Environmental Science & Technology

Article

Cite This: Environ. Sci. Technol. XXXX, XXX, XXX−XXX

pubs.acs.org/est

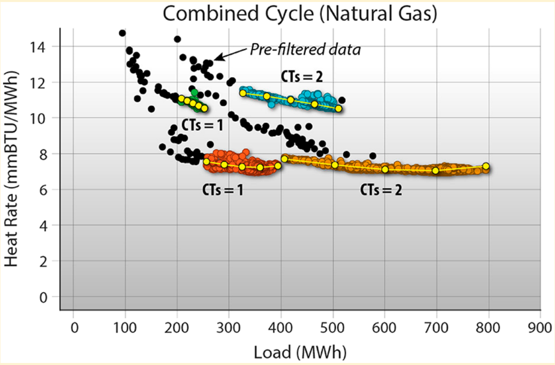
An Analysis of Thermal Plant Flexibility Using a National Generator Performance Database

Michael Rossol,* Gregory Brinkman, Grant Buster, Paul Denholm, Joshua Novacheck, and Gord Stephen

Strategic Energy Analysis Center, National Renewable Energy Laboratory, Golden, Colorado 80401, United States

Supporting Information

ABSTRACT: Grid integration studies are key to understanding our ability to integrate variable generation resources into the power system and evaluating the associated costs and benefits. In these studies, it is important to understand the flexibility of the thermal power fleet, including how thermal plants operate at part load. Without a comprehensive understanding of thermal plant operation, we may over- or underestimate our ability to integrate variable generation resources and thus draw incomplete or inaccurate conclusions regarding their potential economic and environmental effects. The only public data source for understanding many elements of the operational characteristics of the thermal fleet is the U.S. Environmental Protection Agency Clean Air Markets database of historical power plant operation. However, though these data sets have been widely utilized, their use has proven to be difficult, and methods to clean and filter the data are not transparent. Here, we describe the database and a method to clean and filter it. We then use the cleaned database to demonstrate several characteristics of historical plant operation, including frequent part load operation. Finally, we provide a cleaned data set with heat rate curves and describe how to use it in general modeling activities and analysis.



1. INTRODUCTION

A large and growing body of work has evaluated the potential of variable generation (VG) wind and solar resources to make a major contribution to the electric sector.[1,2] Grid integration studies use complex tools that simulate the hour-by-hour (or increasingly subhourly) operation of hundreds or thousands of generators and transmission elements[3−5] to evaluate the impacts of VG resources on system operability. The most detailed studies simulate the commitment and dispatch of every generator in a study area, considering transmission constraints and the need to maintain sufficient operating reserves to address unforecasted changes in demand and system contingencies. These models are fed by large data sets, including data on the performance of the thousands of fossil power plants that currently provide a large fraction of the electricity in the United States and internationally.

A key element of integration studies is understanding the capability of fossil-fueled thermal generators to turn off and on and vary output over multiple time scales.[5,8] This is partly because increased penetration of VG resources results in increased variability of net energy demand (i.e., normal demand minus the contribution of VG). As VG is added to the grid, thermal plants will produce less energy, reducing fossil fuel use and emissions. However, these plants operate at different efficiencies when operated at part load, and increased VG penetration results in greater thermal plant output variability.[3,6] The ability of thermal plants to respond to increased variability, and the impact of this variability on costs and emissions, is often a key element of VG integration studies, so it is important that the operational characteristics of power plants are well represented. Without a comprehensive understanding of thermal plant operation, we may over- or underestimate our ability to integrate VG and thus draw incomplete or inaccurate conclusions regarding the potential economic and environmental benefits of doing so.[7]

In this work, we use historical data from the U.S. Environmental Protection Agency's (EPA) Clean Air Markets (CAM) Division that captures about 71% of the U.S. thermal generation fleet (excluding nuclear) to demonstrate that most plants spend a large fraction of their operation at part load, including traditional baseload generators. We also summarize a method for cleaning and filtering the data using fitted heat rate curves, including a method to isolate curves for combined-cycle power plants that have multiple operating modes. We then describe methods to apply these data sets to models used in grid integration studies. Finally, we provide links to our tools and processed databases for others to use, creating reproducible and transparent data sets that can help improve grid studies that include assessments of the environmental

Received: July 26, 2019
Revised: October 10, 2019
Accepted: October 23, 2019
Published: October 23, 2019

DOI: 10.1021/acs.est.9b04522

Table 1. Summary of Processed Data

	Initial Data		Data Removed				Final Data	
			Step 1		Step 3			
Plant Type	GW	TWh	GW	TWh	GW	TWh	GW	TWh
Boiler (Coal)	270.6	2427.8	4.7	368.6	7.6	29.1	258.4	2030.1
Boiler (NG)	68	125.6	7	29.9	6.3	4.9	54.7	90.1
Boiler (Oil)	15.8	14.9	1.9	2.9	4.0	0.4	9.9	11.6
Boiler (Other Solid Fuel)	0.9	6.3	0.1	1.4	0.0	0.1	0.7	4.9
CT (NG)	117.6	141.9	23	31.1	18	17.4	76.5	93.4
CT (Oil)	20.6	11.0	11.5	7.5	5.2	0.5	3.9	3
CC (NG)	233.0	1976.6	8.3	165.5	39.9	314.2	184.8	1496.9
CC (Oil)	0.6	2.6	0.0	0.1	0.1	0.0	0.5	2.45
Total	727.0	4706.6	56.5	607.0	81.0	366.5	589.5	3733.1

benefits of continued deployment of solar and wind energy resources.

2. MATERIALS AND METHODS

2.1. Thermal Plant Data Needs for Grid Integration Studies. Parameters used to describe thermal power plant performance in grid integration studies include:

- Maximum output and minimum stable level (typically measured in megawatts [MW])
- Efficiency (typically referred to in the United States as heat rate, measured in million British thermal units per megawatt-hour [mmBtu/MWh] as a function of operating point)
- Startup time and fuel requirements
- Ramp rates (typically measured in either %/min or MW/min).

