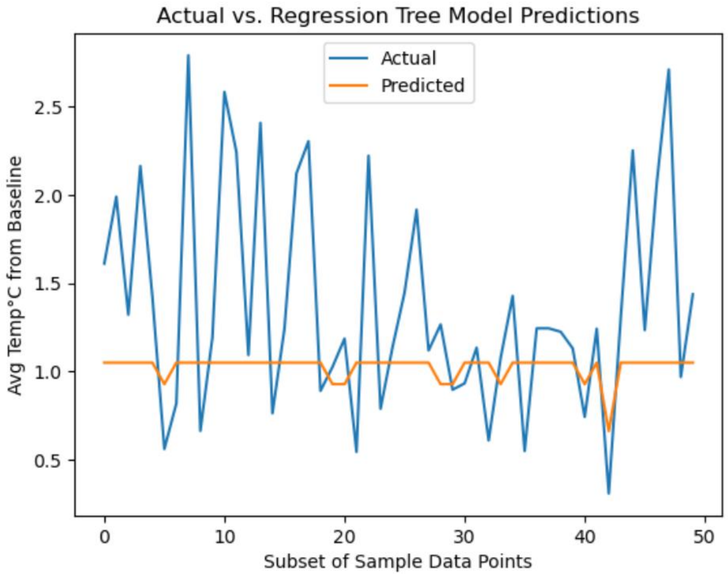



Berkeley
UNIVERSITY OF CALIFORNIA

# Regression Tree

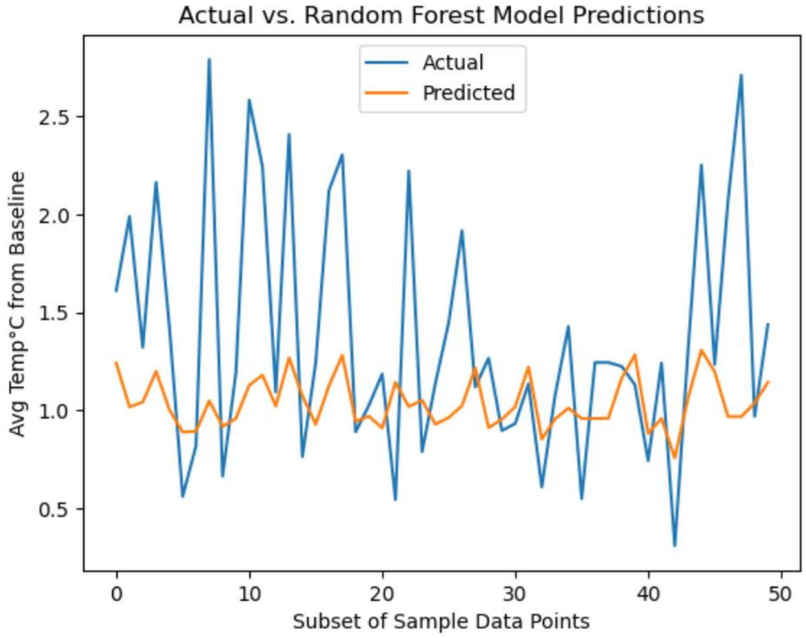
- Potential explainability benefit
- Moderate improvement over baseline model → MAE = 0.519ºC
- Optimal Hyperparameter: Max Depth = 5



