Cornell University

Predicting power plant emissions using public data and machine learning

K. Max Zhang, Jiajun Gu, and Jeffery Sward

Energy and the Environment Research Laboratory

Cornell University

kz33@cornell.edu

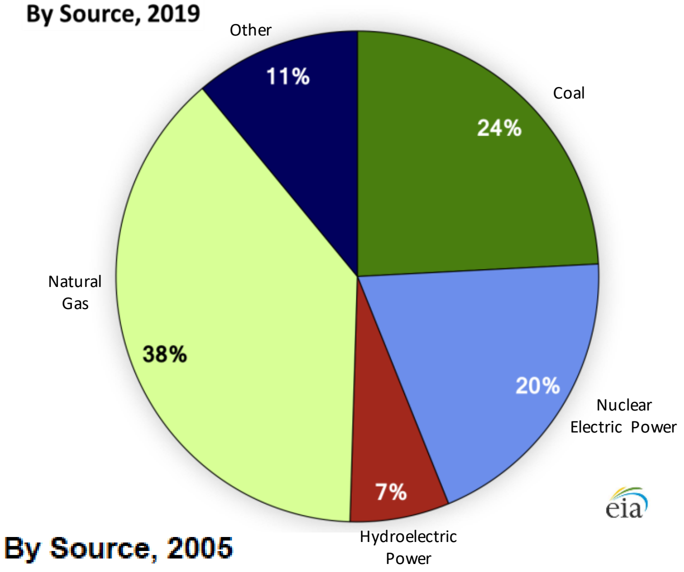
Acknowledgements

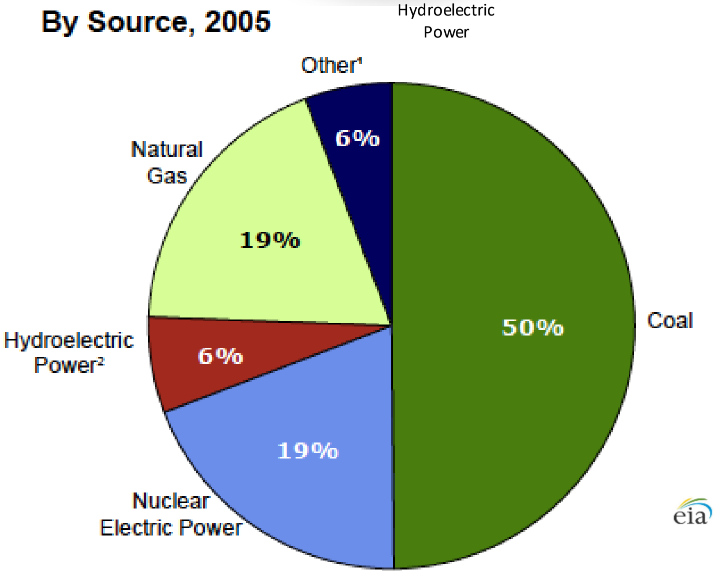
- Winner of the EmPower Air Data Challenge sponsored by USEPA Clean Air Markets Division (CAMD)
- Valuable discussions with:
  - USEPA colleagues including Charles Frushour, Justine Huetteman, Michael Hesse, Travis Johnson, Jeremy Schreifels and Christopher Worley.
  - NYSERDA Project Advisory Committee
  - Ona Papageorgiou and Michael Sheehan at New York State Department of Environmental Conservation (NYSDEC)
  - Ben Cohen at New York Independent System Operator (NYISO)
- Funding support from New York State Energy Research and Development Authority (NYSERDA)

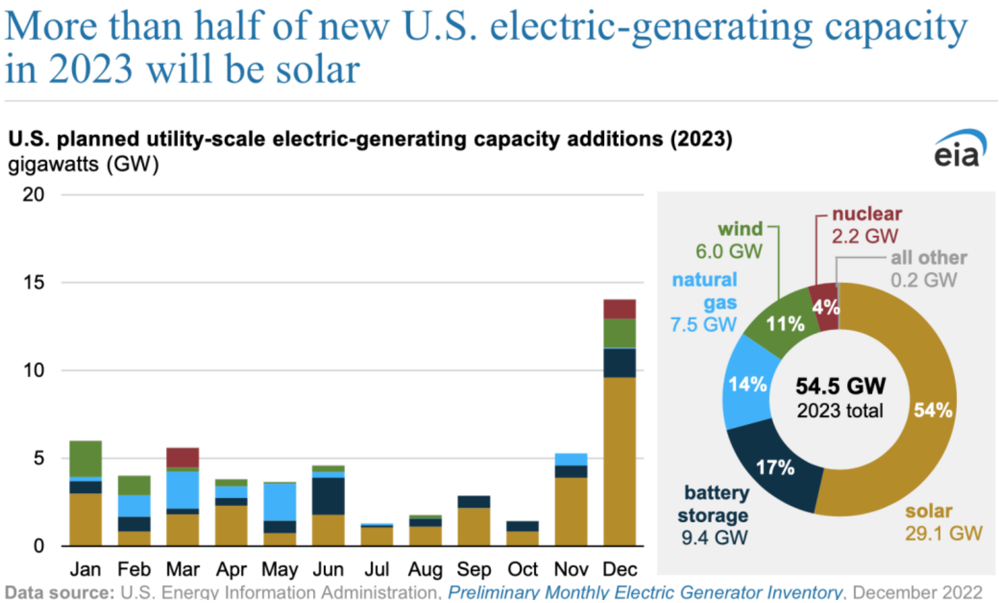
NEW YORK STATE OF OPPORTUNITY.

