Highlights

Towards Objective Evaluation of the Accuracy of Marginal Emissions Factors

Sam Koebrich,Joel Cofield,Gavin McCormick,Ishan Saraswat,Nat Steinsultz

- Describes how Marginal Emissions Factors (MEFs) are a powerful tool to reduce CO_2 through load shifting (e.g. electric vehicle) to the least emitting times; as well as siting new load (e.g. a data center) or new renewables in carbon optimal locations.
- Introduces a taxonomy of MEF models useful for deciding on the signal type that is relevant to a proposed intervention.
- Proposes a series of empirical tests to perform rudimentary validation upon any MEF signal.
- Applies these empirical tests to a variety of public and proprietary MEF models used by industry today.

Towards Objective Evaluation of the Accuracy of Marginal Emissions Factors

Sam Koebrich[a,*], Joel Cofield[a], Gavin McCormick[a], Ishan Saraswat[a] and Nat Steinsultz[a]

[a] *WattTime,*

ARTICLE INFO

Keywords:
demand side management (DSM)
economic dispatch
carbon optimization
marginal emissions rates
causal modeling
life cycle assessment
carbon accounting
renewable energy siting

ABSTRACT

A growing number of policymakers and voluntary actors seek to reduce CO_2 or other emissions by intentionally changing the timing of flexible electricity load. In practice, this requires selecting a model to estimate what is often called the marginal emissions factor (MEF). Multiple different models to estimate MEFs exist, but there is no single objective ground truth MEF dataset with which to compare the relative accuracy of two or more such models. Thus, historically MEF validation has been forced to use subjective tests which rely on the very models they evaluate. Here we propose a technique to instead gain insight into the relative accuracy of different MEF models without relying on the assumptions of any MEF models. It works by isolating indirect model claims that can be objectively tested. This method also requires first applying a taxonomy to clarify the objectives of different types of MEF models, so that accuracy is rigorously defined. We then demonstrate the application of this technique in a comparative analysis of a few sample MEF models in the US, using 2021 data.

1. Introduction

Climate change is an urgent global policy challenge, with its largest driver being emissions from the electricity sector IEA (2023); Freeman, Rouzbeh Kargar, Söldner-Rembold, Ferreira, Couture, Jeyaratnam, O'Connor, Lewis, Hobbs, König, McCormick, Amuchastegui, Nakano, Dalisay, Davitt, Gans, Lewis, Volpato, Gray and McCormick (2022); Nassar, Hill, McLinden, Wunch, Jones and Crisp (2017) Against this backdrop, numerous government policy and corporate sustainability decision makers are interested in maximizing real-world emissions reductions. One emissions management strategy is to shift flexible electric load (such as electric vehicles, batteries, smart thermostats, or data centers) to times that reduce emissions. The times that reduce the most emissions are not necessarily the same times as when electricity is cheapest Holland, Kotchen, Mansur and Yates (2022); Hittinger and Azevedo (2015).

This paper considers the problem facing decision makers who wish to know how a proposed load shift will affect total emissions. The total change in emissions, divided by the total amount of electricity shifted, is called the **marginal emissions factor (MEF)** [1].

A variety of individuals, corporations, and policy makers with voluntary or regulatory emissions reductions targets are already using MEFs to identify opportunities to reduce emissions. For example:

- The Self-Generation Incentive Program (SGIP) in California requires distributed energy storage resources receiving an incentive to measure the emissions impacts of their charging behavior using MEFs SGIP (2023).
- Institutions are using MEFs to decide where to site new renewables Boston University (2021).
- The Inflation Reduction Act directs the US Energy Information Administration (EIA) to begin publishing MEFs Palmer, Prest, Her and Villanueva (2022).

The growing policy relevance of MEFs, as well as a recent proliferation of different MEF models, is raising the importance of validating these models. However, doing so poses some challenges.

The first challenge is definitions. Different authors define MEFs slightly differently, or refer to them by different names. This complicates efforts to compare accuracy across models Voorspools and D D'haeseleer (2000); Ryan, Johnson and Keoleian (2016); EPA (2019); PJM (2021). For example, authors differ on whether an MEF is a direct

sam@watttime.org (S. Koebrich)
ORCID(s): 0000-0002-4970-6309 (S. Koebrich); 0000-0002-0570-2824 (G. McCormick)

[1] Note that "marginal" in this use does not necessarily imply small, but rather an incremental/decremental effect of any size.

emissions measurement (e.g. an emitting generator or a total change in emissions) or a statically significant change in emissions that theoretically *would* occur if certain assumptions held true.

The second challenge is the diversity of use cases for MEFs. For example, changing the timing of a single electric vehicle (EV) may result in a single generator slightly adjusting its output, but adding 10,000 EVs may require multiple generators to respond. The true MEF may therefore differ between these situations even though they occur at the same time and place. Drawing on the econometrics literature, we can think of these changes as **interventions**, each of which could in theory have a different "true" MEF.

A third challenge is the lack of MEF ground truth data. A randomized controlled trial (RCT) could observe the true MEF of a particular intervention by comparing total system emissions having done the intervention, against total system emissions in a counterfactual without the intervention. Unfortunately, it is impossible to directly observe a counterfactual and no RCTs of MEFs have been run to date [2]. Thus, decision makers must inherently rely on modeled MEFs [3], and must find some way to evaluate the accuracy of these modeled MEFs without comparing against a ground truth dataset.

Because of this, MEF users may feel unsure if they have selected an appropriate model, and have no way of knowing if their selected model is accurate. Due to a lack of options, decision makers frequently evaluate an MEF model by comparing it to another MEF model. Not only does this not guarantee accuracy, if the models are intended to measure different interventions, it can generate false comparisons.

We propose a framework rooted in the causal inference literature, and three tools for decision makers to alleviate some uncertainty. These tools are:

1. A definition of MEF accuracy rooted in cause and effect

   Without a shared definition of MEF, it is not possible to ask which model is most accurate. For decision makers interested in maximizing emissions reductions, the most useful MEF definition should focus on effects on real-world total system emissions reductions of whatever policy or action is actually being considered.

   For the sake of this paper, we define a perfectly accurate or true MEF as *the actual difference in total system emissions that is caused by all load shifting interventions, divided by the total amount of load shifted at a given time and place.*

2. A taxonomy of MEFs by relevant intervention

   As this definition makes clear, the accuracy of an MEF model depends on what type of intervention is being measured. We **taxonomize MEF interventions** based on the type of intervention the decision maker seeks to measure. We demonstrate this in Section 3 by highlighting the three most common types of variation, including the **size** of load shifted; the **time frame** over which it is shifted; and the **geographic scale** of the intervention.

   Taxonomizing MEF models enables us to ask two questions. First, which models are *intended* to be applicable for the specific intervention a decision maker is considering? Second, when testing model accuracy, what tests are relevant in determining which models are in fact most accurate for that type of intervention, regardless of modeler intent? Together with the definition above, it now becomes empirically meaningful to even ask which of two *modeled* MEFs is more "accurate".

3. A framework for objectively testing MEFs indirectly

   Because of the lack of ground truth data, it can be difficult to objectively test MEFs by directly comparing them with other MEFs. However, it is possible to objectively evaluate models indirectly, by testing whether they can be falsified by other empirical data. We focus on relevant, public data known to be accurate and available to decision makers.

   In Section 5, we propose a framework of tests that analyze empirical data (described in Section 5) to assess if the values of a particular MEF model at a given time appear feasible. If the model is not aligning with the expected behavior described by the ground truth data, it may be explainable through other externalities, or it may be an indication that the model is flawed.

[2] In economics, political science, and psychology, this is a long-studied problem known as the Fundamental Problem of Causal Inference (see e.g. Rubin 1974)Rubin (1974).

[3] Some grid operators provide a measurement of the MEF for a 1 MW increase in load as an output of their security constrained economic dispatch (SCED) solvers. However, if a user is applying this MEF to control a load shifting intervention with different characteristics, it should be interpreted as a heuristic model rather than the "true MEF" as different interventions may cause different effects.

In Section 6, a variety of MEF models are assessed against these tests to gain insight into their accuracy in measuring medium-size, short-run, regional-scale shifts in load. These tests focus on:

(a) Expected Carbon Intensities to establish bounds for reasonable MEF values.

(b) Empirical Annual Averages to coarsely estimate the accuracy of MEF models.

(c) Expected Net-Demand Temporal Patterns to study how MEFs consider the ramping constraints of different fuel sources, and

(d) Curtailment to examine how MEF models are able to identify periods of excess renewable supply.

Results are also discussed. When there is significant deviance from the expected behavior, we look for a suitable explanation, if available. The method of testable propositions in this paper is generalizable and provides a framework to gain insight into the accuracy of MEFs and measure the effects of other types of interventions. With these results, decision makers who hope to maximize real-world total system emissions reductions can begin to ask: for a possible intervention, which models make claims congruent with ground truth data, and which models are falsified?

2. Literature Review

As early as 1999, studies have examined how load-shifting programs affect emissions using marginal emissions methods. For example, ISO New England (ISO-NE) has conducted an annual marginal emissions rate analysis for more than two decades, initially using a heuristic-based heat rate estimating generator efficiency via locational marginal price (LMP) ISO-NE [4]. CAISO's SGIP is another example of this methodology SGIP (2023).

The U.S. Environmental Protection Agency (US EPA) introduced a more complex statistical approach through the Avoided Emissions and Generation Tool (AVERT) in 2007. AVERT uses historical, hourly unit-level generation and emissions reported through the Clean Air Markets Program Data (CAMPD). For each balancing authority (BA), AVERT bins the total load observed into quantiles. Within each bin, and for each unit in the BA, a Monte Carlo analysis approximates the likelihood of that unit responding to load. AVERT's methodology relies on historical data that is reported quarterly and cannot be used in real-time EPA (2019).