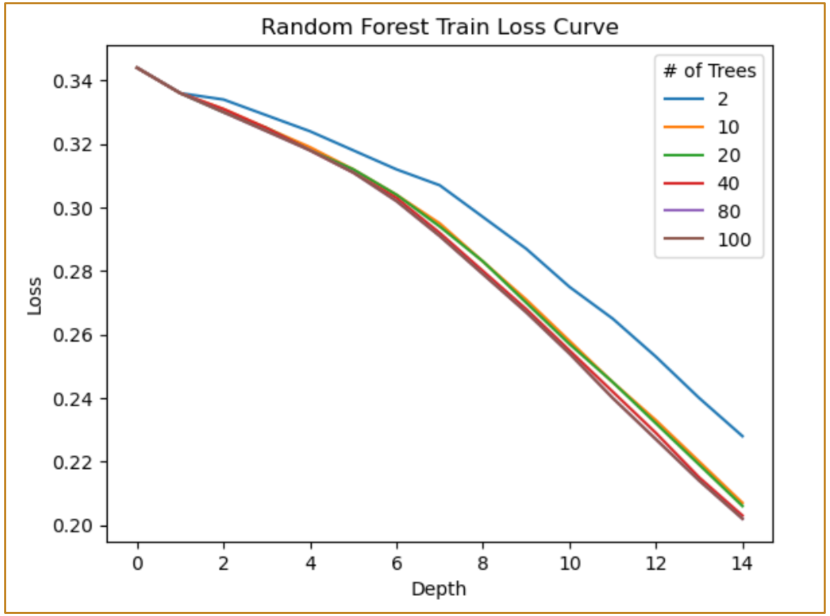
# Random Forest

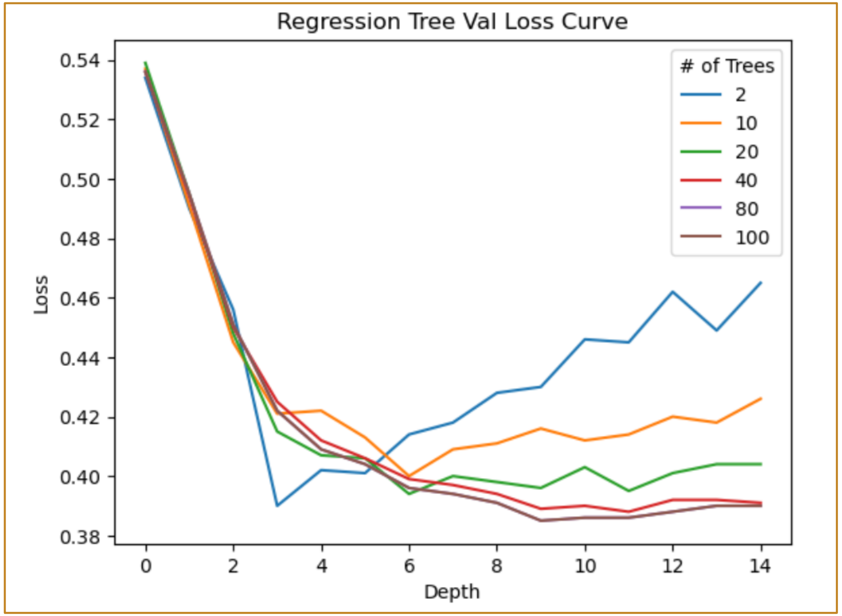
- Ensemble learning with bootstrapping
- Slight improvement over regression tree, but still not a reliable predictor
  - MAE = 0.521ºC
- Optimal Hyperparameters:
  - Max Depth = 9
  - Num Estimators = 80