NYSERDA Supported

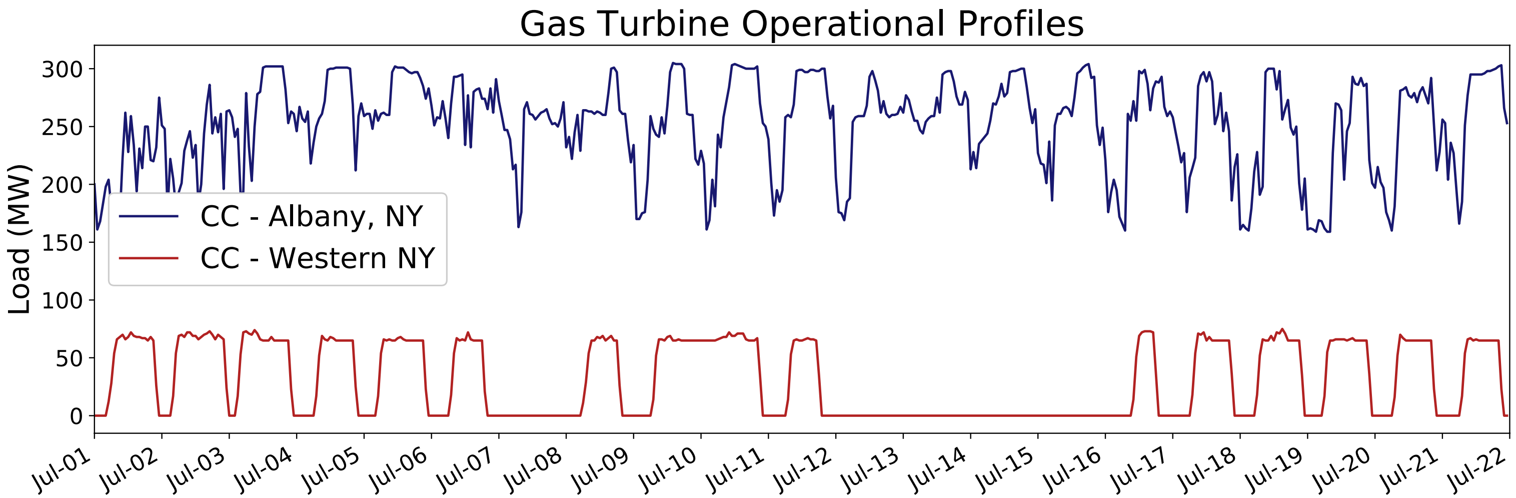
Context







- The U.S. power system is experiencing significant changes.
- 2022: Electricity generated from renewables surpassed coal in the U.S.



