ENVIRONMENTAL RESEARCH
LETTERS

Evaluating the hourly emissions intensity of the US electricity system

ENVIRONMENTAL RESEARCH
LETTERS

CrossMark

OPEN ACCESS

RECEIVED
18 November 2022

REVISED
2 March 2023

ACCEPTED FOR PUBLICATION
3 March 2023

PUBLISHED
3 April 2023

LETTER

Evaluating the hourly emissions intensity of the US electricity system

Gregory J Miller[1,2,*], **Gailin Pease**[2], **Wenbo Shi**[2] **and Alan Jenn**[3]

[1] Energy Systems, Energy and Efficiency Institute, University of California, Davis, Davis, CA, United States of America
[2] Singularity Energy, Inc., Somerville, MA, United States of America
[3] Institute of Transportation Studies, University of California, Davis, Davis, CA, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: grmiller@ucdavis.edu

Keywords: GHG emissions, air pollution, carbon intensity, electricity systems, power generation, climate policy, open data

Abstract

High-quality data for the greenhouse gas and air pollution emissions associated with electricity generation and consumption are increasingly important to enable effective and targeted action to decarbonize the electric grid and to inform research in a range of academic disciplines including environmental economics, lifecycle assessment, and environmental health. To inform the broadest range of use cases, such data should ideally have a high temporal and spatial resolution, be available in as close to real-time as possible, represent the complete power sector, use the highest-quality measured data, have complete historical coverage, and represent both generated and consumed emissions. To date, no published datasets have achieved all of these characteristics. This work is the first to publish a comprehensive, plant-level dataset of hourly-resolution generation, fuel consumption, and direct CO_2, NOx, and SO_2 emissions for the entire U.S. power sector. This data is published as part of the public and open-source Open Grid Emissions Initiative, which also includes hourly, consumption-based emissions intensities for every grid balancing area in the country. Using insights generated by this new dataset, this paper also interrogates how several of the assumptions implicit in the use of existing power sector emissions datasets may under-count or misrepresent the climate and health impacts of air emissions from the U.S. power sector. We envision the Initiative becoming a central repository of, and hub of activity for addressing open research questions related to power sector emissions data, and the go-to source for high-quality, high-resolution data for future research on grid emissions.

1. Introduction

Accurate, comprehensive, and high-resolution data for tracking power sector emissions is increasingly important for climate policy and voluntary climate action. An increasing number of policies and regulations, including New York City's Local Law 97 and the U.S. Security and Exchange Commission's proposed rule on climate risk disclosure, depend on accurately tracking the greenhouse gas (GHG) emissions that each consumer of electricity is responsible for. Additionally, a record number of actors, including corporations, cities, and other institutions, have made voluntary decarbonization pledges that require them to inventory and track their emissions. However, recent research has shown that the emissions factors that have historically been available to inform these efforts may be inadequate to meet the needs of today's electricity consumers. The main public emissions factor datasets for the U.S. include annual-average emissions factors that reflect the emissions intensity of generated electricity, but do not provide any information about how the emissions intensity of consumed electricity varies across time. Consumption-based values are of particular importance for 'Scope 2' inventories of indirect emissions from electricity use. Due to interregional power flows, the emissions intensity of consumed electricity does not always match the average emissions intensity of the regional generation fleet, and imported electricity can account for 20%–40% of the emissions consumed in a region [1, 2]. Additionally, hourly-resolution data is increasingly important to accurately account for consumed emissions due to the temporal variation in both emissions intensity and load [3].

Comprehensive hourly emissions data is also important for academic research. Over the past decade, many studies in the lifecycle assessment and environmental economics literature that assessed the consequential emissions impact of electric vehicles and renewable energy deployment, as well as the broader study of marginal emissions, have relied on hourly generation and emissions data from the EPA Clean Air Markets Division's (CAMD) continuous emissions monitoring system (CEMS) dataset [4–18]. The major limitation of this dataset is that it only covers fossil-fueled generators >25 MW nameplate capacity, some of which only report data for part of the year. The authors of these studies have generally assumed that generators that did not happen to report data to CEMS could not be on the margin, or represented an insignificant proportion of generation in the regions under analysis, although some authors acknowledge that this is not always an appropriate assumption [5, 9, 15]. Another limitation of the CEMS dataset is that it only reports hourly *gross* generation, and not *net* generation. Net generation represents the electricity injected into the grid after accounting for plant-specific losses, which typically range between 3% and 20% of gross generation. Many previous studies using this dataset used the gross generation data directly in their calculations of emission factors [5, 6, 13–15].

Our review revealed no existing published dataset of comprehensive, high-quality, and hourly-resolution emissions data from consumed grid electricity. Existing hourly datasets are either incomplete (CEMS) or are estimates that have not yet been validated based on high-quality measured or reported data. Existing comprehensive and high-quality datasets, such as the EPA's emissions and Generation Resource Integrated Database (eGRID) database, publish low-resolution, annual data. The primary barrier to a dataset that is both comprehensive and high-resolution is that EIA's Form 923, which is used to fill generation and fuel consumption data that is missing from CEMS, is available at the monthly and annual resolutions. Thus, overcoming this challenge requires a robust method for accurately imputing the hourly profile of monthly and annual EIA-923 data.

This paper introduces the Open Grid Emissions (OGE) Initiative, whose goal is to provide accurate, comprehensive, and hourly-resolution public data that represents both direct emissions generated by the power sector and emissions from consumed grid electricity. This paper describes multiple improvements to and novel applications of existing methods for inventorying the generation, fuel consumption, and emissions from the U.S. power sector, as well as a novel set of methods for imputing the hourly values for generators that only report data at the monthly resolution. The initiative uses publicly available data from the U.S. EPA and EIA as inputs and its data, code, and methodological documentation are freely available at the OGE Initiative website, on GitHub, and archived on Zenodo [19–21]. We believe that the OGE dataset is the most complete, most accurate, and highest resolution dataset of U.S. power sector emissions and electricity emissions factors available to date [22]. To our knowledge, it is also the first comprehensive hourly dataset of NO_X, SO_2, CH_4, and N_2O emissions from the U.S. power sector, and of total CO_2 emissions resulting from electricity generation.