Other authors use simulations of merit-order economic dispatch to calculate MEFs. These approaches approximate costs (including fuel, variable operating costs, and maintenance) for each generating unit. The units are arranged in ascending order of cost, and are assumed to dispatch sequentially until the total supply equals a simulated demand value. The next unit on the "load duration curve" is considered marginal and the MEF is equal to its emissions rate. A recent implementation by Deetjen and Azevedo (2019) calculates a rolling average MEF over a load duration curve, while considering reduced-form constraints including minimum run times for coal plants and fuel prices reflective of historical market fluctuations.

More complicated economic dispatch models utilize computationally expensive power system optimization models (such as PLEXOS) to capture grid constraints with greater detail. Constraints include generator ramping, transmission congestion, and interconnection. NREL's Cambium model uses PLEXOS to estimate a short-run and long-run MEF Gagnon, Frazier, Cole and Hale (2021). Simulated economic dispatch models cannot natively model real-time MEFs, as they rely on historical data sets.

While the dispatch models listed above infer an MEF given a hypothetical change in load, a different form of an MEF can also be inferred from security-constrained economic dispatch (SCED) solvers. BAs use these in real-time to determine the economically optimal set of generators to dispatch after considering ramping rates and bid supply curves. An artifact of these programs is the ability to infer how the supply curve would change given a very small (e.g. 1 MW) increase in load in the system. Notably PJM — the U.S. Mid-Atlantic balancing area — provides a five-minute measurement of the MEF inferred through their SCED. While this signal offers a measurement of the MEF for a very small change in load, it is explicitly not extensible to larger changes in load, and contains extreme outliers that may be inappropriate for some use cases PJM (2021).

Lastly, statistical MEF models use empirical data to estimate the causal response expected from an intervention conditioned on known information (e.g., time of day, level of demand, LMP, weather). Hawkes (2010) applied a regression-based approach to Great Britain by fitting a linear regression between the hourly changes in emissions and generation. These regressions are repeated across a number of quantized bins of load in each hour. By conditioning

[4] Since 1999, ISO-NE has changed its marginal emissions methodology numerous times, using statistical models and others.

data upon confounding variables, the MEF is estimated by the slope of the regression in a bin. Zivin et al. (2014) applied a regression model to the U.S. data. Unlike Hawkes, they did not difference hourly emissions and generation data (which introduces an intercept into the regression), and they bin on the month and hour rather than demand level. Siler-Evans et al. (2012) apply a differenced regression model to the U.S. using load bins to predict the marginal fuel mix. Unlike SCED models, these methodologies estimate an MEF based on observations of organic load changes, not changes attributable to a particular intervention.

Previous studies attempt to validate MEFs by comparing with other MEF models. A good example is Elenes et al. (2023), which simulates power grids to estimate the true MEF to directly gain insight about the accuracy of different MEF models. To our knowledge however, no prior paper has proposed a method to validate MEF models solely based on empirical data, using indirect methods. Another notable example is Ryan et al. (2016), which compares different emissions factor models (both marginal and non-marginal) and also proposes a taxonomy. To our knowledge, this paper is the first to propose a taxonomy of marginal emissions models based on the intervention being measured, ensuring accuracy is rigorously defined.

3. Taxonomy of Models

Modeling an MEF depends on specific characteristics of the intervention considered (i.e., a possible increase or decrease in load). For instance, it is expected that shifting 1 megawatt of load will result in a different MEF from shifting 1 gigawatt. True MEFs are likely non-linear across a variety of factors describing the intervention.

The range of possible MEFs is considerable. For example, interventions could have different true MEFs if the intervention is known in advance to the BA versus being a surprise (which could affect planning); a load shift that ramps slowly or quickly (which could affect generators that must cold start differently than fast-ramping units); or one that lasts for minutes or hours each day (which could affect the load duration curve differently). To our knowledge, no intervention-specific models have yet been published. Thus, over time the value of an intervention-based taxonomy may increase. But for all published models, we identified three major dimensions: size, time frame, and the applicable geographic granularity.

Below we provide a table categorizing different marginal emissions models based on the type of intervention one might a priori expect each to best measure, based on how the model was constructed. For example, if an MEF model is primarily derived from statistical analysis of the short-term changes in emissions in a BA between hours, we classify it here as expected to be effective at interventions of roughly this size [5], time frame, and geography.

The size of an intervention ranges from a low-wattage IoT device to multiple GW of virtually aggregated distributed energy resources (DERs). The marginal generators responding to these interventions are expected to differ. Ancillary services will likely handle the former, while the latter may require additional capacity to ramp. Economic dispatch models focus on small, fixed changes in load (1 MW is a common increment), while statistical models identify causal effects that apply at a size of the average hourly difference in load within a grid region.

The *timeframe of effect* assesses the relationship between the intervention and systemic changes in the supply curve. Some models assess the "short-run" impact of marginal load on the existing dispatch stack. These models are well-suited for use cases involving discrete energy consumption decisions focused on the immediate implications of an action (e.g., turning on a thermostat now versus later). Alternatively, "long-run" models seek to capture both the evolution of the grid over many years and the effect of a demand intervention on the grid's structure. Interventions that are repeated for many years or add new demand should consider their long-run impact (e.g., siting a new building). Importantly, an intervention may at first affect short-run operations of the grid, but after time consistent changes in load may spur structural long-run changes in the grid's capacity mix.

Finally, some MEF signals provide higher geographic granularity than others. Some models are nodal, with nodes that may represent physical locations (e.g., a substation) or virtual representations of known transmission constraints. Other models use sub-regions, which may be representative of a group of nodes with highly correlated LMPs. In either case, differences in T&D access can propagate differences in the sources of generation, causing the MEF to differ between nodes [6].

This taxonomy is designed to inspire users to consider the different perspectives that each of these signals can provide. However, it is certainly possible that a model trained on one size, time-frame, or geographic granularity may

[5] Since for many balancing authorities this is in the range of hundreds of megawatts, we refer to this as "medium" size.

[6] Due to present data availability, the testable propositions do not consider T&D constraints outside of periods of curtailment, and only look for a single MEF per balancing authority

Table 1
Taxonomy of MEF Models by Expected Intervention Type

Model	Size	Time-frame	Geographic applicability
eGrid Non-Baseload (Heuristic - Inverse Capacity factor)	Any	Long	Regional (State)
California SGIP (Heuristic - Heat Rate)	Any	Short	Regional (9 California regions)
EPA AVERT (Statistical)	Medium (Avg. hourly load change)	Short	Regional (14 aggregated balancing authorities)
Siler-Evans (Statistical)	Medium (Avg. hourly load change)	Short	Regional (8 NERC regions)
WattTime (Statistical)	Medium (Avg. hourly load change)	Short	Regional (121 U.S. balancing authorities and subregions)
PJM & many ISO Models (SCED)	Small (1 MW)	Short	Local (Nodal)
REsurety (SCED)	Small (1 MW)	Short	Local (Nodal)
Deetjen (Simulated Economic Dispatch)	Medium (Avg. hourly load change)	Short	Regional (8 NERC regions)
NREL Cambium Short (Simulated Economic Dispatch)	Small (1 MW)	Short	Regional (134 balancing authorities)
NREL Cambium Long (Simulated Economic Dispatch)	Small (1 MW)	Long	Regional (134 balancing authorities)

happen to also be accurate at making out-of-sample predictions of other types of interventions. For example, could a model designed to measure changes of one megawatt accurately estimate the effect of hundreds of megawatts, or vice-versa? That is an empirical question, and the only way to answer it is empirical testing. There cannot be a single source of truth for all MEFs, but rather a multitude of discrete MEFs for different interventions. The job of the user is to select a model that most closely aligns with the specifications of their use case. Once a signal is selected, it is prudent to validate that the signal is accurately modeling what it claims to represent. The tests in the following section are designed to assist with this validation task.

In the remainder of this paper, we present tests designed specifically to measure the accuracy of MEF models in measuring the effect of medium-sized (hundred of megawatts), short-run changes to load over regional granularities. In some cases these tests also provide useful insight to the accuracy of MEF models in measuring the effect of other interventions, but further research is needed to develop tests specifically designed for these interventions.

4. Material and methods

We use four U.S. government-managed primary data sources to compare MEF models against. First, the EPA's CAMD includes hourly emissions and gross load measured at the smoke stack for 92.8% of continental US power sector emissions. See Appendix 10 for a detailed analysis and description of how we processed CAMD. Second, EIA-930 provides near-real-time generation by fuel type and total demand, which can be used to calculate net demand at the BA level. Third, EIA-923 contains attributes of every generator's monthly fuel consumption and heat rate. A secondary data source, the EPA's Emission and Generation Resource Integrated Database (eGRID) builds upon these datasets to infer the total emissions and carbon intensity of every unit in the U.S. on an annual basis. Additionally, we gathered curtailment data available from seven major BAs in the U.S., as detailed in Section 5.4.1.

We also selected a group of MEF models that could be publicly obtained and reproduced. We gathered seven models to consider using the framework of tests that will be outlined in Section 5, and applied to the models in Section 6.

- **The 2021 edition of EPA's AVERT model**, which is available via an Excel spreadsheet. To create an MEF profile, we ran AVERT with a flat 100 MW load profile across the entirety of 2021 for seven NERC regions corresponding to U.S. ISO/RTOs.
- **Siler-Evans**, we implemented the methodology described in the 2017 paper, and were able to reproduce results. Then, we re-ran the model using CAMD for 2021.
- **WattTime's v3.2 model**, which is available through an API to request real-time and historical values at a five-minute increment. The WattTime signal can be optionally queried with curtailment, which is applied as a probabilistic de-rate for transmission-constrained subregions. Outside of Sections 5.4 and 6.4, the signal was queried without curtailment.
- **NREL's short-run and long-run Cambium models**, which are produced annually. Cambium produces MEFs for simulated grid data, for even-year increments. In order to approximate 2021 data, we used the 2022 model year from the 2021 release of Cambium. All resource regions contained within a BA were averaged for each hour.
- **PJM's SCED solver**, which is available through an API on the BA's website. The API provides an MEF for approximately 10,000 nodes every five minutes. We accessed node "1", which represents the MEF at the system level, ignoring nodal transmission constraints.
- **CAISO's SGIP signal**, which is available through an API. WattTime maintains the SGIP signal, using a different algorithm (heat rate based) from their statistical model. The MEF is available across nine subregions in California, which have been averaged for this analysis.