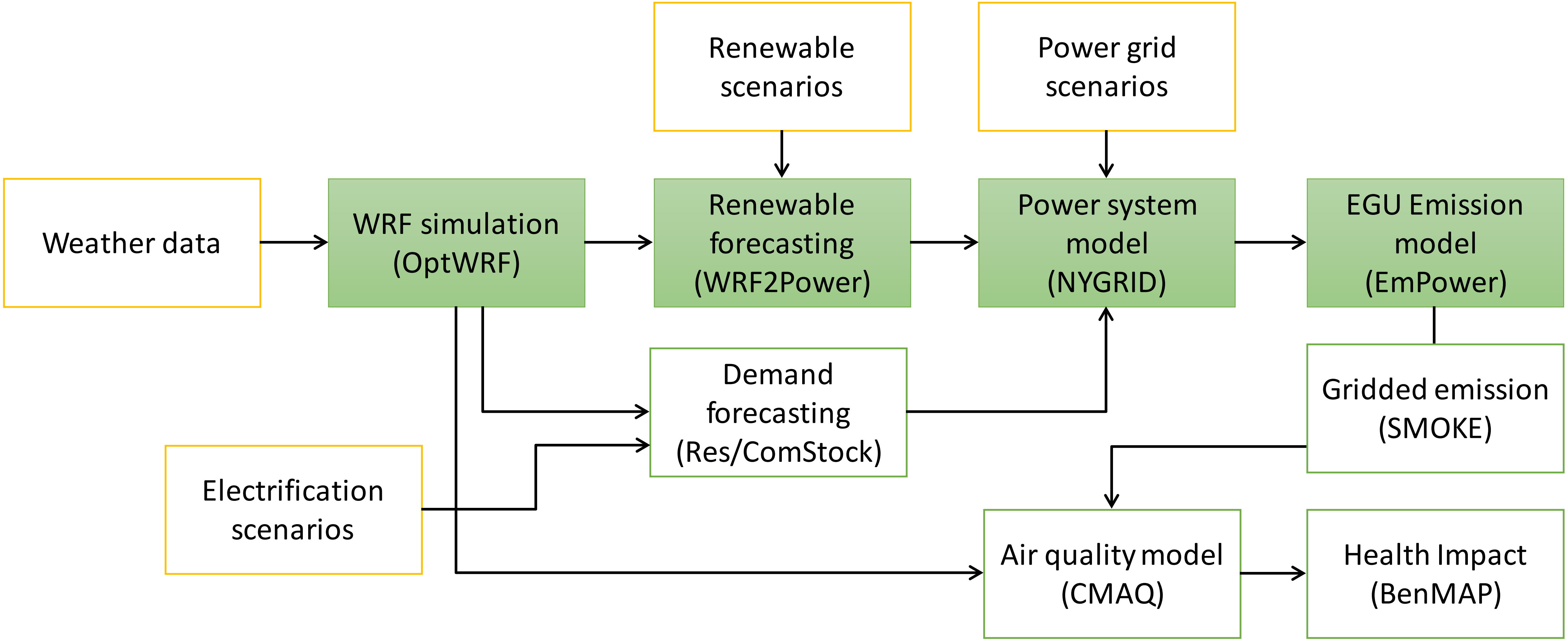
Motivation: Challenges

- Fossil fuels plants experience more and more ramping and do not behave like baseload units.
- Great challenge in estimating the future NOx emissions for air quality planning purposes.
- Continuous Emission Monitoring Systems (CEMS) provide a rich dataset of hourly emissions (NOx, SO_2 and CO_2) and associated characteristics for EGUs larger than 25 MW.
- However, previous efforts to predict EGU emissions from CEMS data using simple regression methods (linear, piecewise linear, etc.) showed mixed results.

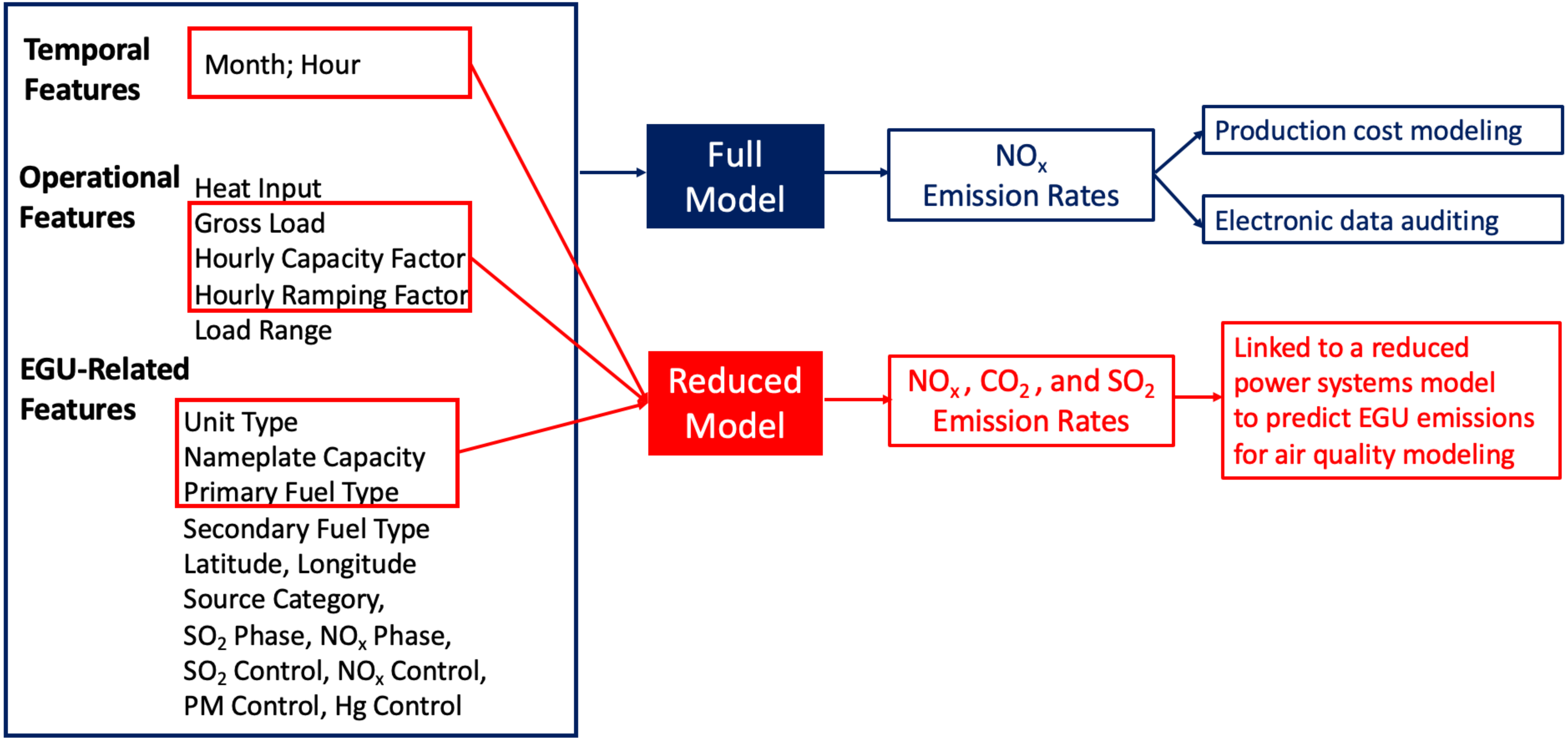
Potential Benefits and Strategies

- There are many benefits from accurately predicting EGU emissions using public datasets
  - Air quality planning
  - Electronically audit CEMS data, identify data anomalies, and enhance data quality.
  - Electric production cost modeling
    - Predicting EGU emissions using data in the public domain is particularly valuable because it makes broader stakeholder engagement possible by avoiding proprietary data internal to power system operators.
- Strategies
  - Employ other public datasets in addition to CEMS data
  - Apply non-linear models (e.g., machine learning techniques) as alternatives to linear models