Berkeley
UNIVERSITY OF CALIFORNIA

# RF Hyperparameter Tuning

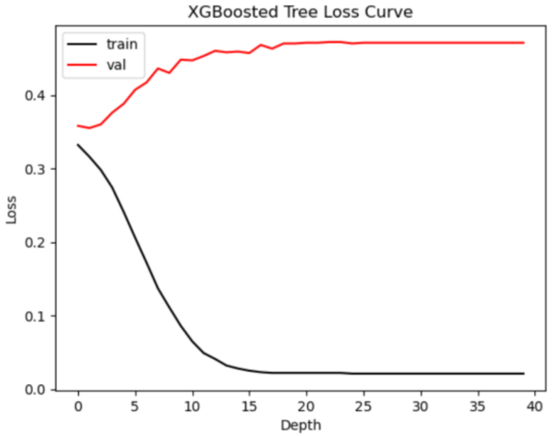


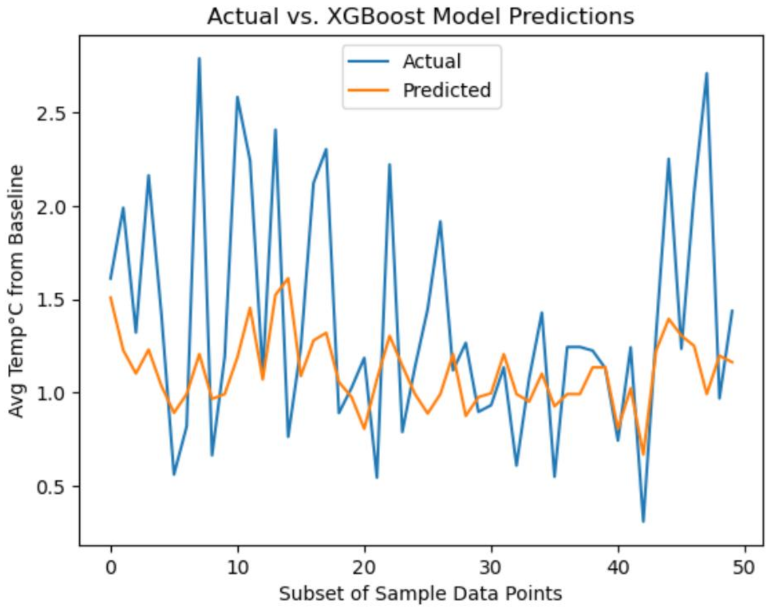




Berkeley
UNIVERSITY OF CALIFORNIA

# XGBoost Tree

- Gradient boosting ensemble learning
- Improvement over random forest, but still not a reliable predictor: MAE = 0.483ºC
- Optimal Hyperparameters: Max Depth = 3