1.1. Background and literature review

To date, most publicly accessible datasets of power sector emissions and electricity emissions factors do not include consumed emissions factors or hourly-resolution data. The U.S. EPA's eGRID is the oldest and most comprehensive dataset of power sector emissions, primarily relying on measured emissions data from CEMS [23]. However, eGRID data is published at an annual resolution. In an attempt to reflect consumption-based emissions, the EPA also aggregates its data into 'eGRID subregions', the boundaries of which are defined to limit the import and export of electricity, but do not explicitly account for power flows between balancing areas [24]. Likewise, the U.S. Energy Information Administration's published 'Emissions by plant and region' dataset, which relies on fuel consumption data reported to EIA Form 923, only includes generated emissions factors at an annual resolution [25]. Because of delay involved in collecting and verifying the data that serve as the inputs to these two datasets, another factor that limits their potential usefulness is that they are released on a 1–2 year lag [24, 25].

Power sector datasets from the EPA and EIA include some data from combined heat and power (CHP) facilities, which produce both electricity and useful thermal output for applications such as district heating or industrial steam. Thus, in order to accurately represent emissions resulting from electricity generation, and provide emission intensities that can be used for inventorying Scope 2 emissions from electricity consumption, these datasets apply an adjustment to exclude the fuel consumed (and thus emissions produced) for non-electricity purposes at CHP plants.

However, many of these existing datasets also explicitly or implicitly include an adjustment that treats the combustion of biomass and biogas as having zero direct atmospheric emissions of CO_2, even though these fuels have direct combustion emission factors (lb CO_2 per mmBTU of fuel) comparable to those of fossil fuels [24]. This biomass adjustment originated in the earliest (1996) version of eGRID based on the assumption that combusting biomass 'do[es] not contribute to global warming,' and since 2014 has continued to be used 'for reasons of consistency' with

Table 1. Comparison of existing publicly accessible sources of power sector emissions and electricity emissions factor data.

		eGRID [23]	EIA emissions by plant & region [25]	Power sector carbon index [26]	U.S. carbon monitor [27, 28]	Grid-emissions project [1, 29]	EIA hourly electric grid monitor [30]	Open grid emissions initiative [20, 21]
Temporal Resolution	Hourly					✓	✓	✓
	Daily				✓		✓	
	Monthly			✓				✓
	Seasonal	✓		✓				
	Annual	✓	✓	✓				✓
EF	Generated	✓	✓	✓	✓	✓	✓	✓
	Consumed					✓	✓	✓
Data Sources	CEMS	Annual		Hourly				Hourly
	EIA-923	✓	✓	✓				✓
	EIA-860	✓	✓	✓				✓
	EIA-930				✓	✓	✓	✓
CO_2 emissions data	Non-biomass electricity & heat		M, I					M
	Total electricity and heat	M						M
	Non-biomass electricity	M, I		M, I	M		M, I	M, I
	Total electricity					M, I (LC)		M, I
Pollutants Tracked	CO_2	✓	✓	✓	✓	✓ (LC)	✓	✓
	CH_4	✓						✓
	N_2O	✓						✓
	CO_2-eq.	✓						✓
	NO_X	✓	✓					✓
	SO_2	✓	✓					✓
	Hg	✓						
Spatial Aggregation	Plant	✓	✓					✓
	Balancing area	✓				✓	✓	✓
	Other region	✓	✓	✓			✓	
	State	✓	✓	✓	✓			
	National	✓	✓	✓	✓		✓	
Approximate data release lag		13–26 months	10–22 months	3–6 months	1–13 months	Several hours	1 d	11–23 months
Historical coverage as of January 2023		1996–2021	2013–2021	2001–September 2022	2019–2021	July 2018–January 2023	July 2018–January 2023	2019–2021

Abbreviations: **M** = mass; **I** = intensity; **LC** = lifecycle.

prior versions, even after the EPA's Science Advisory Board recommended that 'carbon neutrality cannot be assumed for all biomass energy *a priori*,' a finding consistent with the wider academic literature on the topic [24, 31–42]. Many of the other datasets in table 1 have propagated this adjustment in their calculations. However, because most of these datasets are designed to provide data on direct (rather than lifecycle) emissions from electricity generation, the use of this biomass adjustment may be systematically underrepresenting direct CO_2 emissions from power generation and electricity consumption in the U.S. Not only is this adjustment an inaccurate representation of direct CO_2 emissions, but it is also inconsistent with the treatment of all other fuels in these datasets (for which upstream fuel-cycle emissions impacts are not considered). Thus, we believe that OGE is the first and only dataset to date of *total*, direct CO_2 emission factors from electricity generation in the United States.

Several recent academic efforts have improved upon certain limitations of the existing EPA and EIA data. The Power Sector Carbon Index, based on 2018 research by Schivley *et al*, now publishes monthly-resolution power sector CO_2 emissions data on only a 3–6 month lag [26]. Because this dataset is focused on tracking long-term trends in power sector carbon emissions, it does not include consumed emissions factors or data for emissions other than CO_2. Work by Chalendar *et al* in 2019 led to the creation of the 'gridemissions' project, which was the first publicly-available dataset of both generated and consumed, hourly lifecycle CO_2 emissions factors for the U.S., published on only a several-hour lag [1, 29, 43]. Instead of relying on measured emissions data and reported fuel consumption data, the gridemissions project makes use of a relatively new source of near-real-time hourly generation and interchange data available for each balancing authority through EIA Form 930 [44]. The U.S. Carbon Monitor also uses this EIA-930 data to publish daily-resolution estimates of total CO_2 emissions from electricity generation in the U.S [27, 28]. At the end of 2022, the EIA also began to publish near-real-time estimates of generated and consumed CO_2 emissions data as part of its Hourly Electric Grid Monitor, adapting the methodology introduced by the gridemisisons project [30]. There are several additional commercial datasets of hourly, consumed grid carbon intensity that use similar data and methods to the gridemissions project, so are not discussed in this paper [45, 46].

To date, there has been no way to validate how well these near-real-time estimates based on EIA-930 data reflect actual hourly emissions from the power sector. The reported EIA-930 data includes multiple data-quality issues, which have improved over time but continue to affect the accuracy of resulting emissions estimates [47]. These issues include incorrect reporting of the prevailing local time of datapoints, accounting discrepancies in reported interchange values, inconsistent categorization of generators into fuel categories or balancing authorities, and missing data. Additionally, the emissions factors used to convert net generation to emissions are per-fuel, historical annual averages and may not reflect the current, time-varying emissions intensity of specific regional fleets. This means that while such datasets might be useful estimates for real-time operational decision making, they might be of high enough quality to meet the accuracy criteria set by GHG accounting standards for emissions inventories [48].