Integrated modeling framework



Full Model vs Reduced Model

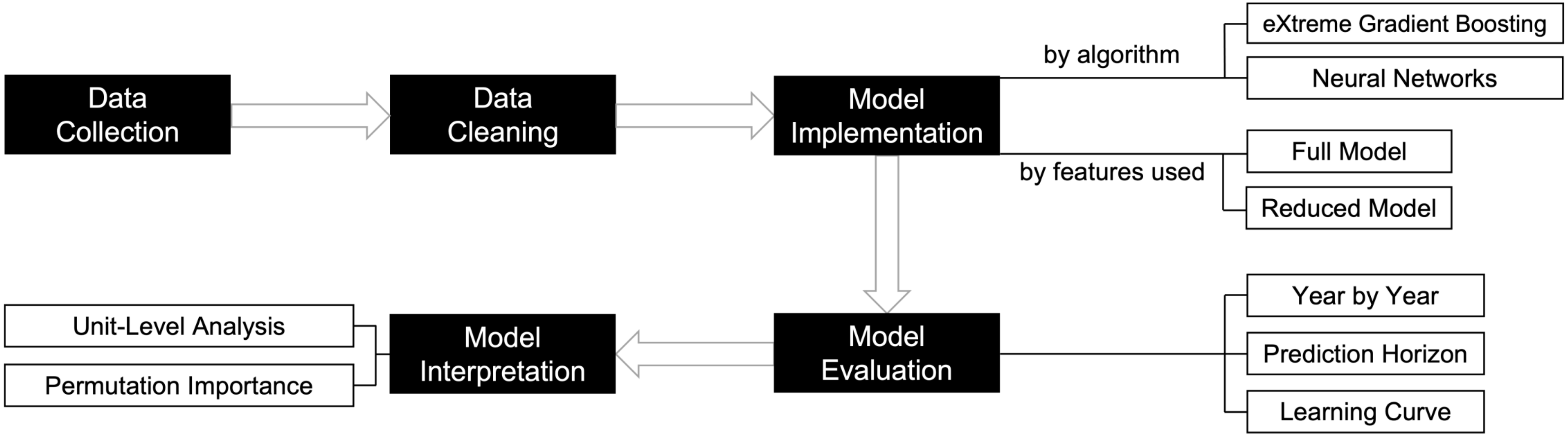


Gu et al., manuscript in revision

Modeling Method

Studied EGUs: All thermal units in New York State (2015-2019)

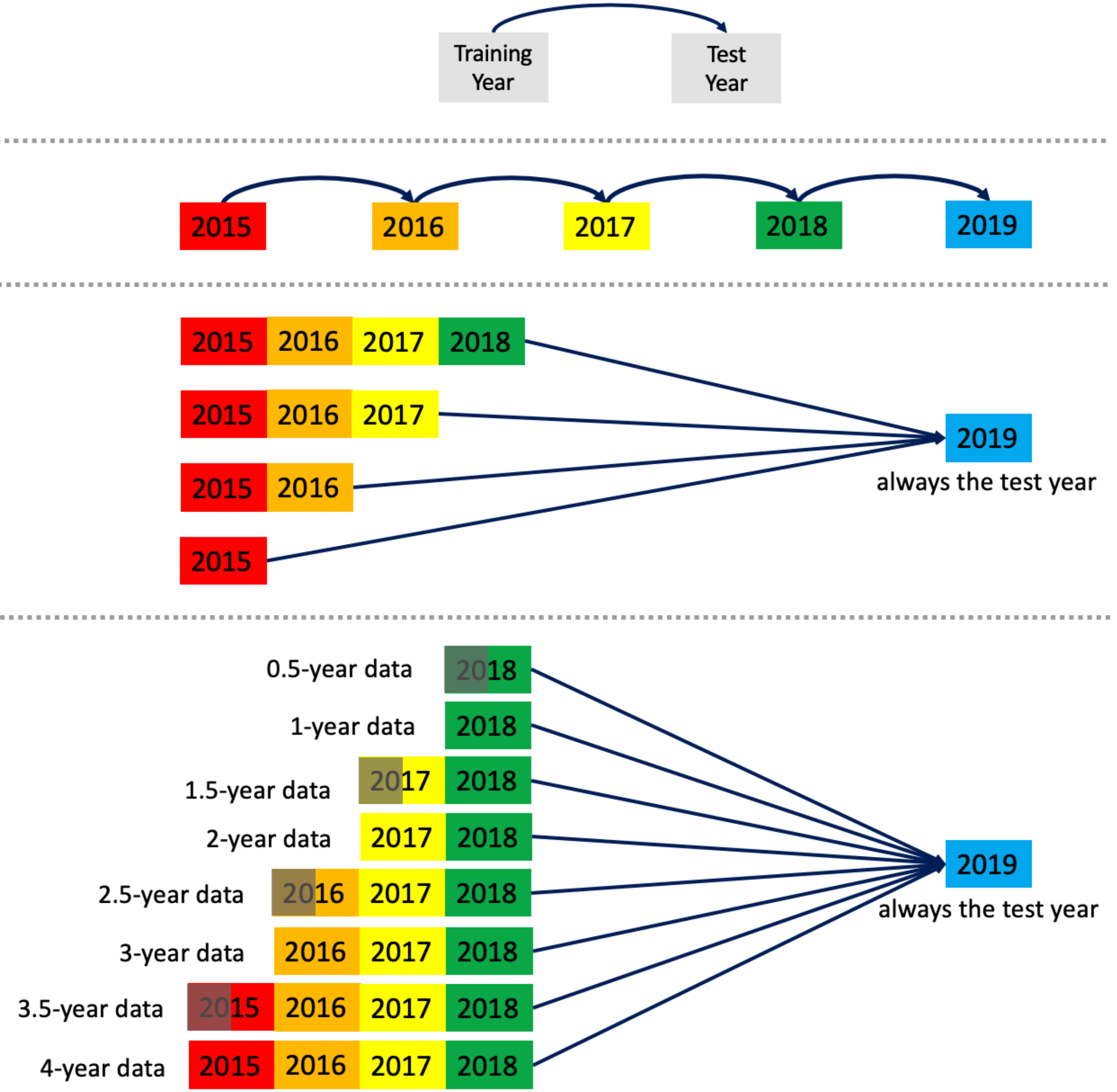
Unit Type	Number of Units	Number of Data Points	Mean	Standard Deviation	Percentile				
					25th	50th	75th	99th	100th
Combined Cycle	56	214,580	19.8	20.3	7.7	12.5	23.9	89.1	621.8
Combustion Turbine	29	30,099	13.0	19.2	3.6	4.1	26.1	43.9	302.0
Tangentially-Fired	21	46,895	129.3	188.1	33.2	81.5	151.6	1116.5	2153.7
Dry Bottom Wall-Fired Boiler	6	3,125	139.0	274.0	18.1	21.8	75.1	1296.3	2350.6



