Singularity

Methodological Documentation for the "Generator REal-Time emissions Assignment" (GRETA) module

Singularity Energy, Inc. December, 2024

1. GRETA Overview

The Generator REal-Time emissions Assignment (GRETA) module is a patent-pending emissions calculation engine that estimates the direct (scope 1) stack emissions of GHGs from an electric generator in real time, based only on data about the net generation output (MW) of a unit. This module meets a need for more accurate accounting of power sector emissions on a (sub-)hourly basis.

The status-quo approach of estimating a generator's emissions using a single annual-average emission factor has two critical limitations when it comes to estimating hourly, rather than annual total emissions:

1. For generators that fuel switch, annual factors will misrepresent the specific emission rate in each hour depending on the fuel being burned.
2. Generator heat rates can change hour to hour, varying based on specific factors

GRETA uses generator-specific, modeled heat rate curves derived from historical operating data from Singularity's Open Grid Emissions (OGE) dataset. This approach represents the most accurate way to estimate hourly emissions in near real time when compared to hourly CO_2 emissions measurements recorded at each plant by continuous emissions monitoring systems (CEMS) and reported to the U.S. Environmental Protection Agency's (EPA) Clean Air Markets Program Division (CAMPD).

Modeled heat rate curves are more accurate at estimating hourly emissions output than annual emissions factors (which assume a static heat rate) because the heat rate of a generator varies based on multiple factors:

- **Operating level / capacity factor**: Generally, generators are more efficient at their full design load than at partial load
- **Operating state**: generators are less efficient during startup than while they are operating, and combined cycle units operating in different gas-to-steam turbine combinations (e.g. 2x1, 2x2, etc) may have different heat rates at the same output level.
- **Ambient air temperature:** generators are generally more efficient when ambient air temperatures are colder, since the air is denser and allows for more complete fuel combustion
- **Season**: Besides temperature differences across seasons, some plants are also subject to stricter emissions control limits during "ozone season" (May-September), and thus may turn on additional emissions control equipment, which reduces generator efficiency.
- **Fuel switching**: Generators that fuel switch may have different heat rates at the same output level for each fuel that they burn.

GRETA consists of two modules: an offline modeling module and an on-demand API module. The modeling module is run infrequently to train the models that the API will use to estimate emissions. The modeling module exclusively uses historical, public datasets for training. GRETA consists of three primary models: 1) a fuel identification/fuel switching model which estimates the fuel type being consumed in each hour, 2) a heat rate curve (HRC) model which translates MW output to fuel MMbtu input, and 3) a CHP model that estimates, for a given month of the year, which proportion of fuel input at CHP plants was used for electric generation. In contrast, the API uses these models to estimate the emissions associated with near-real time generation data that is passed to it on-demand. See Figure 1 for a detailed overview of the entire data pipeline.

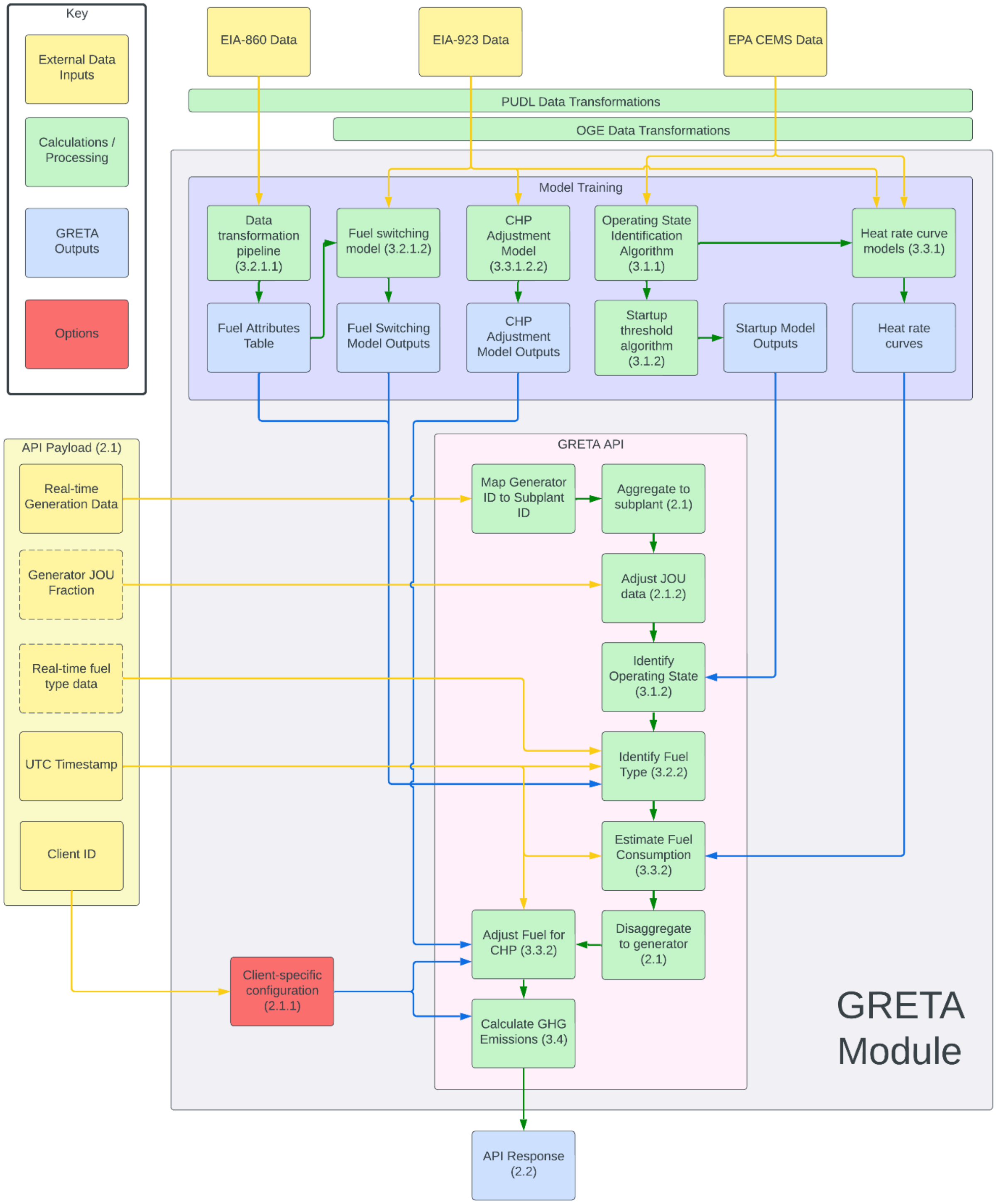

Figure 1: *Overview of the GRETA module (in the grey box). Numbers in parentheses indicate section numbers in this document relevant to each process.*

2. GRETA API Inputs and outputs

2.1 API Inputs

The GRETA API accepts data for a single "generation event" at a time. A generation event consists of generation data for one or more generators in a single operating interval (whether 5 minutes or an hour). Specifically, the GRETA API requires three inputs: the EIA plant ID of the generator, the EIA generator ID of the generator, and the net MW of generation in the interval. The generator ID is not strictly required, but if not provided GRETA will backstop to using a less-precise plant-specific model for estimating emissions.