2. Methods and data

This dataset relies on combining multiple sources of data including the EPA's CEMS data and data from EIA Forms 860, 923, and 930, as illustrated in figure 1. Because these datasets are released in non-standardized formats, sometimes contain incomplete or anomalous data, and can be challenging to cross-link, we build upon several existing open-source projects for standardizing and cleaning these data. The first of these is Catalyst Cooperative's Public Utility Data Liberation (PUDL) project, which provides standardized and unified relational databases of the CEMS, EIA-860, and EIA-923 data [49]. Because EPA and EIA datasets do not always use consistent plant identifier codes or units of analysis, we also rely on the EPA's Power Sector Data Crosswalk (PSDC) to link these datasets together [50]. Finally, the raw EIA-930 data includes data quality issues, so we use a framework from Chalendar and Benson 2021 that reconciles the data with physical energy laws and we adapt methods from Ruggles *et al* to screen anomalous timeseries values [29, 51].

The CEMS dataset reports measured hourly emissions mass for CO_2, NO_X, and SO_2 for large fossil fuel generators. Emissions data for CH_4 and N_2O, and for all pollutants from generators that do not report data to CEMS are calculated based on reported fuel consumption, boiler design parameters, emissions control equipment, and direct (non-lifecycle) emissions factors. All methodologies and assumptions for calculating these air emissions, and for adjusting emissions totals for CHP or biomass generation, are adapted from the U.S. EPA's eGRID methodology [24]. Once emissions totals are calculated, generated emissions factors are calculated by dividing emissions mass by net generation. To convert hourly gross generation data from CEMS to hourly net generation, we calculate a subplant or plant-specific gross-to-net ratio by comparing the monthly net generation reported in EIA-923 to the monthly total gross generation reported in CEMS [9, 17, 52]. Consumption based emissions are modeled using a multi-region input output model, based on calculated generation and emissions rates, and reported hourly interchange values between regions in EIA-930 [1]. While our consumed emission estimates reflect interregional emissions flows, they do not currently reflect the impact of transmission and distribution losses between the point of generation and consumption.

Although this research includes many smaller, incremental improvements to existing methods (documented in the supporting information), its major contributions are methods for accurately crosswalking EIA and EPA datasets at the subplant level and imputing the hourly profile of monthly-resolution EIA data.

2.1. Crosswalking subplant data from multiple sources

Generation, fuel consumption, and emissions data from each input data source is reported at multiple different levels of aggregation: EIA-923 reports some data at the boiler level (where the fuel is combusted and steam produced), some data at the generator level (where the electricity is generated), and some data at the prime-mover level (the type of generator), while the CEMS data is reported at the unit level (which represents a collection of boilers and smokestacks) [50, 53]. Complicating this is that the EIA also uses the term 'unit' to describe multiple generators that operate together, such as in a combined-cycle unit [54]. Sometimes boilers, generators, and EPA units are related in simple one-to-one relationships, but in other cases these units and generators can be configured in complex many-to-many relationships.

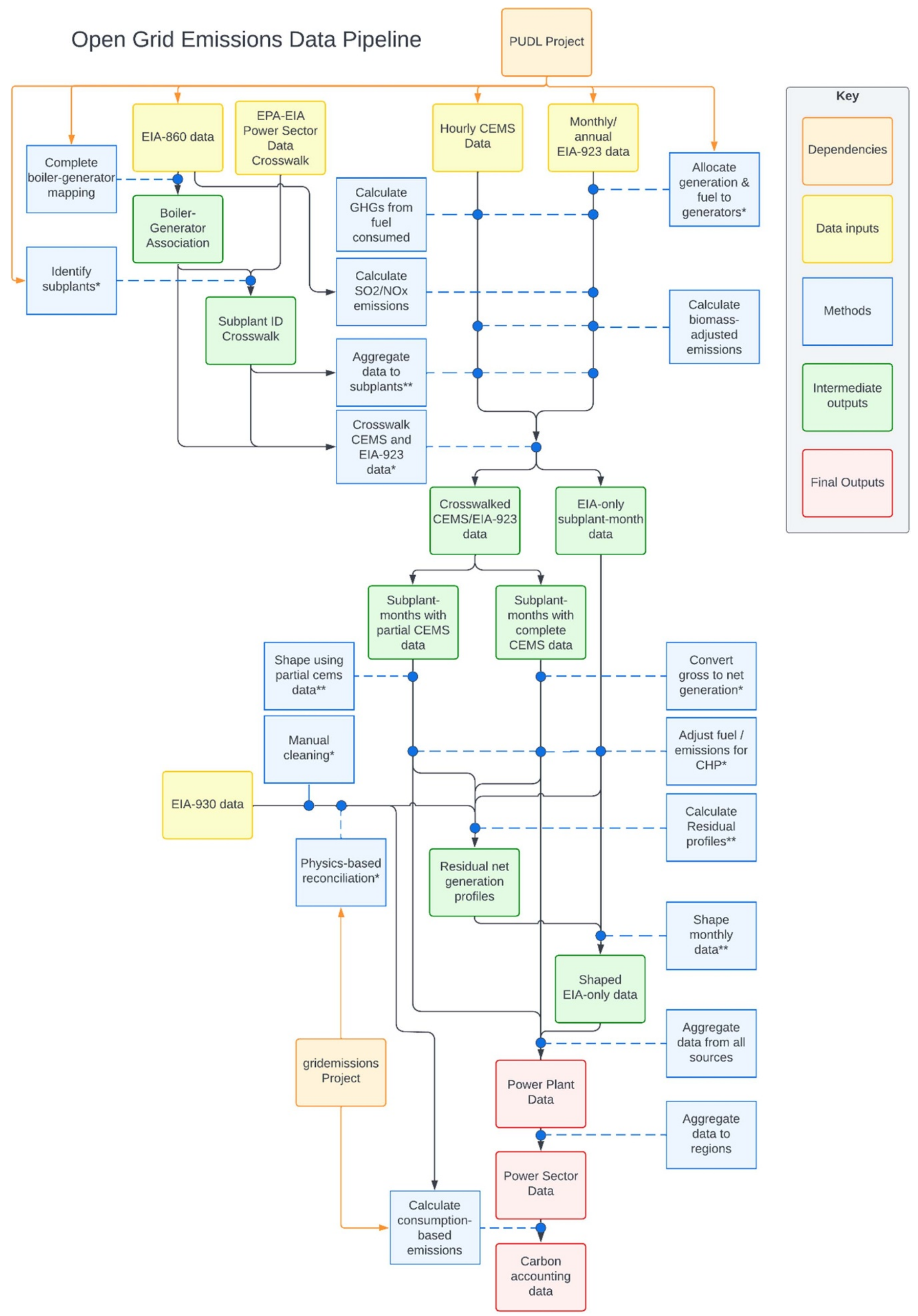

Figure 1. Flow diagram of the major steps of the open grid emissions data pipeline. A method with no asterisk indicates that it was an existing method adapted with no or insignificant modification. A single asterisk indicates an existing method that the authors significantly modified or updated. A double asterisk indicates a novel method introduced for this project.