Without accurate data for each of these parameters, we may under- or overstate the flexibility of individual power plants, limiting our understanding of the overall flexibility of the power system. Particularly critical to understanding our ability to integrate VG resources are the ramp range and decreased efficiency of plants operating at low load levels. Previous analysis has demonstrated a significant correlation between minimum generation levels and increased costs of VG integration.[9] Furthermore, there has historically been some controversy regarding the impact of VG on part-load operation of plants and associated changes in emissions, so it is important to capture part-load heat rates and feasible operating range to verify the emissions-reduction benefits of VG.[10−12]

Data sets from the U.S. Energy Information Administration (EIA) and the Federal Energy Regulatory Commission (FERC) provide some of the data needed to analyze power plant operation, including maximum output and average efficiency.[13−15] However, critical parameters including efficiency as a function of part load are unavailable in any single, publicly available source.

Therefore, a number of studies have utilized historical plant operation data from the EPA CAM database. For example, grid integration studies have used CAM data to generate heat rate curves for generators in the western[3] and eastern[6] United States. Other analyses have used these data sets to estimate marginal emissions factors for various regions of the United States.[16−18] CAM data has also been used in stand-alone analyses of topics including the relationship between temperature and power plant emissions[19] and the relationship between peak electricity demand and air quality.[20]

While CAM data has seen substantial use, it has not been compiled in a publicly available and directly usable format. The data in its raw form is noisy and requires considerable effort to be usable. As part of a larger effort to improve grid modeling (specifically for an evolving grid) in the United States and North America,[21] the goal of this present work is to create transparent and reproducible techniques and data sets for future analyses, including a stand-alone data set of plant-level heat rate curves directly usable for grid integration studies or other applications. We also use the processed data set to demonstrate the importance of evaluating part-load operation in grid planning studies.

2.2. Data Processing. *2.2.1. CAM/CEMS Data.* Data for this study is derived from the EPA CAM data set.[22] Under its authority, the EPA requires that all combustion power plants with capacity greater than 25 MW install and maintain a continuous emission monitoring system (CEMS), which records several operational parameters including fuel input and generation.[23,24] The data set analyzed in this study includes plants fueled by coal, oil, and natural gas, including steam turbines (STs), combustion turbines (CTs), and combined-cycle (CC) power plants. For each generator, we performed four data-processing steps, which we describe as (1) preprocessing and cleaning, (2) generation of heat rate curves, (3) data filtering, and (4) curve fitting for use in grid models. Each data processing step has multiple substeps, as described in the following sections. A detailed flowchart of the process is provided in the Supporting Information (SI), along with links to the code, and the raw and processed data. A summary of the total number of plants evaluated, including those removed in the processing steps, is provided in Table 1 in Results.

2.2.2. Step 1: Data Pre-Processing and Initial Cleaning. The EPA hourly CEMS data is available for download as comma-separated text files. We began by identifying and extracting data associated with plant and boiler ID, time stamp, fuel input, and generator output from the 2016 and 2017 data sets. Variable IDs and detailed descriptions associated with each data element are provided in the SI. These data points are sufficient to generate heat rate curves and minimum generation levels, which are key parameters needed to accurately model thermal power stations in grid integration studies.[3,4]

A unique ID was created for each unit by combining the plant (ORISID) and boiler (BLRID) ID, and then data was cleaned by removing points with missing or nonmeasurement data, as described in the SI. We then removed data points recorded during generator startup and shutdown as our primary goal was to identify operational heat rate curves, and fuel used during startup or shutdown can distort these results. Finally, physically unrealistic heat rate values of either <4.5 mmBtu/MWh or >40 mmBtu/MWh were removed. This step eliminates obviously erroneous data.

Aggregation of Combined-Cycle Units. A CC plant consists of one or more CTs with the waste heat feeding a heat-recovery steam generator (HRSG) driving an ST. The waste heat may often be supplemented with duct burners. The CTs may be operated individually (without the HRSG/ST) or in various combinations depending on the number of turbines. For example, a 2 × 1 plant (two CTs and one ST) could have as many as six different operating conditions (three CT-only combinations and three CT and ST combinations). This greatly complicates estimating the plant heat rate, along with actually implementing this configuration in grid models. The EPA CAM data often reports individual CTs separately, but with the load generated by the ST uniformly distributed between the CTs. To provide a more accurate understanding of operation of the complete CC units, the time-series data reported for all (n) of the CTs associated with the HRSG were combined and a new unique unit ID was created; i.e., for each hour (t_i) of the year:

$$\text{Heat Input}_{\text{CC}}(t_i) = \sum_{1}^{n} \text{Heat Input}_{\text{CT}}(t_i)$$

$$\text{Load}_{\text{CC}}(t_i) = \sum_{1}^{n} \text{Load}_{\text{CT}}(t_i)$$

where Heat Input$_{\text{cc}}$ and Load$_{\text{CC}}$ are the heat input and load for the CC unit, and Heat Input$_{\text{CT}}$ and Load$_{\text{CT}}$ are the heat input and load for the CT. The 710 units reported to CEMS as CC generators were combined to form 504 CC units by grouping individual generators according to their plant and unit codes reported in the Form EIA-860 generator database. This data combination was performed after the initial preprocessing and cleaning described in Generation of Heat Rate Curves.

2.2.3. Step 2: Generation of Heat Rate Curves. After initial cleaning, we used the remaining data to generate heat rate curves. Heat rate is a common U.S. metric of power plant performance that is defined as the amount of fuel required to generate one unit of electrical energy output

$$\text{Heat Rate} = {}^{\text{Heat Input}}\!/_{\text{Net Generation}}$$

The typical units of heat rate are mmBtu/MWh. The U.S. EPA defines heat rate based on the higher heating value (HHV) of the fuel, or the gross energy content, which includes the energy used to vaporize water released or created during the combustion process.[25] Outside the United States, power plants are more commonly defined in terms of thermal efficiency, equal to net generation divided by heat input (typically using the fuel's lower heating value) where both generation and heat input have the same units. The heat rate varies as a function of output level; typically units are more efficient at greater output levels and will vary with different CC operating modes.[26]

Heat input and net generation values were obtained for each hourly interval. Heat input is the thermal energy of the fuel (HTINPUT) reported in millions of Btu (mmBtu). Net generation is the amount of energy actually delivered to the grid and is equal to the gross generation from the plant minus station energy (energy consumed to run the plant). Power stations can consume a significant amount of power for station services.[27] However, net generation is not reported directly in the CEMS database; the CEMS database reports only the gross load, which is calculated by multiplying the power produced during each hour (GLOAD) and the fraction of the hour the unit was operating (OPTIME)

$$\text{Gross Generation} = \text{GLOAD*OPTIME}$$

where GLOAD is in MW and OPTIME is in hours. We eliminated data where OPTIME was less than 1 in order to remove the impact of startups and shutdowns, so for all our data, gross generation = GLOAD.