In addition to these required inputs, GRETA will accept several optional inputs:

- A manually-specified energy source code. GRETA will estimate the energy source code in each interval, so this should only be used in the case of fuel switching generator when the fuel type being burned in each interval is known.
- An adjustment factor for jointly-owned units (JOUs). See section 2.1.2 for details on how this is used.

2.1.1 Client-specific configuration

For each client that uses the GRETA API, several additional one-time configuration options are specified:

- The length, in minutes, of each generation event passed to GRETA. This is used to convert MW inputs into MWh.
- Whether the generation data passed to GRETA represents net generation or gross generation. This is used to adjust the heat rate curves (which use a net generation convention)
- Whether data outputs should be adjusted for combined heat and power (CHP) plants. Generally, if GRETA is used for assessing emissions from electricity generation, this would be set as "True", while if the goal is to estimate total plant emissions (including those used to generate heat), this would be set to "False"
- The IPCC Assessment Report (e.g. AR5, AR6) and horizon (e.g. 20-year, 100-year) to use to determine the global warming potential (GWP) values used for converting CO_2, N_2O, and CH_4 emissions to CO_2-eq

The MISO CarbonFlow application passes 5-minute interval, net generation data to GRETA and uses the most recent (AR6) GWP values with a 100-year horizon for converting GHGs to CO_2-eq emissions.

2.1.2 Data Pre-Processing

Generator-level data is assigned a subplant ID by the GRETA API since plant ID and subplant ID are two of the primary keys for every model output table. If multiple generators belonging to the same subplant are passed in a single generation event, the generation data will be aggregated by subplant in the API since all models are designed to work with aggregated subplant data. The subplant-level results are re-allocated to each generator proportionally to generation prior to returning results.

By default, GRETA assumes that the generation data provided represents 100% of a generator's generation in the interval. However, if a unit is jointly owned, and the net generation MW for only a certain fraction of the unit's generation is passed, this value will be used to adjust the generation value before estimating fuel consumption and emissions, since heat rate curves depend on a generator's total output. For example, if a generation fraction of 0.25 is passed, and the input generation is 10MW, then GRETA will scale the generation to 40MW (10MW / 0.25).

2.2 API Outputs

GRETA returns the following outputs:

- The estimated fuel type in the interval
- The total mmBTU of fuel consumption in that interval
- The total mass (in lb) of CO_2, CH_4, N_2O, CO_2-eq, and non-biogenic CO2 emissions
- The emissions rate (in lb/MWh) of CO_2, CH_4, N_2O, and CO_2-eq.

Non-biogenic CO_2 mass data is provided so that the biogenic and non-biogenic components of a generator's CO_2 emissions can be derived if needed for reporting in compliance with the GHG Protocol Scope 2 Guidance.

3. Detailed Methodology

3.1 Operating State

3.1.1 Modeling operating states

In each operating interval, the operating state of a generator (i.e., whether it is starting up, on, or shutting down) may affect its heat rate and/or fuel type (for generators that use a different fuel during startup). When training models, GRETA uses data about the operating time of each unit in each hour, as reported in CEMS data. For each hour, the CEMS dataset indicates which fraction of an hour each unit was operating. Since a subplant can consist of multiple individual units (see Figure 2), the operating times of all units in the subplant are considered.

This data is used to identify 6 different possible operating states:

- **off**: The minimum and maximum operating time of all units is 0
- **on**: The minimum and maximum operating time of all units is 1
- **partial-on**: the max operating time is 1.0 and the min operating time is less than 1.0. This may indicate different operational configurations for combined cycle units (e.g. 1x1, 2x1, 2x2)
- **startup**: a fractional operating hour where the previous hour's operating time is 0.0, or in a run of hours where the mean operating time is strictly increasing from 0 to 1. For single-unit subplants, this can also include fractional hours when the previous hour operating time was 0.0
- **shutdown**: a fractional operating hour where the next hour's operating time is 0.0, or in a run of hours where the mean operating time is strictly decreasing from 1 to 0. For single-unit subplants, this can also include fractional hours when the previous hour operating time was 1.0
- **cycling**: a run of fractional hours where the start and end operating time are both 0.0, or a single fractional hour between two on hours, or a fractional hour where the previous hour and next hour are both fractional. This indicates rapid cycling on and off at below the hourly resolution.

Figure 3 illustrates how these operating states have been to assigned to a specific subplant, and Figure 4 illustrates how the heat rate curve can vary by operating state.

The current heat rate curve (HRC) model only utilizes data about the on, partial_on, and startup states, and ignores cycling and shutdown data.