In Figure 1 we plot the distributions of these MEF models to better understand the variance and bounds of these signals. Cambium's short run has the most variance of economic dispatch models (aside from PJM's SCED), while AVERT has the most variance of statistical models. Unlike AVERT, neither Cambium's long- nor short-run models contain any negative MEF values. The WattTime and Siler-Evans models tend to be normally distributed with low variance. The CAISO SGIP MEF can approach zero as the LMP approaches zero, which is a heuristic used by that model to consider curtailment. The PJM SCED MEF contains extreme outliers as low as -1.7 million lbs / MWh, and as high as 1.6 million lbs / MWh, a detailed explanation of these outliers is available in Appendix 8.

5. Methodology & Calculation

This section outlines a series of tests that will be later applied (in Section 6) to various MEF models to gain insight into their accuracy.

As noted above, this paper defines a perfectly accurate MEF as the actual difference in total system emissions that is caused by all load shifting interventions, divided by the total amount of load shifted at a given time and place. This means it becomes possible to interpret an MEF model as making a literal statement about what will occur if a set of interventions occur.

Consider a decision maker who wishes to shift 2 MW load in the short-run at one node. In this framework, to say an MEF model is "accurate" for that intervention would mean that adding 2 MW of load at that location, for a short period of time, would increase total system emissions by a number equal to twice the MEF. Applying an MEF model to a particular type of intervention produces a series of specific, empirically verifiable claims. The core methodology of this paper is to examine those claims which are falsifiable to indirectly gain overall insight into the model's overall accuracy. For each test, we will explain why we believe it to be truthful and then use empirical data to justify this claim.

The tests in this paper are designed to test the accuracy of MEF models at estimating medium-size, short-run, regionally granular load shifting interventions. For other intervention types, further research is needed. [7]

[7] Note that it is possible for a model to be accurate at one type of MEF, but not another. For example, a model designed to accurately estimate the emissions effects of small load shifts may also accurately estimate the effects of large size load shifts, or it may not–that is an empirically testable question separate from model intent.

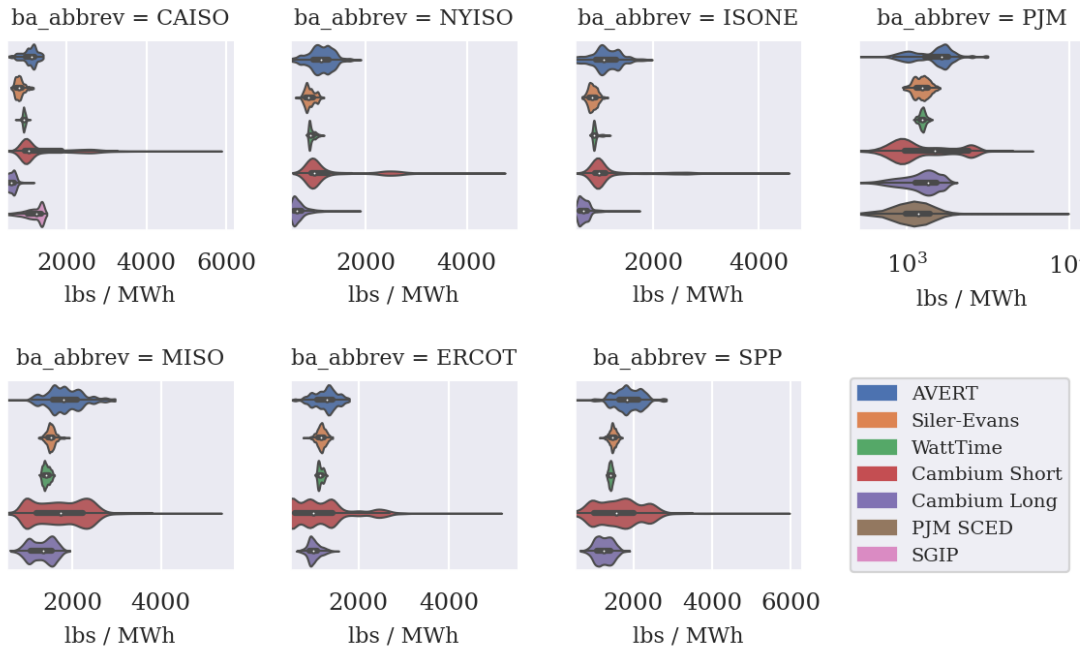

Figure 1: Distribution of MEF Models by Balancing Authority

5.1. Expected Carbon Intensities

The first two tests describe reasonable carbon intensities for an MEF based on observed emissions rates of generators in the CAMD and eGRID data sets. These serve as upper and lower bounds that we will later test MEF signals against.

5.1.1. The MEF is not expected to be negative

Emitting power plants must always be combusting fuel to produce electricity. While it is possible for an individual unit to have a negative net generation (after accounting for idling or parasitic losses), it is not possible for generators to be marginal while in such a state. Similarly, it is possible for a small increase in load to cause a more-intensive generator to ramp down, and a less-intensive generator to ramp up in order to meet the new demand without creating oversupply behind a transmission constraint, however this condition is not expected to occur frequently, or persistently, especially when considering larger interventions. [8] To apply this test, for each model and ISO/RTO we calculate the percent of time in 2021 that the model claims the MEF was negative.

5.1.2. The MEF is not expected to exceed the carbon intensity of the most-carbon-intensive plant measured in CAMD for that hour

Similarly, the MEF values are not expected to exceed the carbon intensity of the most-polluting unit found in CAMD for the corresponding hour and grid region. It is possible for exceptions to occur when plants redispatch due to a change in load, however this condition is not expected to occur frequently or persistently. To apply this test, we take the MEF model values in 2021 for each ISO/RTO, and calculate the percent of times with values exceeding the maximum carbon intensity observed in CAMD for that grid region within that same hour.

[8] A test could look for the lowest-emitting generator at any hour in the CAMD data set. However, many MEF models are probabilistic. If a model predicts a chance of curtailment occurring during a given period, and represents that by de-rating an MEF by the probability of curtailment, this could reasonably produce values below the lowest carbon intensity present at any time in CAMD. Even in the case of an MEF containing curtailment as a component, we would never expect values below zero. Because of this, negative MEF values are not expected.

5.2. Empirical Annual Averages

5.2.1. Across grid regions, the empirical annual average derivative of emissions with respect to load is expected to be equal to annual average MEFs

We use hourly CAMD data to calculate the total change in emissions and the corresponding total change in load between every hour and the next. Dividing the change in emissions by the corresponding change in load creates a simple empirical hourly MEF. However, any single value of this simple model is susceptible to error from confounding variables that are not measured in this experiment. To remove this bias, we fit a linear regression to these values. [9] The slope of this linear regression model will be called an "Empirical Annual MEF," which we expect to approximately equal the annual average of short-run MEF models. The scale of intervention considered by this simple model is the average change in load from one hour to the next. Equation 1 details this regression, where is the intercept of the regression, which is approximately zero, and is an error term Azevedo, Deetjen, Donti, Horner, Schivley, Sergi, Siler-Evans and Vaishnav (2020)

$$\Delta G_{h,ba} = G_{h,ba} - G_{h+1,ba} \quad \text{(MWh)} \tag{1}$$

$$\Delta E_{h,ba} = E_{h,ba} - E_{h+1,ba} \quad \text{(lbs)} \tag{2}$$

$$\Delta E_{h,ba} = \overline{MEF}_{ba} * \Delta G_{h,ba} + \alpha_{ba} + \epsilon_{ba} \tag{3}$$

We demonstrate the Empirical Annual MEF in Figure 2 with a scatter plot showing a linear relationship between this change across all hours of 2021 in ISO-NE. We find the slope (936 lbs / MWh) roughly equals the carbon intensity of combined cycle gas generation. This makes sense, since it is the dominant fossil fuel used in the New England grid region. The high coefficient of determination (0.977) indicates an extremely linear relationship found between changes in load causing changes in emissions. In regions with coal, we find this value to be proportionally higher, as generation from coal is responsive to some amount of changes in load.

In regions with no share of coal in the generation mix (CAISO, NYISO, ISONE) Empirical Annual MEF values are approximately equal to the carbon intensity of natural gas combined cycle — 876, 963, and 936 lbs / MWh, respectively. In regions with coal (MISO, SPP, PJM, and ERCOT) we find comparatively higher values — 1,461; 1,466; 1,264; and 1,159 lbs / MWh, respectively. These values do not come close to the carbon intensities of a single coal plant (upwards of 2,200 lbs / MWh), but represent a mix of both coal and gas in the annual average of the marginal fuel mix. Although different MEF models may capture different perspectives, we would not expect any model to vastly differ from the Empirical Annual MEF, unless they are biased towards a particular fuel source. To apply this test, we will calculate the mean absolute percentage error rate across grid regions between the Empirical Annual MEF values and the average annual values of a particular MEF model.

5.2.2. Across grid regions, the annual average MEF is expected to be positively correlated with the share of coal in the fuel mix

As an extension of Section 5.2.1, we observe a correlation between the share of coal in a grid's fuel mix and an increase in that grid region's Empirical Annual MEF. Coal is typically thought of as "baseload capacity," leading some to propose that gas must be the marginal fuel source in all balancing areas because of the shorter start-up times and lower fuel costs. By demonstrating that the change in emissions relative to the change of load scales proportionally to the share of coal in a grid region, we prove that coal is in fact a marginal fuel source in these grid regions. We fit a linear regression model between the Empirical Annual MEF and percentage share of coal in each grid region's 2021 fuel mix. The fitted model has a slope of an additional 15 lbs / MWh and a correlation of 0.99, indicating that coal is a marginal fuel source during some hours (see Figure 4 in Section 6.2.2 for this test). By extension, annual average MEFs are expected to increase proportionally to the share of coal in a region. We apply this test to MEF models by calculating the correlation between the share of coal generation in each ISO/RTO's fuel mix, and the annual average in each grid region for a particular MEF model. We expect a strong positive correlation, conveying the fact that coal is marginal during some periods.

