EmPOWER Proposal: Hourly Average Emissions Factors for the Emissions & Generation Resource Integrated Database (eGRID)

Background & Motivation

Quantifying greenhouse gas emissions is the foundation for many of the key decisions that need to be made about climate change. When I worked as the Sustainability Manager for Georgetown University, this was made clear to me each year as I completed our annual greenhouse gas inventory, which helped our leadership identify major sources of emissions, prioritize investments, and develop strategies to most quickly and cost-effectively reduce emissions from our operations. For our institution, and thousands of other corporations, cities, policymakers, and researchers across the country, the EPA's Emissions & Generation Resource Integrated Database (eGRID) is the go-to, trusted source for emissions factors for electricity generation. The Greenhouse Gas Protocol, which is the standard used by 9 out of 10 Fortune 500 companies reporting to the Carbon Disclosure Project, recommends using eGRID annual average emission factors for calculating "scope 2" emissions from electricity consumption.

However, as the U.S. electric grid continues to integrate more and more renewable energy, which provides carbon-free power at different times of day, these annual average emission factors may be less useful for accurately accounting the scope 2 emissions of an electricity consumer. In California, for example, where there is a large amount of solar producing in the middle of the day, the average emission rate in the evening can be on average *4.5 times higher* than during mid-day.[1] Thus, as the grid continues to change, the use of annual average emissions factors for accounting could significantly over- or under-estimate the total emissions reported in an emissions inventory, depending on the time of day that an entity is using energy.

These differences between hourly and annual emissions factors can lead to real policy and fiscal impacts. For example, in January 2020, Microsoft announced a commitment to remove from the atmosphere all of the carbon the company had emitted either directly or by electrical consumption since it was founded in 1975. The total amount of emissions that Microsoft will ultimately remove, and thus the total investment they will make, depend on an accurate inventory of emissions. As more and more jurisdictions set carbon neutrality goals, accurately inventorying emissions will be critical for evaluating policy effectiveness, tracking progress, shaping strategies, and ultimately determining when these goals have been achieved.

Since scope 2 emissions from electricity are generally calculated by multiplying electricity consumption by an emissions factor according to some accepted protocol or standard, the broader utilization of hourly emissions factors for GHG accounting will require:

1. Hourly electricity consumption data must be available
2. GHG accounting standards must recognize hourly accounting as a valid method
3. Hourly average emissions factors must be available

In the past, electricity consumers generally only had access to monthly data about electricity consumption, as reported on their monthly bills. However, access to hourly electricity consumption data, generally referred to as "interval data," is becoming more widespread as advanced metering infrastructure (AMI) continues to be deployed. Using AMI, utilities can provide hourly, or even sub-hourly, electricity data to customers through easy to access web

[1] Based on analysis of 2018 & 2019 CAISO data from http://www.caiso.com/TodaysOutlook/Pages/emissions.aspx

portals. According to the U.S. Energy Information Administration, which tracks the national deployment of AMI in its Form 861, a majority of customers now have AMI installed: As of 2018, 57% percent of nationwide residential customers and 54% percent of commercial and industrial customers had AMI installed, and AMI has been growing by 11% per year on average.

The second challenge is that GHG accounting protocols do not currently describe a protocol for accounting using hourly emissions factors, due to the fact that such factors are not widely available. The *GHG Protocol Scope 2 Guidance* notes that "Companies may have access to detailed studies or software solutions linking their facility's time-of-day energy use patterns to the GHG emissions from local generation dispatching during those times. This emission data could be compiled over the course of a year for a consumer to record, match against temporal usage by location, and calculate scope 2 emissions. To date such studies or analyses have not been widely available or used, and have often been contained in proprietary databases with limited consumer access."

Thus, this project proposes to use CAMD and other public data to develop accessible, trusted, and publicly available hourly average emissions factors for the United States. It is our team's intent to apply the existing eGRID methodology, where applicable, to hourly-resolution data to produce a new dataset that can be published alongside the eGRID database in future years. It is our hope that publishing these factors will open many new avenues of academic research, help more accurately account emissions in GHG inventories, and guide effective policy decisions.

Project Approach and Use of CAMD Data

The project proposes to create a database of hourly average emission factors for the United States that can be published alongside the existing eGRID database, help streamline the data pipeline for the production of the existing eGRID database, and analyze how the use of hourly average emission factors may impact GHG inventory calculations.