To estimate the net energy production in each time period, we multiplied gross load by a scaling factor that estimates average station power

$$\text{Net Load}_{\text{ft}} = \text{Gross Load}_{\text{ft}}\text{*Scale factor}_{\text{ft}}$$

where scale factor is equal to the median value of the sum of annual net energy$_{\text{ft}}$ divided by gross energy$_{\text{ft}}$, where ft is the fuel type (i.e., solid or liquid/gas). Annual net energy by fuel type was derived from the Form EIA-923 database[14] while gross generation was obtained from the CEMS data. Because of occasional mismatches or lack of matches between plant-level information in the two data sets, fleet median values were used instead of unit-specific information. The scale factors derived from this process were 0.925 for solid-fuel plants and 0.963 for liquid- and gas-fuel plants. The lower value for solid (mostly coal) plants demonstrates the more energy-intensive processes involved in large coal-fired power plants, including crushers, fans, and emissions controls.[27] We recognize that a more robust conversion from gross to net would be helpful for this data set, but to do so, a more accurate and complete matching between EIA and EPA data sets or alternative data source with net and gross generation for all units is required. For the remainder of the paper and in all figures, load will refer to the net load.

Figure 1 shows an example heat rate curve for a coal-fired power plant (CAM database unit 7097_1, the JK Spruce power plant in Texas) generated with initial, unfiltered data. It illustrates why additional data filtering is needed and why using a heat rate curve is a useful mechanism for additional filtering. The data is very noisy, but a strong visual trend is observable; additional discussion of CEMS equipment and possible

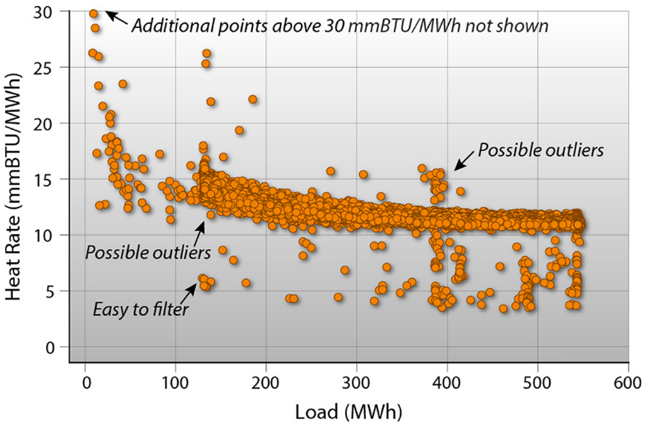


Figure 1. Initial heat rate curves for the CAM database unit 7097_1, JK Spruce power plant. Each data point is an hourly measurement of heat rate as a function of generation.

sources of errors is provided in reference 28. Many points are obviously outliers, being well above or well below normal operating conditions, including many points that are physically impossible for this technology, such as those below 7 mmBtu/MWh. However, simply screening for points above and below certain heat rates may still leave unrealistic data. In this example, a few data points are shown with heat rates that are well within the overall acceptable range but occur at load levels that appear unlikely when compared to the majority of points at this level. Therefore, a screen that considers both heat rate and load level is necessary.

2.2.4. Step 3: Data Filtering and Clustering. After producing the heat rate curves, we applied a series of three filters to eliminate outliers in individual data points as well as entire plants. These three filters, as described below, are (1) minimal operation screening, (2) clustering and filtering, and (3) unit-level filtering.

Minimal Operation Screening. This step removed units with less than 1% of the total available data points. As the CEMS data is provided in hourly timesteps and this analysis focused on data from 2016 to 2017, this corresponded to units with <175 data points. The data for these units was retained in the final data set, but their values were flagged as outliers (see below).

Clustering and Filtering. A set of density-based clustering algorithms was used to isolate the typical operating range of each generator by removing outlying data points. For CC units, an added goal of the filtering procedure was to identify different operating modes using the density-based spatial clustering of applications with noise (DBSCAN) algorithm.[29] DBSCAN determines clusters based on two inputs: a Euclidian distance metric between neighboring points (epsilon) and the minimum number of neighboring points within distance epsilon (minimum samples) for a point to be considered as part of a cluster. This algorithm was selected primarily for its ability to recognize arbitrary numbers of clusters of any shape. The DBSCAN algorithm is used to cluster the CC data using three variables: load, heat rate, and the number of operating CTs. The last variable, the number of operating CTs, provides powerful insight into the different possible operating modes of a unit.

Figure 2 provides the unfiltered heat rate curves (after step 2) for a CC unit (CAM database unit 55441, the Hillabee Energy Center in Alabama). The data provides no field or other simple indication of the actual operating mode (i.e., CC vs CT only). Visually, however, it is easy to recognize the four distinct operational modes of this plant; we have labeled the figure with our interpretation of the actual plant operating configuration. The 2 × 1 mode (the nominal mode) uses both CTs and a steam generator and has the highest efficiency and greatest output. The 1 × 1 mode uses one CT and the steam generator; this represents two possible combinations of plant operation with either CT and the ST. Likewise, the 1 CT mode represents operation of either of the two CTs individually but without the ST. Finally, the 2 CTs mode represents both CTs operating without the ST. The DBSCAN algorithm was able to parse this data into four unique clusters for further analysis and generation of heat rate curves for each operating mode as demonstrated in Results.