[9] We observe that bias across the simple difference based MEFs are normally distributed–so it is removed by considering only the annual average of our linear regression.

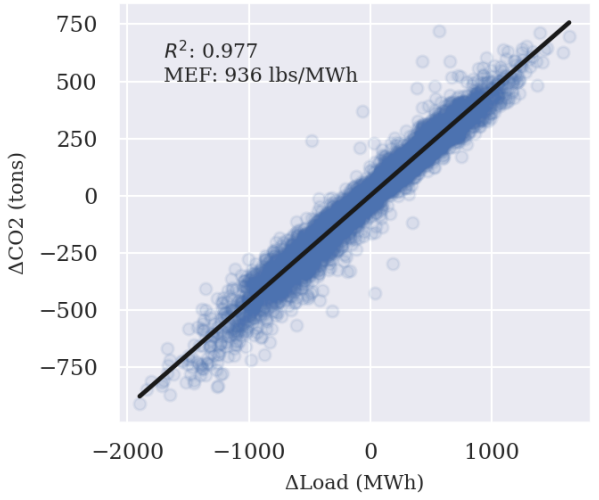

Figure 2: Illustrative Example of $\Delta E_h / \Delta G_h$ in ISONE 2021

5.3. Expected Net-Demand Temporal Patterns

While natural gas is the most-prevalent share of fuel in the U.S. power mix, it is not monolithically the marginal fuel source across all grid regions. Most grid regions in the U.S. utilize a mix of multiple fossil fuel sources to provide reliability and flexibility. The diurnal demands on our power system — typically daily peak load in the early evening, low load during the middle of the night, and then ramping up during the morning — interact with the technical limitations of generators to create variations in the fuel type and carbon intensity of the marginal fuel. As more renewables are added to the grid, using net-demand (demand minus generation from renewables) is an important heuristic to understand the demand on the power grid that is expected to come from dispatchable fuel sources. We expect MEFs to differ during periods of low net-demand compared to periods of high net-demand as market dynamics and supply curves change.

5.3.1. MEFs are expected to be lower during the 20% of peak net-demand compared to the 20% of lowest net-demand in regions with significant coal

In regions with substantial generation from both coal and gas, detecting when the marginal fuel source switches is critical for an MEF to be accurate. Coal is both costlier and slower to start up than gas. The levelized cost of energy (LCOE) of combined cycle gas ($60 / MWh) is nearly half of that of coal ($108) Lazard (2021). Coal plants have start-up times as long as 10 hours, while combined cycle gas can be ready in as little as an hour (International Renewable Energy Agency, 2021). While gas is lower cost and faster to respond, many existing coal plants continue to exist as "stranded assets" that serve daily peak load despite their uncompetitiveness Wilson and Piper (2021). Because coal has such a long start-up time, if these generators desire accessing anticipated higher LMPs at a later time, or capacity market payments, they must stay online through the night during periods of low load. Because of this operating characteristic, we expect to see coal as a marginal fuel source during lower demand hours as gas operates with greater ramping flexibility. Coal is also the most-polluting fuel source on 99.5% of days in PJM, 98.4% of days in MISO and SPP and on 55.1% of days in ERCOT. [10] Since coal generators have low flexibility in operational requirements and relatively high carbon intensities, we expect that coal will be marginal during low net-demand hours, causing relatively higher MEFs during these times Hittinger and Azevedo (2015).

To validate this hypothesis, we take the derivative of emissions with respect to load for each percentile of system net-demand. We call this the "Empirical Percentile MEF." The derivative is calculated from the sum total of emissions and load data from CAMD, while net-demand is calculated using hourly EIA data. In Figure 3, the orange lines represent a cumulative distribution function of system net-demand; the blue lines represent the Empirical Percentile MEF. In the coal-heavy grid regions of MISO, SPP, and ERCOT, the Empirical Percentile MEF begins decreasing somewhere

[10] See Appendix 9 for an analysis of the frequency of various fuel sources being the most polluting on any given day.

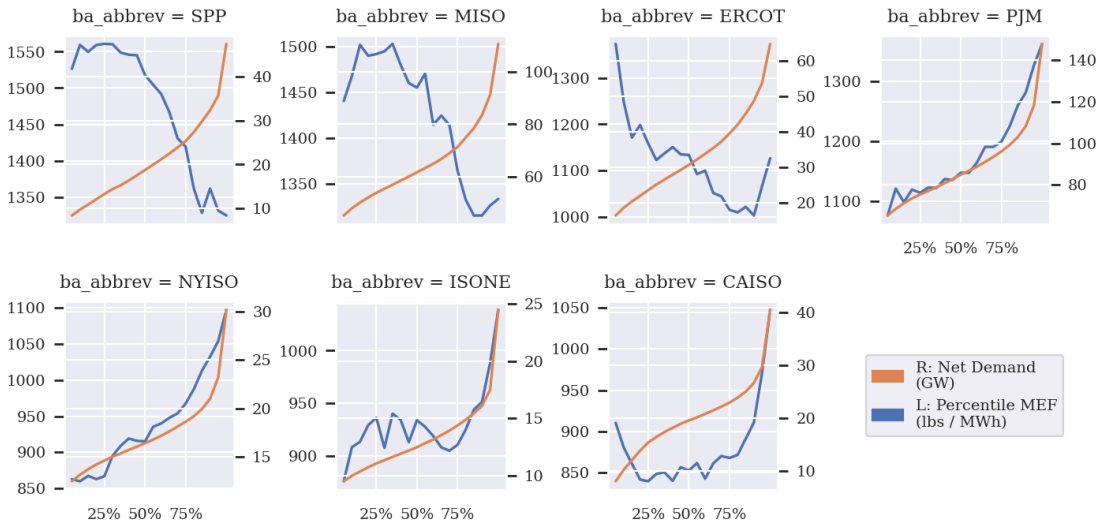

Figure 3: Empirical Percentile MEF vs. Net Demand by Grid Region

around the 80th percentile of system load. It decreases to levels that approach that of combined cycle natural gas generation. This underscores the fact that in regions with significant coal generation it is less carbon intensive to use electricity near peak times (when gas is more likely to be responding) than off-peak (when coal is responding).

PJM presents an exception to this rule, as we see the MEF increase as system net-demand approaches the 80th percentile. A possible explanation for the deviance in PJM's empirical data from this testable proposition is the large volume of net exports from PJM, largely to NYISO and MISO. We calculate that net exports account for 6.5% of PJM's total 2021 generation from all fuel sources, which may be supporting combined cycle natural gas in PJM to stay online during all hours as an exporting fuel source.

To apply this test to MEF models, we align MEF values with system net-demand calculated from EIA data. We then calculate the median MEF value for each model and applicable grid region during two bins: the lowest 20% and highest 20% of net-demand. We expect for the 0–20th percentile median to exceed the 80–100th percentile median in MISO, SPP, and ERCOT, as we know coal is more marginal than gas during these times.

Table 2
Description of ISO/RTO Curtailment Data Sources

	CAISO	NYISO	ISONE	PJM	MISO	ERCOT	SPP	IESO
% time Curtailment	23.8%	25.8%	26.2%	33.2%	25.2%	46.1%	55.4%	37.4%
Resolution	Hourly	Hourly	Hourly	Hourly	Hourly	5 Min.	5 Min.	Hourly
Reported Curtail-ment as	System oversup-ply MW	System oversup-ply MW	System marginal fuel	% nodes with marg. fuel	Regional marginal fuel flag	Plant output capability and actual output	System oversup-ply MW	Plant output capability and actual output

5.3.2. MEFs are expected to be higher during the 20% of peak net-demand compared to the 20% of lowest net-demand in regions without coal

In grid regions without any coal capacity, we observe combined cycle gas generators providing baseload power; less-efficient simple cycle gas generators operate only when needed due to their higher levelized cost of energy around

$170 / MWh Lazard (2021). These flexible 'peaker plants' can have startup times as low as a few minutes, but at a lower efficiency, producing more emissions than combined cycle gas generators Nyberg (2013). Figure 3 also shows the relationship between the Empirical Percentile MEF for each percentile of system net-demand in the three regions where there is negligible or no coal generation: CAISO, NYISO, and ISO-NE. Unlike the coal-heavy regions, there is a positive correlation, as more polluting peakers turn on during the highest percentiles of system demand. We expect MEF models in these grid regions to show the opposite behavior shown in Section 5.3.1, with lower Empirical Percentile MEF values during the lowest 20th percentile of net-demand compared to the highest 20th percentile. This test is applied in the same way as Section 5.3.1, by comparing the median MEF of a model during the 0–20th percentile of lowest net-demand with the median during times with 80–100th percentile net-demand in CAISO, NYISO, and ISO-NE.

5.4. Curtailment

Periods of renewable energy curtailment indicate that a zero-emissions energy source is marginal. This occurs when total supply of generators exceeds the present demand, and transmission constraints limit the ability to export surplus renewable energy. Curtailment is an opportunity for additional demand to be added to the grid within those time windows, while causing zero marginal carbon emissions due to utilizing existing renewable energy that would otherwise be wasted Li, Smith, Yang and Wilson (2017). To prevent this behavior, most grid regions provide ground truth curtailment data indicating when curtailment has occurred in historical periods.

5.4.1. MEFs are expected to include a renewable curtailment component, and assess the accuracy, precision and recall of their predictions.

Table 2 provides the summary of the available data in 2021 across the U.S. ISO/RTOs, with the inclusion of Ontario's IESO. The first row reports the percent of all times with any amount of curtailment in the balancing area. Curtailment is not likely occurring across the whole balancing area during these times, but is likely localized to specific transmission constraints for which data is not publically available. The third row of Table 2 details how renewable curtailment was inferred; in some balancing authorities this is reported directly, while others require imputing it from marginal fuel flags (i.e., solar or wind is reported as a marginal fuel).