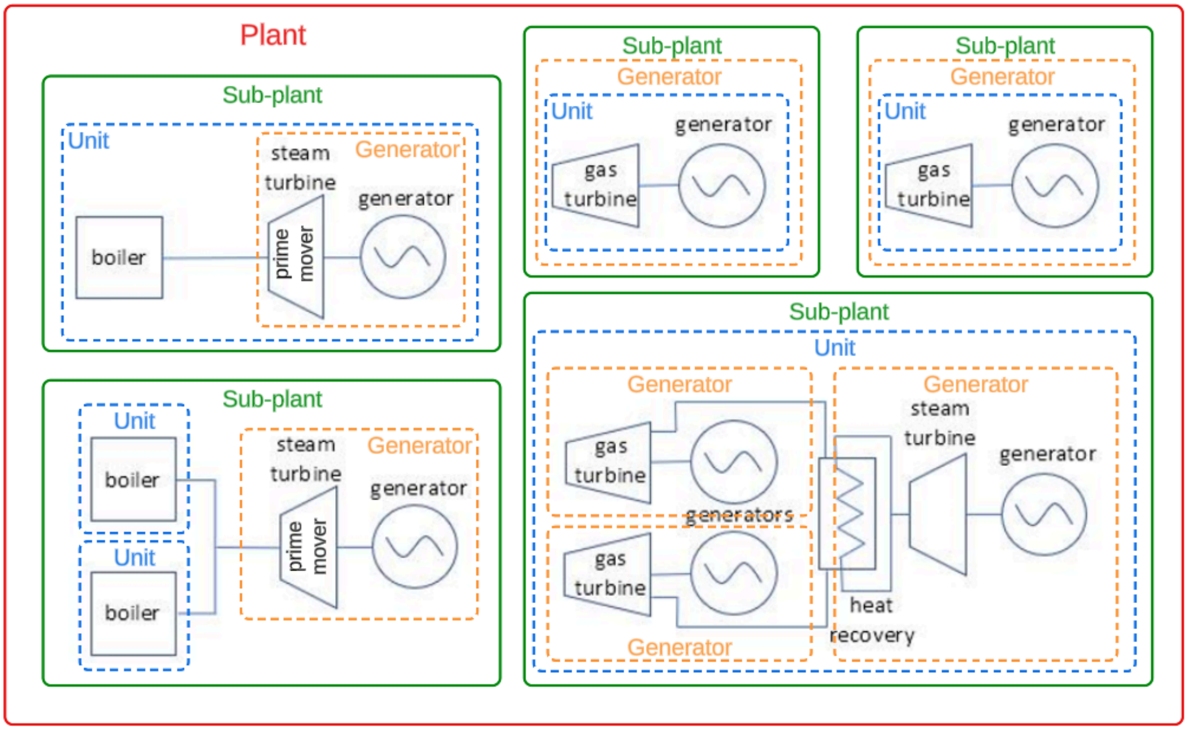

***Figure 2**: Illustration of several potential relationships between units, generators, boilers, prime movers, subplants, and plants for a large hypothetical power plant with 5 distinct subplants. For more information on these relationships see [this documentation](). Illustration uses images adapted from the [documentation]() for the EPA-EIA Power Sector Data Crosswalk.*

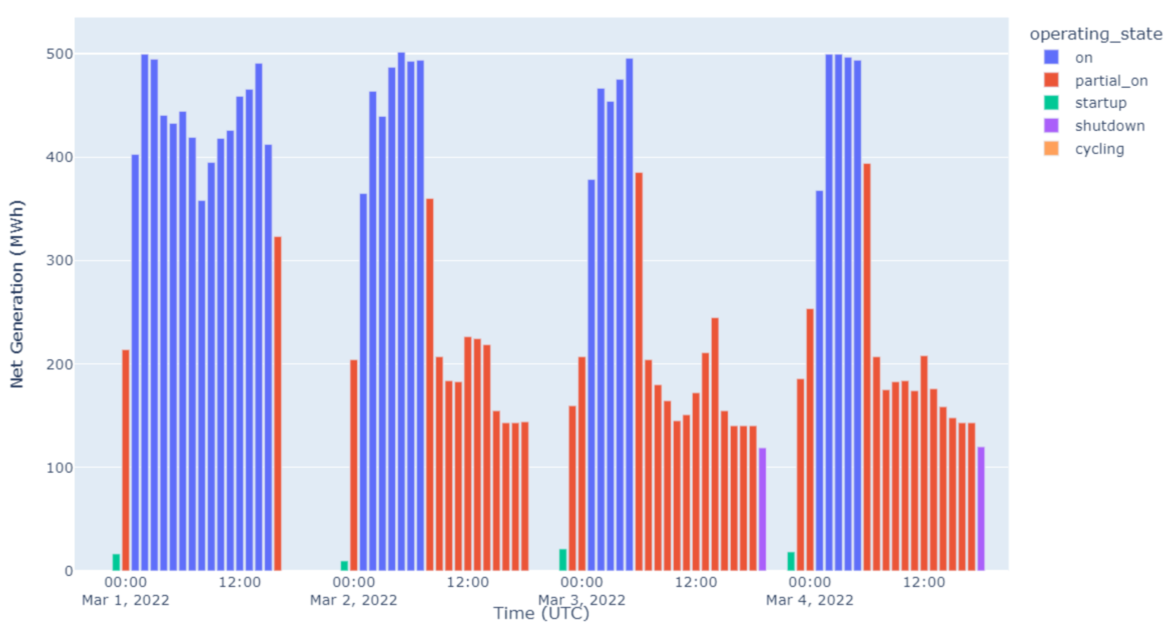

***Figure 3**: Illustration of the operating states assigned to generation for the CC1 combined cycle unit at the Dynergy Moss Landing Power Plant in California for 4 days in March 2022.*

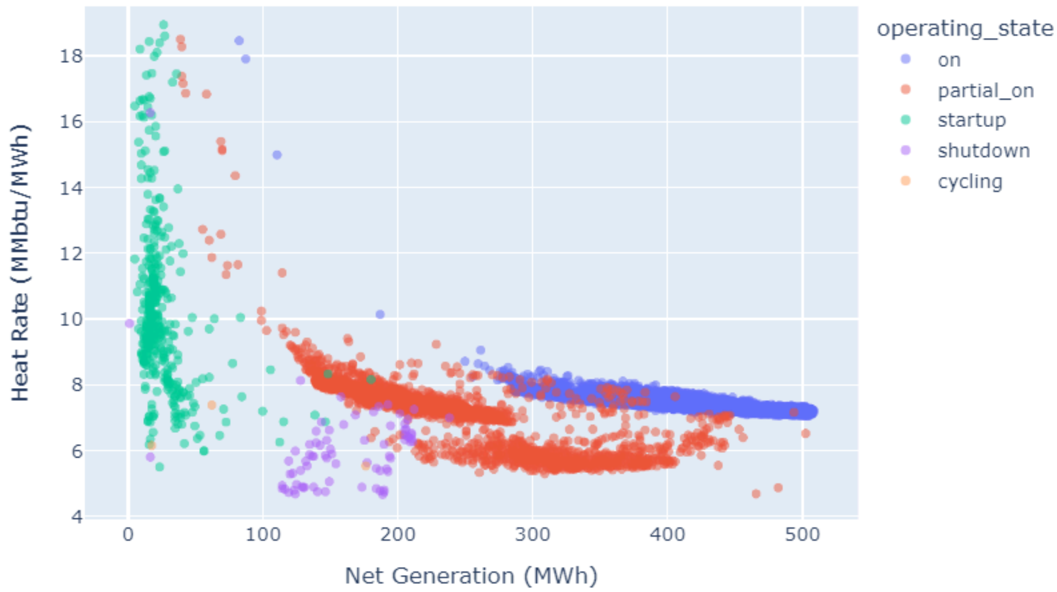

Figure 4*: Illustration of how the heat rate curves can vary by operating state for the same subplant, in this case the CC1 combined cycle unit at the Dynergy Moss Landing Power Plant in California.*

3.1.2 Operating State Assignment

Identifying most of these operating states involves analysis of generation data over multiple time periods to determine, for example, if a generator is transitioning from off to on. However, in the API, GRETA only consumes one timepoint at a time, and currently is configured such that all timepoints are independent of each other. This means that currently, the GRETA API cannot identify the operating state of a unit in the same way that it is identified during model training.