Accurately matching these data from each source is crucial for identifying data gaps, calculating conversion factors (such as gross-to-net generation ratios), and ensuring that we are not double-counting data. Simply aggregating the data to the plant level is not always a good approach since data for certain parts of a plant may be missing from a dataset, or because different parts of a plant have different operational characteristics (such as a backup diesel generator located at a nuclear power plant).

In order to perform the most granular and accurate crosswalking of the data from each source, we assign each plant part a unique 'subplant ID' based on its relationship to all other units, generators, and boilers at a plant, based on the EPA unit to EIA generator associations in the PSDC, the boiler-generator associations in EIA-860, and supplemental associations added in the PUDL dataset [55]. The PUDL Project introduced a method that uses network analysis to group boilers and generators into 'PUDL units' by identifying groups of connected subgraphs based on their relationships in EIA-860 [56]. We expand this method to use the more complete relationships between boilers, generators, EIA units, and EPA units to identify subplants, which represent the smallest unit of analysis that can be used to confidently crosswalk data from each source. This work also improves upon existing approaches by crosswalking the data at the monthly temporal resolution, rather than the annual resolution, which allows us to more accurately crosswalk data for plants that only report to CEMS for part of a year.

2.2. Imputing hourly profiles for monthly-reported data

A novel contribution of this work is a method for imputing the hourly profile of the monthly-resolution data for plants that do not report to CEMS. This method includes a set of three broad approaches that are applied to the monthly data depending on the most specific observed hourly data that is available for each subplant.

The first two approaches apply to fossil-fueled plants where an incomplete subset of units report hourly data to CEMS. If only a subset of units within a subplant report hourly data to CEMS, we use that incomplete hourly data to shape the complete monthly data for the complete subplant data from EIA-923. If a subset of subplants within a plant report hourly data to CEMS, we use the combined hourly profile of all CEMS-reporting subplants to shape the monthly EIA-923 data for the subplant(s) that do not report to CEMS. These two approaches assume that the operational profile of different units within a single subplant, or of different subplants within a single plant will be similar. While believe that this approach is generally more accurate than using fleet-average profiles, this assumption may not always be accurate [57]. We found that in 2020, the average correlation between the hourly profiles of different CEMS units in the same subplant was 0.67, and between different subplants at the same plant was 0.39.

For plants that do not report any hourly data to CEMS (which generally includes all clean and renewable generators, as well as any plants smaller than 25 MW nameplate capacity) or for large emitting plants that only report data to CEMS during ozone season (May–September), we can reasonably estimate their aggregate hourly generation profile using observed fleet-wide hourly generation data. Since 2018, the U.S. EIA has collected hourly net generation data by plant primary fuel type (coal, natural gas, petroleum, nuclear, hydro, wind, solar, and other) for each balancing authority in the U.S. as part of their Hourly Balancing Authority Operations Report (Form 930).

To calculate the hourly net generation profile for all subplants that do not report to CEMS in each month, we aggregate the hourly CEMS net generation profile for each fuel category in each region and subtract it from the total hourly net generation for that regional fleet as reported in EIA-930, as shown in figure 2. This residual hourly profile represents the aggregate hourly generation profile of all plants that do not report data to CEMS. The hourly residual profile for each fleet-month is then normalized as a percentage of monthly total net generation for that fleet, and used to shape the monthly total net generation, fuel consumption, and emissions data for all plants in that fleet that do not report to CEMS.

Because this method relies on observed data, we believe this to be the best available method to estimate the hourly profile of these plants. However, one limitation of the EIA-930 data is that it may inconsistently categorize individual plants into fuel categories or balancing areas, sometimes resulting in the aggregated hourly generation from CEMS exceeding the total reported generation from EIA-930. In this case, we shift the CEMS profile down so that the EIA-930 generation is greater than or equal to CEMS generation in all hours, then re-calculate the residual. This approach prevents the residual profile from including negative net generation, while preserving residual shape of the two profiles as much as possible.

For fuel categories that do not report data to CEMS (solar, wind, hydro, nuclear), and in cases where a high-quality residual profile cannot be calculated for a regional fleet, we use the total EIA-930 regional fleet profile, which represents the generation-weighted average profile of all generators in the fleet.

In the small percentage of cases that none of the above methods can be used (see table 2), we fall back on a set of less robust methods. If there is no EIA-930 data, but there is CEMS data that represents at least three different plants in a fleet, the CEMS average profile is used as a proxy. For missing regional wind or solar data, we impute the profile by averaging the generation profiles from neighboring regions in the same time zone, or using national-average data if neighboring region data is not available. This method performs reasonably well for solar (median correlation of $\sim$0.9 with cross-validated profiles for both neighboring and national imputation), but is less robust for wind (median correlation of 0.45 for neighboring