With curtailment occuring upwards of 45–55% of the time somewhere in the wind-heavy regions of ERCOT and SPP, there is significant opportunity for an MEF model to identify times when electricity usage can be optimized to reduce or eliminate marginal CO_2. However, of the data sources identified in Table 2, none provide this data in real-time, with delays ranging from several hours to several weeks. Fortunately, with ample ground truth data and predictable causal factors (e.g., low levels of net demand), predicting curtailment in real-time is an achievable task with traditional machine learning techniques, or by considering the merit-order of renewables in a simulated economic dispatch model. It is expected that MEF models include curtailment as a component to identify periods when renewables are marginal. When this component can be isolated (as a boolean signal, or a probability of curtailment occurring), it is appropriate to calculate traditional error metrics including accuracy, precision, and recall of the predicted curtailment signal against ground truth data. We will apply these metrics for all available curtailment models in Section 6.4.

6. Results and Anomalies

For each test described in Section 5, we apply the test to various MEF models. When a model claims something that is unexpected, we seek out an explanation based on the factors and types of interventions that particular MEF model is considering.

6.1. Reasonable Carbon Intensities

6.1.1. PJM SCED & AVERT unexpectedly report negative MEFs in a small number of hours

We applied the test described in Section 5.1.1 to all MEF models across the seven U.S. ISO/RTOs. We find that only AVERT and PJM's SCED claim a negative MEF during any time in 2021. AVERT claims the MEF is negative during 0.09% of the time in NYISO, 0.07% in MISO, and 0.1% in ERCOT. PJM's SCED claims a negative MEF 0.83% of the time in their balancing area. All other models do not claim that an MEF is negative during any hours. As explained in Appendix 11, SCED models such as PJM can legitimately produce negative MEF values, however these are not expected to be frequent or persistent. We find that the average duration of negative MEF values in the PJM model to be 6 minutes, and the maximum observed duration is 35 minutes. We can not readily find an explanation for the small number of hours when AVERT is claiming the MEF is negative.

Table 3
Percent of Time MEFs are Greater Than Maximum CAMD

	CAISO	NYISO	ISONE	PJM	MISO	ERCOT	SPP
Siler-Evans	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
AVERT	0.29%	3.15%	2.67%	0.73%	2.71%	0.00%	1.70%
WattTime	0.12%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Cambium Short	27.71%	20.45%	13.71%	2.83%	2.62%	3.88%	4.68%
Cambium Long	0.00%	1.23%	0.07%	0.00%	0.00%	0.00%	0.00%
PJM SCED	—	—	—	1.09%	—	—	—
CAISO SGIP	0.96%	—	—	—	—	—	—

6.1.2. AVERT, Cambium, & PJM SCED unexpectedly exceed CAMD in greater than 1% of hours.

We applied the methodology described in Section 5.1.2 to test if modeled MEF values exceed the carbon intensity of the most-polluting generator in CAMD for every hour. We find that several models violate this constraint, as shown in Table 3 PJM SCED exceeds this bound by around 1% of all hours in 2021, which is understood to be an effect of considering transmission constraints affected by a small intervention, as explained in Appendix 11. The Cambium Short Run model violates the maximum constraint frequently in several balancing areas, with mean errors ranging between 245 lbs / MWh in ERCOT to 920 lbs / MWh in NYISO. However, this is explainable as Cambium's models run using simulated (not historical) load and weather data. There is no direct temporal connection with the CAMD data. CAISO's SGIP model also has a small number of violations with a mean error of 115 lbs / MWh.

The AVERT, Siler-Evans, and WattTime models are all based upon CAMD, and so they do not consider generators excluded from this dataset (such as those smaller than 25 MWs), which would otherwise be a possible explanation for their anomalies. The likely explanation for the small number of violations from these models is that they are statistical representations of historical data, and thus it is explainable if the model is overfit, or grid operations during specific observed periods are not well represented in the historical data set. Users of these models should be aware that they are occasionally claiming the MEF is greater than is reasonably expected. AVERT's mean error ranges between 121 lbs / MWh in NYISO, and 281 lbs / MWh in PJM, while Siler-Evans' error is 13 lbs / MWh in CAISO and WattTime's is 40 lbs / MWh there.

6.2. Empirical Annual Averages

6.2.1. Error rates between the annual averages of MEF models and Empirical Annual MEFs range between 0.2% and 64.5%

For each MEF model and grid region, we take the annual average MEF value, and calculate the error rate with the Empirical Annual MEF. We expect low error rates if a model is unbiased, however, significant error rates can inform us if a model appears to be unexpectedly biased to underestimate or overestimate MEFs. Our results in Table 4 show that the AVERT model averages 21.0% higher across grid regions than the empirical value. We are not aware of an explanation for AVERT's anomalous behavior. Cambium's Short-Run MEF model appears to be biased to overestimate the MEF by an average of 25.8%, while the Long-Run MEF underestimates the MEF by -16.7%. The Long-Run model's performance is explainable as long-run models are inherently considering systemic changes in the future power system, which we expect to decrease in emissions intensity.

Table 4
Empirical Annual Average MEF Compared to Modeled Annual Averages

	CAISO	NYISO	ISONE	ERCOT	PJM	SPP	MISO	Avg. Error	Std. Dev. Error
Empirical Annual MEF (lbs / MWh)	876	963	936	1159	1264	1466	1461	—	—
EPA AVERT Annual Avg. MEF	1081	1096	1072	1269	1606	1866	1825	20.1%	6.7%
	23.4%	13.8%	14.5%	9.5%	27.1%	27.3%	24.9%		
Siler-Evans Annual Avg. MEF	826	880	850	1181	1247	1482	1505	-2.7%	4.7%
	-5.7%	-8.6%	-9.2%	1.9%	-1.3%	1.1%	3.0%		
WattTime Annual Avg. MEF	941	923	896	1161	1249	1421	1402	-1.3%	3.9%
	7.4%	-4.2%	-4.2%	0.2%	-1.2%	-3.1%	-4.0%		
Cambium Short Annual Avg. MEF	1442	1290	1182	1161	1672	1534	1730	25.8%	19.9%
	64.5%	34.0%	26.3%	0.1%	32.3%	4.6%	18.4%		
Cambium Long Annual Avg. MEF	619	691	689	1020	1324	1251	1296	-16.7%	11.3%
	-29.3%	-28.2%	-26.4%	-12.0%	4.7%	14.6%	-11.3%		
PJM SCED Annual Avg. ME	—	—	—	—	1227	—	—	-2.9%	—
	—	—	—	—	-2.9%	—	—		
CAISO SGIP Annual Avg. MEF	1166	—	—	—	—	—	—	33.0%	—
	33.0%	—	—	—	—	—	—		

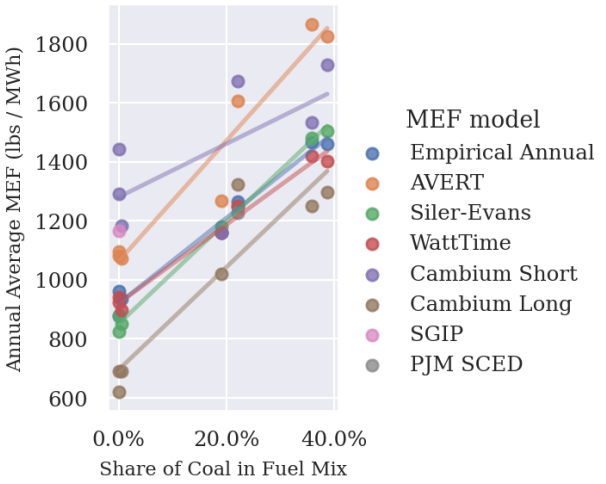

Figure 4: MEF vs. Share of Coal in Fuel Mix

However, it suggests that (unsurprisingly) Cambium's Long-Run MEF model is not accurate at estimating the effect of short-run interventions. The Short-Run model's error is concentrated in the regions of CAISO, NYISO, ISO-NE, and PJM, all of which use natural gas as the primary marginal fuel source. Cambium considers imports from neighboring regions, so one possible explanation is that Cambium is claiming imports from neighboring grid regions with coal substantially increasing the short-run MEF in these regions. CAISO's SGIP model also overestimates the annual MEF by 33%, the SGIP model estimates the heat rate of the marginal generator by using the LMP less any operating costs to determine the amount of natural gas that could be purchased at the natural gas spot price. Perhaps confounding factors such as profit taking by generators, or overestimating actual price paid for natural gas, are causing this model to bias towards less-efficient generators being marginal than the data supports. The Siler-Evans (-2.7%), PJM SCED (-2.9%), and WattTime (-1.3%) models have far lower average error rates.

6.2.2. All models are correlated with the share of coal in the fuel mix.

Across U.S. ISO/RTOs, we find that all MEF models produce annual average values that are highly correlated with the percent of coal in the grid. Pearson correlations range from 0.67 for Cambium Short-Run to 0.997 for Siler-Evans. Figure 4 plots this relationship for each model, along with linear models fit between data points.

6.3. Expected Net-Demand Temporal Patterns

6.3.1. Siler-Evans & Cambium unexpectedly model that median MEFs during peak net-demand periods are greater than low net-demand periods in regions with coal

We applied the methodology described in Section 5.3.1 to test if MEFs are higher during off-peak net-demand periods in MISO, ERCOT, and SPP compared with on-peak periods. We expect MEF models to find coal to be more marginal during lower net-demand periods, resulting in a higher MEF. We see an average difference of -12.5% across the Empirical Annual MEF, compared with a -15.1% reduction in the AVERT MEF, -7.0% from WattTime, -6.0% in the Cambium Long-Run model. Unexpectedly, we see a 0.5% average increase from the Siler-Evans model, and a 17.4% increase from the Cambium Short-Run model.

Cambium's unexpected increase in MEF during peak net-demand in coal-heavy regions can be explained for the same reasons outlined in Section 6.2.1; Cambium is based on simulated data, and as such observed net-demand is not temporally linked to the data. The Siler-Evans anomaly is driven by a 4.4% increase in peak vs. off-peak MEFs in MISO. However, this is not expected or easily explainable. Table 5 contains the off-peak and on-peak average MEF values for each model.