The GRETA API can currently identify 3 different operating states using the data available:

- off: if generation_mw is 0
- startup: if generation_mw is > 0 and < the maximum startup threshold
- on: if generation_mw is >= the maximum startup threshold

This maximum startup threshold is determined by attempting to identify a unique range of MW values between 0 and X for each subplant where the subplant is most likely to be starting up rather than on. GRETA uses a binning approach, identifying for each MW bin in the generator's operating range whether it is more likely to be starting up or on (based on the frequency of historical observations in each bin). In some cases, these ranges can overlap (i.e., the highest bin in which startup is most likely is greater than the lowest bin in which on is most likely). In this case, GRETA takes a recursive approach to identify the best point within this overlapping range to set a non-overlapping boundary between startup and on. If the overlapping range contains a higher probability of startup events, the minimum on MW bin is increased by 1. Conversely, if the range contains a higher probability of on events, the maximum startup MW is decreased by 1. This process repeats until convergence on a single number. See Figure 5 for an illustration of this process.

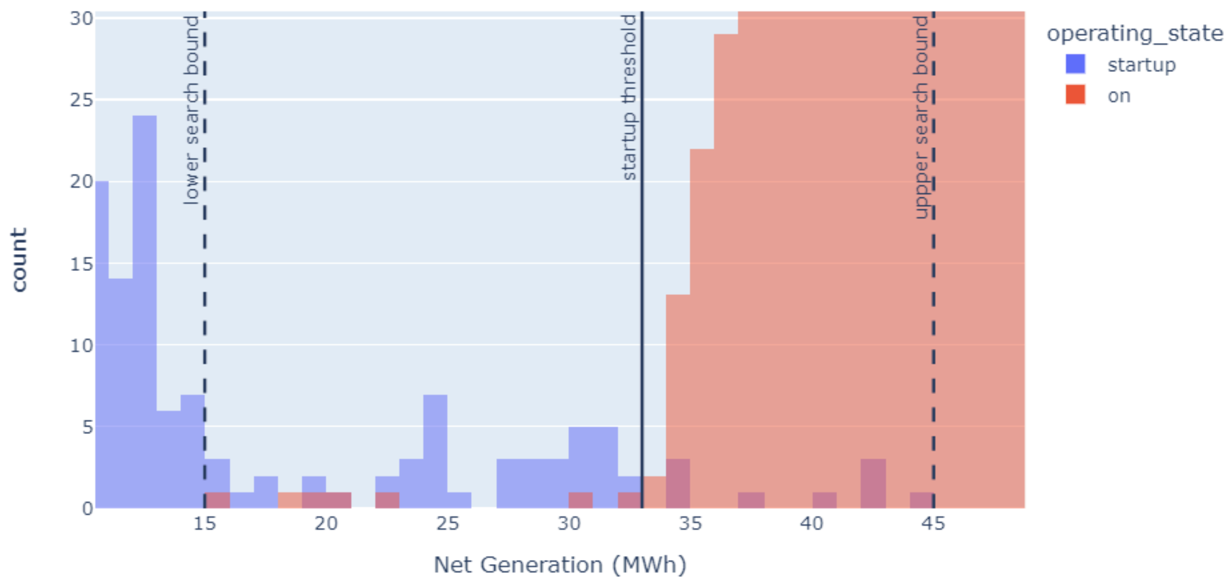

Figure 5*: Histogram of the number of "startup" or "on" hours at each generation level for the Saguaro Plant in Arizona that illustrates how the startup threshold is calculated. The algorithm searched between 15MWh (the lowest value for the "on" state) and 45MWh (the highest value for the "startup" state). In this initial window, the "on" state is most probable, so the upper bound of the search window is decreased from 45MWh to 44MWh. This continues until the search range converges on 33 MWh.*

3.2 Fuel Type

For each interval, GRETA identifies the fuel being consumed, which may change from one interval to the next for generators that use different fuels during startup, fuel switch, or co-fire fuels. This assignment is performed dynamically in the API based on a generator's publicly-reported fuel attributes and models trained based on historical fuel consumption data.

For the purposes of fuel assignment, GRETA recognizes four categories of generators:

1. **Single-fuel generators**, which can only utilize a single fuel type to generate electricity
2. **Fuel-switching generators**, which can switch between using 2 or more different fuels
3. **Co-firing generators**, which can use two or more different fuels simultaneously

4. **Different startup fuel generators**, which use a fuel that differs from their primary fuel only during startup

These categories are non-exclusive; for example, a generator may fuel switch and use a different fuel during startup (see Table 1).

Table 1*: The percentage of subplants in the U.S. that fall into each fuel consumption category. Note that fuel switching, co-firing, and different startup fuel generators are subsets of "multi-fuel" generators, but are not mutually exclusive so will not add up to the mutli-fuel subtotal.*

Fuel Consumption Category	Percent of U.S. Subplants
Single-fuel generators	84.1%
Multi-fuel generators	15.9%
Fuel-switching generators	11.0%
Co-firing generators	6.3%
Different startup fuel	8.1%

3.2.1 Fuel type modeling

3.2.1.1 Determining fuel attributes

The fuel attribute category of a generator is determined based on data that is reported for each generator in EIA-860 and EIA-923.

The "Fuel Switching Data" reported in Schedule 3 of EIA-860 identifies for certain generators whether they consume multiple fuels, whether they co-fire fuels, the two primary fuels they switch between, and the (up to) 6 fuels they cofire, as well as additional data about the time to switch and other limits on switching. To ensure complete and accurate data, GRETA checks for internal consistency (i.e. sometimes a generator is identified as fuel switching, but not multi-fuel), consistency with the fuel types reported in the "Generator Data" table of EIA-860, and the fuel consumption data reported in EIA-923. In the case that there are inconsistencies, (i.e. the generator is not marked as a multi-fuel generator, but it reports consuming multiple fuels in EIA-923) GRETA uses the data from EIA-923 and other parts of EIA-860 to correct the fuel attributes where necessary. GRETA also utilizes multiple recent years of this reported data to check for consistency. Where the data is inconsistent across years (e.g. a generator is identified as fuel switching in one year but not another), the maximum Boolean value is used (i.e. if True in at least one year, it as True in all years).