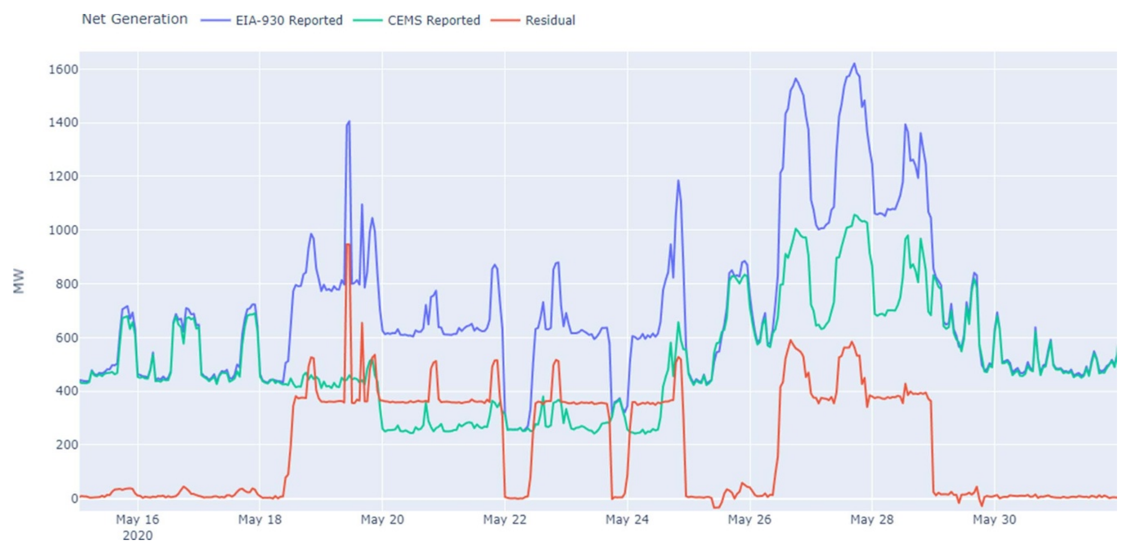

Figure 2. By comparing generation data reported to EIA-930 (blue line) with data reported to CEMS (green line), we are able to determine the generation profile of all plants that do not report data to CEMS (red line). Examining this example of the natural gas fleet in the Balancing Authority of Northern California (BANC) reveals that operational patterns for plants that report to CEMS (in green) differ substantially from plants that do not report to CEMS (in red). The residual profile shown by the red line was used to shape the May generation, fuel consumption, and emissions totals reported in EIA-923 for the natural gas plants in BANC that did not report data to CEMS.

Table 2. Breakdown of the methods used to impute the hourly profile of each data output in 2021, ranked in order from highest quality to lowest quality. Note that the topmost category (CEMS reported) represents actual reported hourly data, not imputed data.

Quality	Source/imputation method	Net generation	Electricity emissions CO_2	NO_X	SO_2
Highest	CEMS reported (not imputed)	58%	93%	65%	89%
↑	Partial CEMS	0%	1%	3%	2%
↓	Residual EIA-930 profile	39%	3%	14%	2%
Lowest	EIA-930 profile	0%	0%	0%	0%
	CEMS-avg profile	0%	1%	5%	2%
	Imputed wind/solar profile	0%	N/A	N/A	N/A
	Assumed flat (monthly avg)	2%	3%	13%	5%

and 0.11 for national), for which there is more spatial variation in profiles. In the case that no hourly data is available for a specific fleet, we apply a flat profile, which is equivalent to using the monthly average value for all hours.

3. Results and discussion

3.1. Validation of OGE data against previous datasets

To validate our results, we compared the annual total results of the OGE dataset for year 2021 with the results of previous datasets, as shown in table 3. Our results for generation and CO_2 emissions are generally consistent with previous estimates. Our SO_2 results are consistent with the EIA's estimate (about 2% higher) but 10% lower than the eGRID estimate for SO_2 emissions. This discrepancy primarily results from the use of different SO_2 emissions factors by the EIA and eGRID (OGE uses the EIA factors).

Our NO_X emission result is 5% higher than both the eGRID and EIA estimates, which we believe primarily results from the ability of our subplant crosswalking method to identify data gaps in CEMS and use EIA-923 data to fill them. For example, both eGRID and the EIA emissions dataset report that the Hardee Power Station in Florida emitted about 2740 lb of NO_X emissions in 2021. However, the crosswalk process revealed that this CEMS data only corresponded with one of five generators at the plant, and that if the emissions from the other four generators are included, the plant's actual NO_X emission total for 2021 is closer to 2.3 million lb—over 85 000% higher than the NO_X emissions reported by any other dataset. These missing emissions can have significant, real-world implications on our understanding of the environmental and health impacts of power generation in specific communities in the U.S.

3.2. CEMS data alone may not be representative of the U.S. power sector

The quality of our output data can be partially judged based on evaluating what portion of the data is derived from input data with the highest temporal

Table 3. Comparison of total annual results from each emissions dataset for 2021. Each cell shows the total value and the percentage difference from the OGE total (in bold).

Metric		OGE	eGRID	EIA annual	U.S. carbon monitor	Power sector carbon index	EIA hourly electric grid monitor
Net generation (TWh)		**4.131**	4.120 (−0%)	4.108 (−1%)	N/A	4.157 (+1%)	3.951 (−4%)
CO_2 (trillion lb)	Power sector	**4.092**	4.052 (−1%)	N/A	N/A	N/A	N/A
	Electricity	**3.608**	N/A	N/A	N/A	N/A	N/A
	Biomass-adj. electricity	**3.502**	3.512 (+0%)	N/A	3.388 (−3%)	3.493 (−0%)	3.343 (−5%)
NO_X (billion lb)	Power Sector	**2.880**	2.746 (−5%)	2.763 (−5%)	N/A	N/A	N/A
	Electricity	**3.388**	N/A	N/A	N/A	N/A	N/A
	Biomass-adj. electricity	**2.380**	2.161 (−9%)	N/A	N/A	N/A	N/A
SO_2 (billion lb)	Power sector	**2.638**	2.898 (+10%)	2.575 (−2%)	N/A	N/A	N/A
	Electricity	**2.104**	N/A	N/A	N/A	N/A	N/A
	Biomass-adj. electricity	**2.103**	2.189 (+4%)	N/A	N/A	N/A	N/A

Table 4. The left part of the table shows the percentage of OGE 2021 output data from each input source, in descending order of data quality from top to bottom. The percentage of net generation is shown for both total net generation and net generation from fuel combustion (Comb.). The column 'n' represents the percentage of the number of data values (subplant-hours) in the dataset. The right part of the table reflects a regional breakdown of the percentage of total data that is included in the CEMS dataset.

Percentage of national data by input source						
		Net Generation		Electricity Emissions		
Data Source	n	Total	Comb.	CO_2	NO_X	SO_2
Hourly CEMS	11%	58%	92%	93%	65%	89%
Monthly EIA-923	21%	34%	7%	6%	18%	9%
Annual EIA-923 only	67%	8%	1%	1%	16%	2%
Mixed CEMS & annual 923	1%	<1%	< 1%	<1%	<1%	<1%

Percentage of regional data reported in CEMS						
		Net Generation		Electricity Emissions		
ISO	n	Total	Comb.	CO_2	NO_X	SO_2
CAISO	6%	36%	69%	64%	5%	7%
ERCOT	18%	58%	89%	93%	61%	99%
ISONE	5%	54%	87%	72%	15%	22%
MISO	10%	65%	91%	94%	71%	90%
NYISO	11%	48%	94%	88%	34%	69%
PJM	14%	59%	96%	95%	69%	95%
SPP	14%	54%	97%	98%	81%	98%

resolution. As shown in table 4, most of the emissions data in the OGE dataset come from measured hourly CEMS data. The remaining gap in this coverage is filled mostly by monthly-resolution data from EIA-923, for which we impute an hourly profile using the method described above. A smaller but significant portion of the data (especially net generation and NOx data) is derived from annually-resolution EIA-923, which as a default is assigned a monthly profile by the EIA based on the profile of similar plants. A very small portion of the results are a mix of CEMS data for certain months, and annual-resolution EIA-923 for other months, which raises the possibility that data for these plants could either be double-counted or under-counted if the annual data is not accurately disaggregated to individual months.