Table 5
Coal Off Peak (0-20th Percentile Demand) and On Peak (80-100th) MEFs

		MISO	ERCOT	SPP
Empirical Percentile MEF	Off-Peak Median MEF	1521	1294	1634
	Peak Median MEF	1387	1101	1412
	Percent Difference	-8.8%	-14.9%	-13.6%
AVERT	Off-Peak Median MEF	1780	1440	2140
	Peak Median MEF	1620	1200	1720
	Percent Difference	-9.0%	-16.7%	-19.6%
Cambium Short	Off-Peak Median MEF	1689	973	1384
	Peak Median MEF	1680	1260	1706
	Percent Difference	-0.5%	29.6%	23.2%
Cambium Long	Off-Peak Median MEF	1425	951	1321
	Peak Median MEF	1062	1145	1152
	Percent Difference	-25.5%	20.4%	-12.8%
Siler-Evans	Off-Peak Median MEF	1464	1205	1482
	Peak Median MEF	1528	1179	1473
	Percent Difference	4.4%	-2.2%	-0.6%
WattTime	Off-Peak Median MEF	1445	1204	1444
	Peak Median MEF	1318	1116	1374
	Percent Difference	-8.8%	-7.3%	-4.8%

6.3.2. All models estimate that MEFs are higher in peak net-demand periods than during off-peak periods in regions without coal.

We find that all models demonstrate the expected behavior of the test described in Section 5.3.2 across all regions tested. For this test we compare the median MEF during the 0–20th percentile of net-demand, with the median MEF during the 80–100th percentile net-demand. The Empirical Percentile MEF increases by a mean of 16.4% across CAISO, NYISO, and ISO-NE. Comparatively, AVERT's MEF increases by 28.4%, the Cambium Long-Run model increases by 24.3%, Siler-Evans increases by 15.8%, the Cambium Short-Run increases by 15.0%, and WattTime's model increases by 10.1%. The CAISO SGIP model only increases by 0.4%, compared to an expected 8.4% increase from the Empirical Percentile MEF.

Table 6
Gas Off Peak (0-20th Percentile Demand) and On Peak (80-100th) MEFs

		CAISO	NYISO	ISONE
Emp. Percentile MEF	Off-Peak Median MEF	899	846	881
	Peak Median MEF	974	1078	928
	Percent Difference	8.4%	27.4%	5.4%
AVERT	Off-Peak Median MEF	1060	1020	1020
	Peak Median MEF	1180	1320	1300
	Percent Difference	11.3%	29.4%	27.5%
CAISO SGIP	Off-Peak Median MEF	1236	—	—
	Peak Median MEF	1241	—	—
	Percent Difference	0.4%	—	—
Cambium Short	Off-Peak Median MEF	1057	930	933
	Peak Median MEF	1087	1126	1017
	Percent Difference	2.9%	21.1%	9.0%
Cambium Long	Off-Peak Median MEF	529	576	625
	Peak Median MEF	673	758	730
	Percent Difference	27.3%	31.8%	16.9%
Siler-Evans	Off-Peak Median MEF	818	811	804
	Peak Median MEF	872	991	881
	Percent Difference	6.6%	22.1%	9.5%
WattTime	Off-Peak Median MEF	960	876	865
	Peak Median MEF	940	1007	911
	Percent Difference	-2.1%	15.0%	5.3%

6.4. Curtailment

6.4.1. Only the WattTime and PJM SCED models include curtailment in their real-time signal

Of the MEF models considered, only the WattTime and PJM models include curtailment as a component of the MEF signal. WattTime models curtailment using machine learning models trained on BA specific historical curtailment data, and using model inputs including load, LMP, time of day, and weather. Table 6 includes accuracy, precision, and recall metrics for when WattTime's curtailment model predicts a probability of curtailment higher than 50% anywhere in the grid region. Accuracy is the percent of times where the correct boolean value was predicted; precision is the proportion of true-positives out of true and false positives; recall is the proportion of true positives out of true positives and false negatives.[11] Low recall in WattTime's PJM, MISO, and ISO-NE models indicate that curtailment is rarely predicted as occurring in these regions.

Table 7
Accuracy, Precision, and Recall of WattTime's v3.2 Curtailment Models

	CAISO	ISONE	PJM	MISO	ERCOT	SPP	IESO
Accuracy	91.5%	73.8%	67.3%	74.8%	73.3%	86.7%	69.1%
Balanced Accuracy	85.3%	50.8%	50.8%	50.0%	71.1%	86.3%	58.8%
Precision	88.9%	52.9%	100.0%	0.0%	96.8%	86.7%	99.3%
Recall	73.4%	2.4%	1.5%	0.0%	43.5%	89.6%	17.6%

PJM includes curtailment endogenously in their MEF SCED solver by including renewable resources in their supply curves. When the PJM SCED solver finds a renewable resource marginal at the system level, the MEF value will be 0.[12] We find only 0.2% of hours in the PJM SCED data report an MEF of exactly zero, which we infer as system-level curtailment occurring. Of these 0.2% of times, 80.4% of them are during a period when the PJM curtailment data indicates that some subregion of the grid experienced curtailment.

WattTime's MEF model localizes curtailment by clustering nodal LMPs to understand where congestion exists. PJM's SCED solves an individual MEF for each pricing node . When transmission constraints are present within a node, causing more than one generator to be marginal, PJM's MEF for each pricing node is the average of the system-level marginal fuel, and the marginal fuel for each transmission constraint impacting the node (PJM, 2021). Of the balancing authorities that report ground truth curtailment data, none publicly report isolated curtailment information at a nodal level. Without this data, it is impossible to validate how the curtailment components of MEF models are performing at a nodal level.

7. Conclusion

The growing use of MEF models by decision makers actively shifting load to reduce emissions, combined with numerous available MEF models, is creating a need to evaluate the relative accuracy of MEF models and establish standards for their application. Current methods to do so must rely in large part on other MEF models themselves, without a means to validate those models in turn. This paper describes and demonstrates a universal conceptual framework to instead evaluate MEF models objectively based on obtainable data.

This paper demonstrates one suite of such tests applied to medium-scale (hundreds of megawatts), short-run load shifts at the regional level: the most commonly used application for MEFs today. Additionally, the framework is easily extensible to other types of load shifting interventions, and further research is needed to develop related tests for these applications–particularly medium-scale, long-run load shifts at the regional level: the second most commonly used application for MEFs today.

These tests can provide valuable insight into MEF accuracy, based on publicly available data for the United States. Further research using nonpublic data, randomized controlled trials, or non-US data would also provide more detailed insight into the relative accuracy of different MEF models for this purpose. Decision makers interested in reducing their emissions using MEFs can use this paper as a guide to first help them decide on the characteristics a model should have

[11] A model with higher precision than recall is comparatively under-predicting curtailment, while a model with higher recall than precision is over-predicting.

[12] For pricing node 1, which is PJM's representation of the system-wide MEF without considering any transmission constraints.

to align with their specific intervention, then apply these tests as validation criteria to better understand the accuracy of different MEF models for that type of intervention.

If a model violates various testable propositions, it does not necessarily discredit the model; but it does provide an obligation for further research to search for an explanation for the unexpected behavior. The power grid is vastly complicated, and changing constantly. There will always be edge cases, actors behaving uneconomically, and overlapping policy landscapes that present outliers in data that are not easy to represent in computational models. Users ought to recognize this and instead look for models that offer the strongest causal mechanisms, while considering performance on metrics that we can measure.

We invite the academic community to assess other MEF models against these testable propositions, and to suggest new ones. This paper is intended to be the first in a series from the authors, gradually introducing more rigorous tests, identifying natural experiments, and alternative data sources that can be used to validate MEF models.

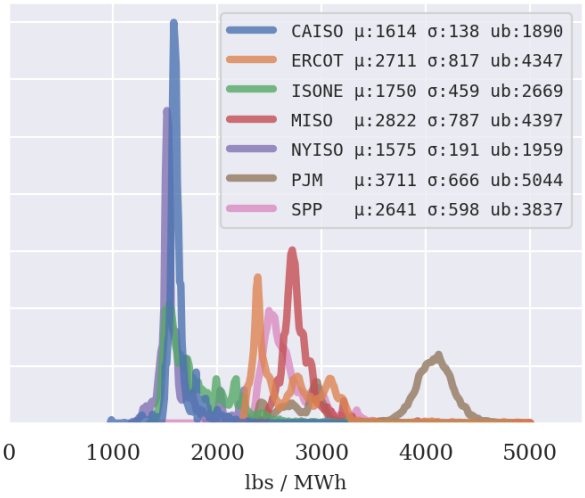

Figure 5: Distibution of Maximum Daily CI by BA

8. Appendix: Maximum Expected Carbon Intensities

In Sections 5.1.2 and 6.1.2 we consider the maximum hourly carbon intensity present in CAMD, and compare MEF models against this as a reasonable upper bound. In some applications, this may not be possible due to the delay in publishing CAMD quarterly, or a desire to inspect the 7.2% of power sector emissions not included in CAMD. In order to provide a "rule of thumb" estimate for a maximum expected carbon value of a statistical MEF model during any hour, we sought out an empirical approach. The distribution of maximum daily carbon intensities is not normal for most balancing authorities, as it is truncated at a lower bound greater than zero, and skewed towards a long right tail of inefficient peaker plants. Because the data is not normally distributed, we bootstrap 99th confidence intervals of the standard deviation for each BA and then calculate the upper bound as two standard deviations away from the mean. Figure 5 displays kernel density estimations of these densities, along with reporting the 99th confidence intervals of standard deviation, and the population mean for the major US balancing authorities. If an MEF model is claiming marginal emissions rates are greater than this upper bound, there may be explainable factors such as dispatch models considering leveraged transmission effects. If such an explanation can not be found it is likely that the model is incorrect.

9. Appendix: Fuel Source of Most Carbon Intensive Generators

In Section 5.3.2 we expect low net-demand MEFs to be higher than high net-demand MEFs due to the low flexibility operating characteristics, and higher carbon intensities of coal. We undertook the following analysis to better understand how often a given fuel type is the most carbon intensive in each grid region. First, we calculated the daily average carbon intensity of each generator in CAMD, only considering units that operated at a greater than 20% capacity factor for a given day, and less than 5% of heat output was used for combined heat and power. Next, we selected the generator that was operating at the highest carbon intensity in each BA, for each day. We then rank these peak carbon intensities in a cumulative distribution function to show the proportion of days where the most carbon intensive generator was below a certain level, this is shown in Figure 6.