Gu et al., manuscript in revision

Model Evaluations

- Year by Year
- With Different Prediction Horizons
- With Different Amounts of Training Data (Learning Curve)



Gu et al., manuscript in revision

Overall performance for NOx emission rates

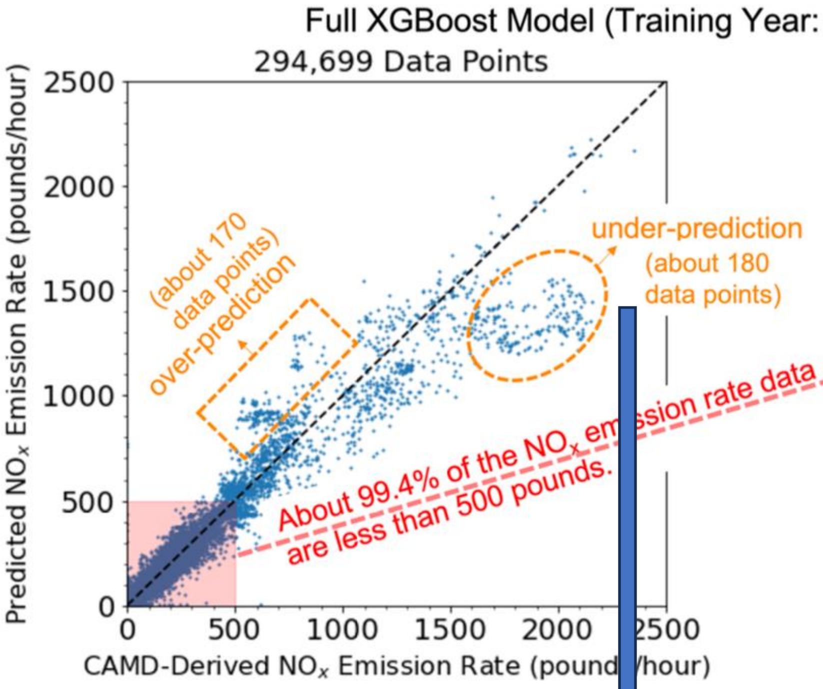
Table 2 The LR, XGBoost, and NN predictive performance in terms of R^2, RMSE (pounds/hour), and nRMSE of full models on NO_x emission rates (trained on the previous year's data and tested on the following year's data, year-by-year from 2015 to 2019).