These results are also helpful for understanding to what extent previous studies that relied solely on CEMS data may be mis-characterizing emissions from the U.S. power sector. Our results, shown in table 4, show that although CEMS data covers approximately 90% of all combustion-based generation, CO_2 emissions, and SO_2 emissions nationally, it represents less than 60% of total generation, and less than two-thirds of all NO_X emissions. This data coverage can be substantially worse in specific regions. Many of the largest and most widely studied balancing areas in the U.S., shown on the right side of table 4, exhibit significant data gaps. In CAISO, for example, CEMS data represents less than two-thirds of CO_2 emissions, only one-third of generation, and less than 10% of all NO_X and SO_2 emissions. In 2020, we found that the average 'nonbaseload factor' of generators that do not report data to CEMS was 86%, suggesting that these generators are very likely to be operating on the margin and responding to changes in load. This suggests that previous studies of marginal emissions, many of which relied solely on CEMS data, could be mischaracterizing the consequential emissions impact of electricity consumption by ignoring this subset of generators.

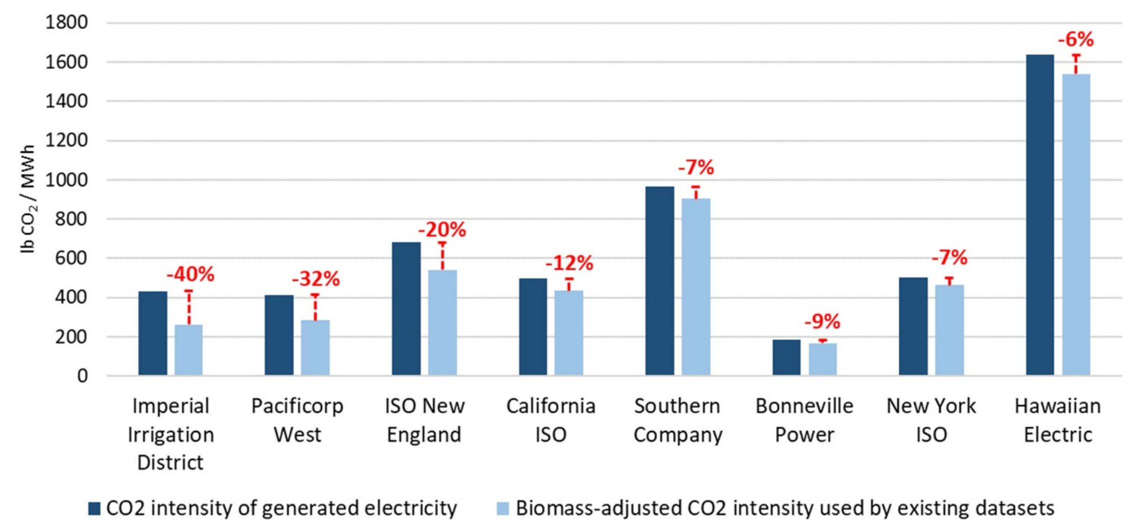

Figure 3. The default biomass-adjusted CO_2 emissions factors published by existing datasets substantially underrepresent the emissions intensity of generated electricity in some of the largest balancing areas in the U.S. This table shows the difference between total electricity emission factors and biomass-adjusted emission factors from OGE in 2021.

3.3. Existing datasets substantially underrepresent CO_2 emissions by treating biomass as carbon-neutral

Our results show that the existing use of the biomass adjustment in existing datasets that excludes CO_2 emissions from biomass combustion underestimates direct CO_2 emissions from electricity generation by 3% nationally. However, the impact can be much larger in specific regions, including several of the largest balancing areas in the U.S., as shown in figure 3. Besides being widely used in academic literature, biomass-adjusted emissions factors from eGRID are widely used for policies, markets, and emissions tracking systems. These uses include the ENERGYSTAR Portfolio Manager (the most-used energy tracking tool for commercial buildings), fueleconomy.gov (the official U.S. government source for vehicle fuel economy information), and any GHG inventories that use eGRID factors [24]. Because biomass-adjusted factors systematically underrepresent direct CO_2 emissions from electricity generation and consumption, this work suggests that their use should be discontinued in these contexts.

4. Conclusion

The research presented in this paper has potentially far-reaching implications for future academic research, GHG accounting, policymaking, and voluntary decarbonization efforts. The OGE Initiative dataset includes hourly, monthly, and annual-resolution data which cover a wide variety of potential use cases: consumed emissions factors, regional power sector generation and generated emissions, and individual power plant data. The consumed hourly emissions factors are applicable to scope 2 GHG accounting, attributional lifecycle assessment studies, and validation of near-real-time estimates of consumed emissions factors. The regional power sector emissions and generation data can be used by policymakers and regulators to track progress toward climate goals, for calculating state or national emissions inventories, or as part of next-generation energy attribute certificates. Finally, the individual power plant data can enable more complete academic research and modeling of the power sector and could be useful to environmental justice advocates for pinpointing hourly point sources of air pollutants in local communities.

The OGE dataset includes many known issues, documented on GitHub, which we anticipate will be addressed over time. However, there are two fundamental limitations of this dataset that may not be possible to address without changes to the data sources on which it relies: its historical coverage and its data lag. Because our hourly imputation method and consumed emission calculations depend on the EIA-930 dataset, which does not include data prior to mid-2018, it would be challenging to extend the dataset's historical coverage of hourly data prior to 2019. However, pre-2019 data could be added to the dataset at monthly and annual resolutions, and at an hourly resolution for a less-complete subset of generation. A second limitation of the dataset is its substantial data availability lag. This lag results from the dataset's reliance on reported data from EIA Forms 860 and 923, final versions of which are published on a 10–22 month lag (i.e. data for all of 2021 was published in October 2022). Although monthly versions of these EIA forms are published on only a 2 month lag, these monthly versions of the data are incomplete because some plants only report data once per year [58].