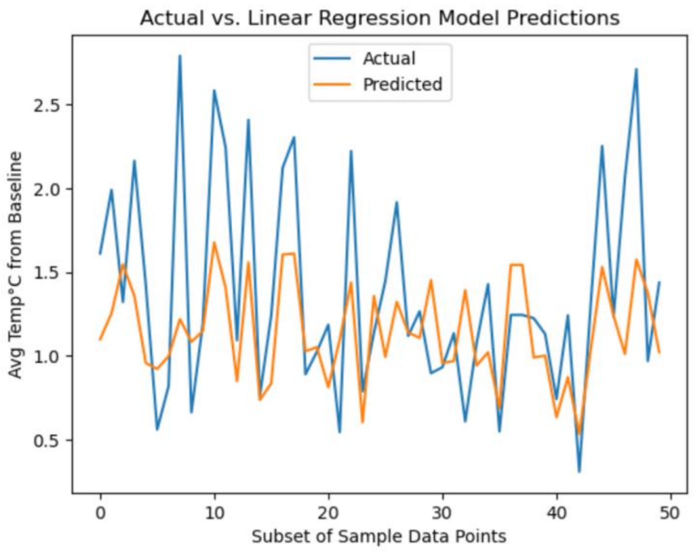


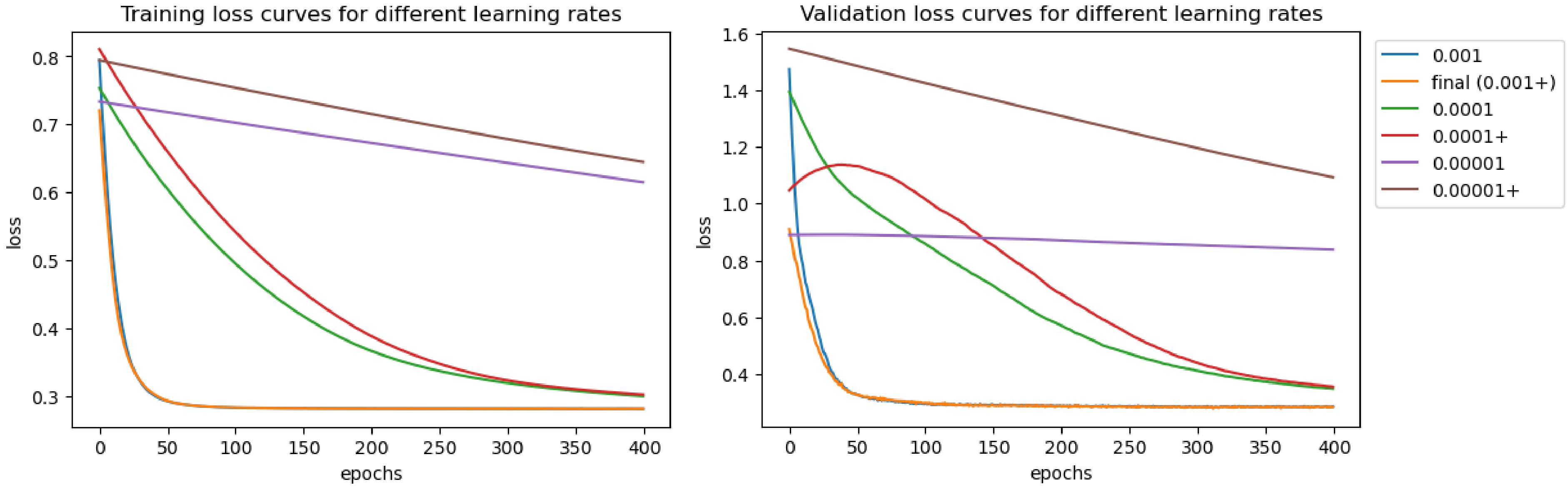

Berkeley
UNIVERSITY OF CALIFORNIA

# Linear Regression

- Linear regression with optimized feature selection
- Significant improvement over baseline model and moderate improvement over tree variants
- MAE = 0.382ºC
- Optimal Hyperparameters:
  - Initial LR: 1e-3
  - LR Schedule: Exponential Decay
  - Epochs: 150
  - Batch Size: 400



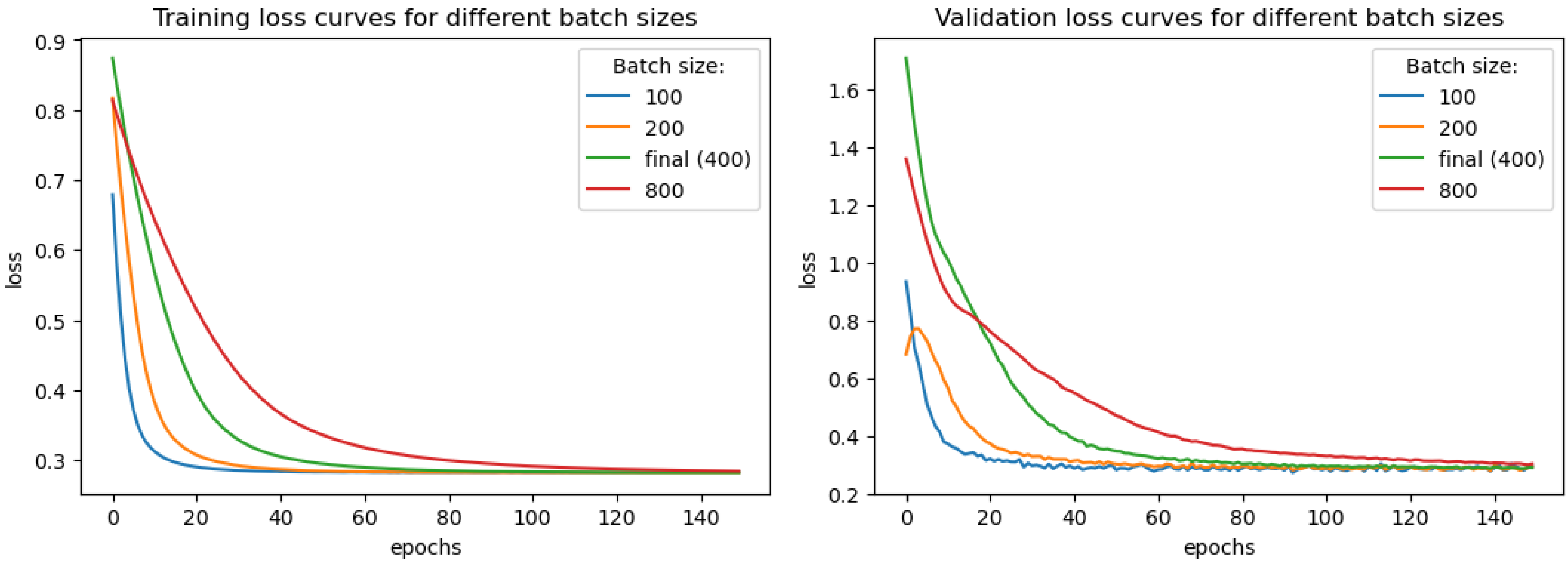
# Linear Regression Hyperparameter Tuning