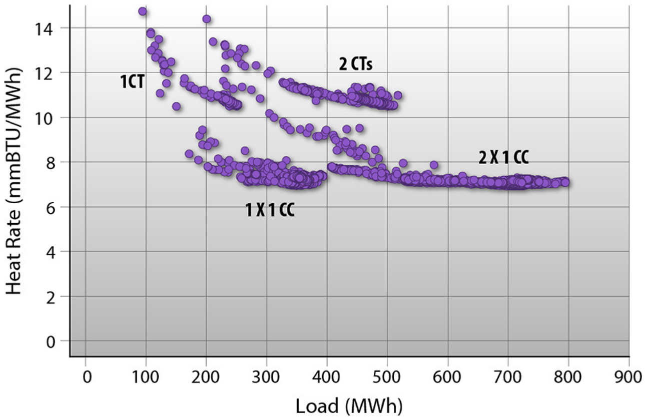


Figure 2. Illustration of unfiltered CC CEMS data for unit 55411_CC1.

For non-CC units, where a single cluster is desired, DBSCAN was found to produce either anomalously small or large clusters. This is because DBSCAN must produce contiguous clusters, and the silhouette score interprets the noise as a cluster. Therefore, a modified version of DBSCAN was used to isolate the typical operating range (the densest regions of the data) while excluding outlying points. In this modified clustering approach, the k-nearest neighbor distances were computed for every point where k is equal to the number of minimum samples. Next, points at which any of the k-nearest neighbor distances is greater than epsilon are designated as noise. Additional details on the DBSCAN algorithm and the modified version used for non-CC units are provided in the SI.

We have provided the results of the filtering procedure for download; see the SI for URLs. For completeness, all data points present after the initial cleaning (step 1) and removal of unrealistic heat rate values are provided. Points removed during the minimal operation screen and during clustering and filtering are flagged with a cluster value of −1. For non-CC units, the points retained after filtering are flagged with a cluster value of 0. For CC units, unique operating modes are grouped by unique positive cluster values.

Unit-Level Filtering. The previous two steps eliminated individual data points. The third and final step was to eliminate entire plants. Namely, we identify remaining plants that have systematically erroneous heat rates. An example is shown in Figure 3. Figure 3a shows data from unit 8906_51RH, a natural gas boiler associated with the Astoria Generating Station. The data from this unit has relatively little noise; however, the average heat rate (about 5 mmBtu/MWh) is physically impossible for this technology class, implying a systematic bias in measurement, reporting, or data recording. Figure 3b shows a plant with an excessively high heat rate (unit 2092_3, Ralph Green Station). Each figure also shows a band of area representing two standard deviations above and below the mean minimum heat rate for all generators of this type. Plants outside this band were removed. This filter was not performed on groups with less than 100 units (oil-fired units and boilers labeled as "other") as these groups could not be well represented by a Gaussian distribution, making the filter overly conservative.

For CC natural gas units, we used a slightly modified approach; distributions of heat rates are provided in the SI. Due to multimode operation, and because the EPA does not require CC operators to report steam cycle generation to CAM, the distribution of minimum heat rate values is bimodal. Units with minimum heat rate values >9 mmBtu/MWh (the

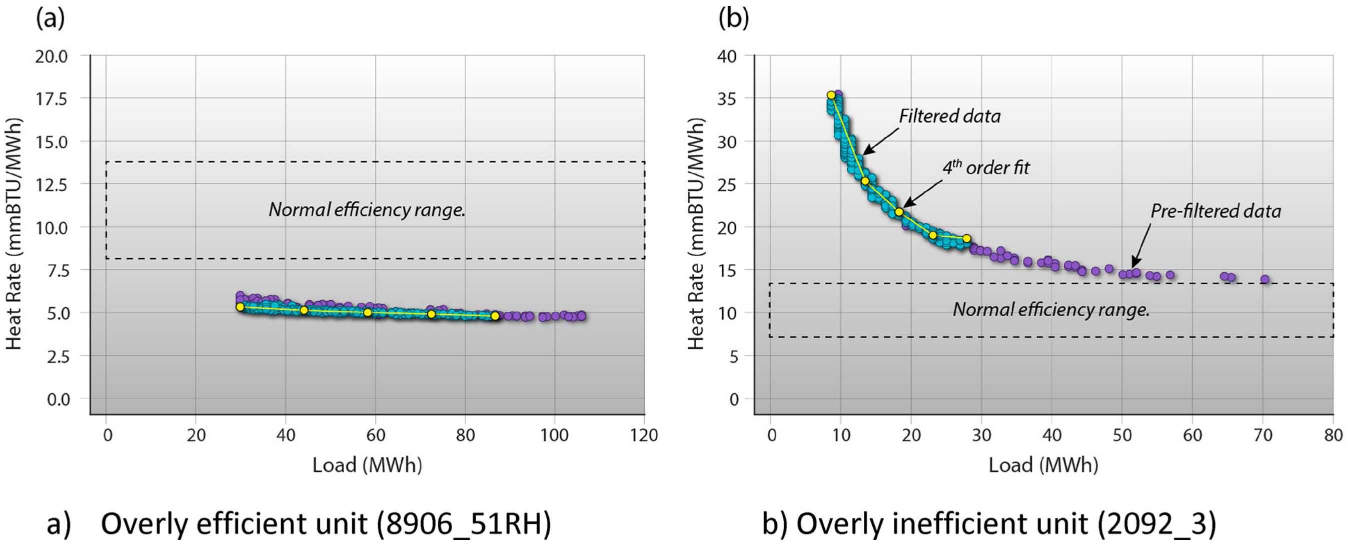


Figure 3. Standard deviation unit filter.

intersection of the two distributions) were deemed to have not reported steam generation and were thus excluded. Furthermore, the most efficient CC units were also removed using the methodology provided above and the bounds indicated in the SI.

Units identified during this step were excluded from our analysis in Results but remain available in the published data.