Any generators not identified as multi-fuel generators (i.e., do not fuel switch or cofire fuels) are considered to be single-fuel generators. At the plant level, if any subplant is multi-fuel, the entire plant is considered multi-fuel.

Along with the fuel attribute category, GRETA also identifies (sub)plant-specific primary fuel from the OGE dataset. This primary fuel is determined from data reported in EIA survey forms 860 (EIA-860) and 923 (EIA-923) using the methodology described here.

Treatment of Pumped Storage Hydro (PSH) facilities

In EIA-860, the energy source of pumped storage hydro (PSH) facilities is identified as “water” (WAT) which is the same as conventional hydroelectric facilities. However, these facilities operate as energy storage facilities, which are otherwise identified using the “electricity” (MWH) energy source code in EIA. To ensure that PSH facilities are treated consistently with other energy storage facilities, GRETA re-codes the energy source of PSH facilities from WAT to MWH.

3.2.1.2 Fuel switching and cofire model

Since a fuel-switching generator can switch fuels over time, GRETA needs to be able to predict which fuel a generator is most likely to be consuming in any given interval. In the case of a cofire generator, GRETA needs to predict in what ratio the cofired fuels are being consumed in a given interval. Currently, this model utilizes a single variable to predict the fuel type: the month of the year. This model was selected as a starting point partially because historical, generator-specific fuel consumption data is only available at the monthly resolution in EIA-923.

While this approach is better than assuming a single fuel type or a constant mix of fuels throughout the year, this current model has several substantial limitations:

- It assumes that fuels switch only monthly, and at the beginning of each month. In reality, fuels can switch more frequently (multiple switches in a single month) and at any time.
- Other non-fixed effects that may affect fuel switching, such as the price of fuel or the operating level of the generator, are not considered
- The model assumes a consistent seasonal fuel switching pattern across years (i.e. that a plant will always use a certain fuel every July), but this may not be the case for all plants.
- Although all data in EIA-923 is structured as monthly data, some plants only report annual data, and the EIA allocates this data to months based on the fuel consumption

patterns of similar plants. However, this allocated annual data may not represent actual fuel switching patterns for a specific generator.

To model the probability that a given fuel is being consumed in a specific month at each subplant, GRETA uses monthly fuel consumption data reported in EIA-923 for the most recent 3 years of reported data (at the time of writing, this included data from 2020-2022). To avoid training on data that is no longer relevant due to a major operational change at a generator, GRETA first removes any data prior to a major operational change from the training set on a subplant-specific basis (see the more detailed discussion of this approach in section 3.3.1.1). GRETA uses a cleaned version of this data from OGE, which allocates reported fuel consumption at the prime-mover level to each generator, before aggregating the data to subplants. The detailed methodology for this allocation process can be found here.

Using this dataset, GRETA trains a monthly fixed effects model for all multifuel subplants to predict what percentage of the total fuel consumed (in mmBTU) came from the primary versus secondary fuel. To prepare the data for training, GRETA creates a ranked list of energy source codes for each subplant, where *energy_source_code_1* is the fuel that was most consumed at the subplant in the training period, *energy_source_code_2* is the fuel that was second-most consumed, and so on. For each subplant, the model then only keeps fuel consumption data associated with the top two energy source codes, and ignores any other fuels consumed. While this results in some data loss, it simplifies the model by enabling identification of the fuel percentage of the two most-used fuels with a single response variable. For each subplant-month, GRETA calculates an *energy_source_code_1_fraction*, which represents the percentage of total fuel consumption from the primary source code in that month, such that (1 - *energy_source_code_1_fraction*) is the percentage of fuel from the second fuel. This process is repeated at the plant level using plant-aggregated data.

Finally, GRETA fits a linear regression in the form:

energy_source_code_1_fraction ~ C(month)

For each model, if the adjusted R2 value is greater than 0.1, the monthly fixed effects model is kept; otherwise, GRETA will return an annual average *energy_source_code_1_fraction* for all months, since there is no clear seasonality in the data.

This model is used for both fuel switching and cofiring generators. In the case of fuel switching generators, the monthly coefficient is used to determine the single most probable fuel that is being consumed at that (sub-)plant in that month. In the case of cofiring generators, the monthly coefficient is used to determine in which ratio the two primary fuels are being burned simultaneously.

3.2.2 Fuel type assignment

In the GRETA API, fuel types are assigned to generators in a multi-step process. If the EIA generator ID of the generator is known, GRETA assigns the fuel type based on a subplant-specific mapping; otherwise, if only the EIA plant ID is known, it will be assigned based on a potentially less-precise plant-specific mapping.

1. **Manual identification**: If a fuel type is manually specified by the input data, GRETA will use the specified fuel type. This would only apply in the case that the fuel type for fuel-switching generators is known in real time by the end user. In MISO's case, these fuel types are not known.
2. **Single-fuel generators**: If the (sub)plant only ever uses a single fuel type, GRETA assigns the (sub)plant-specific primary fuel identified in the OGE dataset. This primary fuel is determined from data reported in EIA survey forms 860 (EIA-860) and 923 (EIA-923) using the methodology described here.
3. **Identify startup fuel**: If the generator uses a different fuel during startup, and GRETA determines that the generator's operating state is "startup," GRETA assigns the primary startup fuel type as reported in EIA-860.
4. **Handle fuel-switching and co-firing generators**: If the generator has fuel switches or co-fires and has not already been assigned a fuel type in a previous step, GRETA assigns the most likely primary fuel based on the month of the year. For fuel switching generators, GRETA will assign a single fuel type and calculate emissions based on that fuel. For co-firing generators, GRETA assigns a single fuel type, but calculates emissions based on the modeled cofire ratio for each fuel (see section 3.4).

Assigning fuel switching

In the production data pipeline, for subplants that fuel switch, if the monthly coefficient for energy_source_code_1_fraction is greater than or equal to 0.5, it is assigned the (sub-)plant_specific energy_source_code_1 (identified by the fuel switching model). Otherwise, if the fraction is less than 0.5, it is assigned energy_source_code_2. Emissions are calculated based on whichever of the two energy source codes is assigned.