We find that a coal generator was the most carbon intensive generator on over 98% of days in MISO, SPP, and PJM in 2021. In ERCOT, a small number of natural gas outliers in the 3000 - 7000 lbs / MWh range occurred during the February 2021 winter grid shut down, possibly as a result of operating at extremely inefficient temperatures or heat rates. Besides these outliers, it is notable that natural gas is the most polluting fuel source in ERCOT some 44.9% of days, operating at carbon intensities around 3000 lbs / MW–far more polluting than gas operates in other balancing authorities. Oil powered generators operate infrequently due to their high costs, and have carbon intensities similar to that of simple-cycle gas generators when they do operate. Figure 7 displays supply curves of the annual carbon intensities, by fuel type and generator. This plot helps us identify the steep right increase in carbon intensity amongst

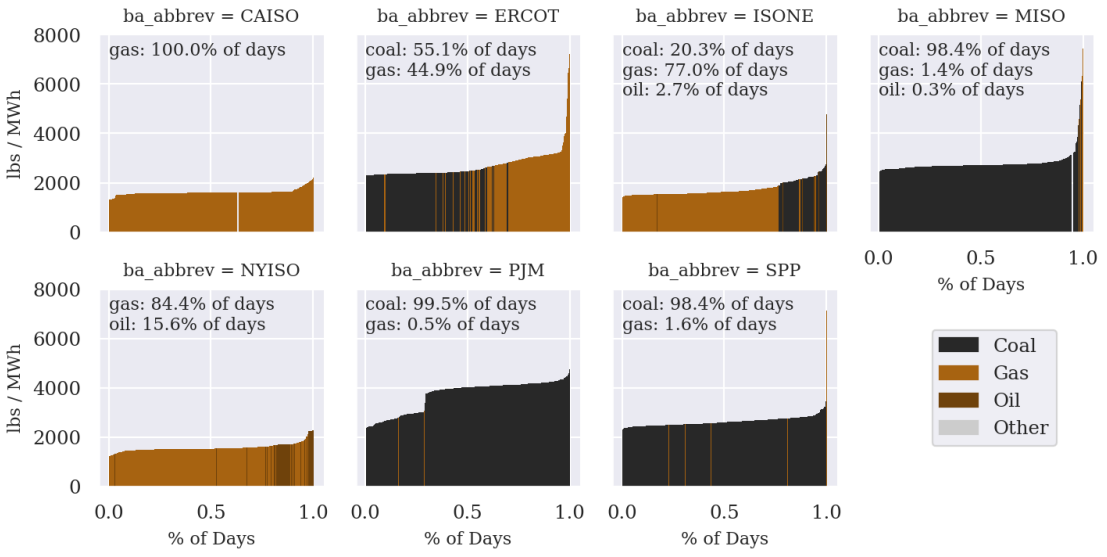

Figure 6: CDF of Unit with Maximum Daily CI, by Fuel and BA

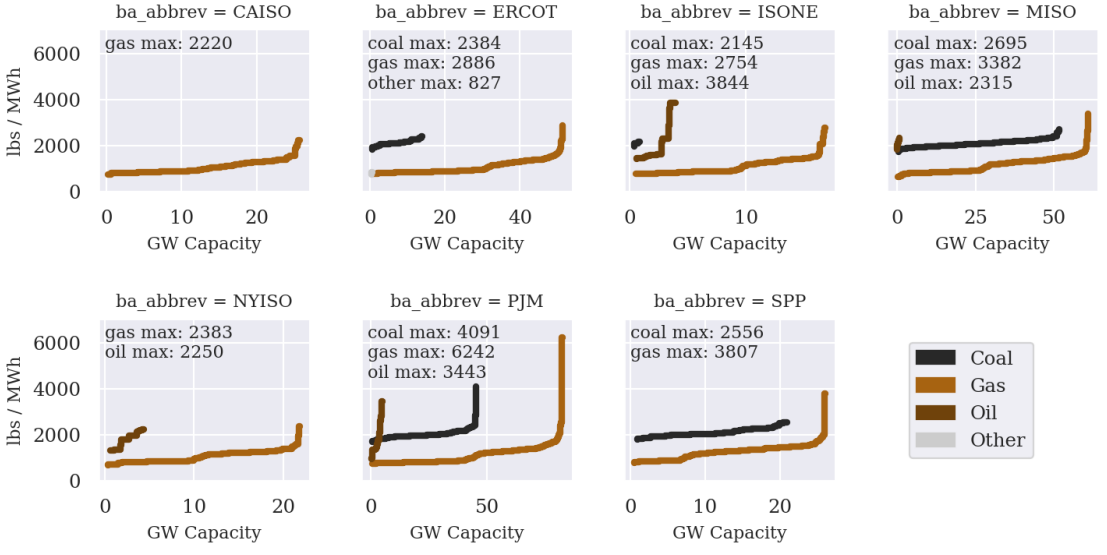

Figure 7: Supply Curves of Capacity by Annual Carbon Intensity

some of the peaker plants, however, these only comprise a very small number of plants in most BAs. Additionally, although oil peaker plants can be considerably polluting, the majority of their curves are below those of simple-cycle gas.

10. Appendix: Description of Data Processing and Analysis of the Clean Air Markets Data (CAMD)

CAMD's unit-level emissions and generation data serves as a crucial tool for validating MEF models, and in fact many are built on top of it. However, some generating units are not required to report hourly data to CAMD. These include simple cycle generators built before 1990, generators smaller than 25 MW in size, and certain independent

power producers. Our first research question was to better understand the scale of missing units and emissions from CAMD to confirm it as a representative data source. By comparing data against EIA-923 and eGRID we find that CAMD contains hourly data for 92.8% of 2021 continental U.S. power sector CO_2 emissions. This is consistent with recent work from the Open Grid Initiative, which finds 93% of CO_2 from the US power sector is represented in CAMD Miller, Pease, Shi and Jenn (2023). While the behavior of some small peaker plants is under-represented in CAMD, we do not find evidence that these are responsible for the bulk of the missing emissions from CAMD. Therefore, we are confident that CAMD is well representative of the carbon intensities and dispatch patterns of nearly all power sector emissions in the United States. CAMD data includes some rare outliers with extremely high emissions rates caused by down-ramping and observation misalignment. CAMD reports hourly emissions as cumulative totals, while load is estimated by multiplying a discrete observation (MW output recorded at some point during the hour) with the operating time percentage from that hour. We observe large outliers in hours when units are ramping down, explained by small denominator division caused by this measurement misalignment. In rare cases a very small load change can therefore cause a significant change in emissions. But it is not possible for the same effect to occur at scale, because the number of megawatt-hours over which this can occur is a fraction of the output of a single unit. Thus, for the purpose of evaluating MEFs over any significant size load change, we only consider generating units in CAMD with hourly capacity factors greater than 20% and units who provide less than 5% of their output to combined heat and power on an annual basis, as reported in eGRID. Some MEF models produce values at a five-minute increment (WattTime, PJM SCED). For these, we compare each five-minute value against the corresponding hourly CAMD value.

11. Appendix: Explanation of Outliers in PJM's SCED MEF

PJM reasons that marginal load added to one location can cause a generator to turn off at another. For instance, adding 1 MW of load at Node A could cause a gas plant at Node B to ramp down by 499 MW, while a coal plant at Node C turns on to provide 500 MW. In this exaggerated case, PJM's methodology would assign the entire emissions of the 500 MW coal plant, minus the emissions from the 499 MW gas plant to the single 1 MW change in load PJM (2021).

An illustrative example of extreme values in PJM's marginal emissions model:

$$\begin{aligned}\text{PJM MEF} &= (500 \times \Delta\text{CoalCI}) - (499 \times \Delta\text{GasCI}) \\ &= (500 \times 2200) - (499 \times 1000) \\ &= 601,000 \text{ lbs/MWh}\end{aligned}$$

Economic dispatch approaches are constrained by the technical limitations of generators, such as start-up and ramping delays, varying efficiencies at different production levels, and minimum output levels. These constraints often cause discontinuous effects in the marginal emissions signal when models are considering very small changes in load, as is the case in the illustrative example. Statistical approaches would model the example above by causally linking a change of load under certain grid conditions with a change of emissions in certain plants, the stepwise outliers average out, and the marginal emissions rate would likely be the carbon intensity of coal — 2,200 lbs / MWh — rather than the entire burden of a coal plant turning on. In any case, users of the PJM data should not interpret carbon intensities above 3,500 lbs / MWh as peaker plants, but rather as a result of the economic dispatch model adjusting output from multiple marginal units.

Acknowledgment

This work was funded in-part by Meta and Microsoft. We appreciate the thoughtful review of Nikky Avila, Julius Kusuma, Brian White, and John DeAngelis at Meta; Will Buchanan, Avi Allison, and Jason Oppler at Microsoft; Eric Hittinger at the Rochester Institute of Technology; as well as Harrison Fell and Jeremiah Johnson at North Carolina State University. Alexander Barajas-Ritchie assisted with formatting.

Declaration of Competing Interest

The authors are all employees of WattTime, a 501(c)(3) non-profit which developed one of the models considered in this paper. WattTime also hosts the website that publicly provides the CAISO SGIP model, but did not develop that model.

12. Word Count

7935

CRediT authorship contribution statement

Sam Koebrich: Data Curation, Formal Analysis, Software, Visualization, Writing - Original Draft. **Joel Cofield:** Conceptualization, Methodology, Supervision, Writing - Review & Editing. **Gavin McCormick:** Conceptualization, Funding Acquisition, Writing - Review & Editing. **Ishan Saraswat:** Data Curation, Writing - Original Draft. **Nat Steinsultz:** Data Curation, Writing - Original Draft.