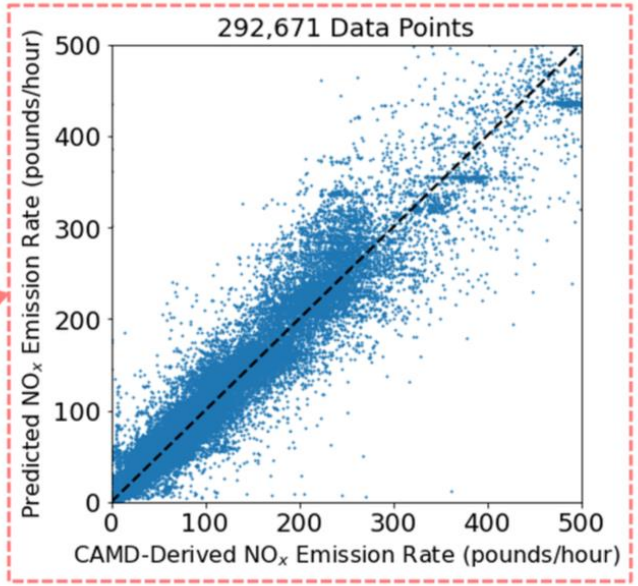
Training Year	Test Year	Full Model								
		LR			XGBoost			NN		
		R^2	RMSE	nRMSE	R^2	RMSE	nRMSE	R^2	RMSE	nRMSE
2015	2016	0.91	26.2	0.011	0.96	17.7	0.007	0.96	18.5	0.008
2016	2017	0.89	26.3	0.012	0.96	16.0	0.007	0.96	16.3	0.007
2017	2018	0.90	29.4	0.012	0.95	21.0	0.009	0.95	19.6	0.008
2018	2019	0.82	23.9	0.011	0.96	11.0	0.005	0.96	11.8	0.005

Table 3 The LR, XGBoost and NN predictive performance in terms of R^2, RMSE (pounds/hour), and nRMSE of reduced models on NO_x emission rates (trained on the previous-year data and tested on the following-year data, year-by-year from 2015 to 2019).

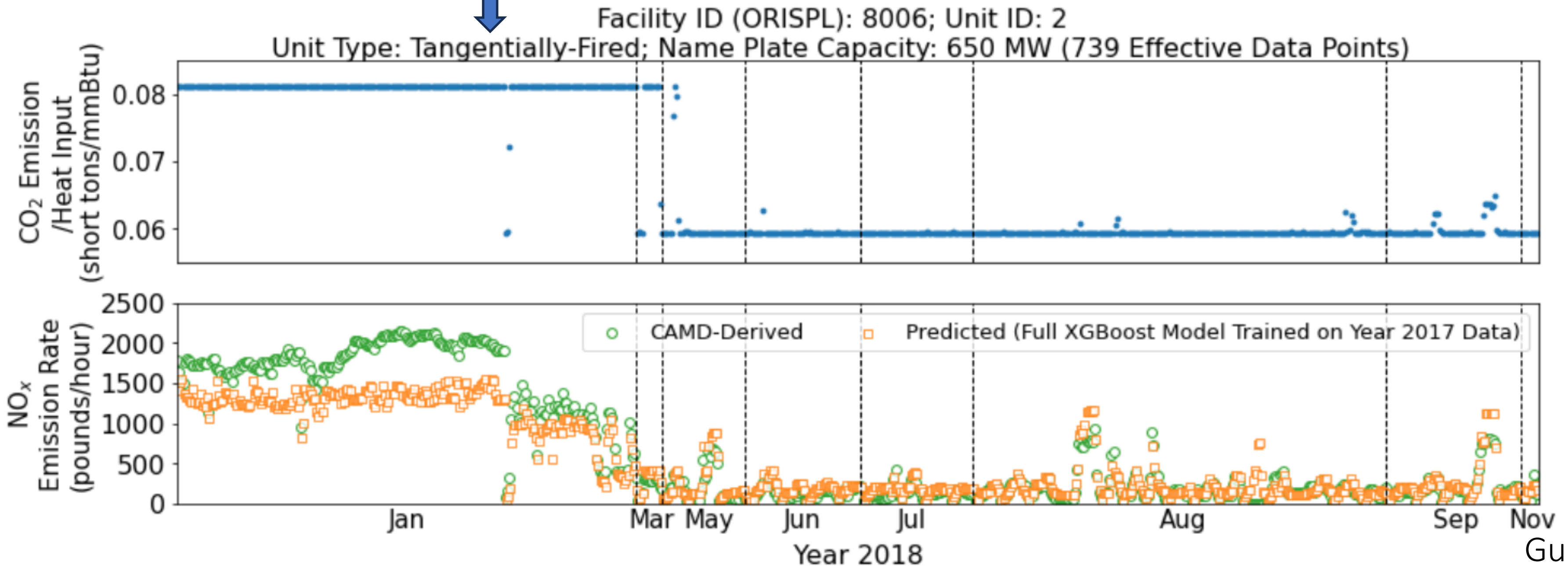
Training Year	Test Year	Reduced Model								
		LR			XGBoost			NN		
		R^2	RMSE	nRMSE	R^2	RMSE	nRMSE	R^2	RMSE	nRMSE
2015	2016	0.54	59.7	0.025	0.93	22.9	0.010	0.90	27.4	0.011
2016	2017	0.39	62.9	0.028	0.93	21.8	0.010	0.90	25.2	0.011
2017	2018	0.38	72.7	0.031	0.86	34.2	0.015	0.86	34.2	0.015
2018	2019	0.29	47.5	0.021	0.91	16.4	0.007	0.90	17.8	0.008