“+” indicates learning rate with exponential decay

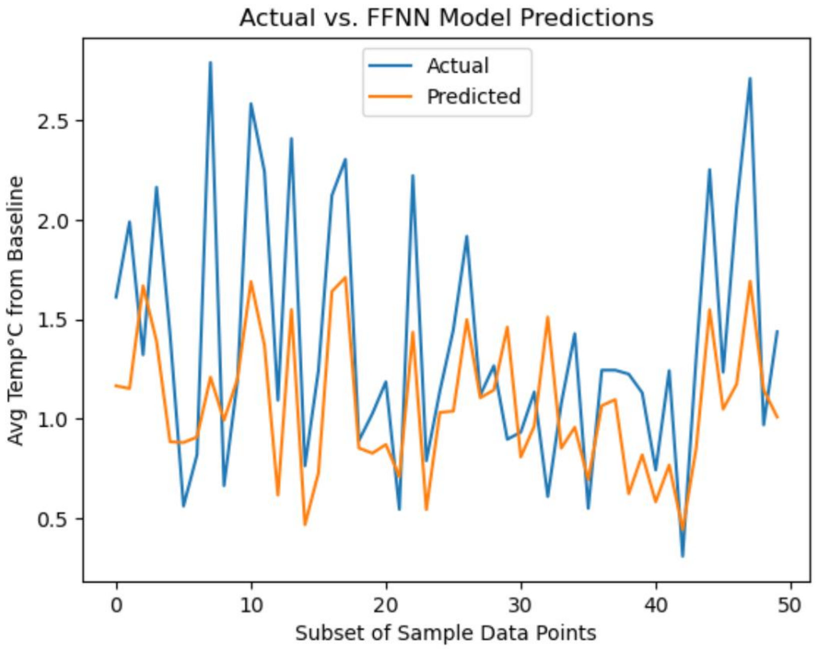
Berkeley
UNIVERSITY OF CALIFORNIA

# Linear Regression Hyperparameter Tuning



# FFNN

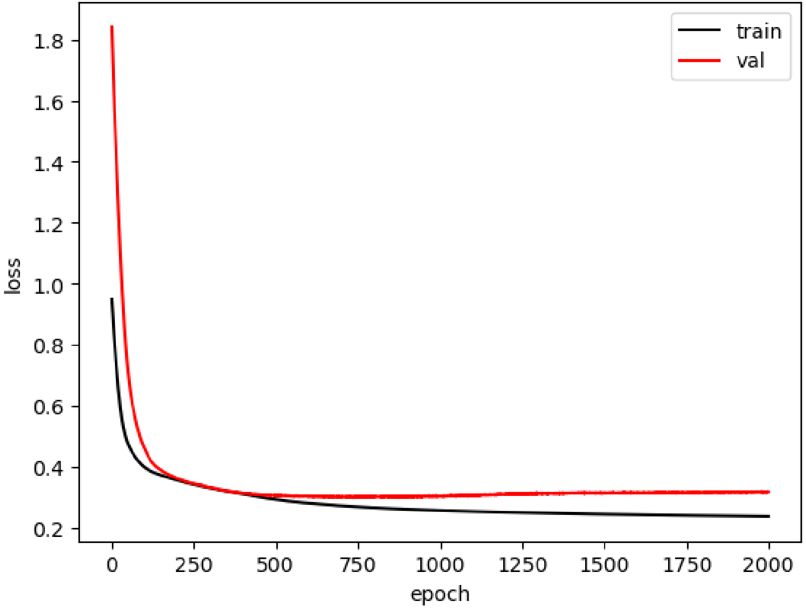
- Comparable to linear model, but much higher computational cost
- Results indicate few non-linear relationships
    - MAE = 0.411ºC
- Optimal Hyperparameters:
    - 2 hidden layers (128 units each)
    - Initial LR: 1e-5
    - LR Schedule: Exponential Decay
    - Epochs: 2000
    - Batch Size: 500



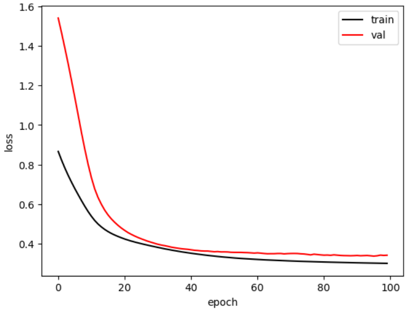

Berkeley
UNIVERSITY OF CALIFORNIA

# FFNN Hyperparameter Tuning

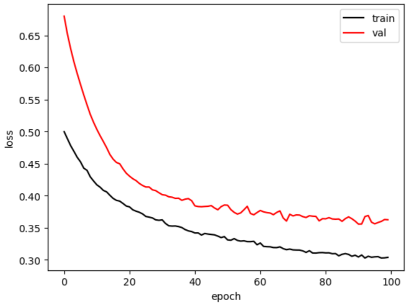
Final Model (2 layers, 128 units, no dropout):