2.2.5. Step 4: Curve Fitting for Use in Grid Models. The heat rate curves generated in this project are for use in a production cost model environment that uses mixed-integer linear programs. Many grid simulation tools are formulated as linear or mixed-integer mathematical programming problems. These tools require a piecewise-linear representation of each generator's heat input function (mapping net load to heat input), which can be scaled linearly to quantify fuel or emissions costs.

We used the Magnani and Boyd heuristic[30] to generate piecewise-linear heat input function fits for one-, two-, and three-linear-segment functions for each generator. Using multiple pieces can add significant computational burden, particularly when simulating large systems.[6] In many cases, the heat rate curve is sufficiently linear that only a single piece can be used. To aid in minimizing the computation complexity of mixed-integer models using the data (while retaining as much accuracy as possible), we have also provided a data set that offers a recommended number of pieces. This number was determined using the Akaike Information Criteria[31], which is described in the SI. An example application of a piecewise-linear input/output curve for a gas-fired CT generator, along with the corresponding heat rate curve, is also provided in the SI.

In addition to the raw and processed data, and piecewise-linear fits, we also fit a fourth-order polynomial to the filtered data set for each generator.[32] While these curves are not intended to represent the fundamental physics of generator performance, they may be useful for users who want a continuous function.

Finally, we also generated a set of generic heat rate curves that can be used to represent plants with missing or poorly represented data, or new plants for modeling future scenarios that include new builds. For each unit of a specific type, we took the median of 15 evenly spaced points along the normalized heat rate curve for all units (from minimum to maximum). These median heat rate values were used to generate a composite shape for a plant of a specific class. The heat rate curve can then be adjusted up or down to fit a specific point (such as the heat rate at full output) on the new plant. These generic curves are also posted with the full data set described in the SI.

We have posted cleaned data and processed results at a repository with details described in the SI. Data for each generator includes minimum and maximum generation levels, piecewise-linear fit parameters, and coefficients for our fitted fourth-order polynomial heat rate curve. The minimum and maximum generation levels were taken as the minimum and maximum load point after preprocessing and filtering. The heat rate curve can be translated directly into a carbon-dioxide emissions curve by multiplying the heat content of the fuel by its carbon content.

3. RESULTS AND DISCUSSION

3.1. Results. Table 1 summarizes the processed data and the effect of the various screens. It includes the total capacity and the 2-year generation (energy) value for each generator type.

The data processing retained between 65% and 95% of all plant types except the oil-fired CTs, which typically run very infrequently. Additional summary statistics of the number of plants and data points removed are provided in the SI. Overall, the cleaned and processed data set captures about 71% of thermal units with greater than 25 MW of capacity based on the 2017 EIA 860 data. Our data set also captures about 71% of annual generation from all combustion generation sources based on 2016−2017 Form EIA-923 data. One contributing factor is the elimination of a large number of CC plants that did not report steam-cycle generation and were therefore eliminated. Overall, limitations of the CC data, including the challenges of mapping CC units between EPA and EIA data sets, introduced a number of uncertainties and reduced the amount of data we could process with a high degree of certainty.

Figure 4 shows example processed heat rate curves for six units that demonstrate some of the range of results in data quality; results for all units can be observed via the links

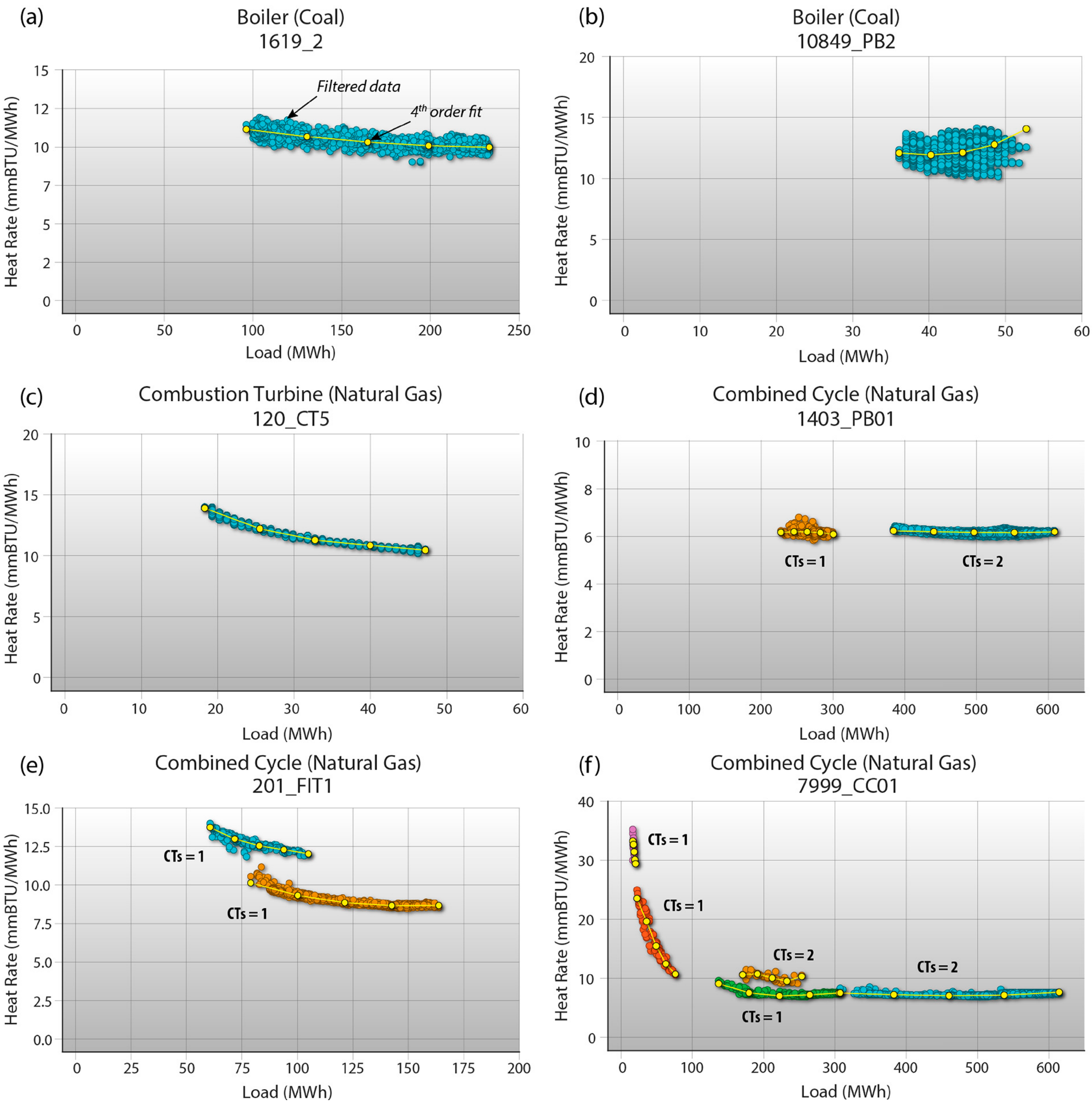


Figure 4. Example of processed heat rate curves.