Data Sources

The data used for this project will be a mashup of many of the existing sources of data used by eGRID—including the EPA's Continuous Emissions Monitoring System (CEMS) data, the EIA's Form 860, and EIA Form 923—but will also integrate new sources of data, namely the EIA's Hourly Electric Grid Monitor (EIA Form 930). Unfortunately, most of these datasets are released in non-standardized formats, are difficult to cross-link, and are not always machine-readable. In addition, these datasets often contain missing or incomplete data that must be cleaned before they can be used for analysis. These challenges are well-documented in eGRID's technical support document, and well-familiar to any researchers who have worked with these data before. Standardizing, linking, and cleaning these datasets takes valuable time and resources away from actually using, analyzing, and drawing meaningful insights from these data.

Building upon the Public Utilities Data Liberation (PUDL) Project

In 2017, work began on the Public Utilities Data Liberation (PUDL) project, an open-source project that takes the original spreadsheets, CSV files, and databases from the EPA, EIA, and FERC, and turns them into unified tabular data packages that can be used to populate a database, or read with Python, R, Microsoft Access, and many other tools. In the three years since, the project has grown extensively, both in the scope of data sources it has integrated and in the size of its user community. The PUDL Python package, available on Github, has been downloaded over 3,000 times, and its finished data packages are freely available to download from Zenodo, an open data archive run by CERN. While the project has an active community of contributors from institutions like Carnegie Mellon University and the University of California who are continually refining PUDL, the project is spearheaded by a team of data scientists and energy experts at Catalyst Cooperative, a worker-owned consultancy committed to making energy and climate data more open and usable. In 2019, Catalyst Cooperative was able to significantly expand PUDL with support of the Alfred P. Sloan Foundation and contributions from its active community of users.

More information about PUDL can be found at:

- The PUDL repository on GitHub,
- PUDL documentation on Read The Docs
- The PUDL data release on Zenodo
- The description of PUDL on Catalyst Cooperative's website

For this project, we will build upon PUDL's existing methods for cleaning and standardizing these datasets, and leverage its existing codebase and user community to develop the hourly emissions dataset. The PUDL database and software use liberal open data and open source licensing, making this code and data free to access, easy to audit, and easy to use and update. This could enable a broad community of stakeholders, including academic researchers, journalists, local policymakers, and students to more easily use the data. All of the work that goes into further improving the data linkages between EPA and EIA datasets, improving the data quality and usability of these datasets, and calculating hourly average emissions factors will be integrated into PUDL's open-source codebase for all users of these datasets to benefit from.

The methods we will use for estimating emissions, determining plant primary fuels, and aggregating plant-level data will follow the established eGRID methodology as closely as possible. In the cases where the existing methodology does not apply to hourly-level data or EIA-930 dataset that we will be integrating, we will consult the academic literature and engage with EPA staff to use well-accepted methods. In the case that new methodologies would be developed, we would plan to publish these methods in a peer-reviewed academic journal.

Some of the existing academic literature and code on which we plan to build includes (but is not limited to):

- Schivley, Greg, Inês Azevedo, and Constantine Samaras. "Assessing the Evolution of Power Sector Carbon Intensity in the United States." *Environmental Research Letters* 13, no. 6 (June 2018): 064018. https://doi.org/10.1088/1748-9326/aabe9d.
- Chalendar, Jacques A. de, John Taggart, and Sally M. Benson. "Tracking Emissions in the US Electricity System." *Proceedings of the National Academy of Sciences*, November 27, 2019. https://doi.org/10.1073/pnas.1912950116.
- Tranberg, Bo, Olivier Corradi, Bruno Lajoie, Thomas Gibon, Iain Staffell, and Gorm Bruun Andresen. "Real-Time Carbon Accounting Method for the European Electricity Markets." *Energy Strategy Reviews* 26 (November 2019): 100367. https://doi.org/10.1016/j.esr.2019.100367.