References

, 2021. Innovation landscape brief: Flexibility in conventional power plants. Technical Report. International Renewable Energy Agency. URL: https://www.irena.org/-/media/Files/IRENA/Agency/Publication/2019/Sep/IRENA_Flexibility_in_CPPs_2019.pdf?la=en&hash=AF60106EA083E492638D8FA9ADF7FD099259F5A1.

, 2022. Locational Marginal Emissions: A Force Multiplier for the Carbon Impact of Clean Energy Programs. Technical Report. REsurety. URL: https://resurety.com/locational-marginal-emissions/. (Accessed April 28, 2023).

Azevedo, I., Deetjen, T., Donti, P., Horner, N., Schivley, G., Sergi, B., Siler-Evans, K., Vaishnav, P., 2020. Electricity emissions & damage factors: Quantifying electricity emissions and health, environmental and climate change damages for the u.s. power sector. URL: https://drive.google.com/file/d/1ZhJU4k-Y65oYnFs2CtpPrHt9waf0cORJ/view.

Boston University, 2021. 2020 sustainability annual report. URL: https://www.bu.edu/sustainability/files/2021/02/BU-Sustainability-Annual-Report-2020.pdf.

CAISO, 2022. 2021 production and curtailments data. Managing Oversupply. URL: http://www.caiso.com/informed/Pages/ManagingOversupply.aspx.

Callaway, D., Fowlie, M., McCormick, G., 2018. Location, location, location: The variable value of renewable energy and demand-side efficiency resources. Journal of the Association of Environmental and Resource Economists 5, 39–75.

Deetjen, T., Azevedo, I., 2019. Reduced-order dispatch model for simulating marginal emissions factors for the united states power sector. Environmental Science & Technology 53, 10506–10513.

EIA, 2021. EIA-930 Data. Hourly Electric Grid Monitor. https://www.eia.gov/electricity/gridmonitor/about. (Accessed April 28, 2023).

Elenes, A., Williams, E., Hittinger, E., Goteti, N., 2022. How well do emission factors approximate emission changes from electricity system models? Environmental Science & Technology 56, 14701–14712.

EPA, 2019. Avoided emissions and generation tool (avert) user manual, version 2.3. Office of Air and Radiation Climate Protection Partnerships Division.

EPA, 2021. Power sector emissions data guide. Office of Atmospheric Programs, Clean Air Markets Division. URL: https://www.epa.gov/system/files/documents/2022-07/CAMD%27s%20Power%20Sector%20Emissions%20Data%20Guide%20-%2007182022.pdf.

EPA, 2022. The emissions & generation resource integrated database technical guide with year 2020 data. Office of Atmospheric Programs, Clean Air Markets Division.

EPA, 2023. Avoided emissions and generation tool (avert). v3.2 excel edition. URL: https://www.epa.gov/avert/download-avert. (Accessed April 28, 2023).

ERCOT, . 60-day sced disclosure reports. https://www.ercot.com/mp/data-products/data-product-details?id=NP3-965-ER. (Prior data was received through correspondence with ERCOT).

Fell, H., Johnson, J., 2021. Regional disparities in emissions reduction and net trade from renewables. Nature Sustainability 4, 358–365.

Freeman, J., Rouzbeh Kargar, A., Söldner-Rembold, I., Ferreira, A., Couture, H., Jeyaratnam, J., O'Connor, J., Lewis, J., Hobbs, M., König, H., McCormick, C., Amuchastegui, N., Nakano, T., Dalisay, C., Davitt, A., Gans, L., Lewis, C., Volpato, G., Gray, M., McCormick, G., 2022. Power Sector: Electricity Generation Methodology. Technical Report. Climate TRACE. URL: https://github.com/climatetracecoalition/methodology-documents/blob/main/Power/Power%20sector-%20Electricity%20Generation%20Methodology.pdf.

Gagnon, P., Frazier, W., Cole, W., Hale, E., 2021. Cambium Documentation: Version 2021. Technical Report NREL/TP-6A40-81611. National Renewable Energy Lab (NREL), Golden, CO (United States).

Hawkes, A., 2010. Estimating marginal co2 emissions rates for national electricity systems. Energy Policy 38, 5977–5987.

Hawkes, A., 2014. Long-run marginal co2 emissions factors in national electricity systems. Applied Energy 125, 197–205.

Hittinger, E., Azevedo, I., 2015. Bulk energy storage increases united states electricity system emissions. Environmental Science & Technology 49, 3203–3210.

Holland, S., Kotchen, M., Mansur, E., Yates, A., 2022. Why marginal co2 emissions are not decreasing for us electricity: Estimates and implications for climate policy. Proceedings of the National Academy of Sciences 119, e2116632119.

IEA, 2023. Co2 emissions in 2022. URL: https://www.iea.org/reports/co2-emissions-in-2022.

IESO, . Generator output and capability. http://reports.ieso.ca/public/GenOutputCapability/. (Accessed April 28, 2023).

ISO-NE, . 1999 nepool marginal emission rate analysis. https://www.iso-ne.com/system-planning/system-plans-studies/emissions. (Accessed April 28, 2023).

ISO-NE, 2021. Dispatch fuel mix data. Operations Reports. URL: https://www.iso-ne.com/isoexpress/web/reports/operations/-/tree/gen-fuel-mix. (Accessed April 28, 2023).

Lazard, 2021. Levelized cost of energy analysis–version 15. URL: https://www.lazard.com/media/sptlfats/lazards-levelized-cost-of-energy-version-150-vf.pdf. (Accessed April 28, 2023).

Li, M., Smith, T., Yang, Y., Wilson, E., 2017. Marginal emission factors considering renewables: A case study of the us midcontinent independent system operator (miso) system. Environmental Science & Technology 51, 11215–11223.

Lorenzen, M., Scer, M., 2020. More than a megawatt: Embedding social & environmental impact in the renewable energy procurement process. URL: https://c1.sfdcstatic.com/content/dam/web/en_us/www/assets/pdf/sustainability/sustainability-more-than-megawatt.pdf.

Miller, G., Pease, G., Shi, W., Jenn, A., 2023. Evaluating the hourly emissions intensity of the us electricity system. Environmental Research Letters 18, 044020.

MISO, 2015. Miso real-time fuel on the margin report reader's guide. URL: https://misodocs.blob.core.windows.net/marketreports/Real-Time%20Fuel%20on%20the%20Margin_Real-Time%20Fuel%20on%20the%20Margin%20Report%20Readers%20Guide.pdf.

Monitoring Analytics, 2009. Pjm marginal fuel postings data description. URL: http://www.monitoringanalytics.com/data/docs/data-description.pdf.

Monitoring Analytics, 2021. 2021 pjm marginal fuel posting data. URL: http://www.monitoringanalytics.com/data/marginal_fuel.shtml. (Accessed April 28, 2023).

Nassar, R., Hill, T., McLinden, C., Wunch, D., Jones, D., Crisp, D., 2017. Quantifying co2 emissions from individual power plants from space. Geophysical Research Letters 44, 10–045.

Nyberg, M., 2013. Thermal efficiency of gas–fired generation in California: 2012 update. Technical Report. California Energy Commission.

NYISO, . 5b 2021 hourly wind curtailments. URL: https://www.nyiso.com/documents/20142/29607069/2021%20Hourly%20Wind%20Curtailments.xlsx. (Accessed June 1, 2023).

Oliveira, T., Varum, C., Botelho, A., 2019. Econometric modeling of co2 emissions abatement: Comparing alternative approaches. Renewable and Sustainable Energy Reviews 105, 310–322.

Palmer, K., Prest, B., Her, S., Villanueva, S., 2022. Options for EIA to publish CO2 emissions rates for electricity. Technical Report. Resources for the Future (RFF).

PJM, 2021. Marginal emission rates–a primer. https://www.pjm.com/-/media/etools/data-miner-2/marginal-emissions-primer.ashx. (Accessed April 28, 2023).

Rubin, D., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66, 688.

Ryan, N., Johnson, J., Keoleian, G., 2016. Comparative assessment of models and methods to calculate grid electricity emissions. Environmental Science & Technology 50, 8937–8953.

SGIP, 2023. Handbook: Provides financial incentives for installing clean, efficient, on-site distributed generation. Technical Report. Center for Sustainable Energy; Southern California Edison; Pacific Gas and Electric Company; SDGE; SoCalGas.

Siler-Evans, K., Azevedo, I., Morgan, M., 2012. Marginal emissions factors for the us electricity system. Environmental Science & Technology 46, 4742–4748.

Siler-Evans, K., Azevedo, I., Morgan, M., Apt, J., 2013. Regional variations in the health, environmental, and climate benefits of wind and solar generation. Proceedings of the National Academy of Sciences 110, 11768–11773.

SPP, . Ver curtailments. URL: https://marketplace.spp.org/pages/ver-curtailments#%2F2021. (Accessed April 28, 2023).

Steinberg, D., Brown, M., Wiser, R., Donohoo-Vallett, P., Gagnon, P., Hamilton, A., Mowers, M., Murphy, C., Prasana, A., 2023. Evaluating Impacts of the Inflation Reduction Act and Bipartisan Infrastructure Law on the U.S. Power System. Technical Report NREL/TP-6A20-85242. National Renewable Energy Laboratory.

Voorspools, K., D D'haeseleer, W., 2000. The influence of the instantaneous fuel mix for electricity generation on the corresponding emissions. Energy 25, 1119–1138.

WattTime, 2022. Watttime's approach to modeling and validation. URL: https://www.watttime.org/app/uploads/2022/10/WattTime-MOER-modeling-20221004.pdf.

Wilson, A., Piper, S., 2021. A nationwide push for green energy could strand $68b in coal, gas assets. S&P Global Market Intelligence URL: https://cdn.roxhillmedia.com/production/email/attachment/880001_890000/6c9c2b7774f900a0df3eaf4a0a4fb8c42519f749.pdf. (Accessed April 28, 2023).

Zivin, J., Kotchen, M., Mansur, E., 2014. Spatial and temporal heterogeneity of marginal emissions: Implications for electric cars and other electricity-shifting policies. Journal of Economic Behavior & Organization 107, 248–268.