provided in the SI. The coal unit in Figure 4a (1619_2, Brayton Point) shows a good characteristic shape and fairly tight band around the fitted curve, while the coal unit in Figure 4b (10849_PB2, Silver Bay) shows a limited operating range and large band resulting in a less clear data fit. The example CT in Figure 4c (120_CT5, Yucca) shows one of the better fits with little deviation from the trend. We show two 1 × 1 CC plants, with the first (Figure 4d, 1403_PB01, Nine Mile Point) demonstrating operation only in CC mode and the second (Figure 4e, 201_FIT1, Thomas Fitzhugh) showing both CC and single-cycle operation 1 × 1 operation. Finally, we show a 2 × 1 CC plant showing multiple modes of operation (Figure 4f, 7999_CC01, Grays Harbor).

Results from the data demonstrate the importance of considering the impacts of part-load operation on generator (and system) performance. Figure 5 summarizes power plant operating regime by type for all final processed units (the final column in Table 1). For each plant, we divided operation into quartiles, with each quartile defined as one-fourth of the range between minimum and maximum output, as determined by our data processing method, with Q1 representing the quartile with lowest output and Q4 the highest. We then calculated the fraction of annual energy produced in that quartile shown in the left column and the fraction of operating hours in that quartile shown in the right column. Results are weighted by generation for energy and by capacity for the percentage of hours. For example, as seen in Figure 5a, the results show that

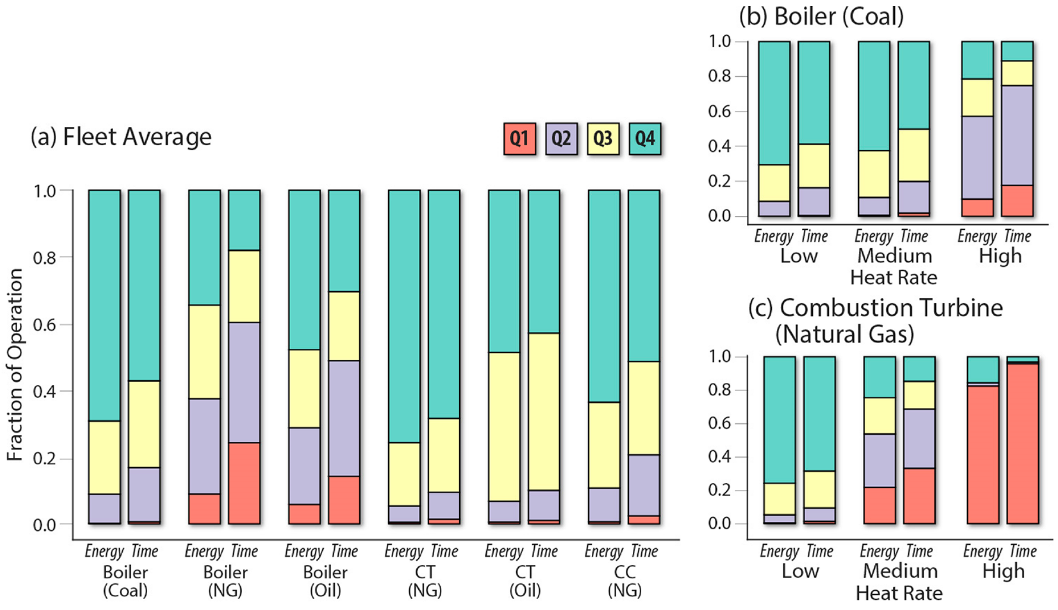


Figure 5. Summary of operating modes. First (left) bars indicate fraction of energy produced, while second (right) bars indicate fraction of operating time in each load quartile.

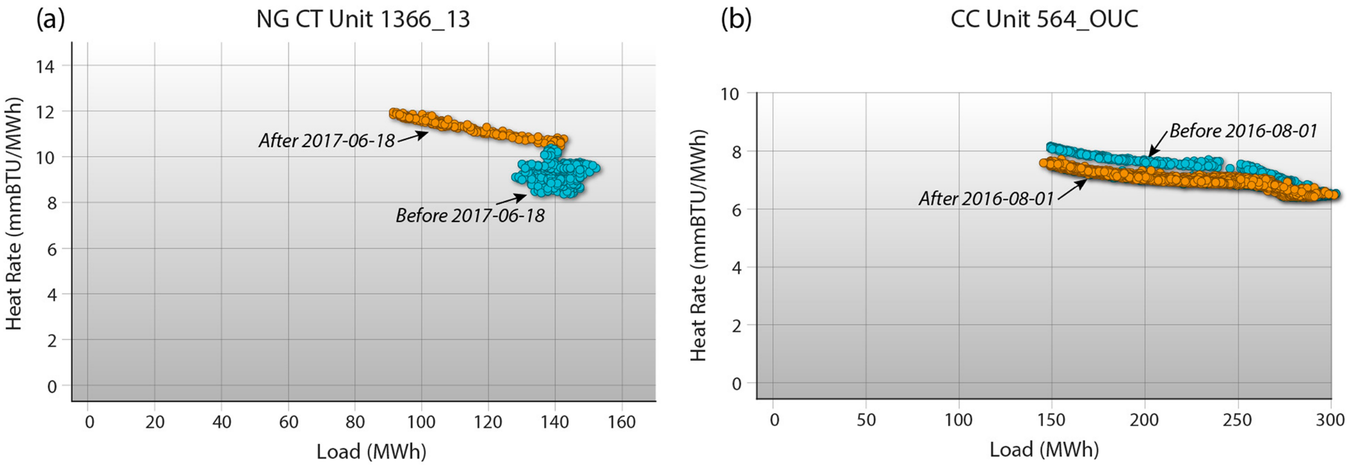


Figure 6. Data indicating changes to plant operation or measurement.

even traditionally "baseload" coal plants spend less than half of their operating hours generating within 25% of full output (Q4), although they produce about 55% of their energy operating at or near full output. Alternatively, gas-fueled combustion turbines only produce 30% of their energy operating in nearly full output.