For subplants that cofire fuels, GRETA uses the same model that is used for fuel switching, but the model outputs are interpreted slightly differently. GRETA still assigns a single energy_source_code using the same approach as for fuel switching subplants, but emissions are calculated differently (see section 3.4).

3.3 Fuel consumption

Heat rates (in mmBTU/MWh) describe the efficiency of a power generator as the ratio of fuel input (mmbtu) required to generate a single MWh of electricity output, and thus can be used to convert generation data (MWh) to fuel consumption (mmBTU) by multiplication. However, the heat rate of a generator can vary depending on multiple factors, including the capacity factor of the generator, so it is necessary to use modeled heat rate curves (HRCs) to first identify the appropriate heat rate to use at a given output level, time of year, and operational state.

3.3.1 Heat Rate Curve Modeling

Heat rate curves are calculated for each subplant by fitting a polynomial regression to historical generation and fuel data.

GRETA uses historical hourly CEMS data to train heat rate curves for the subplants that report these data to the EPA's CAMPD. For subplants that do not report these data, GRETA uses historical monthly data reported to EIA-923 to calculate an average heat rate (rather than a heat rate curve). For cases where generation data passed to the GRETA API is not mapped to a specific generator, GRETA also calculates plant-level heat rate curves using CEMS and EIA data aggregated to the plant level to use as backstop values.

GRETA is currently trained on three years of historical data (2020-2022) and the models are updated annually as new CEMS and EIA data becomes available.

3.3.1.1 Preparing CEMS data for training

After raw, unit-level CEMS data is loaded, it is prepared for training by removing anomalous and extraneous values. Specifically, the following steps are taken:

- Removal of extraneous or unusable data
  - Remove any observations where the generator is off (in which it reports zero fuel consumption, generation, and emissions)
  - Remove any observations where the fuel consumption data is missing, since heat rates cannot be calculated for these hours
  - Remove observations with positive fuel consumption but zero gross generation. These mostly represent plants that report only steam load and not gross MW.
- Removal of physically impossible observations
  - Remove any observations where the reported gross generation, fuel consumption, or CO_2 data is negative

    - Remove observations if the thermal efficiency (gross generation converted to mmBTU divided by fuel consumption in mmBTU) is >= 100%.
- Removal of outlier data
    - Remove observations where the gross heat rate in that hour is < 4.5 or > 40 mmbtu/MWh. These cutoff values were used in Rossol et al 2019, and confirmed to be reasonable through independent evaluation of the CEMS data
    - Remove observations where the reported gross generation is > 125% of the nominal nameplate capacity of the subplant (as reported in EIA-860).

The unit level data is then aggregated to the subplant level.

Accounting for major operational changes in the training data

Although GRETA currently uses up to 3 years of historical data for training, there are cases when it would not make sense to use the full history of data. When selecting training data for each subplant, it is important to use historical data for years when the operational configuration of the subplant is similar to what it is today. Thus, the training dataset will only include data starting from the most recent major operational change at the subplant. A "major operational change" is defined as an instance when 1) a new generator is added to the subplant, 2) an existing generator retires from the subplant, and/or 3) the technology description of the generator reported in EIA-860 changes (e.g. from "Conventional Steam Coal" to "Natural Gas Steam Turbine").

Gross to net generation conversion

Because CEMS data only includes gross generation (and not net), training a heat rate curve based on net generation requires converting the CEMS data to net. GRETA follows the gross-to-net (GTN) conversion methodology described here. However, instead of using year-specific gross-to-net (GTN) ratios, which can artificially shift the heat rate curve across years (which results in a poorer fit), GRETA calculates a gross to net ratio.

3.3.1.2 Fitting heat rate curves

Using the prepared CEMS data, GRETA then fits a 2nd-order polynomial heat rate curve to the data of the form:

$$MMBtu = A_0 + (A_1 \times MW) + (A_2 \times MW^2)$$

Currently, for each subplant, separate HRCs are fit for:

- Operating state: one HRC is trained for hours when the operating state is "on" or "partial on", and another HRC is trained for hours when the operating state is "startup"

- Ozone season: one HRC is trained for hours in ozone season (May 1 - Sept 30) and a separate curve is trained for hours not in ozone season. Additionally, non-seasonal (annual) heat rate curves are also trained in case the subplant only reports CEMS data for part of the year but operates year round. Seasonal heat rate curves are only trained for the "on" operating state since this seasonality does not appear to affect startup significantly.

***Table 2**: The full set of possible heat rate curves trained for a given subplant*

Available Data	Curve Type	Operating State	Aggregation	Season
CEMS	2nd-order polynomial	Startup	Subplant	Annual
		On	Subplant	Ozone
				Non-ozone
				Annual
			Plant	Ozone
				Non-ozone
				Annual
EIA	Linear	On	Subplant	Annual
			Plant	Annual

To ensure that a fit HRC would not return any physically impossible (i.e. negative) fuel consumption values, GRETA plugs a range of values into the fit curve to check that none of the returned values would be negative. If it would, GRETA re-runs the regression forcing the intercept through zero. The curve is then re-tested, and if it would still possibly return a negative value, the regression is once again re-run, forcing a linear fit with the intercept through zero.

The output table containing these heat rate curves also includes the adjusted R-squared value for the fit, the number of observations, and the range of validity (the min and max generation values used to train the curve). Theoretically, these curves should not be applied to generation data outside of the range of validity, as that would represent an extrapolation of the fit. Thus, instead of applying the heat rate curve to generation values outside of the training range, GRETA also calculates the heat rate at the lowest end of the training range and the heat rate at the highest end of the training range. If a generation value passed to the GRETA API is less than or greater than the training range, GRETA will use the heat rate value at the range minimum or maximum, respectively. This essentially assumes that the heat rate gets no worse at low generation values, and no better at high generation values.

During training, certain subplants can be forced to use linear rather than polynomial HRC fits. This would only be necessary in the case of JOUs where the ownership share is unknown or variable, and thus cannot be used to determine an output-specific heat rate. This case is not relevant for the MISO implementation.

3.3.1.2.1 Heat rate curves for subplants that do not report to CEMS

For plants that do not report hourly data to CEMS, the only available data to train heat rate curves is monthly net generation and fuel consumption data from EIA-923. For subplants that only report data to EIA, only a single average heat rate is calculated, and the operating state and season are disregarded due to the smaller sample size and coarser resolution of the data.