- Ruggles, Tyler H., & Farnham, David J. "EIA Cleaned Hourly Electricity Demand Data (Version v1.1)" [Data set]. Zenodo (February 2020). http://doi.org/10.5281/zenodo.3690240
- *Electricity Life Cycle Inventory*. Python. 2017. Reprint, U.S. Environmental Protection Agency (EPA) and National Energy Technology Laboratory (NETL), 2020. https://github.com/USEPA/ElectricityLCI.
- *AVoided Emissions and geneRation Tool (AVERT) [Version 2.3].* U.S. EPA, 2019. https://www.epa.gov/statelocalenergy/avoided-emissions-and-generation-tool-avert
- *PyISO*. Python. 2014. Reprint, WattTime, 2020. https://github.com/WattTime/pyiso.

Using EIA-930 data to fill missing hourly emissions data

The previous challenge of developing hourly emissions factors was the lack of available data at the hourly level. While the EPA CEMS data is published at the hourly level, it only contains data for a subset of generators nationwide, generally those with a nameplate capacity >25MW. While those generators account for a significant majority of generation and emissions, there remain a significant portion of generators for which hourly emissions data are not reported. In eGRID these emissions data are filled based on fuel type and heat input data reported in EIA-923. However, the EIA-923 data is reported at a monthly level, so there was no direct way to estimate how those monthly emissions are distributed on an hourly basis.

To infer this hourly profile of emissions from non-reporting generators, we will use EIA-930's hourly data on net generation by fuel type, reported for each balancing area in the U.S. While the specific methodology needs to be refined during the project period, our general approach is as follows: to calculate the hourly generation pattern for all generators not reporting to CEMS, we would subtract the hourly net generation of reporting units of a specific fuel type (calculated from CEMS) from the total hourly net generation of that fuel type for the entire balancing area (reported by EIA-930). We could then distribute monthly the heat input of those generators from EIA-923 to this hourly profile. Performing this calculation will require converting hourly gross generation from CEMS into hourly net generation by calculating a gross-to-net linear model for each plant in the U.S. This methodology would likely build on the methodology used in the EPA's Avoided Emissions and Generation Tool (AVERT) for converting gross to net generation.

Another possible method would be to simply calculate a monthly average heat rate for all generators of a specific fuel type within each balancing area, based on EIA-923 data, and multiply by it by the hourly net generation by fuel type reported by EIA-923. What this method gains in simplicity, it potentially loses in accuracy, as it ignores hourly fluctuations in heat rate. Part of our analysis will compare the estimates of this simplified method with the estimates from the method described in the previous paragraph, which we expect to be more accurate.

Anticipated challenges with hourly calculations

Our team anticipates several potential challenges that we will need to address throughout this project. First, the hourly net generation by fuel type data contained in EIA-930 only became available in July 2018, meaning that the first full year of complete data is 2019. This data may possibly be available for earlier dates through the Open Access Same-Time Information System (OASIS) to which all independent system operators (ISOs) and Regional Transmission Operators (RTOs) report such data. ISOs and RTOs, however, only cover about 2/3 of U.S. electricity users.

The other challenge is that EIA-930 only contains data for the continental U.S., and thus excludes Alaska and Hawaii. Another challenge may be dealing with combined heat and power (CHP) units. In eGRID, the emissions from CHP units are allocated to the electrical generation and thermal output based on thermal output data. To date, our team is not aware of any thermal output data that is available on an hourly basis. Overcoming this challenge may require conferring with CHP plant operators or looking in the academic literature to determine whether a monthly ratio could be applied, or whether the ratio between thermal and electrical output fluctuates significantly at an hourly level. One final challenge may be calculating the hourly emissions factors for all of the same regional boundaries used in eGRID. The eGRID database reports annual average emissions at the state, balancing area, eGRID region, and NERC region scales. However, the hourly EIA-930 data is reported only at the balancing area and “region” level (which appear to differ from both the eGRID region and NERC region boundaries). Further analysis will be required to determine whether state-level, eGRID-level, or NERC-region level emissions factors can be produced.

Developing “eGRID2019” and a streamlined eGRID production process

Since a complete year of hourly EIA-930 data for 2018 is likely not available, we will be unable to produce an hourly dataset for 2018 that could be directly compared to eGRID2018. Such a comparison would be useful for answering questions such as the relative impact of using annual verses hourly average emission factors for GHG inventories or policy analysis. Since such analysis is critical for evaluating the usefulness and application of hourly average emission factors, it would be helpful to have eGRID2019 data to which to compare it. Due to resource constraints, however, eGRID is only published biannually for even-numbered years.