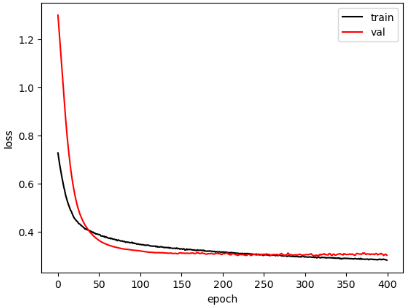

1 layer:



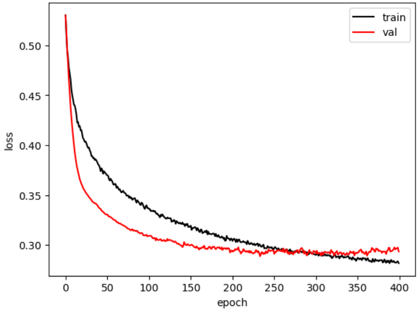

2 layers, 50 units:




3 layers:



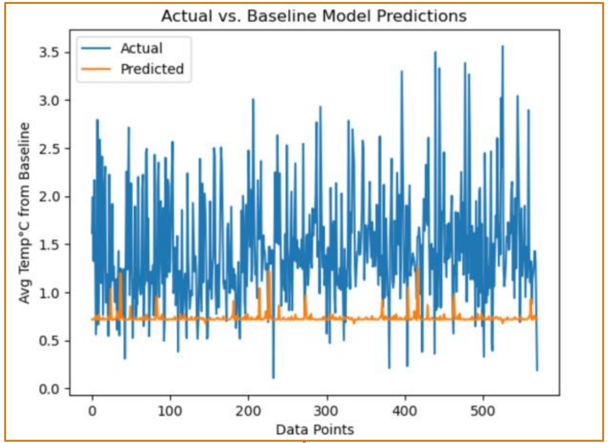

2 layers, dropout layer:

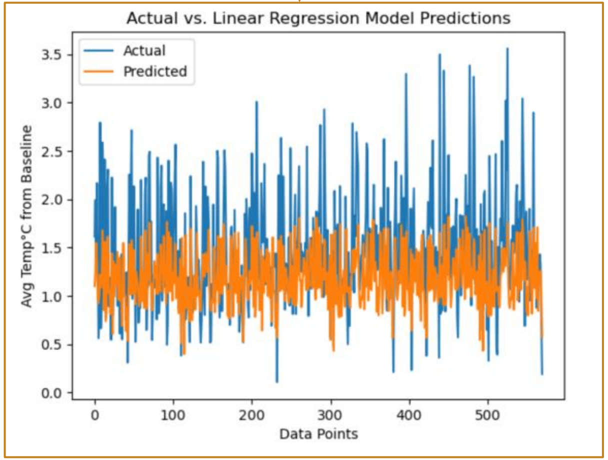



Berkeley
UNIVERSITY OF CALIFORNIA

# Conclusion & Future Considerations

- Best Performance: Linear regression model
- 51% reduction in loss compared to baseline
- For future: More modeling with LSTM for better recognition of patterns in temperature changes over extended periods
- Overall: unable to accurately predict recent temperature extremes from agricultural emissions alone




Berkeley
UNIVERSITY OF CALIFORNIA

# Questions?

Berkeley
UNIVERSITY OF CALIFORNIA

# Contributions

| | Domain Research | Pre-processing/ Feat. Eng | EDA | Decision Tree & Variants | Linear Regression | FFNN | Slides |
|---|---|---|---|---|---|---|---|
| Rachel | X | | X | X | | | X |
| Darya | X | | | | X | X | X |
| Julia | | X | | | X | X | X |
| Faye | | X | X | X | X | X | X |

# Github Repository

https://github.com/rachtripoli/DATASCI207_finalproject_Likhareva_Titchenal_Tripoli_Zhao/tree/main

# Data Sources/References


1. https://www.kaggle.com/datasets/alessandrolobello/agri-food-co2-emission-dataset-forecasting-ml/data
2. https://psl.noaa.gov/enso/mei/
3. https://ourworldindata.org/greenhouse-gas-emissions-food#:~:text=The%20specific%20number%20that%20answers,we%20include%20all%20agricultural%20products.
4. https://www.ipcc.ch/2021/08/09/ar6-wg1-20210809-pr/
5. https://www.gfdl.noaa.gov/news/noaa-scientists-harness-machine-learning-to-advance-climate-models/