The EIA-923 data is prepared for training by:

- Allocating monthly fuel- and prime-mover specific data from the “Monthly Generation and Fuel Consumption Time Series File” to individual generators (using the methodology described here), and then aggregating the data to the subplant level.
- Removing data associated with clean (non-thermal) fuels since these will not have heat rate curves
- Remove observations where net generation or fuel consumption is zero or negative
- Remove observations if the thermal efficiency (calculated as net generation converted to mmBTU divided by fuel consumption in mmBTU) is physically impossible (>=100%).
- Remove observations with a gross heat rate < 4.5 or > 40 mmbtu/MWh, consistent with the thresholds used for CEMS data described in section 3.3.1.1

The heat rate curves for the EIA-only subplants are calculated as a 1st-order polynomial (linear) fit with the intercept through zero (MMBtu = A1 x MW). GRETA takes this approach because of the inconsistency of the resolution of the training data (monthly) and the resolution of the predictions that the GRETA API needs to make (hourly or subhourly). GRETA uses a linear heat rate because the EIA generation data represents total monthly MWh, but the input to a heat rate curve is instantaneous MW. Monthly MWh cannot be accurately converted to MW (by dividing by the total number of hours in the month) because this would assume a constant output rate throughout the month. Furthermore, the intercept of this linear fit is forced through zero because otherwise the intercept would represent the total monthly fuel consumption at 0 MWh of total monthly generation, not the fuel consumption for a specific hour.

Plant-level heat rates are then calculated using the same approach using plant-aggregated data.

3.3.1.2.2 Adjusting Heat Rate Models for CHP

Because combined heat and power (CHP) plants produce both heat and power, not all of the fuel they consume is used to produce electricity. To accurately characterize emissions estimates specifically from electricity generation, the generator's total fuel consumption represented by the heat rate curve needs to be adjusted.

The methodology for how CHP adjustments are generally performed in historical data is described here. The ratio used to adjust total fuel consumption to only fuel consumption used for electricity production is called the electric allocation factor (EAF). EAFs can change throughout the year if the CHP plant adjusts the amount of heat produced relative to the amount of electricity generated. Sometimes these EAFs have clear seasonal patterns, while in other cases they do not. GRETA trains a seasonal fixed effects model on historical EAF data to determine how EAFs change in each month and predict the EAF for data passed to the GRETA API.

This model is trained using processed historical EAF data from OGE, for years 2020-2022. To prepare the data for training, the following steps are taken:

- Remove missing data
- Remove any observations where the EAF is 0 (indicating no fuel consumption for that month)
- Remove data for any subplants where the EAF is always 1 (indicating that the subplant is not a CHP unit)

This subplant-level data is also aggregated to the plant level to calculate plant-level models.

For each (sub)plant, GRETA calculates an ordinary least squares (OLS) regression in the form:

electric_allocation_factor ~ C(month)

If the adjusted R-squared value of this regression is above a certain threshold (0.4), the model outputs are kept. Otherwise, GRETA calculates a mean EAF across all observations and assigns that value to all months. The mean EAF is also used if there is only a single year of training data for the (sub)plant or if there is no data for one or more months across all years of data.

3.3.1.2.3 Calculating backstop heat rate curves

It may not be possible to calculate subplant-specific HRCs for some units if they do not have any historical data to train on, for example in the case of newly-constructed or proposed generators. Ideally, a generator would have at least a year of data available in order to calculate seasonal heat rate curves, and an additional year of history to validate the model on an out-of-sample dataset. In these cases, GRETA calculates a technology-specific HRC as a backstop value.

This approach calculates a "typical" HRC for similar subplants and applies the fit curve to the subplant with no historical data. Similar subplants are defined as subplants that share the same technology type and prime mover (e.g. combustion turbine, internal combustion engine, steam turbine). The EIA defines several dozen distinct technology types, including labels such as "Natural Gas Steam Turbine", "Conventional Steam Coal", and "Petroleum liquids" (see EIA-860 data for a full enumeration). Some of these technologies (like "Conventional Steam Coal") are associated with a single prime mover type (e.g. “Steam Turbine” (ST)), while others (like "Petroleum Liquids") may be associated with several different prime movers. Because combined cycle units can be represented using multiple different prime mover codes (CA, CT, CS, CC), these codes are consolidated to all be "CC" so that all parts of a combined cycle unit are grouped together when training HRCs (see this table for a description of all prime mover codes used by the EIA).

Because individual generators of a technology type have different nameplate capacities, operating ranges, and efficiency levels, heat rate curves cannot simply be calculated using the raw generation and fuel consumption data combined from multiple generators. However, all else equal, GRETA currently assumes that two generators with the same technology of different sizes operating at 100% of their capacity have similar heat rates. Thus, before calculating the backstop heat rate curves, the data for each generator is normalized by its nameplate capacity. If the normal HRC regression takes the general form:

$$Fuel_t \sim generation_t \text{ (represented here in linear form for simplicity)}$$

then backstop regressions take the form:

$$(fuel_t \,/\, nameplate\_capacity) \sim (generation_t \,/\, nameplate\_capacity)$$

The *generation/nameplate_capacity* parameter is just the generator's capacity factor, so this could also be represented as:

$$(fuel_t \,/\, nameplate\_capacity) \sim capacity\_factor_t$$

This would represent a technology-specific regression of normalized fuel consumption on capacity factor. Raw generation data (in MW) could not be plugged into the regression to get total fuel consumption (in mmBTU). Instead, using these technology-specific curves requires calculating an interval-specific capacity factor, which is plugged into the HRC, and then multiplied by the generator's nameplate capacity to get fuel consumption. In other words, if a typical HRC is represented as *f(g)* with g representing generation, then calculating fuel consumption from the technology-specific curve requires calculating f(g/c)*c where c is the generator nameplate capacity.