Since this project is already using the same data sources and methodology as the existing eGRID database, it would not be too much of an extra lift to develop an open-source version of the eGRID data pipeline in Python. Thus, in order to enable meaningful comparisons of hourly and annual emissions factors, this project would generate a 2019 version of the eGRID database using an open-source Python module built on the PUDL data pipeline. In order to calibrate our open-source eGRID python module, we would run it using 2018 data, making sure that the outputs matched the published eGRID2018 dataset.

In addition to enabling comparison between the hourly and annual emissions factors for 2019, we anticipate that this module could significantly automate and streamline the several-month-long eGRID production process, potentially enabling the publication of eGRID on an annual basis, and reducing cost to the Federal Government for production of the database. Resources would still need to be budgeted for updating the eGRID methodology and producing the technical support document, but this python module could in theory automate the entire processing step, from downloading the data to outputting the spreadsheet. For example, the EPA may choose to still only update the methodology on a biannual basis, but in the odd years, the eGRID python module could simply be run to produce a full eGRID database, or perhaps just an “eGRID lite” that would only include certain emissions data. Conversations with EPA staff would be needed to fully scope how realistic such an annual publication might be, but we hope to create the tools that could facilitate further streamlining.

Summary of Project Products

- Publication of "eGRID-hourly" dataset for 2019, in csv or Excel format. This will contain an "8760 time series" of average emissions factors (one factor for each of the 8,760 hours in a non-leap year) for each of the greenhouse gases tracked by eGRID (CO2, CH4, N2O, and CO2e), and other pollutants where possible (SO2, NOx, Hg). These emissions time series will be published for each of the regions at the balancing area level and regional (whether eGRID region, EIA region, and/or NERC region) level
- Technical documentation explaining the data sources, methodology, and outputs of the eGRID-hourly production process, which could be integrated as an appendix to the eGRID technical documentation or published as a separate document
- Publication of eGRID2019 database based on eGRID2018 methodology, in Excel format
- Open-source python code, published on GitHub, used to produce eGRID2019 and eGRID-hourly 2019 datasets.
- Two manuscripts submitted to peer-reviewed academic journals: one methods paper regarding the development of hourly average emission factors, and another analyzing the implications of using annual or hourly emissions factors for GHG inventories

How the Project Addresses Challenge Objectives

We believe that our project approach and outcomes will advance the knowledge, use, and understanding of CAMD data by: **analyzing data, enhancing communications, developing technology and data mashups, and improving data quality**.

By integrating CAMD data with existing and new (EIA-930) EIA datasets, and by using new technological platforms (the python-based, open-source PUDL project) to streamline the data integration pipeline, we see this work as a novel **technology and data mashup**. Leveraging the algorithms used by PUDL to electronically audit and clean CAMD data, we hope to further **improve the data quality** of these datasets. This project develops a new way to present CAMD data to the public (on an hourly basis) to **enhance communications** about the data and encourage dialogue about how the data should be applied. Finally, this project will **analyze these data** to determine the impacts of hourly average emissions factors on greenhouse gas inventories.

Team Bio

This team is unique in that it is not only led by a core team of researchers, data scientists, and policy wonks from UC Davis and Catalyst Cooperative, but it leverages the skills and expertise of the broad community of contributors to the open-source PUDL project.

Greg Miller | LinkedIn | GitHub | Google Scholar | Greg will serve as the project manager and lead researcher on this project. Greg is an Energy Systems PhD student in the Energy Graduate Group at the University of California, Davis. His research focuses on the engineering, economic, and social dimensions of transitioning the electric grid to 100% clean and renewable energy. His current research focuses on developing metrics to quantify the impact of industrial energy demand flexibility on the integration of renewable energy and reduction of emissions from the power sector. He currently uses data science tools in Python to work with EPA, EIA, and ISO data, and simulate energy demand and grid operation. His previous professional experience is in corporate sustainability, greenhouse gas accounting, and local environmental policy. He is a Scholar with the Renewable Energy Buyers Alliance (REBA) and is a former Fellow with the Clean Energy Leadership Institute (CELI). He has a Bachelor of Science in Foreign Service, focused on International Energy Security, from Georgetown University.