Gu et al., manuscript in revision

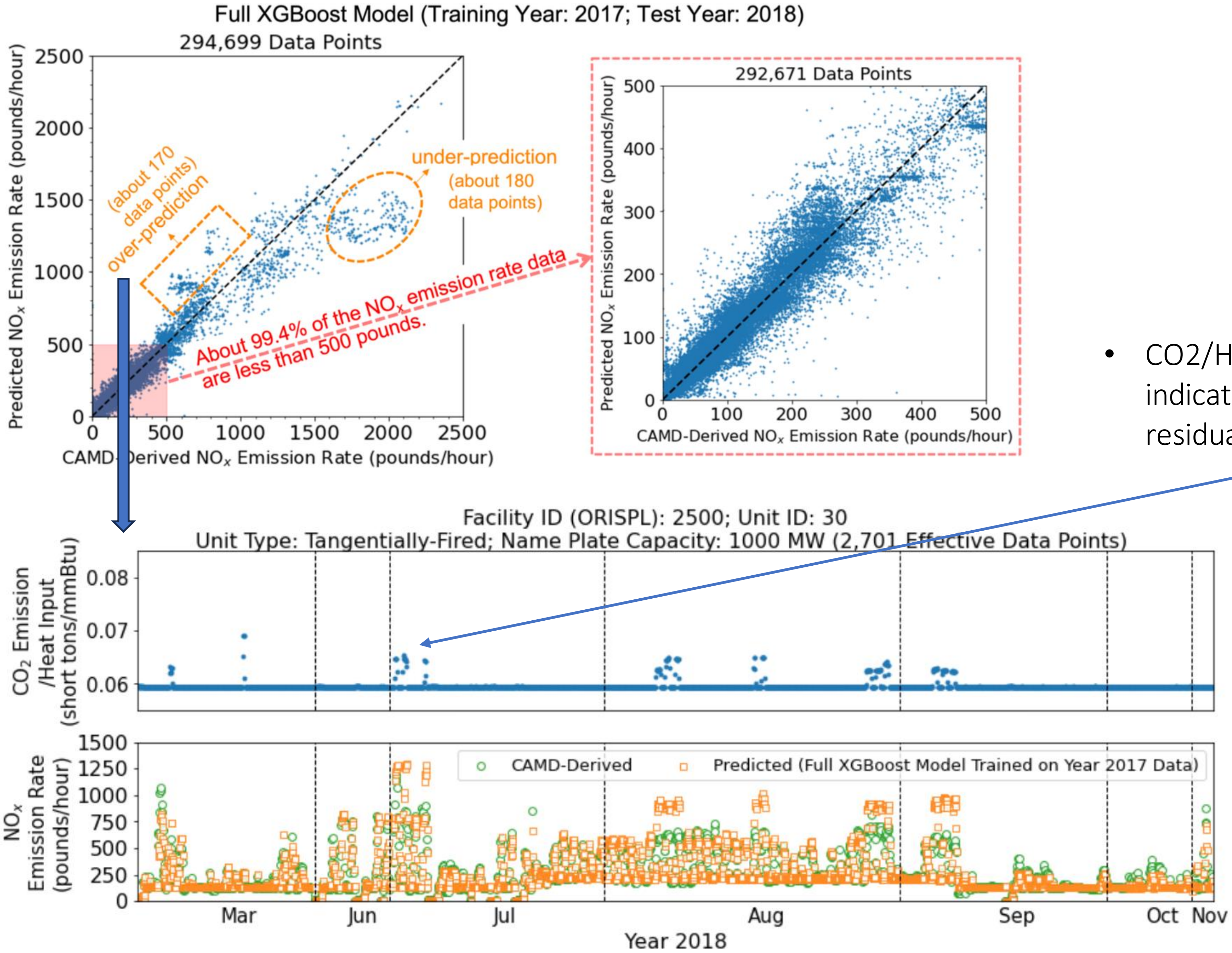




- Northeast Cold Snap: Dec 26 through until Jan 7 (13 days)
- Winter Peak at 25,081MW on Jan 5, 2018, very close to the all-time winter peak.
- CO2/Heat Input ratios indicating fuel switching to residual oil in Jan 2018.