Although this data lag means that this dataset may not be relevant for real-time management of emissions, the OGE dataset could be used as a benchmark dataset to validate and improve near-real-time estimates of grid emissions [59]. If it could be demonstrated that such real-time estimates are accurate enough, they could adopted more widely to enable more timely analysis, reporting, and regulation of consumed emissions from electricity consumption. Validated and highly-accurate real-time estimates of grid emissions could be particularly useful for managing episodic peak emission events, enabling corporations to disclose and be held accountable for their climate footprint in a more timely manner, and enable carbon-informed demand response.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.5281/zenodo.7689791.

Acknowledgments

The authors would like to thank Travis Johnson and Justine Huetteman of the U.S. Environmental Protection Agency's Clean Air Markets Division for sharing their expertise and helping advance the creation of this dataset through the EmPOWER Air Data Challenge. We would also like to thank Christina Gosnell, Zane Selvans, Austen Sharpe, and Trenton Bush of Catalyst Cooperative for their work on the Public Utility Data Liberation Project, their partnership on the EmPOWER project proposal, and their ongoing partnership on this work. We would like to thank Singularity Energy for funding this research and releasing the initiative as an open-source project. We would like to thank Milo Knowles of Singularity Energy for his contributions to the OGE code on GitHub.

Funding

This research was funded by Singularity Energy, Inc.

Conflict of interest

Author GM received financial compensation for this research from Singularity Energy as an independent contractor, and is currently employed as an employee of Singularity Energy. Author WS is the CEO of Singularity Energy, and author GP is an employee of Singularity Energy.

Supporting Information

Detailed methodological documentation for the dataset can be found at https://docs.singularity.energy/docs/open-grid-emissions-docs

The code repository can be found at https://github.com/singularity-energy/open-grid-emissions and is archived at https://zenodo.org/record/7600707.

The dataset is available to download at https://singularity.energy/open-grid-emissions/ and is archived (including intermediate data outputs) at https://zenodo.org/record/7600513.

ORCID iDs

Gregory J Miller https://orcid.org/0000-0003-3750-9292
Alan Jenn https://orcid.org/0000-0003-4232-0697

References

[1] Chalendar J A, De, Taggart J and Benson S M 2019 Tracking emissions in the US electricity system *Proc. Natl Acad. Sci.* **116** 25497–502

[2] Tranberg B, Corradi O, Lajoie B, Gibon T, Staffell I and Andresen G B 2019 Real-time carbon accounting method for the European electricity markets *Energy Strategy Rev.* **26** 100367

[3] Miller G J, Novan K and Jenn A 2022 Hourly accounting of carbon emissions from electricity consumption *Environ. Res. Lett.* **17** 044073

[4] Holland S P and Mansur E T 2008 Is real-time pricing green? The environmental impacts of electricity demand variance *Rev. Econ. Stat.* **90** 550–61

[5] Siler-Evans K, Azevedo I L and Morgan M G 2012 Marginal emissions factors for the U.S. electricity system *Environ. Sci. Technol.* **46** 4742–8

[6] Siler-Evans K, Azevedo I L, Morgan M G and Apt J 2013 Regional variations in the health, environmental, and climate benefits of wind and solar generation *Proc. Natl Acad. Sci.* **110** 11768–73

[7] Kaffine D T, McBee B J and Lieskovsky J 2013 Emissions savings from wind power generation in Texas *Energy J.* **34** 155–75

[8] Cullen J 2013 Measuring the environmental benefits of wind-generated electricity *Am. Econ. J. Econ. Policy* **5** 107–33

[9] Graff Zivin J S, Kotchen M J and Mansur E T 2014 Spatial and temporal heterogeneity of marginal emissions: implications for electric cars and other electricity-shifting policies *J. Econ. Behav. Organ.* **107** 248–68

[10] Archsmith J, Kendall A and Rapson D 2015 From cradle to junkyard: assessing the life cycle greenhouse gas benefits of electric vehicles *Res. Transp. Econ.* **52** 72–90

[11] Novan K 2015 Valuing the wind: renewable energy policies and air pollution avoided *Am. Econ. J. Econ. Policy* **7** 291–326

[12] Holland S P, Mansur E T, Muller N Z and Yates A J 2015 Environmental benefits from driving electric vehicles? (available at: www.nber.org/papers/w21291)

[13] Callaway D S, Fowlie M and McCormick G 2018 Location, location, location: the variable value of renewable energy and demand-side efficiency resources *J. Assoc. Environ. Resour. Econ.* **5** 39–75

[14] Deetjen T A and Azevedo I L 2019 Reduced-order dispatch model for simulating marginal emissions factors for the United States power sector *Environ. Sci. Technol.* **53** 10506–13

[15] Donti P L, Kolter J Z and Azevedo I L 2019 How much are we saving after all? Characterizing the effects of commonly varying assumptions on emissions and damage estimates in PJM *Environ. Sci. Technol.* **53** 9905–14

[16] Fell H and Johnson J X 2021 Regional disparities in emissions reduction and net trade from renewables *Nat. Sustain.* **4** 358–65

[17] Bushnell J and Novan K 2021 Setting with the sun: the impacts of renewable energy on conventional generation *J. Assoc. Environ. Resour. Econ.* **8** 713249

[18] Thind M P S, Wilson E J, Azevedo I L and Marshall J D 2017 Marginal emissions factors for electricity generation in the midcontinent ISO *Environ. Sci. Technol.* **51** 14445–52

[19] Open Grid Emissions Singularity (available at: https://singularity.energy/open-grid-emissions)

[20] Miller G J, Pease G, Knowles M and Shi W 2023 Open grid emissions dataset 2019–2021 (v0.2.2) *Zenodo* (https://doi.org/10.5281/zenodo.7689791)

[21] Miller G, Pease G, Knowles M and Shi W 2023 Open grid emissions initiative v0.2.2 GitHub (https://doi.org/10.5281/zenodo.7692997)

[22] Ryan N A, Johnson J X and Keoleian G A 2016 Comparative assessment of models and methods to calculate grid electricity emissions *Environ. Sci. Technol.* **50** 8937–53

[23] United States Environmental Protection Agency (EPA) 2023 Emissions & Generation Resource Integrated Database (eGRID) 2021 (available at: www.epa.gov/egrid)

[24] U.S. Environmental Protection Agency 2023 eGRID technical guide with year 2021 data (available at: www.epa.gov/system/files/documents/2023-01/eGRID2021_technical_guide.pdf)