Catalyst Cooperative |The Catalyst Cooperative team will serve in two main roles: 1) in an advisory role regarding data wrangling and method design, and 2) as editors and reviewers of the code.

> **Christina Gosnell** | Christina is the president and co-founder of Catalyst Cooperative, where she enjoys working at the intersection of public data, utility policy, and database design. Outside of her work at Catalyst, she tracks electricity regulation with E9 Insight, helping renewable energy companies, nonprofits and national research labs navigate the changing regulatory landscape. Christina studied Environmental Sciences at University of Colorado Boulder, where she worked at the CU Environmental Center, and co-founded the CU Environmental Studies Club and the divestment group Fossil Free CU. After graduating, Christina assisted research on high pressure gasification of biochar at the National Renewable Energy Lab and worked at Clean Power Finance, contributing to the National Solar Permitting Database.
>
> **Zane Selvans, PhD** | After a couple of years as Director of Research and Policy at the nonprofit Clean Energy Action, Zane became one of the founding members of Catalyst Cooperative, where he now leads the technical side of the team, developing an open platform for public data related to US electric utilities. Zane Selvans received his B.S. in Engineering from Caltech, focused on computer science and machine learning. After a brief stint in Silicon Valley he returned to Caltech as scientific support staff, working on a variety of NASA planetary exploration projects, before heading to the Laboratory for Atmospheric and Space Physics at the University of Colorado in Boulder for graduate school, where he developed computational techniques for studying the surface geology of icy moons in the outer solar system. Since earning his Ph.D. in 2009, Zane has focused on climate and energy policy.
>
> **Steven Winter** | Steve is the Chief Financial Officer of Catalyst Cooperative. In managing operations and business development at Catalyst, Steve enjoys bringing together his interest in climate and energy policy with his passion for social enterprise.

The PUDL Community | Although not formally a part of the project team, multiple other users regularly contribute to the PUDL project. Given that this project will be developed alongside the existing PUDL codebase, we anticipate that over the course of this yearlong project, other researchers may contribute to the development and success of this project.

Project schedule and milestones

We have divided this project into 12 key milestones, each of which will have multiple tasks that must be completed. The individual tasks are too numerous to list in this document, but each task will follow the same basic approach: 1) review of literature and method design 2) code “minimum viable product” 3) refine code functionality to generate complete and accurate output 4) test, edit, and package the code 5) merge code with PUDL “master” branch. Greg will lead steps 1-3 of each task, and the Catalyst Cooperative team will lead steps 4-5.

If awarded the project, our first step would be to set up a scoping meeting with Justine Huetteman, the EPA point of contact for the CAMD power sector data and EmPOWER challenge, and Travis Johnson, the EPA point of contact for the eGRID database, to establish a schedule for regular check-ins throughout the challenge period. Other important stakeholders to this project could include: the eGRID contractor (ABT Associates); the EIA teams who manage the Forms 860, 923, and 930 data that we will be using; the World Resources Institute (developers of the Greenhouse Gas Protocol); and any other eGRID stakeholders whose input may be valuable for the development of hourly emission factors. We will work with the EPA team to determine if, when, and how it would be appropriate to engage with any of these stakeholders. Through regular communication with the EPA team, and the open nature of the GitHub platform that we will be using to develop and publish our code, our progress toward the following milestones will be well tracked and transparent to all project stakeholders.

	Project Milestone	**Target Completion**
1	Integrate EIA-930 data into PUDL and clean	July 2020
2	Crosslink included EPA and EIA datasets, including net-to-gross generation conversions	July 2020
3	Clean CEMS data using PUDL according to eGRID2018 methods	August 2020
4	Complete python version of eGRID2018 production pipeline and validate results with published eGRID2018	September 2020
5	Generate eGRID2019 once all 2019 EIA datasets are available	October 2020
6	Estimate hourly emissions time series for non-reporting generators	December 2020
7	Complete eGRID-hourly production pipeline	January 2021
8	Complete technical documentation	February 2021
9	Publish datasets on Zenodo	February 2021
10	Submit methods paper to peer-reviewed journal	March 2021
11	Submit analysis paper to peer-reviewed journal	June 2021
12	Deliver summary report and presentation to EPA and work with EPA to publish datasets on eGRID webpage	June 2021