Figure 5b,c examines the general relationship between plant efficiency and frequency of part-load operation for the coal and gas CT plants by dividing the fleet into three equal-capacity heat rate bins. These results demonstrate that the most efficient (i.e., lowest heat rate) units spend a higher fraction of their operating hours at higher output. As indicated in Figure 5b, the most efficient coal generators (which would be near the bottom of the dispatch stack after zero- or very-low-fuel cost plants) generate close to 70% of their energy (light bars) while operating close to full output.

Figure 5 also shows the results for CC generators, demonstrating that they spend a smaller fraction of their time at very low output (see quartile 1). However, this result must be placed in the context that this quartile typically represents operation using only CT generation. We also examined the CC data to identify operation in its nominal CC operating mode (i.e., all CTs operating plus ST), and the fleet average generation in nominal operation was about 86% of annual energy and 79% of operating hours. While CC plants generate most of their energy in CC mode, the frequency of operating in CT mode, combined with the flexibility of CT mode operation (including rapid startup and ramp rates), indicates the importance of capturing this option in studies of increased VG penetration. Previous studies demonstrate greater cycling and part-load operation as a function of VG deployment, which could include greater use of CT-only operation in CC plants.[3] This points to the value of improved accuracy in capturing part-load operation across the spectrum of planning and simulation tools used to analyze future grid operations, particularly under scenarios of greater net load variability.

3.2. Discussion. Energy modelers in the United States are fortunate to have a large volume of free, publicly available data sets describing many characteristics of the power plant fleet.

To date, however, power plant efficiency curves have been an important missing component of the data. These curves can be reproduced from the EPA CAM database, but it is noisy and includes clearly erroneous data. Furthermore, the CAM database can be used for a variety of applications beyond generating heat rate curves, and the data has been used in several scholarly works. Some of this work has policy implications, so it is important that the data be used in a transparent and reproducible manner. Our goal was to mine the CAM database for historic operation data and provide the energy modeling community with a processed database, along with a documented tool to remove nonrepresentative data or errors. While we focus on heat rate curves, the processed data set and tools could be used for additional applications such as improved analysis of nitrogen oxides and sulfur dioxide emissions rates in relationship with pollution control equipment, fuel type, and part-load operations. Additional work could examine emissions associated with startup operations, while considering the accuracy of the CEMS data under these conditions. The cleaned data sets can also be used to track trends in power plant chronological operation in response to fuel prices, emissions controls, power plant fleet changes (including renewable deployment), and other policies and market conditions. They also provide an additional method to screen data sets for measurement errors.

As an example, the processed data sets make it easier to identify either real operational changes or potential measurement errors during certain time periods. Figure 6 shows the final processed data for two units with unusual shapes. Figure 6a shows the data for a CT unit (1366_13, Paddys Run), which shows a cluster of data that clearly differs from the general trend. Upon further analysis, all of this cluster occurs before June 16, 2018, suggesting a possible issue with the CEMS equipment before this date and thus suggesting avoiding use of this part of the data set. Alternatively, some units show more subtle differences that still suggest changes in either measurement or actual operation but without an easy "algorithmic" method to reject or further filter. For example, Figure 6b (564_OUC, Stanton Energy Center) shows the data for a 1 × 1 CC unit showing two distinct bands of operation in blue (the orange dots obscure the lower band of blue) but transitioning to a single band on August 1, 2016. Both curves follow a well-defined shape, but the average heat rate after August 1, 2016 is about 10% lower than before. This might suggest a power plant improvement, or other changes, although the change occurs from 1 day to the next (without a multiday outage), precluding a major plant upgrade. We found similar shifts for other CC units in which heat rates measurably change (both up and down) after a certain time. These differences all suggest careful consideration of the raw data or use of simple averages to project future conditions and warrant further study to identify best practices for use of historical CAM database for analysis of the evolving electric sector.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.est.9b04522.

> CAM data and DBSCAN processing methods, description of processed data files, and links to raw and processed data (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: michael.rossol@nrel.gov.

ORCID

Michael Rossol: 0000-0001-9351-8404

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was authored by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Wind Energy Technologies Office. Additional thanks to Charlton Clark and Aaron Bloom for support, Devonie McCamey for editorial assistance, and Billy Roberts for artistic assistance. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government.

REFERENCES

(1) Cochran, J.; Mai, T.; Bazilian, M. Meta-analysis of high penetration renewable energy scenarios. *Renewable Sustainable Energy Rev.* **2014**, *29*, 246−253.

(2) Helisto, N.; Kiviluoma, J.; Holttinen, H.; Lara, J. D.; Hodge, B.-M.Including operational aspects in the planning of power systems with large amounts of variable generation: A review of modeling approaches*Wiley Interdiscip. Rev.: Energy Environ.*20198

(3) Lew, D., Brinkman, G., Ibanez, E., Florita, A., Heaney, M., Hodge, B.-M., Hummon, M., Stark, G. *Western Wind and Solar Integration Study Phase 2*. National Renewable Energy Laboratory: Golden, CO, 2013. NREL/TP-5500−55588; https://www.nrel.gov/docs/fy13osti/55588.pdf. 5 Dec 2018.