[25] U.S. Energy Information Administration (EIA) 2021 Emissions by plant and by region (available at: www.eia.gov/electricity/data/emissions/)

[26] Schivley G, Azevedo I and Samaras C 2018 Assessing the evolution of power sector carbon intensity in the United States *Environ. Res. Lett.* **13** 064018

[27] Liu Z *et al* 2020 Carbon monitor, a near-real-time daily dataset of global CO_2 emission from fossil fuel and cement production *Sci. Data* **7** 392

[28] Hong C H *et al* 2021 Near-real-time and state-level monitoring of U.S. CO_2 emissions (available at: http://eartharxiv.org/repository/view/2233/)

[29] de Chalendar J A and Benson S M A 2021 Physics-informed data reconciliation framework for real-time electricity and emissions tracking *Appl. Energy* **304** 117761

[30] U.S. Energy Information Administration Hourly electric grid monitor (available at: www.eia.gov/electricity/gridmonitor/index.php)

[31] U.S. Environmental Protection Agency 2003 *Technical support document: emissions & generation resource integrated database, data years 1996–2000* (available at: www.epa.gov/system/files/other-files/2023-01/egrid_historical_files_1996-2016.zip)

[32] U.S. Environmental Protection Agency 2014 Framework for assessing biogenic CO_2 emissions from stationary sources (available at: https://19january2017snapshot.epa.gov/sites/production/files/2016-08/documents/framework-for-assessing-biogenic-co2-emissions.pdf)

[33] Cherubini F, Peters G P, Berntsen T, Strømman A H and Hertwich E 2011 CO_2 emissions from biomass combustion for bioenergy: atmospheric decay and contribution to global warming *GCB Bioenergy* **3** 413–26

[34] Downie A, Lau D, Cowie A and Munroe P 2014 Approaches to greenhouse gas accounting methods for biomass carbon *Biomass Bioenergy* **60** 18–31

[35] Johnson E 2009 Goodbye to carbon neutral: getting biomass footprints right *Environ. Impact Assess. Rev.* **29** 165–8

[36] Nian V 2016 The carbon neutrality of electricity generation from woody biomass and coal, a critical comparative evaluation *Appl. Energy* **179** 1069–80

[37] Haberl H *et al* 2012 Correcting a fundamental error in greenhouse gas accounting related to bioenergy *Energy Policy* **45** 18–23

[38] Sterman J D, Siegel L and Rooney-Varga J N 2018 Does replacing coal with wood lower CO_2 emissions? Dynamic lifecycle analysis of wood bioenergy *Environ. Res. Lett.* **13** 015007

[39] Brandão M, Kirschbaum M U F, Cowie A L and Hjuler S V 2019 Quantifying the climate change effects of bioenergy systems: comparison of 15 impact assessment methods *GCB Bioenergy* **11** 727–43

[40] Brander M 2017 Comparative analysis of attributional corporate greenhouse gas accounting, consequential life cycle assessment, and project/policy level accounting: a bioenergy case study *J. Clean. Prod.* **167** 1401–14

[41] Rudek J and Schwietzke S 2019 Not all biogas is created equal *Environmental Defense Fund Energy Exchange* (available at: https://blogs.edf.org/energyexchange/2019/04/15/not-all-biogas-is-created-equal/)

[42] Paolini V, Petracchini F, Segreto M, Tomassetti L, Naja N and Cecinato A 2018 Environmental impact of biogas: a short review of current knowledge *J. Environ. Sci. Health* A **53** 899–906

[43] Jacques D C Gridemissions: tools for power sector emissions tracking (available at: https://energy.stanford.edu/gridemissions)

[44] U.S. Energy Information Administration Real-time operating grid *Hourly Electric Grid Monitor* (available at: www.eia.gov/electricity/gridmonitor/index.php)

[45] Singularity Energy Carbonara [API] (available at: https://carbonara.singularity.energy)

[46] Electricity Maps Live 24/7 CO_2 emissions of electricity consumption (available at: https://app.electricitymaps.com/map)

[47] U.S. Energy Information Administration About the EIA-930 data *Hourly Electric Grid Monitor* (available at: www.eia.gov/electricity/gridmonitor/index.php)

[48] Sotos M 2015 *GHG protocol scope 2 guidance: an amendment to the GHG protocol corporate standard* (World Resources Institute) (available at: https://ghgprotocol.org/scope_2_guidance)

[49] Selvans Z *et al* 2022 The Public Utility Data Liberation (PUDL) project (2022.11.30) *Zenodo* (https://doi.org/10.5281/zenodo.5677623)

[50] Huetteman J, Tafoya J, Johnson T and Schreifels J 2021 EPA-EIA power sector data crosswalk (available at: www.epa.gov/airmarkets/power-sector-data-crosswalk)

[51] Ruggles T H, Farnham D J, Tong D and Caldeira K 2020 Developing reliable hourly electricity demand data through screening and imputation *Sci. Data* **7** 155

[52] Cicala S 2022 Imperfect markets versus imperfect regulation in US electricity generation *Am. Econ. Rev.* **112** 409–41

[53] United States Environmental Protection Agency (EPA) 2022 *CAMD's power sector emissions data guide* (available at: www.epa.gov/system/files/documents/2022-07/CAMD%27s%20Power%20Sector%20Emissions%20Data%20Guide%20-%2007182022.pdf)

[54] U.S. Energy Information Administration 2022 *Form EIA-860 Instructions* (available at: www.eia.gov/survey/form/eia_860/instructions.pdf)

[55] Gosnell C 2018 Boiler generator associations from EIA 923 and 860 *Catalyst Cooperative* (available at: https://catalyst.coop/2018/08/07/boiler-generator-associations/)

[56] Catalyst Cooperative pudl.analysis.epacamd_eia documentation (available at: https://catalystcoop-pudl.readthedocs.io/en/latest/autoapi/pudl/analysis/epacamd_eia/index.html)

[57] Rossol M, Brinkman G, Buster G, Denholm P, Novacheck J and Stephen G 2019 An analysis of thermal plant flexibility using a national generator performance database *Environ. Sci. Technol.* **53** 13486–94

[58] U.S. Energy Information Administration (EIA) 2022 Form EIA-923 detailed data with previous form data (EIA-906/920) (available at: www.eia.gov/electricity/data/eia923/)

[59] Pease G 2022 Validating real-time electricity emissions rates with an hourly historical benchmark *Singularity* (available at: https://medium.com/singularity-energy/validating-real-time-electricity-emissions-rates-with-an-hourly-historical-benchmark-a9990a2c9049)