Because the HRC is transformed using only nameplate capacity, which is a fixed value for each specific generator, the technology-specific regression coefficients can actually be transformed into generator-specific coefficients like a standard HRC using the following proof:

If a third-order polynomial HRC is f(g), where g is generation (MW), then f(g) can be written out as:

$$fuel = a0 + a1*g + a2*g*g + a3*g*g*g$$

And technology-specific HRC is f(g/c), which can be written out as:

$$fuel/c = a0 + a1*(g/c) + a2*g/c*(g/c) + a3*(g/c)*(g/c)*(g/c)$$

Then calculating total fuel consumption from the technology-specific curve requires multiplying both sides by c, or f(g/c)*c, which can be written out as:

$$(fuel/c)*c = a0*c + a1*(g/c)*c + a2*(g/c)*(g/c)*c + a3*(g/c)*(g/c)*(g/c)*c$$

Cancelling terms results in:

$$fuel = a0*c + a1*g + a2*(g/c)*g + a3*(g/c)*(g/c)*g$$

Then, re-arranging terms to group c with each coefficient becomes:

$$fuel = (a0*c) + (a1)*g + (a2*(1/c))*g*g + (a3*(1/c)*(1/c))*g*g*g$$

This formula is the same as the standard HRC f(g), except with a0 multiplied by c, a2 multiplied by (1/c) and a3 multiplied by $(1/c)^2$

Thus, technology-specific HRCs can be transformed into generator-specific HRCs by transforming each coefficient using the generator-specific nameplate capacity:

- Multiply the technology-specific *a0* by subplant-specific *c*
- The technology-specific a1 stays the same
- Multiply the technology-specific a2 by subplant-specific (1/c)
- Multiply the technology-specific a3 by subplant-specific $(1/c)^2$

Transforming these into subplant-specific HRCs enables directly incorporating these HRCs into the HRC table and using the same downstream logic, rather than needing a separate logic to handle these backstop values.

3.3.2 Fuel consumption calculation

In the GRETA API, fuel consumption is calculated by plugging in the provided MW value into the appropriate heat rate curve, then applying any relevant adjustments.

Adjusting MW values
Because HRCs expect as an input total subplant MW output, the raw MW values passed to the GRETA API are first adjusted using several steps:

- If the generator is a jointly owned unit, the raw MW value is divided by the JOU generation fraction to get the total MW
- MW values provided for multiple generators at the same subplant are aggregated to the subplant level

Calculating fuel consumption
Using the provided EIA IDs, the operating state that was identified using the process described in section 3.1.2, and the season (identified using the timestamp associated with the data), GRETA selects the appropriate HRC model to use for each generator in the generation event. If for some reason there is no available HRC for the given season or operational state, GRETA will default to using the annual and/or "on" heat rate curve as a backstop. Then, the adjusted, subplant-level MW values are plugged into the HRC to calculate fuel consumption for that interval.

At this stage, several adjustments can be made to calculated fuel consumption to remove anomalous values:

- If calculated fuel consumption is negative, the fuel consumption is replaced with 0
- If calculated fuel consumption results in a heat rate above 40 mmBTU/MWh or below 4.5 mmBTU/MWh, replace it with a fuel consumption value as if the heat rate were 40 mmBTU/MWh or 4.5 mmBTU/MWh, respectively (see sections 3.3.1.1 and 3.3.1.2.1 for more discussion of these thresholds)
- If the generation value is 0 MW, ensure that the returned fuel consumption is 0

For generation events where the reporting interval is 5 minutes, the calculated mmBTU/hr value is converted to total mmBTU by dividing by 12 (a 5-minute interval is 1/12th of an hour).

Adjusting fuel consumption for CHP

All calculated fuel consumption values are then adjusted for CHP. If a subplant has a CHP model trained for it, the electric allocation factor for the appropriate month is selected. If no model is trained, the subplant is assumed to not be a CHP unit and is assigned an EAF of 1. The EAF is then multiplied by fuel consumption to get a CHP-adjusted fuel consumption value.

3.4 Emissions output calculation

The ultimate emissions calculation relies on attributes identified (e.g., fuel type) and values calculated (fuel consumption) in previous steps. The methodology for calculating GHG emissions mass from fuel consumption is described in detail here, but generally involves multiplying the fuel consumption by a fuel-specific combustion emission factor that represents the mass of CO_2, CH_4, or N_2O released to the atmosphere when a single mmBTU of a fuel is combusted.

For generators that cofire two fuels at once, the emissions are calculated using a weighted-average combustion fuel emission factor based on the cofire fraction identified by the fuel type assignment described in section 3.2.2. Emissions are calculated as:

*total_fuel_mmbtu * [(fuel_1_fraction * fuel_1_EF) + ((1 – fuel_1_fraction) * fuel_2_EF)]*

Component GHG emissions are converted to CO_2-eq values using the methodology described here.

After all fuel and emissions calculations are complete, the calculated subplant-level totals are disaggregated back to each generator specified in the inputs before being returned by the API.

4. Future work

GRETA currently has multiple known limitations that are planned to be addressed as part of a future roadmap:

- **Monthly resolution of the fuel-switching model**: As described in section 3.2.2.2, there are several limitations to the current monthly fixed effects model used to predict fuel switching. In the future, options for improving the temporal granularity and accuracy of this model could be explored: identifying the fuel type of reported CEMS data to use for training, using external predictors (such as fuel price), and improving handling of annually-reported EIA-923 data.
- **Improved CEMS data cleaning/preparation for training**: There are several ways that the quality of the CEMS training data could be improved, including accounting for subplants that only report partial data to CEMS and accounting for plants that only report partial subplant data to CEMS. In addition, improving the definition of “major

operational changes" at a plant, which is used to improve the relevance of the training data, could be done by incorporating data about the latest uprate/derate, the addition of new pollution control equipment, or other operational changes described in EIA survey forms.

- **Allowance for utilization of gross generation or mixed gross/net generation as API inputs**: Currently GRETA only accepts net generation data as an input. However, in certain cases, only gross generation data may be available, or net generation data may only be available for certain generators and not others. Additional flexibility here could help improve the accuracy of the outputs.
- **Heat rate curve fit**: Exploring whether a 3rd order polynomial HRC could result in more accurate outcomes (without over-fitting) could be undertaken. Additional regressors such as ambient air temperature could also be explored.
- **Improved fleet-level backstop heat rate curve modeling**: When calculating fleet-level backstop heat rate curves, the fit of the regression model on the data is much worse than when calculating a the fit for a single generator. The fit may be improved by considering additional variables when training the model, such as the age of the generator, installed pollution controls, altitude (which affects air density), and other factors.

5. Glossary / Definitions

Acronyms

- API: Application Programming Interface
- CAMPD: Clean Air Markets Program Division
- CEMS: continuous emissions monitoring system
- CHP: Combined Heat and Power
- EAF: electric allocation factor
- EIA: U.S. Energy Information Administration
- EIA-860: EIA Survey Form 860
- EIA-923: EIA Survey Form 923
- EPA: U.S. Environmental Protection Agency
- GHG: greenhouse gas
- GRETA: the Generator REal-Time emissions Assignment module
- HRC: heat rate curve
- IPCC: Intergovernmental Panel on Climate Change
- JOU: jointly-owned unit
- mmBTU: million british thermal units
- MW(h): Megawatt(-hour)
- OGE: Open Grid Emissions
- UTC: Universal Coordinated Time