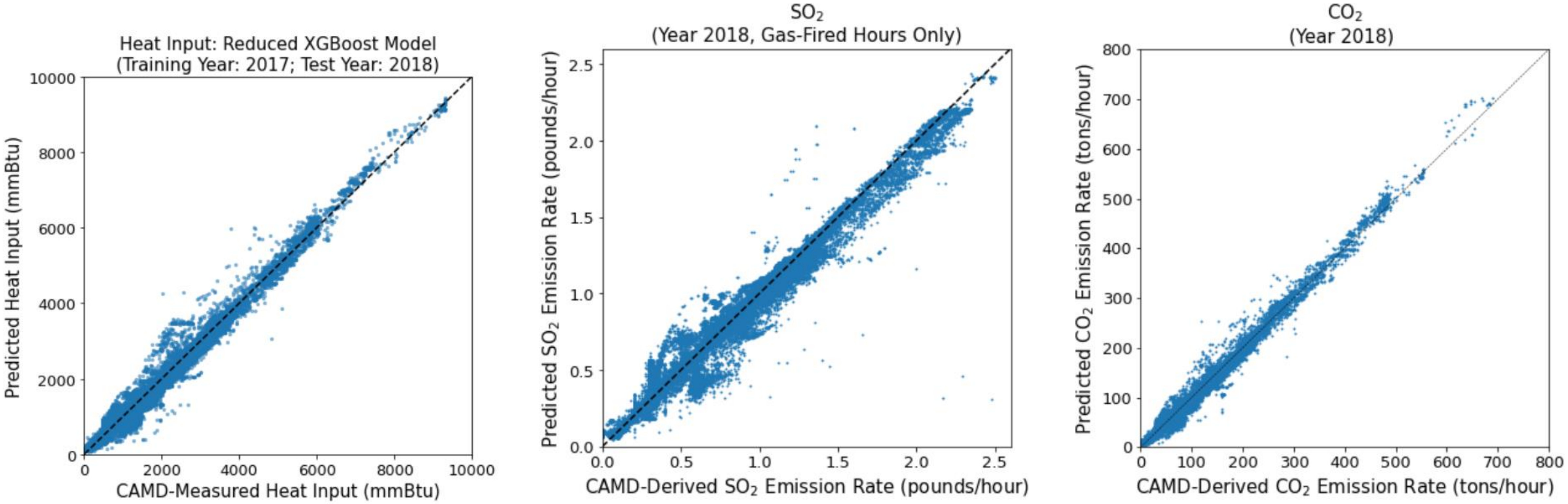
Gu et al., manuscript in revision



- CO2/Heat Input ratios indicating the supplement of residual oil in this period.

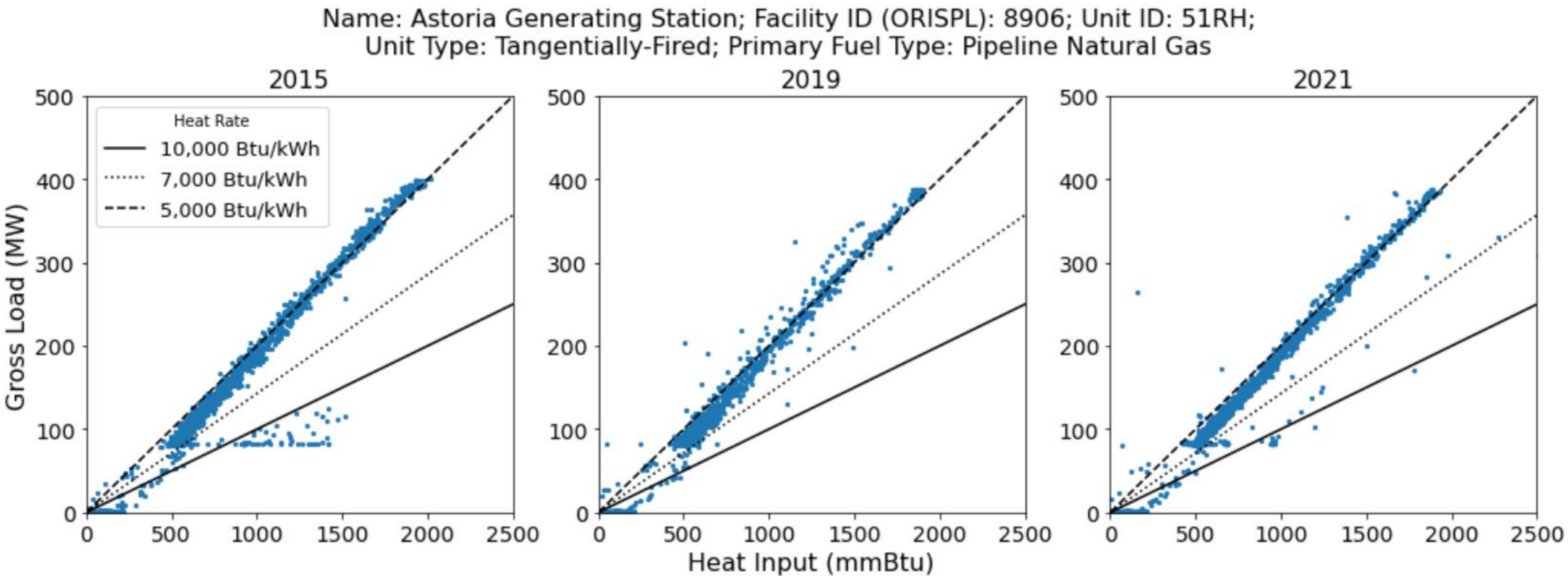
Gu et al., manuscript in revision

Results on Heat Input, SO_2, and CO_2



Gu et al., manuscript in revision

Data anomaly



- Title V permit from NYSDEC shows that this unit is a twin-furnace boiler that exhausts emissions through two stacks, counted as two units (Unit 51RH and Unit 52SH).
- Dividing the gross load for the full boiler by the heat input for each individual furnace would result in unrealistically low heat rates.
- Implication: Stricter enforcement of the EGU data reporting procedure

Gu et al., manuscript in revision

Remarks

- Non-linear models such as XGBoost and NN were shown to outperform the Linear Regression (LR) model consistently and significantly
  - Especially in reduced models with a limited number of features available.
- We found the EPA Field Audit Checklist Tool (FACT) to be very useful to supplement CEMS data.
- We recommend:
  - Stricter enforcement of the EGU data reporting procedure, providing emission control operational information,
  - Obtaining EGU-related data from multiple sources in the public domain
- Overall, using multiple public datasets and machine learning techniques can reliably predict EGU emissions.