(4) Holttinen, H. *Expert Group Report on Recommended Practices 16. Wind Integration Studies*. IEA Wind. 2013.https://community.ieawind.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=35029087-52c3-52eb-c2d5-3a60b458ba5c&forceDialog=0. 1 July 2019.

(5) Schill, W.-P.; Pahle, M.; Gambardella, C. Start-up costs of thermal power plants in markets with increasing shares of variable renewable generation. *Nat. Energy* **2017**, *2*, 17050.

(6) Bloom, A., Townsend, A., Palchak, D., Novacheck, J., King, J., Barrows, C., Ibanez, E., O'Connell, M., Jordan, G., Roberts, B., Draxl, C., Gruchalla, K. *Eastern Renewable Integration Study*. National Renewable Energy Laboratory: Golden, CO, 2016. NREL/TP-6A20−64472; https://www.nrel.gov/docs/fy16osti/64472.pdf. 5 July 2019.

(7) Denholm, P.; Brinkman, G.; Mai, T. How Low Can You Go? The Importance of Quantifying Minimum Generation Levels for Renewable Integration. *Energy Policy* **2018**, *115*, 249−257.

(8) Lew, D.; Brinkman, G.; Kumar, N.; Lefton, S.; Jordan, G.; Venkataraman, S. Finding Flexibility: Cycling the Conventional Fleet. *IEEE Power and Energy Magazine* **2013**, *11*, 20−32.

(9) Bistline, J. E. Turn Down for What? The Economic Value of Operational Flexibility in Electricity Markets. *IEEE Transactions on Power Systems* **2019**, *34* (1), 527−534.

(10) Katzenstein, W.; Apt, J. Air Emissions Due to Wind and Solar Power. *Environ. Sci. Technol.* **2009**, *43* (2), 253−258.

(11) Mills, A.; Wiser, R.; Milligan, M.; O'Malley, M. Comment on "Air Emissions Due to Wind and Solar Power. *Environ. Sci. Technol.* **2009**, *43* (15), 6106−6107.

(12) Katzenstein, W.; Apt, J. Response to Comment on "Air Emissions Due to Wind and Solar Power. *Environ. Sci. Technol.* **2009**, *43* (15), 6108−6109.

(13) Form EIA-860 detailed data with previous form data (EIA-860A/860B); https://www.eia.gov/electricity/data/eia860/. 9 Dec. 2018.

(14) Form EIA-923 detailed data with previous form data (EIA-906/920); https://www.eia.gov/electricity/data/eia923/. 11 Dec. 2018.

(15) FERC Form 1 - Electric Utility Annual Report - Data (Current and Historical); https://www.ferc.gov/docs-filing/forms/form-1/data.asp. 10 Dec. 2018.

(16) Li, M.; Smith, T. M.; Yang, Y.; Wilson, E. J. Marginal Emission Factors Considering Renewables: A Case Study of the U.S. Midcontinent Independent System Operator (MISO) System. *Environ. Sci. Technol.* **2017**, *51* (19), 11215−11223.

(17) Thind, M. P. S.; Wilson, E. J.; Azevedo, I. L.; Marshall, J. D. Marginal Emissions Factors for Electricity Generation in the Midcontinent ISO. *Environ. Sci. Technol.* **2017**, *51* (24), 14445−14452.

(18) Siler-Evans, K.; Azevedo, I. L.; Morgan, M. G. Marginal Emissions Factors for the U.S. Electricity System. *Environ. Sci. Technol.* **2012**, *46* (9), 4742−4748.

(19) Abel, D.; Holloway, T.; Kladar, R. M.; Meier, P.; Ahl, D.; Harkey, M.; Patz, J. Response of Power Plant Emissions to Ambient Temperature in the Eastern United States. *Environ. Sci. Technol.* **2017**, *51* (10), 5838−5846.

(20) Farkas, C. M.; Moeller, M. D.; Felder, F. A.; Henderson, B. H.; Carlton, A. G. High Electricity Demand in the Northeast U.S.: PJM Reliability Network and Peaking Unit Impacts on Air Quality. *Environ. Sci. Technol.* **2016**, *50* (15), 8375−8384.

(21) U.S. Department of Energy: Grid Modernization Initiative; https://www.energy.gov/grid-modernization-initiative. 15 July 2019.

(22) Environmental Protection Agency (EPA): Air Markets Program Data; https://ampd.epa.gov/ampd/. 18 July 2019.

(23) Part 75 of Volume 40 of the Code of Federal Regulations (CFR); https://www.ecfr.gov/cgi-bin/retrieveECFR?gp=1&SID=287870523535af49d3562aec528d94c9&ty=HTML&h=L&n=40y17.0.1.1.4&r=PART. 5 Dec. 2018.

(24) *Plain English Guide to the Part 75 Rule*. U.S. Environmental Protection Agency. Clear Air Markets Division: Washington, DC. June 2009.

(25) Environmental Protection Agency (EPA): Greenhouse Gas Inventory Guidance Direct Emissions from Stationary Combustion Sources. 2016; https://www.epa.gov/sites/production/files/2016-03/documents/stationaryemissions_3_2016.pdf. 1 July 2019.

(26) El-Wakil, E. *Powerplant Technology*. McGraw-Hill: October 2002.

(27) Nowling, U. Understanding Coal Power Plant Heat Rate and Efficiency. *Power Magazine* **2015**; https://www.powermag.com/understanding-coal-power-plant-heat-rate-and-efficiency/. 10 Dec. 2018.

(28) Korellis, S.; Dene, C. Evaluating the Use of CEMS for Accurate Heat Rate Monitoring and Reporting. *Power Magazine* **2016**; https://www.powermag.com/evaluating-use-cems-accurate-heat-rate-monitoring-reporting/. 12 July 2019.

(29) Ester, M.; Kriegel, H. P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 1996.

(30) Magnani, A.; Boyd, S. P. Convex piecewise-linear fitting. *Optim Eng.* **2009**, *10*, 1−17.

(31) Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19* (6), 716−723.

(32) Numpy.polyfit. The SciPy community. 2019; https://docs.scipy.org/doc/numpy/reference/generated/numpy.polyfit.html. 20 July 2019.

DOI: 10.1021/acs.est.9b04522